Estimating Dynamic Treatment Effects from Project STAR*

Weili Ding
University of Michigan

Steven F. Lehrer
Queen's University

October 2003

Abstract

This paper considers the analysis of data from randomized trials which offer a sequence of interventions and suffer from a variety of problems in implementation. In experiments that provide treatment in multiple periods ($T > 1$), subjects have up to $2^T - 1$ counterfactual outcomes to be estimated to determine the full sequence of causal effects from the study. Traditional program evaluation and non-experimental estimators are are unable to recover parameters of interest to policy makers in this setting, particularly if there is non-ignorable attrition. We examine these issues in the context of Tennessee's highly influential randomized class size study, Project STAR. We demonstrate how a researcher can estimate the full sequence of dynamic treatment effects using a sequential difference in difference strategy that accounts for attrition due to observables using inverse probability weighting M-estimators. These estimates allow us to recover the structural parameters of the small class effects in the underlying education production function and construct dynamic average treatment effects. We present a complete and different picture of the effectiveness of reduced class size and find that accounting for both attrition due to observables and selection due to unobservables is crucial and necessary with data from Project STAR.

Comments Welcome

We are grateful to Petra Todd for helpful discussions and encouragement at the initial stages of this project. We would also like to thank Alan Krueger for generously providing a subset of the data used in the study.

1 Introduction

Recent years have seen an interdisciplinary resurgence of interest that examines the economics and econometrics of broken randomized trials.[1] These studies focus on the estimation of various causal parameters in the presence of a variety of implementation problems in single period programs where participants either comply fully with their assignment or choose not to comply at all. Yet many randomized trials in social science and clinical medicine involve repeated or multiple stages of intervention,when it is possible that the participation of human subjects in the next stage is contingent on past participation outcomes. The study of causal effects from a sequence of interventions is limited even in the case of perfect compliance.[2] Only recently in economics, Lechner and Miquel (2002) and Miquel (2002,2003) examine the identification of dynamic treatment effects under alternative econometric approaches when attrition is ignorable. This paper concerns itself with randomized trials that provide a sequence of interventions and suffer from various forms of noncompliance including selective attrition.

We examine these issues in the context of Tennessee's highly influential class size experiment, Project STAR. The experiment was conducted for a cohort of students with refreshment 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). Most published results from this study have reported large positive impacts of class size reduction on student achievement, which has provided much impetus in the creation of large-budget class size reduction policies in many states and countries.[3] Several of these studies have noted and attempted

[1] Comprehensive surveys of recent developments in the economics literature can be found in Imbens and Rubin (1997) and Heckman, LaLonde and Smith (2001). See Yau and Little (2001) and Frangakaris and Rubin (2002) for developments in biostatistics and statistics.

[2] 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, Greenwald and Hu (1999). In these papers, subjects are required to be re–randomized each period to identify the counterfactual outcomes.

[3] See Finn et al., 2001 and the references within for an updated list of STAR papers. The United States Congress set aside $1.3 billion for class-size reduction in 2000-01, while individual states spend additional dollars. California enacted legislation in 1996 that reduced K-3 class sizes by roughly ten students per class at an annual cost of over $1 billion; the cost in 2002 was $1.6 billion. Minnesota and Nevada's proposed budget reduction recommend for $237 million and $80 million respectively. In Florida, estimates have shown the class-size initiative could cost the state as much as $27.5 billion through 2010. The positive results have influenced education policies in other countries such as Canada, where in 1997 the Education Improvement Commission in Ontario argued that in order to achieve the modest gains that were witnessed in Project STAR funding would have to be increased by

to address complications due to missing background and outcome data and noncompliance with the randomly assigned treatment that occurred during implementation.[4] However, to the best of our knowledge, an examination of the data as the result of a sequence of treatment interventions with various non-compliance issues has not been formally explored.

A variety of complications arise in experiments involving human subjects. These include subjects exiting the experimental sample (attrition bias), not taking the treatment when assigned (drop-out bias), or receiving the treatment or similar treatments when not assigned (substitution bias). Faced with these complications researchers often report either an 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. In his highly influential examination of Project STAR, Krueger (1999) follows the latter approach using initial random assignment to class type as an instrument for current class size to uncover the causal effect of reduced class size on student achievement. The IV estimate is simply the ratio of the ITT estimates of the effect of being assigned to different class types on outcomes to that on program participation. The IV estimate obtains a causal interpretation provided a series of assumptions detailed in Angrist, Imbens and Rubin (1995) are satisfied,[5] and the resulting parameter estimate is often referred to as a local average treatment effect (LATE) in the economics literature or as a complier average causal effect (CACE) in the statistics literature.[6] Further, without stronger assumptions we generally cannot identify from the population who those compliers are, which barely shed light on the corresponding policy questions.

In multi-period randomized experiments with noncompliance but ignorable attrition, estimators employing initial assignment as instruments provide estimates of the cumulative effects of a

57%. The Ontario government passed the Education Quality Improvement Act in 1997 that placed a maximum on average class sizes. The government provided school boards with $1.2 billion over three years to reduce class sizes. In 2001, Quebec began spending $137 million annually to fund a four year class size reduction program. Other provinces including British Colombia and Alberta have similar programs.

[4] In his analysis, Krueger (1999) presents instrumental variable estimates to correct for biases related to deviations from assigned class type. Nye Hedges and Konstanioiulos (1999) show that the attrition patterns were similar across small and large classes. Ding and Lehrer (2003) find that these attrition patterns by class type differ by school type. Specifically, students initially assigned to small classes were significantly less likely to leave the sample from schools where class size reductions were beneficial. About 25% of the sample schools in Kindergarten saw their small classes perform better academically than their regular classes, while 50% of the schools saw things the other way around.

[5] Without these assumptions which are detailed in footnote 12, the IV estimator has no interpretation as a causal effect.

[6] Our use of "complier" here follows Angrist, Imbens and Rubin (1995), which defines complying individuals to be those who would only receive the treatment when assigned.

program only for those compliers who conformed with their initial assignment in all subsequent years of the program. However, in the presence of non-ignorable attrition, ITT estimators are biased and IV estimators are distorted from a causal effect even with a randomized instrument.[7]

The scope of non-compliance in Project STAR is extensive. Approximately ten percent of the subjects switch class type annually and over half of the subjects who participated in kindergarten sample attrited. These attritors differed significantly in their initial behavioral relationships. Attritors received half of the average benefit of attending a small class in kindergarten. Further, the pattern of attrition differed markedly between class types within and across schools. By treating attrition as random and ignorable past studies may have overstated the benefits of reduced class size.

In multi-period experiments, implementation problems proliferate as subjects may exit in different periods or switch back and forth in between the treatment and control groups across time. To estimate the average treatment effects of reduced class size in a multi-period setting, the researcher must compute counterfactual outcomes for each potential sequence of classroom assignment. In the context of Project STAR this yields 16 possible paths for the kindergarten cohort in grade three. Note that even if the experiment perfectly re-randomized subjects annually, an instrumental variables approach would be unable to estimate the full sequence of causal effects since the number of randomized instruments is less than the number of counterfactual outcomes.

To estimate the average treatment effects of reduced class size in a multi-period setting, we consider a sequential difference in difference strategy. We account for non-ignorable attrition using inverse probability weighting M-estimators. Our parameter estimates have a direct structural interpretation since our underlying model allows cognitive achievement to be viewed as a cumulative process as posited by economic theory. Further, we allow the effects of observed inputs and treatment receipt on achievement levels to vary at different grade levels. The structural parameter estimates permit us to construct estimates of the full sequence of dynamic treatment effects to present a more complete picture of the effectiveness of reduced class size.

We find there are benefits to attending a small class initially in all subject areas in kindergarten and grade one. However, there does not exist additional benefits from attending small classes in both years in grade one. Further, we find there are no significant dynamic benefits from continuous treatment versus never attending small classes in all subjects in grades two and three.

[7] Frangakis and Rubin (1999) demonstrates that neither standard ITT analyses (i.e. analyses that ignores the discrepancy between assigned treatment and actual treatment) or standard IV analyses (i.e. analyses that ignores the interaction between treatment and attrition) will obtain valid estimates of the ITT and LATE respectively.

Attendance in small classes in grade three is significantly negatively related to performance in all subject areas. The data suggests that the decreasing returns to small class attendance is related to significantly greater variation in incoming academic performance in small classes relative to regular classes. The weakest incoming students in mathematics in each classroom experienced the largest gains in achievement, which is consistent with the story of teaching towards the bottom. Finally, specification tests indicate that accounting for attrition due to observables and controlling for selection due to unobservables is crucial and necessary with data from Project STAR.

2 Parameters of Interest

We begin by providing a brief overview of the parameter estimates and the effect of several sources of implementation biases in a one period model of treatment.[8] In the context of the STAR class size experiment, we refer to being in small classes as being in the treatment group and otherwise in the control group. A student is initially assigned to a small class, $M = 1$ or a regular class, $M = 0$ when she enters a school in the STAR sample.[9] Due to the non-mandatory compliance nature of this social experiment, each year the actual class type a student attends may differ from the initial assignment. We use $S_t = 1$ to denote actually being in a small class in grade t and $S_t = 0$ as being in a regular class. At the completion of each grade t, she takes exams and scores A_t (potential outcomes, A_{1t} if attending a small class and A_{0t} if attending a regular class). Notice that we cannot observe A_{1t} and A_{0t} for the same individual. Some subjects leave the STAR sample over the four years, let $L_{t+1} = 1$ indicates that a subject leaves a STAR school and attends a school elsewhere after the completion of grade t, if she remains in the sample for the next period $L_{t+1} = 0$.

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. Thus, in a single period experiment the relevant parameter is the average treatment effect (ATE) $\triangle_{ATE_t} = E(A_{1t} - A_{0t})$ or in its conditional form $E(A_{1t} - A_{0t}|X_t)$ where X_t are characteristics

[8] See Heckman, LaLonde and Smith (2001) for a comprehensive overview of the economics and econometrics of program evaluation. Detailed discussions of dropout bias, substitution bias and attrition bias can be found in Heckman Smith and Taber (1999), Heckman, Hohmann Smith and Khoo (2001) and in a special issue of The Journal of Human Resources Spring 1998 respectively.

[9] Students were added to the sample in later years because either kindergarten was not mandatory, they had previously failed their grade and had to repeat it, switched from a private school or recently moved to the school district that contained a participating school.

that affect achievement.

Project STAR was designed to use random assignment to circumvent problems result from selection in treatment. By randomly assigning subjects to class types the researcher is assured that the treatment and control groups are similar to each other (i.e., equivalent) prior to the treatment and any difference in outcomes between these groups is due to the treatment, not complicating factors. In implementation, however, if people self select outside of their assigned treatment, risks rise that the groups may no longer be equivalent prior to a period of treatment and the standard experimental approach identifies alternative parameters of interest in a single period model of treatment intervention.

2.1 Sources of Bias in a Single Period Intervention

Self selection has given rise to three categories of bias in the economics literature: dropout bias, substitution bias and attrition bias. The first two biases involve noncompliance with treatment assignment while the last term deals with missing data. In the context of Project STAR, dropout bias occurs if an individual assigned to the treatment group (small class) does not comply with her assignment and attends a regular class ($M = 1, S = 0$). In total, 88.0% of the subjects who were initially assigned to small classes and completed all four years of the experiment attended small classes in all the years.[10] Correspondingly substitution bias arises if members of the control group transfer to small classes ($M = 0, S = 1$).[11] Of those subjects assigned to regular classes in kindergarten, only 83.3% comply with their assignment in all four years of the experiment.

In the presence of noncompliance with treatment assignment, the standard experimental impact which compares means of the outcome variable between individuals assigned to the treatment and the control group is an estimate of the intention to treat (ITT). The ITT effect can be defined as

$$\widehat{ITT} = \bar{A}_{M=1} - \bar{A}_{M=0} \tag{1}$$

where $\bar{A}_{M=1}$and $\bar{A}_{M=0}$ are the sample mean achievements of individuals assigned to small and regular classes respectively. Thus, the researcher carries out an "as randomized" analysis in place of an "as treated" analysis. The approach ensures that if randomization is violated, factors associated with drop-out or substitution do not corrupt the interpretation of causal effects. ITT

[10] Of the 12% who dropped out, slightly more than half (68 students) were moved to regular classes in grade 1 after being termed incompatible (Finn and Achilles (1990)) with their classmates in Kindergarten. 18 of those students returned to small classes after grade 1.

[11] Parental actions would result in substitution bias. It would also occur if members of the control group find a close substitutes for the experimental treatment through the use of services such as private tutoring.

is appropriate if one is interested in estimating the overall effects of *treatment assignment*. Since education policies on class sizes are concerned with the actual experience of students in different class sizes, the ITT estimates are not valid for cost benefit analysis of policies that mandate caps on class size for every student.

Standard IV analysis that makes use of initial random assignment as an instrument for current class size recovers an alternative parameter that is referred to in the statistics literature as the complier average causal effect (CACE). Angrist, Imbens and Rubin (1996) list a series of assumptions that if satisfied, allow IV estimates to be interpreted as average treatment effects for compliers.[12] Complying individuals are those who would only receive the treatment when assigned.[13] The identification of a group of compliers is not straightforward in general. The CACE can be defined as

$$\widehat{ITT}^{IV} = \bar{A}^{c}_{M=1} - \bar{A}^{c}_{M=0} \tag{2}$$

where $\bar{A}^{c}_{M=1}$and A^{c}_{M0t} refer to the sample mean potential achievement outcomes of complying individuals if assigned to small and regular classes respectively.

The CACE estimate obtained using an IV approach implicitly re-scales the experimental impact. Even with experimental data, non-experimental assumptions (see footnote 12) are required to identify the CACE in the presence of drop-out bias or substitution bias. With dropout, the CACE estimate is given as

$$\widehat{CACE}_1 = \frac{\bar{A}_{M=1} - \bar{A}_{M=0}}{\Pr(S_t = 1 | M_t = 1)} \tag{3}$$

The experimental impact is re-scaled by the sample proportion of compliers in the treatment group and implicitly assumes that those who dropout received a zero impact from the intervention. With both substitution and dropout the IV estimate recovers an alternative CACE given as

$$\widehat{CACE}_2 = \frac{\bar{A}_{M=1} - \bar{A}_{M=0}}{\Pr(S_t = 1 | M_t = 1) - \Pr(S_t = 1 | M_t = 0)} \tag{4}$$

[12] The assumptions inlude random assignment of the instrument, strong monotonicity of the instrument (i.e. instrument affects probability of treatment receipt in only one direction), instrument affects outcomes only through the endogenous teatment regressor (i.e. exclusion restriction) and the stable unit value treatment assumption which posits that there are no general equilibrium effects. Without these assumptions, the IV estimator is simply the ratio of intention-to-treat estimators with no interpretation as an average causal effect.

[13] In other words, these individuals were induced to switch classes by the instrument (complied with initial assignment). This parameter is also referred to as a local average treatment effect (LATE). Since different instruments exploit different sources of variation in the data, the use of alternative instruments result in different LATE parameters.

which re-scales the experimental impact by the difference between the sample proportion of compliers in the treatment group and the sample proportion of non-compliers in the control group. The estimator implicitly assumes that those who drop out and those who substitute in received a zero impact from the intervention as the dropouts would never have attended a small class and the substitutes would have attended a small class in the absence of the experiment.

While an intent-to-treat analysis is robust to the problem of students changing class types, there still remains the problem of students being lost to follow-up. Attrition bias is a common problem researchers face in longitudinal studies when subjects non-randomly leave the study and the remaining sample for inference is no longer random but choice based. For example, only 48.77% of the kindergarten sample participated in all four years of the STAR experiment. The ITT and CACE estimates presented above are not robust to attrition bias.[14]

More formally, assume that we are interested in the conditional population density $f(A_t|X_t)$ but in practice we observe $g(A_t|X_t, L_t = 0)$ since A_t is observed only if $L_t = 0$. Additional information is required to infer $f(*)$ from $g(*)$. Assuming that attrition occurs when $L_{t+1} = 1\{L^*_{t+1} > 0\}$ where L^*_{t+1} is a latent index that is a function of observables (X_t, A_t) and unobservable components. Only when attrition is completely random (i.e. $Pr(L_{t+1} = 0|A_t, X_t) = Pr(L_{t+1} = 0|X_t) = Pr(L_{t+1} = 0)$) would traditional experimental analysis that compares outcomes of the treatment and control groups recover unbiased parameter estimates.

Attrition may be due to selection on observables and / or selection on unobservables. Fitzgerald, Gottschalk and Moffitt (1998) provide a econometric framework for the analysis of attrition bias and describe specification tests to detect and methods to adjust estimates in its presence. Econometric solutions require one to determine the factors leading to non-random attrition. Selection on observables is not the same as exogenous selection since selection can be made on endogenous observables such as past academic performance (lagged dependent variables) that are observed prior to attrition. If only selective attrition 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 to achieve unbiased estimates and $f(*)$ can be inferred from $g(*)$.

To test for selection on observables, we examine whether individuals who subsequently leave

[14] The majority of the literature that has examined the STAR data for issues related to non-compliance considers attrition patterns between class types. Past studies have presented results from simple t-tests indicating that there are significant differences between attritors and non-attritors in critical variables. In contrast, we consider regression based tests as a simple comparision of means between subsamples of those lost to follow up and those who remained in the STAR experiment, may be misleading regarding the extent of significant association of these characteristics with sample attrition once the full set of educational inputs are controlled.

the STAR experiment are systematically different from those who remain in terms of initial behavioral relationships. We estimate the following contemporaneous specification of an education production function in kindergarten by subject area

$$A_{ij} = \beta' X_{ij} + \beta'_L L_{ij} X_{ij} + v_j + \varepsilon_{ij} \tag{5}$$

where A_{ij} is the level of educational achievement for student i in school j , X_{ij} is a vector of school, individual and family characteristics, L_{ij} is an indicator for *subsequent* attrition ($L_{ij} = L_{it+s}\ for\ s = 1...T-1$), v_j is included to capture unobserved school specific attributes and ϵ_{ijT} captures unobserved factors. The vector β'_L allows for both a simple intercept shift and differences in slope coefficients for future attritors. Selection on observables is non-ignorable if this coefficient vector is significantly related to scaled test score outcomes at the point of entry (completion of kindergarten) conditional on the individuals characteristics and educational inputs at that point of the survey.[15]

The results are presented in table 1 and Wald tests indicate that the β'_L coefficient vector is significantly different for attritors and non-attritors in all subject areas. The attrition indicator is significantly negatively related to test score performance in all three subject areas indicating that the levels of performance for subsequent attritors is significantly lower in kindergarten. In all subject areas, the joint effect of attrition on all student characteristics and class type is significantly different from zero. Students on free lunch status that left scored significantly lower than free lunch students who remained in the sample in mathematics only. Interestingly female attritors out performed female non-attritors in kindergarten in all subject areas but the magnitude is small. Finally, in both mathematics and word recognition attritors received half the gain of reduced class sizes suggesting that non-attritors obtained the largest gains in kindergarten which may bias future estimates of the class size effect upwards. These results provide strong evidence that selection on observables exists and is non-ignorable. Correcting for selection on observables in the panel will reduce the amount of residual variation in the data due to attrition and likely reduce the biases due to selection on unobservables.[16]

[15] This test was originally developed in Becketti, Gould, Lillard and Welch (1988). Fitzgerald et al. (1998) demonstrate that this test is be a simply the inverse of examining whether past academic performance significantly affects the probability of attrition. Note, in this paper we subsequently estimate attrition logits to create weights to account for non-compliance. As shown in table 3, past academic performance is also significantly related to attrition further indicating that selection on observables is not ignorable.

[16] This occurs if the biases due to observables did not previously offset the biases due to unobservables. We are unable to directly or indirectly test for selection on unobservables as this requires an auxillary data source or a rich set of instruments. In our empirical approach we account for the possibility that attrition is due to

In a single period intervention the estimated intent to treat and CACE parameter is distorted from a causal effect unless the research accounts for the additional complications presented by attrition which complicates the interpretation of past estimates from Project STAR. Moreover, as we discuss in the next section it is important to treat the data as if it were from a multi-period intervention.

3 Multi-Period Intervention

The STAR project occurred for students in kindergarten through grade three. Answers to many hotly debated questions, such as when class size reductions are most effective or whether small classes provided any additional benefits in grades two and three, can be properly answered in a multi-period intervention framework. For policy purposes, one may be interested in determining whether or not the benefits of small class attendance persist in subsequent grades or which treatment sequence yields the largest benefits. In this context, the relevant parameters of interest are the full sequence of dynamic average treatment on the treated parameters that we define in the next section.

We begin by considering a two period case with constant effects, perfect compliance, no attrition bias and no refreshment samples. A_{ij2} takes one of two possible values depending on which treatment sequence $[(S_{i2} = S_{i1} = M = 1)$ or $(S_{i2} = S_{i1} = M = 0)]$ an individual was assigned to. A standard economic model of individual achievement would postulate that both current and lagged inputs affect current achievement. Equation 6 is a linearized representation of the cumulative education production function at period two

$$A_{ij2} = \beta'_{x2} X_{ij2} + \beta'_{S2} S_{i2} + \beta'_{x1} X_{ij1} + \beta'_{S1} S_{i1} + v_j + \varepsilon_{ij1} \tag{6}$$

where A_{ij2} is the level of educational achievement for student i in school j in year 2, X_{ijt} is a vector of current school, individual and family characteristics in year t, v_j is included to capture unobserved school attributes and ϵ_{ijt} captures unobserved factors in year t. Consider estimation of the following contemporaneous specification of an education production function in period two

$$A_{ij2} = \gamma' X_{ij2} + \gamma'_S S_{i2} + v_j + w_{ij2} \tag{7}$$

where w_{ij2} may include lagged inputs if they affect current achievement. In this case, γ'_S presents an estimate of the cumulative effect $(\beta'_{S2} + \beta'_{S1})$ of being in a small class for two periods.

unobserved factors.

It is not possible to separately identify β'_{S2} and β'_{S1} by estimating equation 6 since $S_{i2} = S_{i1}$ (perfectly colinear). With annual estimates of equation 7, one could examine the evolution of the cumulative effect, β'_S. With the exception of the initial year of randomization one would not be able to estimate the effect of being in a small class in that particular year without invoking strong assumptions. These assumptions are similar to those that underlie education production function studies (value added models) in that one must assume how lagged inputs affect future achievement. For instance, if the impacts are assumed to depreciate at a constant rate (as in a linear growth or gains specification in the education production function literature), it is straightforward using repeated substitution to recover estimates of the effect of being in a small class in a particular year.

If compliance was not perfect then individual achievement outcomes in period 2 would take one of four possible sequences $[(S_{i2} = 1, S_{i1} = 1), (S_{i2} = 1, S_{i1} = 0), (S_{i2} = 0, S_{i1} = 1), (S_{i2} = 0, S_{i1} = 0)]$. While this may break up the collinearity problem, unbiased estimates would be obtained only if individuals switched class type exogenously. If these transitions were due to observed test performance, individual characteristics (observed or unobserved), unobserved parental education tastes, corresponding econometric solutions are required to address these selection issues. Further, determining the causal effect of class size for each individual requires the calculation of three counterfactuals as the effect of being in a small class in the first year (S_{i1}) on second period achievement (A_{ij2}) may interact in unknown ways with second year class assignment (S_{i2}). For example, class size proponents argue that teaching strategies differ in small versus large classes (i.e. "on-task events" versus "institutional events" (e.g., disciplinary or organizational)). The effect of the current class may differ due to past learning experiences as well as incoming knowledge or foundation.

In contrast to claims in Finn, Gerber, Achilles and Boyd-Zaharias (2001) that "with few exceptions students were kept in the same class grouping throughout the years they participated in the experiment", simple summary statistics indicate that 15.20% of the students who participated in the experiment all four years switched class type at least once.[17] Further, fewer than half of the kindergarten students participated in all four years of the experiment (3085 out of 6325 students). The full set of transitions for the cohort of students who participated in

[17] Our comparision is small classes versus regular or regular with aide classes. As many schools contained multiple classes of the same class type there is likely to be even more transitions between classes of the same class type as well as switches between regular classes with and without teacher aides. Note that this pooling was also undertaken in Krueger and Whitmore (2001) and Finn, Gerber, Achilles and Boyd-Zaharias (2001) since the results are not significantly different between these two groups.

Project STAR in kindergarten is shown in figure 1. Notice that excluding attrition in grade two, there is support for all eight sequences and fourteen of the sixteen possible sequences in grade three. Accounting for this large number of transitions further motivates treating the data as a multi-period intervention.

4 Empirical Approach

Our approach builds on Miquel (2003), which demonstrates that a conditional difference-in-differences approach can nonparametrically identify the causal effects of sequences of interventions.[18] We consider a sequential linear difference in difference estimator which provides estimates of the full sequence of dynamic average treatment effects for the treated. In a single period intervention, a treatment effect for the treated estimates the average gain from treatment from those that select into treatment and is the relevant parameter for policies that are voluntary. Dynamic versions compare alternative sequences as individuals determine at the end of each grade whether they wish to alter their participation sequence and are defined below.

For ease of exposition we consider a two period model and temporarily ignore the role of attrition and school effects. An individual outcome at the conclusion of the second period is given by

$$A_{i2} = S_{i1}S_{i2}A_i^{11} + (1-S_{i1})S_{i2}A_i^{01} + S_{i1}(1-S_{i2})A_i^{10} + (1-S_{i1})(1-S_{i2})A_i^{00} \tag{8}$$

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}) assuming the four paths are all the sequences an individual can take.

As posited by a standard economic model we allow cognitive achievement to be viewed as a cumulative process. We linearize the production function at each time period allowing us to express an individual's achievement outcome in period one as

$$A_{i1} = \upsilon_i + \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \tag{9}$$

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

$$A_{i2} = \upsilon_i + \alpha_2' X_{i2} + \alpha_1' X_{i1} + \alpha_{S2}' S_{i2} + \alpha_{S1}' S_{i1} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + \varepsilon_{i2} \tag{10}$$

[18] Miquel (2002) proves that instrumental variable strategies are unable to identify the full set of dynamic treatment effects.

and t_2 reflects common period two effects. First differencing the achievement equations generates the following system of two equations

$$\begin{aligned} A_{i2} - A_{i1} &= \alpha_2' X_{i2} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + (\alpha_1 - \beta_1)' X_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + \varepsilon_{i2}^* \quad (11) \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1}^* \end{aligned}$$

where $\varepsilon_{i2}^* = \varepsilon_{i2} - \varepsilon_{i1}$ and $\varepsilon_{i1}^* = \upsilon_i + \varepsilon_{i1}$. Consistent estimates of the structural parameters of the education production function in equations 9 and 10 are obtained from this system of equations via full information maximum likelihood provided that the off-diagonal elements of the variance-covariance matrix are restricted to equal zero to satisfy the rank condition for identification.[19] Consistent structural estimates of β_{S1}' and of the teacher characteristics in the X_{i1} matrix are obtained since subjects and teachers were randomized between class types in kindergarten and to the best of our knowledge compliance issues did not arise until the following year. A subset of the structural parameter estimates for the X_{i1} matrix may not be identified since they may be correlated with ε_{i1}^*.[20]

This implementation allows the effects of observed inputs and treatment receipt on achievement levels to vary at different grade levels. This is also more flexible than other commonly used empirical education production function specifications in that it does not restrict the depreciation rate to be the same across all inputs in the production process. However, by assumption the effect of unobserved inputs are restricted to be constant between successive grades.

The full sequence of dynamic effects can be estimated as follows

$$\begin{aligned} \tau^{(1,1)(0,0)}(1,1) &= \alpha_{S1}' + \alpha_{S2}' + \alpha_{S12}' \quad (12) \\ \tau^{(1,1)(1,0)}(1,1) &= \alpha_{S2}' + \alpha_{S12}' \\ \tau^{(0,1)(0,0)}(0,1) &= \alpha_{S2}' \end{aligned}$$

where $\tau^{(x,y)(v,w)}(x,y)$ presents the dynamic average treatment effect for the treated for an individual who participated in program x in period 1 and program y in period 2 and compares her actual sequence (x,y) with potential sequence (v,w). The parameters presented in (12) are of policy interest. For example, $\tau^{(1,1)(0,0)}(1,1)$ provides an estimate of the average cumulative

[19] Note it is possible to exploit cross-equation restrictions by accounting for the error-component structure of the residual but requires the assumption that υ_i is uncorrelated with the regressors. We discuss extensions in the concluding section of the paper.

[20] Since outcome data prior to kindergarten was not collected by the STAR research team alternative approaches that explicitly allow for pre-kindergarten inputs are not possible and prevent obtaining consistent estimates of the non teacher characteristic elements of X_{i1} matrix.

dynamic treatment effect for individuals who received treatment in both periods, $\tau^{(1,1)(1,0)}(1,1)$ provides 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.

It is straightforward to extend the above two period regression example to T periods. Miquel (2003) proves that the full sequence of causal effects are estimated under the straightforward assumptions of common trend, no pretreatment effects and a common support condition.[21] Intuitively, the idea builds upon classical difference in difference analysis which uses pre-intervention data to remove common trends between the treated and controls. In this setting, data between periods of the interventions is used in addition to remove common trends between individuals on alternative sequences.

While concerns regarding substitution bias and dropout bias can also be addressed through the individual fixed effect under the plausible assumption that substitution or dropout reflect some time invariant unobservables such as parental concern over their child's development over this short time period, attrition bias may contaminate the results.[22] As shown in the preceding subsection it is possible to reweight the data to account for attrition due to selection on observables. We consider estimating the following attrition logit

$$Pr(L_{it+1} = 0|A_{it}, X_{it}) = 1\{\alpha' Z_{it} + w_{it} \geq 0) \tag{13}$$

where t is the period being studied and Z_t is a matrix of variables that are observed conditional on $L_t = 0$ and may include lagged dependent variables; A_{t-s}. The predicted probability of staying in the sample ($\hat{p}_{it}$) are then constructed

$$\hat{p}_{it} = F_w(\hat{\alpha}' Z_{it}) \tag{14}$$

where F_w is the logistic cumulative distribution function.

[21] The common support assumption ensures that there are comparable individuals in each of the counterfactual sequence. The latter assumptions affect conditional expectations and are taken for a full sequnce. In a one period case, the common trend assumption assumes that the sole difference before and after is due to treatment across groups as in the absence of treatment both groups would have in expectation similar gains in academic performance. Finally, the pretreatment assumption is that there is no effect of the treatment on outcomes at any point in time prior to actual participation. The extension to multi-period is not complex as described in Miquel (2003).

[22] Note that the individual fixed effect can also account for attrition due to selection on unobservables provided permanent unobserved heterogeneity is the driving force. Thus, the term captures both initial achievement and parental concern that is assumed fixed between two consecutive grades.

Table 3 presents results from a series of logistic regressions for the determinants of remaining in the STAR experiment. The sample for each time period is restricted to units that were in the sample in the previous period. Notice that subjects who scored higher on their most recent mathematics examination are more likely to remain in the sample at each grade level. The significance of earlier test score performance in the different subject areas further demonstrates that attrition due to observables is not ignorable.

Returning to our two period example, we now assume a random sample in period one and non-random attrition due to observables at the end of period one after removing the permanent unobservable factors affecting attrition. We calculate the probability of remaining in the sample for period two $\hat{p}_{i1}$, and following Wooldridge (2002) use it to reweight observations in estimating equation (11) as follows

$$\begin{aligned} \frac{A_{i2} - A_{i1}}{\hat{p}_{i1}} &= \frac{\alpha_2' X_{i2} + (\alpha_1 - \beta_1)' X_{i1} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + t_2 + \varepsilon_{i2}^*}{\hat{p}_{i1}} \quad (15) \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \end{aligned}$$

This method provides consistent $\sqrt{N}$ asymptotic normal estimates. However, the asymptotic variance is conservative since it ignores the fact that we are weighting on the estimated and not the actual $\hat{p}_{i1}$.[23]

We estimate equation 15 for grade one as well as corresponding versions for grade two and grade three with the kindergarten sample. Attrition is an absorbing state and the weights used in estimation for grades two and three ($\hat{r}_i^2$ and $\hat{r}_i^3$) are simply the product of all past estimated probabilities

$$\begin{aligned} \hat{r}_i^2 &= \hat{p}_{i2} * \hat{p}_{i1} \quad (16) \\ \hat{r}_i^3 &= \hat{p}_{i3} * \hat{p}_{i2} * \hat{p}_{i1} \end{aligned}$$

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 $s-1$.[24] Note, it is trivial to add school effects to the estimating equations, however, identification of school effects will only come from the limited number of school switchers.

[23] The asymptotic variance matrix that adjusts for first stages estimates is smaller. See Wooldridge (2002) for details and a discussion of alternative estimation strategies.

[24] The assumption that attrition is an absorbing state holds in the STAR sample used in our analysis and allows the covariates used to estimate the selection probabilities to increase in richness over time. See Wooldridge (2002) for a discussion.

Finally, in the above analysis we treat attrition as leaving the sample permanently and assume other missing data problems are at random. That is if a student only has reading and mathematics scores in the dataset we assume that she randomly missed the word recognition test. Selective test completion would be simple to correct for in this setting replacing the L_{it+1} indicator with a subject specific missing data indicator L^s_{it+1} and following the same estimation strategy assuming that test completion in kindergarten is random. The advantage of this approach is that we can use more observations for each subject area. We implement this approach as a robustness check on our earlier results.

5 Data

In our analysis, we include only the sample of students who participated in the STAR experiment starting in kindergarten. Pooling the kindergarten sample with the refreshment samples (students who joined the experiment after kindergarten) rests on two assumptions. First, individuals leave the sample in a random manner. Second, subsequent incoming groups are conditionally randomly assigned (based on seat availability/capacity constraint) within each school. We have shown in section 2.1 the selective attrition pattern. The second claim can be examined through simple regressions of the random assignment indicator (RA_{ijT}) on individual characteristics and school indicators as follows

$$RA_{ijT} = \gamma' X_{ijT} + v_j + e_{ijT} \tag{17}$$

for each group of students entering the experiment in year T. The results are presented in the top panel of table 2.

The results clearly demonstrate that incoming students were not conditionally randomly assigned in grades one and three. The incoming students in grades one and three as well as the full samples (bottom panel) in grades one, two and three have a significantly higher percentage of students on free lunch status in the control groups. Since the incoming subjects are not conditionally randomly assigned in grade one and grade three this invalidates the use of initial random assignment as an instrument for these cohorts of students.[25]

[25] A linear probability model is used to assess conditional random assignment in Krueger and Whitmore (2001) for the full sample of incoming students with year of entry indicators. This approach simply weights the data across grades and schools over three times as much weight on the kindergarten sample than the grade one sample. Note the statistical signifigance of the results does not change if a logit was estimated in place of a linear probability model. We did not consider checking whether extrinsic measures of teacher quality were randomly assigned since they are known to have minimal correlation with actual teacher quality.

Our outcome measures are total scaled scores from the Reading, Mathematics, Word Recognition sections of the Stanford Achievement test. The Stanford Achievement Test is a norm-referenced multiple choice test designed to measure how well a student performs in relation to a particular group, such as a sample of students from across the nation. The scaled scores are calculated from the actual number of items correct adjusting for the difficulty level of the question to a single scoring system across all grades.[26] Ding and Lehrer (2003) demonstrate that transformations of scaled scores to other outcome measures such as percentile scores or standard scores either reduce the information contained in the outcome data or require assumptions that are likely to be violated by the underlying data. We treat each test as a separate outcome measure because subjects are not comparable and one may postulate that small classes may be more effective in some subject areas such as mathematics where classroom instruction is used as opposed to group instruction for reading.

6 Results: Dynamic Treatment Effects

Our structural estimates of the causal effects of reduced class size are provided in table 4. For example, S_{i1} captures the unique regression adjusted contribution of attending a small class in grade one on achievement at different points in time. Thus alternative sequences at a given time (*i.e.* $S_{iK}S_{i1}S_{i2}$ versus $S_{iK}S_{i1}(1-S_{i2})$) are restricted to receiving the same common effect of S_{i1}.

Several interesting patterns emerge from these estimates. In kindergarten and grade one small class attendance ((S_{iK}) *and* (S_{i1})) has a positive and significant effect in all subjects areas. However, there does not exist additional (nonlinear) benefits from attending small classes in both years ($S_{iK}S_{i1}$) in grade one. Moreover, Ding and Lehrer (2003) find that the positive effect of small class in kindergarten is driven by 25% of the schools in the STAR sample, which show positive effects of small class in all three subjects; while 50% of the schools in kindergarten experienced either significantly negative or statistically insignificant small class effects in all three subjects.

After grade one, no significantly positive effect of small class exists ($P(t) \leq 10\%$) except

[26] The raw score is simply the number of correct responses a student gives to test items. Total percent scores divide the raw score by the total number of items on the test. Raw scores are converted to scaled scores by use of a psychometric technique called a Rasch model process. The Rasch model developed by George Rasch in 1960, is a one parameter logistic model that examines how performance relates to knowledge as measured by items on a test. Intuitively the idea is that the probability that an exam taker of a certain ability level answers a question correctly is based solely on the difficulty level of the item. The estimated coefficient is on the ability continuum where the probability of a correct response is 50%.

for grade two math. In the higher grades nearly all of the estimated structural parameters are statistically insignificant. Thus, the structural estimates do not lend much support for positive effect of small class attendance beyond grade one. In fact, the average small class effect in grade three (S_{i3}) is significantly ($\leq 10\%$) negatively related to contemporaneous achievement in all three subject areas.

Estimates of the dynamic average treatment effect for the treated are presented in table 5 and are calculated with the structural parameter estimates discussed above using the formulas presented in equations 12. A maximum of 1, 6, 28, and 120 effects can be calculated for each grade. However, due to lack of support of some treatment paths only 78 effects can be calculated for grade 3. We present evidence comparing sequences with the largest number of observations. These treatment effects can also be interpreted as policy simulations explaining how much one would increase achievement by switching sequences conditional on your full history of student, family and teacher characteristics.

In grade one, the set of dynamic treatment effects suggest that the largest gains in performance in all subject areas occur for students who attended small classes in either 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 significance of attending a small class in grade one alone is slightly larger in all subject areas than attendance in kindergarten alone (i.e. $\tau^{(0,1)(0,0)}(0,1) > \tau^{(1,0)(0,0)}(1,0)$), there does not exist a significant difference between either sequence ($\tau^{(0,1)(1,0)}(0,1)$). From a policy perspective the results support class size reductions, but only a single dose of small class treatment instead of continuing treatment.

These estimates provide a more complete picture of the structure and source of the gains in small class reductions. In kindergarten there was a significant effect driven by a subset of schools. Following kindergarten there are positive effects in grade one for students who made a transition between class types. Both students who substituted into small classes and dropped out of small classes scored significantly lower than their grade one classmates in each kindergarten subject[27] and received a significantly greater improvement in grade one achievement compared to their grade one classmates.[28] It is possible that teachers were targeting the weaker students

[27] These results are from within classroom regressions controlling for grade one student, family and teacher characteristics.

[28] These results are from within classroom regressions controlling for kindergarten and grade one student and teacher characteristics.

in the class. Further, these growth rates were significantly larger than those achieved by their kindergarten classmates who did not switch in grade one.[29] These tests are possible since scaled scores are developmental and can be used to measure growth across grades since within the same test subject area. The Stanford Achievement Tests use a continuous scale from the lowest to the highest grade levels of the tests. Thus a one point change from 50 to 51 is equivalent to a one point change from 90 to 91.[30]

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. This result contrasts sharply with prior work (Finn et al., 2001) that find the benefits of small classes persisting in later grade and increasing the longer an individual stayed in small classes. Moreover, the economic significance of these dynamic benefits from continuous treatment are smaller in magnitude than $\tau^{(1,1)(0,0)}(1,1)$. Together, this suggests a erosion of the early gains in later grades. The raw data supports these findings as simple t-tests between these two groups of students (always versus never attended small classes) indicate that the growth in performance in each subject area was significantly higher for students who never attended small classes in higher grades.[31] Multiple regression results further demonstrate that students who never attended small classes experienced larger growth in mathematics both from grade one to grade two and grade two to grade three. These students also had greater gains in reading from grade one to grade two.[32]

[29] It is worth noting that those students who substituted into small classes in grade one scored significantly higher than their classmates on kindergarten reading and word recognition.

[30] Other test score measures such as percentile scores, grade equivalent scores, raw scores or standard scores do not offer these benefits in interpretability.

[31] Students who never attended small classes has greater growth in performance from grade one to two in mathematics and reading than those always in small classes (t = 2.3068 with P > t = 0.0106 on one-sided test in math and t = 2.1296, P > t = 0.0166 on one-sided test in reading. The hypothesis is that gains for those never attended small classes is greater than gains for those always in small classes.), with no significant differences in word recognition (t = 0.9905, P > |t| = 0.3220). From grade two to three, never attenders gained more than always attenders in math (t = 1.6844, P > t = 0.0461 in one sided test) with no significant differences in reading and word recognition (t = -0.1373, P > |t| = 0.8908, t = 0.0024, P > |t| = 0.9981 two-sided test respectively) between these groups.

[32] The regressions include school incators as well as student and teacher characteristics. The regreesor if interest is an indicator variable set equal to 1 if $S_{iK} = S_{i1} = S_{i2} = 1$ and set to 0 if $S_{iK} = S_{i1} = S_{i2} = 0$. Individuals whose treatment history are on alternative paths are not included in the regressions. The effect (and standard error) of this regressor 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

Krueger (1999) reports that students received large benefits the first year they spent in a small class. Our results support this finding in all subject areas in grade one and in grade two mathematics. Grade two reading and word recognition have insignificantly small effects ($\tau^{(0,0,1)(0,0,0)}(0,0,1)$). In grade three, first time entrants ($\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$) had significantly negative returns from small class attendance 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 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 $\tau^{(0,1)(0,0)}(0,1)$.

The dynamic treatment effects for the treated for students who switched class types for the first time motivated a closer examination of their behavior and changes in performance. We find that switching to small classes yielded benefits to students who had significantly lower past performance in math. We compared students who dropped out of or substituted into small classes with their new classmates based on prior performance on examinations by subject area. In all subject areas and grades, students who joined small classes scored significantly lower than their new classmates with the exception of reading for those who substituted in grade two. Yet, only in mathematics did these students receive significantly greater growth in performance between grades for each period.

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. To investigate this claim which may explain what we have found in Project STAR, we identified the five students in each grade one class who had the weakest subject area performance in kindergarten. 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 teacher indicators and the full history of teacher, family and student characteristics. We found that being a "weak" student in the classroom in any subject area led to significantly higher growth in mathematics. Further, being a "weak" student in any subject area significantly reduced growth in reading.[33] At all grade levels we found that being one of the "weakest" students in the classroom in mathematics and word recognition led to significantly larger gains in performance within the classroom in the respective subject areas.

attending a small class in reading and word recognition which explains the dynamic benefits at that time.

[33] These results are robust to several alternative definitions of being a "weak" student. The results in word recognition varied by definition of a "weak" students. Relative to classmates growth, the "weak" students experienced i) significantly larger gains in word recognition, ii) significantly smaller gains in mathermatics and iii) no significant difference in performance gains in reading.

The benefits occurring to students who made transitions between class types following kindergarten runs counter to the hypothesis that students benefit from environmental stability. We conduct a more detailed examination of small classes in grade one. In each grade one small class, we identified members of the largest subgroup of students who were taught by the same teacher in kindergarten. We then ran regressions of growth in performance by subject area on this indicator controlling for school indicators and the full history of student and teacher characteristics. Members of this largest subgroup had significantly smaller gains than their classmates in mathematics (coeff.=-6.129, s.e. 2.714) and word recognition (coeff.=-4.524, s.e. 3.008) and no significant differences in readings. Multiple regressions using the number of your classmates who were taught by the your 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. These results do not support arguments for environmental stability.[34] Neither do they directly contradict the stability hypothesis since peer groups (classmates) were no longer exogenously fromed after kindergarten.

An additional effect of these transitions is they substantially increased the variation of background subject knowledge within small classrooms in higher grades. The variation in past performance was twice as large in grade two and three than grade one in reading and word recognition. In higher grades, small classes had significantly more variation in past performance in mathematics and reading than regular classes.[35] Faced with relatively less variation in the incoming knowledge of students, regressions indicate students in regular classes were able to achieve significantly larger gains in mathematics and reading between grades one and two and in mathematics from grade two to three.[36] As regular classes gained more, the dynamic benefits of small class attendance vanished. There were no significant differences in the variation of prior performance on word recognition tests between class types in higher grades nor significant

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

[35] T-tests on the equality of variances in incoming test scores indicate significantly larger variation in small classes in mathematics in grades two (P < F_obs = 0.04) and three (P < F_obs = 0.11) and in grade two reading (P < F_obs = 0.06). Variation may influence student performance through teaching methods as having a more diverse classroom may lead to increased difficulties for instructors at engaging the different levels of students. Note that in grade one, we believe heterogeneity in the class room is driven by the incoming students some of which did not attend kindergarten.

[36] Regressions including school indicators demonstrate that gains in reading between grades one and two (coefficient =-2.54, std. err.=1.05) and gains in mathematics between between grades one and two (coefficient =-2.22, std. err.=1.11) and between grades two and three (coefficient =-2.21, std. err.=0.88) were significantly lower in small classes.

differences in gains in performance on word recognition examinations between class types in grades two and three. While the patterns exhibited in higher grade may be explained by the existence of a trade-off between variation in incoming student performance and class type, more investigation is needed and the underlying economic model must be expanded to include peer effects to directly test this hypothesis.[37]

Overall, the patterns of the results does not provide systematic evidence of positive small class size effect. It is interesting to see when small classes work and when it fails by comparing growth rates in performances between alternative sequences. Yet, the evidence clearly finds that small classes do not work unconditionally.

6.1 Specification Tests

This studies 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 non-compliance are due to unobservables. The importance of accounting for attrition due to observables is examined using a test proposed by DuMouchel and Duncan (1983). The test evaluates the significance of the impact of sampling weights on unweighted estimation results by including first order interactions between the covariates and the weighting variable. Weighted and unweighted estimates are significantly different if the F test on these additional covariates is significant. In the absence of sample selection bias, unweighted estimates are preferred since they are more efficient than the weighted estimates. Test results are presented in table 6 and demonstrate that weighted estimates are preferred in all subject areas and grade levels at conventional levels in reading and mathematics and below the 20% level in word recognition.

[37] A discussion of peer effects estimation is beyond the scope of the current paper. Since students switch class types, refreshment samples may be non randomly assigned to class type there are a variety of selection issues that need to be considered. An attempt at peer effect estimation with this data can be found in Boozer and Cacciola (2001) who examine peer effects in class type and not actual class attended and use an instrumental variables procedure to overcome the myriad of selection issues where initial class type assignment is used as the instrument under the assumption that initial assignment in each year of the program was random. Note that the hypothesis is also 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 Conneticut and peer group composition affects achievement in Texas respectively. Further, international evidence from the TIMMS study finds grade four Korean students who are ability streamed in classrooms were the only country to significantly outperform the US in both science and mathematics had the largest teacher-pupil ratio of the countries that participated in the study (28.6 pupils per teacher in Korea versus 17.1 pupils per teacher in the US; OECD (1997)).

Assuming there does not exist selection on unobservables permits direct estimation of the structural equations 9 and 10. This approach is implicitly undertaken in past studies using STAR data (even those that include school fixed effects) since υ_i is assumed to be both uncorrelated with the regressors and equal to zero.[38] A likelihood ratio test can be conducted to test whether the individual intercept effects can be restricted to equal zero. Under the Null, the restriction is valid and the efficient estimator is least squares estimation without differencing. Table 7 present results of this specification test. In all subject areas and grades the Null hypothesis is strongly rejected supporting the presence of unobserved heterogeneity and the estimation of equation 11. Finally, it is worth noting that DuMouchel and Duncan (1983) tests confirm that weighted estimates are preferred for these direct estimates of the structural equations further indication that ignoring selective attrition in past studies leads to inconsistent parameter estimates.[39]

6.2 Robustness Checks

To check the robustness of our structural parameter estimates, we estimate a simpler attrition model by subject area with only the most recent lagged test score is used as an explanatory variable to predict whether the subject completes the examination in the next period. This has the advantage of substantially increasing the sample for analysis by over one thousand observations per subject area. In each attrition model, the lagged dependent variable entered significantly demonstrating that selection on observables is not ignorable. We present weighted structural parameter estimates in Table 8.[40]

There are a few minor differences between the samples in the structural parameters. For example, in grade one, the combined effect of being in treatment both years is significantly neg-

[38] Past studies have not directly estimated the structural parameters of the education production function without imposing additional assumptions. For example, Krueger (1999) estimates a contemporaneous version assuming past inputs do not affect achievement and also considered alternative specifications that restricted the manner in which past inputs affect current achievement. Ding and Lehrer (2003) present evidence that these assumptions are rejected by the underlying data and these alternative empirical education production function models do not recover the structural parameters.

[39] Structural parameter estimates that do not account for either selection or unobservables or attrition due to observables are available from the authors by request. Not surprisingly, these estimates yield alternative policy reccomendations that is more supportive of past conclusions drawn from Project STAR. Finally, the significance of the results does not change if we compare inverse probability weighted estimates of the likelihood functions. These are not presented as the likelihood for weighted MLE does not fully account for the "randomness" of the weighted sampling and is not a true likelihood.

[40] Unweighted estimates that correspond to the same sample are available from the authors by request. Note the DuMouchel and Duncan (1983) test suggest that the weighted estimates are preffered for this sample.

ative in both mathematics and word recognition. The larger sample also permits identification of additional parameters in grade three such as $S_{i1}S_{i2}S_{i3}$. Our focus is on the impact of changes in these estimates on the dynamic treatment effects. We find few changes in the statistical significance of the dynamic treatment effects presented in table 5. In higher grades, we find the the dynamic benefits of substituting into a small class in grade two become significantly smaller in mathematics. Further, substituting in to small classes in grade three ($\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$) becomes insignificant in all subject areas.

In grade one, the results continue to lend increased support to only a single dose of class size reductions. The economic significance of kindergarten increases and $\tau^{(0,1)(0,0)}(0,1) < \tau^{(1,0)(0,0)}(1,0)$. However, ($\tau^{(0,1)(1,0)}(0,1)$) remains statistically insignificant. The trade-off between small class attendance in kindergarten versus grade one is settled when examining 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. Attendance in small classes in grade one (S_{i1}) is either negatively related or unrelated to performance in grades two and three.

Overall, these results suggest that the benefits of attending a small class early may extend only in reading and word recognition. Following grade one, receiving additional treatment does not accrue any additional benefits and it remains a subject for future research to pin point why the benefits of small class instruction do not grow and actually declined in the STAR study. Further, it remains a subject of further study to understand why the benefits of early small class attendance do not persist in mathematics. We find evidence that students with the lowest entry scores gained the most within the classroom in mathematics and it remains open the exact mechanism that led to this result. In conclusion, the results suggest from a policy perspective that the single dose of small class treatment should be received in kindergarten to yield persistent positive benefits in reading at all grade levels and benefits in word recognition in kindergarten and grades one and three.

7 Conclusion

This paper considers the analysis of data from randomized trials which offer a sequence of interventions and suffer from a variety of problems in implementation. In this setting, neither traditional program evaluation estimators or non-experimental estimators recover parameters of interest to policy makers, particularly if there is non-ignorable selective attrition. Our approach is applied to the highly influential randomized class size study, Project STAR. We discuss how a researcher could estimate the full sequence of dynamic treatment effects for the treated using

a sequential difference in difference strategy that accounts for attrition due to observables using inverse probability weighting. These estimates allow us to recover the structural parameters of the small class effect in the underlying education production function and construct dynamic average treatment effects.

The evidence presented in this study (and our companion paper) presents a more complete picture of the effectiveness of reduced class sizes. Past estimates generally treat the data as if it were from a single period intervention, ignore the influences of past educational inputs and recover parameters not of interest to policy makers. Further, by ignoring selective attrition on observables past estimates are likely to be upward biased since attritors received half the benefits of reduced class size in kindergarten. Past estimates generally treat other forms of non-compliance as random whereas we find strong evidence for selection due to individual unobserved heterogeneity. Finally, estimates of conditional random assignment demonstrate that analysis with any sample above the kindergarten year may require further bias corrections.

We find that small class attendance is most effective in kindergarten. The benefits of attending a small class in early years does not have lasting impacts in mathematics and some lasting impact in reading and word recognition. This result is surprising, since in practice, teachers generally divide the full class of students in to small groups for reading whereas they teach the full class mathematics. The dynamic treatment effects indicate that there were no significant benefits of receiving instruction in small classes in the current and all prior years of the experiment as compared to never being in a small class in mathematics and above grade two in reading and word recognition. Finally, we present evidence that teachers are able to identify weak students in mathematics and boost their achievement relative to their classmates and in higher grades a trade-off between variation in background knowledge and class size may account for decreasing small class achievement gap.

While this paper presents compelling new evidence to one of the hotly debated education policy areas several methodological limitations remain. First, for identification we assume that the variance covariance matrix is diagonal and that there is no serial correlation after controlling for person-specific fixed effects and grade effects. If this assumption is valid, more efficient estimates can be obtained by exploiting the zero covariance restrictions via nonlinear GMM as proposed by Hausman, Newey and Taylor (1987). However, serial correlation would exist if unobserved factors affect achievement in a different manner each period. Ding and Lehrer (2003) propose a simple specification test based on an instrumental variables procedure to test if the growth rate of unobserved factors (i.e. innate ability) is constant between periods. If the growth rate is not equal to one, the achievement equations could be quasi-differenced and instrumental

variables regression techniques used to obtain consistent estimates of the structural parameters. Third, this study considers a weighting strategy rather than an imputation method to deal with attrition or selective test completion. To the best of our knowledge studies have yet to examine which of these approaches performs better with panels that have a triangular structure. Fourth, translating the benefits of alternative sequences of small classes to later academic and labor market outcomes is of importance for policy purposes. Krueger and Whitmore (2001) present strong evidence that being initially assigned to a small class increased the likelihood that a subject took the SAT or ACT college entrance examination using the full sample. Fifth, a more complete understanding of the trade-off between increased student variability, class size and teaching methods is needed to see if this hypothesis accounts for the reduced class size benefits in higher grades and larger benefits to low achieving students in mathematics. Data on teaching practices has been collected by the original STAR researchers but has yet to be made available to the general research community. Answers to these and other questions present an agenda for future research.

References

[1] Angrist, Joshua D., Guido W. Imbens and Donald B. Rubin (1996), "Identification of Causal effects Using Instrumental Variables," *Journal of the American Statistical Association,* vol. 91, pp. 444 - 455.

[2] Becketti, Sean, William Gould, Lee Lillard, and Finis Welch (1988), "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," *Journal of Labor Economics,* vol. 6, no. 4, pp. 472 - 92.

[3] Boozer, Michael A., and Stephen Cacciola (2001), "Inside the 'Black Box' of Project STAR: Estimation of Peer Effects Using Experimental Data," *Economic Growth Center Discussion Paper 832*, Yale University.

[4] Coleman, James S. (1992), "Some Points on Choice in Education," *Sociology of Education*, vol. 65, no. 4, pp. 260 - 262.

[5] Ding, Weili and Steven F. Lehrer (2003), "New Evidence on Education Production Functions and the Class Size Debate," *mimeo*, University of Michigan.

[6] DuMouchel, William H. and Greg J. Duncan (1983), "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples," *Journal of the American Statistical Association,* vol 78, pp. 535 - 543.

[7] Finn, Jeremy D., Susan B. Gerber, Charles M. Achilles and Jayne Boyd-Zaharias (2001), "The Enduring Effects of Small Classes," *Teachers College Record*, vol 103, no. 2, pp. 145-183.

[8] Finn, Jeremy D., and Charles M. Achilles (1990), "Answers about Questions about Class Size: A Statewide Experiment," *American Educational Research Journal*, vol. 27, pp. 557 - 577.

[9] Fitzgerald, John, Peter Gottschalk and Robert Moffitt (1998), "An Analysis of Sample Attrition in Panel Data: The Michigan Panel Study of Income Dynamics," *Journal of Human Resources,* vol. 33, no. 2, pp. 300 - 344.

[10] Frangakis, Costas E., and Donald B. Rubin (2002), "Principal stratification in causal inference," *Biometrics*, vol. 58, pp. 21 - 29.

[11] Frangakis, Costas E., and Donald B. Rubin (1999), "Addressing complications of intention-to-treat analysis in the presence of all-or-none treatment-noncompliance and subsequent missing outcomes," *Biometrika,* vol. 86, pp. 365 - 379.

[12] Hanushek, Eric A. (1999), "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis*, vol. 21, no. 2, pp. 143 - 163.

[13] Hausman, Jerry A., Whitney K. Newey, and William E. Taylor (1987), "Efficient Estimation and Identification of Simultaneous Equation Models with Covariance Restrictions," *Econometrica*, Vol. 55, no. 4., pp. 849 - 874.

[14] Heckman, James J., Robert Lalonde, and Jeffrey Smith (2001), "The Economics and Econometrics of Active Labor Market Programs," Handbook of Labor Economics, Volume 3, Ashenfelter, O. and D. Card, eds., Amsterdam: Elsevier Science.

[15] Heckman, James J., Neil Hohmann, Michael Khoo and Jeffrey Smith (2000), "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, vol. 115, pp. 651 - 690.

[16] Heckman, James J., Jeffrey Smith and Chris Taber (1998), "Accounting For Dropouts in the Evaluation of Social Experiments," *Review of Economics and Statistics*, vol. 80, no. 1, pp. 1 - 14.

[17] Hoxby, Caroline M. (2000a), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics*, vol. 115, pp. 1239 - 1285.

[18] Hoxby, Caroline M. (2000b), "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.*

[19] Imbens, Guido W. and Donald B. Rubin (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies,* vol. 64, pp. 555-574.

[20] Krueger, Alan B. and Diane Whitmore (2001), "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal,* vol. , pp. 1 - 28.

[21] Krueger, Alan B. (1999), "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics,* vol. 114, no. 2, pp. 497 - 532.

[22] Lechner, Michael, and Ruth Miquel (2002), "Identification of Effects of Dynamic Treatments by Sequential Conditional Independence Assumptions," *mimeo*, University of St. Gallen.

[23] Miquel, Ruth (2003), "Identification of Effects of Dynamic Treatments with a Difference-in-Differences Approach," *mimeo*, University of St. Gallen.

[24] Miquel, Ruth (2002), "Identification of Dynamic Treatment Effects by Instrumental Variables," *mimeo*, University of St. Gallen.

[25] Mosteller, Frederick (1995), "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children: Critical Issues for Children and Youths,* V (1995), 113 - 127.

[26] Nye, Barbara, Larry V. Hedges and Spyros Konstantopoulos (1999), "The Long-Term Effects of Small Classes: A Five-Year Follow-Up of the Tennessee Class Size Experiment," *Educational Evaluation and Policy Analysis*, vol. 21, no. 2, pp. 127 - 142.

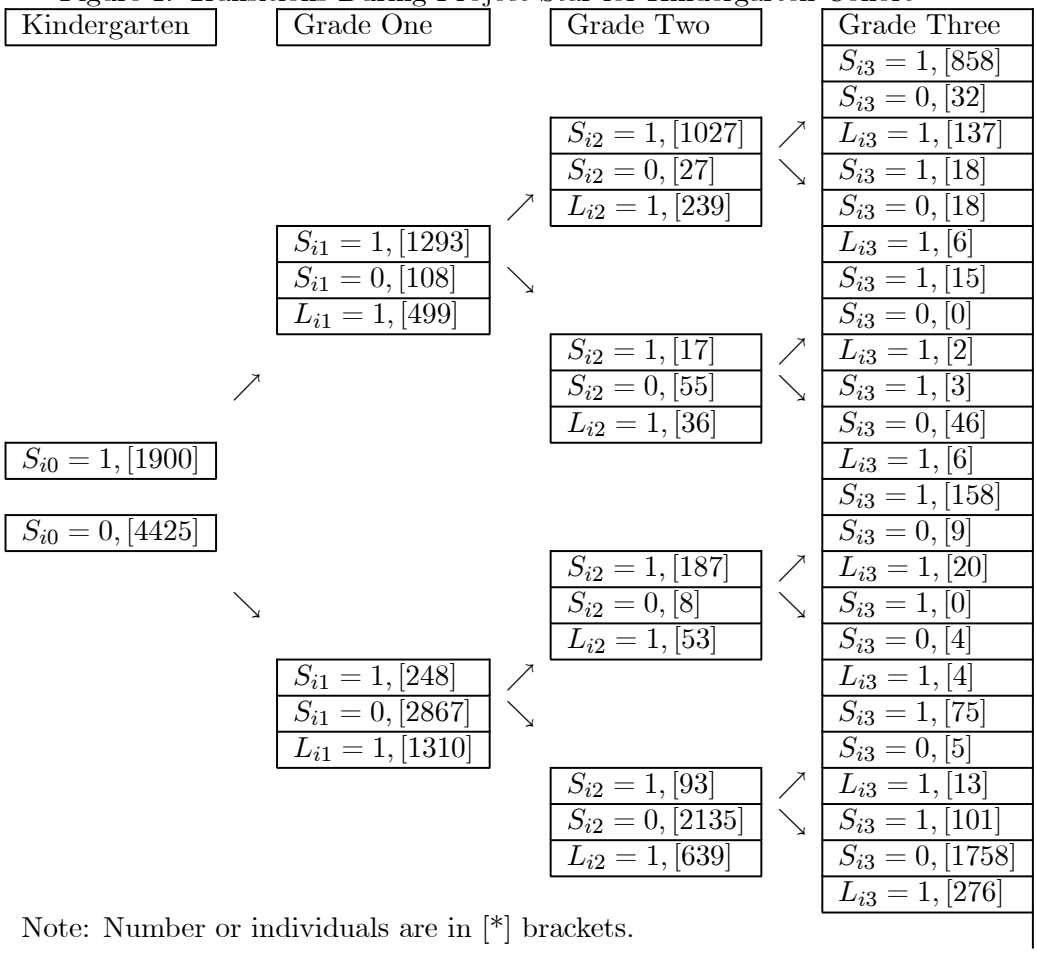
[27] Robins, James M., Sander Greenland, S. and Fu-Chang Hu (1999), "Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome," *Journal of the American Statistical Association - Applications and Case Studies*, vol. 94, pp. 687-700.

[28] Robins, James M. (1986). "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Mathematical Modelling*, vol. 7, pp. 1393 - 1512, with 1987 Errata to "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Computers and Mathematics with Applications,* vol. 14, pp. 917 - 921; 1987 Addendum to "A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Computers and Mathematics with Applications*, vol. 14, pp. 923-945; and 1987 Errata to "Addendum to 'A new approach to causal inference in mortality studies with sustained exposure periods - Application to control of the healthy worker survivor effect," *Computers and Mathematics with Applications*, vol. 18, pp. 477.

[29] Wooldridge, Jeffrey M. (2002), "Inverse Probability Weighted M-estimators for Sample Selection, Attrition and Stratification," *Portuguese Economic Journal,* vol. 1, no. 2, pp. 117 - 139.

[30] Yau, Linda and Roderick J. A. Little (2001), "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,* vol. 96, pp. 1232 - 1244.

Figure 1: Transitions During Project Star for Kindergarten Cohort



Note: Number or individuals are in [*] brackets.

Table 1: Are Attritors Different from Non-attritors

Subject Area	Mathematics	Reading	Word Recognition
Kindergarten Class Size	-1.252	-0.795	-0.902
	(0.306)	(0.182)	(0.209)
White or Asian Student	20.183	8.446	8.300
	(2.769)	(2.004)	(2.526)
Female Student	2.578	3.341	2.478
	(1.365)	(1.074)	(1.296)
Student on Free lunch	-13.688	-12.233	-13.895
	(1.695)	(1.191)	(1.483)
Years of Teaching Experience	0.334	0.262	0.337
	(0.220)	(0.124)	(0.135)
White Teacher	-1.425	-1.927	-1.945
	(4.423)	(3.116)	(3.556)
Teacher has Master Degree	-1.962	-1.506	-0.820
	(2.396)	(1.412)	(1.719)
Attrition Indicator	-32.800	-20.236	-23.016
	(7.221)	(4.583)	(5.754)
Attrition Indicator Interacted with	0.670	0.285	0.431
Kindergarten Class Size	(0.310)	(0.198)	(0.240)
Attrition Indicator Interacted with	-3.622	-0.117	-0.968
White or Asian Student	(2.756)	(1.829)	(2.377)
Attrition Indicator Interacted with	5.552	2.915	3.720
Female Student	(2.079)	(1.455)	(1.734)
Attrition Indicator Interacted with	-5.301	-0.544	0.468
Student on Free lunch	(2.400)	(1.561)	(1.897)
Attrition Indicator Interacted with	0.190	0.079	-0.059
Years of Teaching Experience	(0.211)	(0.130)	(0.164)
Attrition Indicator Interacted with	1.495	2.421	0.783
White Teacher	(3.520)	(2.150)	(2.700)
Attrition Indicator Interacted with	-1.095	1.042	1.701
Teacher has Master Degree	(2.513)	(1.589)	(1.879)
Number of Observations, R-Squared	5810, 0.304	5729, 0.294	5789, 0.258
Joint Effect of Attrition on Constant	42.22	33.19	26.28
and Coefficient Estimates	[0.000]	[0.000]	[0.000]
Joint Effect of Attrition on all	3.08	1.39	1.58
Coefficient Estimates but not constant	[0.003]	[0.207]	[0.135]
Effect of Attrition	20.63	19.50	26.28
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.

Table 2: 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	-0.275	-0.061	7.63*10E-4
	(0.012)	(0.193)	(0.041)	(0.063)
Female Student	0.012	0.199	-0.020	-0.017
	(0.019)	(0.126)	(0.021)	(0.028)
Student on Free lunch	-8.74*10E-3	-0.262	0.013	-0.057
	(0.017)	(0.167)	(0.022)	(0.037)
Joint Test of Student	0.29	1.83	1.24	1.01
Characteristics	[0.831]	[0.150]	[0.301]	[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	-0.003	-0.008	-0.021
	(0.012)	(0.021)	(0.025)	(0.027)
Female Student	0.012	0.007	0.004	0.008
	(0.019)	(0.009)	(0.009)	(0.009)
Student on Free lunch	-8.74*10E-3	-0.038	-0.030	-0.044
	(0.017)	(0.016)	(0.016)	(0.016)
Joint Test of Student	0.29	2.05	1.38	2.98
Characteristics	[0.831]	[0.114]	[0.255]	[0.037]
Number of Observations	6300	6623	6415	6500
R Squared	0.318	0.305	0.328	0359

Note:Regressions include school indicators. Standard errors corrected at the school level are in ( ) parentheses. Probability > F are in [ ] parentheses.

Table 3: Logit Estimates of the Probability of Remaining in the Sample

	Grade One	Grade Two	Grade Three
Kindergarten Reading	.00720 (.00322)	.00230 (.00494)	.00041 (.00597)
Kindergarten Mathematics	.00865 (.00116)	-.00152 (.00189)	.00126 (.00252)
Kindergarten Word	-.00035 (.00242)	-.00061 (.00369)	-.00546 (.00464)
Grade One Reading	*	.00189 (.00293)	.00053 (.00397)
Grade One Mathematics	*	.01262 (.00222)	-.00494 (.00307)
Grade One Word	*	.00834 (.00260)	.00834 (.00258)
Grade Two Reading	*	*	.00868 (.00404)
Grade Two Mathematics	*	*	.00728 (.00289)
Grade Two Word	*	*	-.00195 (.00292)
Log likelihood	-2755.54	-1239.39	-743.39
Number of Observations	5703	3127	2452

Note: Specifications include the complete history of teacher characteristics, free lunch status and class size. Specifications also includes school indicators, child gender and child race. Standard errors corrected at the teacher level in parentheses.

Table 4: 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.

Table 5: 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)

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

Table 6: 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 7: 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.

Table 8: Structural Estimates of the Treatment Parameters in Education Production Functions using Simpler Attrition Model to Account for Test Completion

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}	12.794 (4.742)***	11.221 (5.088)***	12.580 (5.433)***
S_{i1}	10.322 (2.798)***	4.032 (2.962)	9.282 (3.568)***
$S_{iK}S_{i1}$	-12.748 (5.461)***	-3.164 (5.914)	-10.514 (6.603)
Grade Two			
S_{iK}	8.993 (7.063)	17.40 (8.054)***	-1.690 (4.068)
S_{i1}	-15.755 (11.672)	-37.592 (16.710)***	-23.035 (16.522)
S_{i2}	9.001 (4.839)**	-2.471 (4.4149)	7.278 (8.297)
$S_{iK}S_{i1}$	0.437 (15.122)	-0.044 (22.636)	0.061 (21.173)
$S_{iK}S_{i2}$	-0.933 (8.931)	-19.001 (11.704)	-10.165 (21.262)
$S_{i1}S_{i2}$	14.477 (12.686)	43.044 (17.248)***	29.128 (17.002)**
$S_{iK}S_{i1}S_{i2}$	-7.712 (16.250)	8.050 (24.184)	9.189 (28.858)
Grade Three			
S_{iK}	2.512 (11.252)	12.487 (9.726)	20.241 (11.072)**
S_{i1}	7.347 (11.921)	3.743 (19.584)	3.533 (27.390)
S_{i2}	32.700 (25.589)	-14.059 (11.435)	-16.140 (8.272)**
S_{i3}	-2.991 (3.932)	-3.547 (3.411)	-5.491 (4.815)
$S_{iK}S_{i1}$	-2.424 (19.982)	-14.738 (27.662)	-18.626 (33.645)
$S_{iK}S_{i2}$	42.515 (28.165)	-19.929 (26.944)	-49.423 (35.623)
$S_{iK}S_{i3}$	-9.926 (26.641)	20.363 (23.145)	29.862 (26.369)
$S_{i1}S_{i2}$	-30.957 (29.537)	6.710 (27.010)	-3.718 (36.282)
$S_{i1}S_{i3}$	-34.354 (28.549)	-45.065 (25.648)**	-65.591 (29.914)***
$S_{i2}S_{i3}$	-27.291 (25.802)	13.957 (11.755)	25.368 (9.699)***
$S_{iK}S_{i1}S_{i2}$	-43.321 (34.722)	38.333 (40.920)	94.618 (53.809)**
$S_{i1}S_{i2}S_{i3}$	66.369 (39.566)**	46.807 (31.803)	69.728 (38.514)**
$S_{iK}S_{i1}S_{i2}S_{i3}$	8.646 (28.371)	-34.171 (28.758)	-72.552 (36.493)***

Note: Corrected standard errors in parentheses. The sequences $S_{iK}S_{i1}S_{i3}$ and $S_{iK}S_{i2}S_{i3}$ lack unique support to permit identification in grade 3.