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

Weili Ding
Queen's University

Steven F. Lehrer*
Queen's University and NBER

November 2006

* We wish to thank seminar participants at Harvard University, McGill University, NBER Economics of Education Fall 2003 meetings, New York University, Queen's University, University of Calgary, University of California - Riverside, University of Florida, Université Laval, Economics and Education Development Conference at the Federal Reserve Bank of Cleveland, 2005 CEA meetings and the 2004 NASM of the Econometric Society for comments and suggestions. We are grateful to Petra Todd for helpful discussions and encouragement at the initial stages of this project. We would also like to thank Alan Krueger for generously providing a subset of the data used in the study. Lehrer wishes to thank SSHRC for research support.

Abstract

Randomized trials in the social sciences and clinical medicine often provide a sequence of interventions and tend to suffer from a variety of implementation problems. In this setting, neither traditional program evaluation estimators nor non-experimental estimators recover the full set of causal parameters of interest to policy makers, particularly if there is non-ignorable selective attrition. We introduce an estimation strategy based on an underlying economic model of a cumulative production process to estimate treatment effects in such a setting. This approach is applied to the highly influential randomized class size study, Project STAR. Estimates from our model are combined to recover the full set of dynamic treatment effects, presenting a complete and different picture on the effectiveness of reduced class size.

1 Introduction

Many consider randomized experiments to be the gold standard of evaluation research due to among other factors the ease in which causal effects can be explained to policymakers and lay audiences.[1] The validity of using a simple comparison of outcomes across treatment and control groups to recover causal parameters relies critically on the experiment being implemented as designed. In practice, researchers regularly confront violations to the randomization protocol. Experiments that lack perfect compliance are often referred to as broken randomized trials and the interpretation of traditional estimation strategies in this setting is no longer straightforward.[2] Further, numerous randomized trials in social sciences and clinical medicine involve repeated or multiple stages of intervention, where the participation of human subjects in the next stage may be contingent on past participation outcomes. The study of causal effects from a sequence of interventions is limited even in the case of perfect compliance.[3] Only recently in economics, Lechner and Miquel (2005), Lechner (2004) and Miquel (2002, 2003) examine the identification of dynamic treatment effects under alternative econometric approaches when attrition is ignorable. This paper concerns itself with randomized trials that provide a sequence of interventions and suffer from various forms of noncompliance including nonignorable attrition and selective switching in between treatment and control groups.

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 findings 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.[4] Several of these studies have noted and attempted to address complications due to missing background and outcome data and noncompliance with the randomly assigned treatment that occurred during implementation.[5] However, to the best of our knowledge, an examination of the data as the result of a sequence of treatment interventions with various noncompliance issues has not been explored.

Since governments continue to move towards developing education policies based on scientific evidence, understanding how to interpret findings from randomized experiments that suffered from various forms of noncompliance is important. The scope of noncompliance in Project STAR is extensive. Approximately ten percent of the subjects switch class type annually, by grade three over 50% of the subjects who participated in kindergarten left the STAR sample and 4 schools left the experimental study. It is worth stating explicitly that in an experiment in which subsets of participants in both the treatment and control groups have selected to not comply fully with their assignment, standard methods to analyze experimental data may be insufficient. Faced with noncompliance researchers using data from single period experiments often report either the intent to treat (ITT) parameter that compares outcomes based on being assigned to rather than actual receipt of treatment or undertake an instrumental variables strategy. The IV estimate generally uses the randomized treatment assignment as an instrumental variable for actual treatment receipt and the resulting estimate is interpreted as a local average treatment effect (LATE).[6] Yet, Balke and Pearl (1997) demonstrate that in the face of imperfect compliance that these estimates are potentially misleading as they may lie entirely outside the theoretical bounds for an average causal effect of the intervention. Further, Frangakis and Rubin (1999) demonstrate that if the randomized intervention suffers from non-ignorable attrition where subjects leave the study in a non-random manner, the ITT estimators is biased and the IV estimator is **distorted** from a causal effect even with a randomized instrument. Taken together, this casts doubts on the usefulness of traditional approaches in the presence of noncompliance to estimate impacts that are of direct use to policymakers.

Multi-period randomized trials have the potential to address additional questions of interest to policymakers that extend beyond simply whether the intervention was successful. For instance one could determine how did the impact of the intervention evolve over time? In how many periods should the treatment be offered? Are program impacts heterogeneous with respect to demographic characteristics? Unfortunately even if attrition is random, these questions cannot be addressed with ITT analysis as it does not provide information on the actual treatment experienced, particularly when compliance is an issue. Similarly the IV approach that uses initial assignment as an instrument for treatment receipt provides an estimate of the cumulative effects of a program only for compliers. Interpretation of this parameter requires even stronger assumptions to become relevant for policy analysis as one must assume that the subpopulation that remains responsive to the instrument does not change over time. More generally, 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 if attrition is random. 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.

This study extends the burgeoning literature in the economics of education that seeks to use experimental data to estimate the effectiveness of inputs to the education production process by introducing a strategy for policy evaluation with multi-period experimental data that permits a direct link between the structural parameters from an underlying economic model of education production to dynamic treatment effect parameters. Past studies approach policy evaluation with experimental data by generally undertaking either a structural approach or a treatment effect approach. Since the goals of the literature on structural equation estimation and on the estimation of treatment parameters are often different, few studies discuss how the parameters from these approaches are related with the same data.[7] Specifically we consider estimation of the structural parameters from an underlying model that allows cognitive achievement to be viewed as a cumulative process as posited by economic theory. A sequential difference in difference strategy that allows for time varying effects of education inputs on achievement levels is adopted. We account for non-ignorable attrition over time using inverse probability weighting M-estimators. We describe how the structural parameter estimates permit us to construct estimates of the full sequence of dynamic treatment effects, presenting a richer picture of the effectiveness of reduced class size. The framework that we introduce could also be applied to evaluate interventions that affect other human capital outcomes that are assumed to be a function of cumulative process such as health.

Our empirical analysis reaches three major conclusions:

1) Accounting for non-random attrition is important in Project STAR. The pattern of attrition differed markedly between class types within and across schools and cannot be ignored by the analyst. By treating attrition as random and ignorable, we demonstrate that earlier studies have overstated the benefits of reduced class size since those who withdrew from the study differed significantly in their initial behavioral relationships; receiving half of the average benefit of attending a small class in kindergarten. Further, IV and ITT estimates occasionally lie outside of the range of tight bounds on the average treatment effect that account for non-random implementation failures.

2) We present a more complete picture of the effectiveness of class size reductions. We find benefits from small class attendance 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. Attendance in small classes in grade three is significantly negatively related to performance in all subject areas. We conduct several robustness checks and demonstrate that these results are unlikely to be due to statistical power. 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. Further, 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.

3) Specification tests indicate that controlling for selection on unobservables is crucial and necessary to estimate treatment effects with data from Project STAR.[8] A handful of subjects each year self-selected outside of their assigned groups, leading the groups to no longer be equivalent in observed covariates prior to subsequent treatment. Further, we find that in several subsequent periods, new students that entered the experiment through refreshment samples were not assigned with equal probability between the treatment and control groups, exacerbating biases from outside confounding factors.

The rest of the paper is organized as follows. In Section 2, we provide a summary of the causal inference literature on estimating treatment effects in single period interventions with and without implementation failures. We subsequently describe the causal parameters of interest in multi-period experiments. To estimate these causal parameters we introduce an empirical framework that builds on the standard economic model of human capital accumulation in section 3. The assumptions underlying our identification strategy are discussed and the estimation approached is described in this section. Section 4 presents a description of the data used in our analysis. Our results are presented and discussed in Section 5. A concluding section summarizes our findings and discusses directions for future research.

2 Causal Parameters of Interest

2.1 Single Period Interventions

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. Existing studies using Project STAR data treat the experiment as a single period intervention and we begin by providing a brief overview of the parameter estimates and the effect of several sources of implementation biases in a single period model of treatment.[9]

In the context of the STAR class size experiment, we refer to being in small classes as being in the treatment group and regular classes 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.[10] 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). The evaluation problem occurs since we cannot observe A_{1t} and A_{0t} for the same individual.

In a single period class size reduction experiment the relevant parameter of policy interest is the average treatment effect (ATE) $\triangle_{ATE_t} = E(A_{1t} - A_{0t})$ or in its conditional form $E(A_{1t} - A_{0t}|X)$ where X are characteristics that affect achievement. Project STAR was designed to use random assignment to circumvent problems result from selection in treatment. If subjects were random across class types the researcher is assured that the treatment and control groups are similar to each other 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 groups, risks rise that the groups may no longer be equivalent prior to treatment and the experimental approach is not able to identify the ATE.

2.1.1 Sources of Bias in a Single Period Intervention

Self-selection has given rise to three categories of bias: 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, 12.0% of the subjects who were initially assigned to small classes and completed all four years of the experiment dropped out of treatment.[11] Correspondingly substitution bias arises if members of the control group transfer to small classes ($M = 0, S = 1$).[12] In contrast to claims in Finn et al. (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.[13]

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 estimand is

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

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. ITT is appropriate if one is interested in estimating the overall effects of *treatment assignment*. The approach ensures that if randomization is violated, factors associated with dropout or substitution will not corrupt the interpretation of ITT. Since education policies on class sizes are concerned with the actual experience of students in different class sizes and the average effects of treatment received, 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 a local average treatment effect (LATE). Angrist, Imbens and Rubin (1996a) list a series of assumptions that if satisfied, allow IV estimates to be interpreted as average treatment effects for compliers.[14] Complying individuals are those who would only receive the treatment when assigned.[15] The identification of a group of compliers is not straightforward in general. The LATE estimand is

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

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

The LATE estimate obtained using an IV approach implicitly re-scales the experimental impact. Even with experimental data, non-experimental assumptions (see footnote 15) are required to identify the LATE in the presence of dropout bias or substitution bias. With dropout, the LATE estimand becomes

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

The ITT 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 LATE estimand given as

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

which re-scales the ITT from equation (1) by the difference between the compliance rate in the original treatment group and noncompliance rate in the original 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.

Past studies using Project STAR data treat the experiment as a single period intervention. They report either a ITT or a LATE parameter using initial assignment as an instrument for class attended. However, Frangakis and Rubin (1999) demonstrates that IV and ITT analyses recover parameters that are distorted from a causal effect if selective attrition is present. In this situation, bounds that place a range under which causal parameters including the ATE lies can be more informative. These methods not only can shed some light on the parameter of policy interest that are robust to non-random attrition but also have the advantage of placing weaker assumptions about treatment selection and outcomes than methods used to recover the LATE parameter. In our analysis, we will consider several alternative strategies that place bounds on ATE and contrast them with estimates obtained from ITT and IV. Yet, analysis of Project STAR data should not only account for non-random attrition but also consider the multiple periods of interventions that occurred. As we discuss next, by explicitly accounting for the dynamic structure of the intervention researchers are able to estimate a wider range of causal parameters that can address a more complete series of policy relevant questions.

2.2 Multi-Period Experiments

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 later grades, can be properly answered in a multi-period intervention framework. For policy purposes, one may be interested in 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.

Dynamic average treatment effect for the treated parameters are the average difference between two alternative sequences of treatment received. Implicitly it calculates the average difference between groups of individuals with different historical paths of treatment received for an individual who has selected one of these paths. Following Lechner (2004), we formally define $\tau^{(x,y)(v,w)}(x,y)$ as the dynamic average treatment effect for the treated parameter, which for individuals who participated in program x in period 1 and program y in period 2 and measures the average difference in outcomes between their actual sequence (x,y) with potential sequence (v,w). The number of potential sequences in multi-period experiments depends not only on the number of stages where treatment was offered but also on the degree of compliance at each stage.

To illustrate, consider a two period case with constant effects, perfect compliance, no attrition bias and no refreshment samples. For each individual, A_{ij2} takes one of two possible outcomes depending on which treatment sequence $[(S_{i2} = S_{i1} = M = 1)$ or $(S_{i2} = S_{i1} = M = 0)]$ they were assigned to. A standard economic model of individual achievement would postulate that both current and lagged inputs affect current achievement. Equation (5) is a linearized representation of the cumulative education production function in period two

$$A_{ij2} = \beta'_{S2} S_{i2} + \beta'_{S1} S_{i1} + \varepsilon_{ij2} \tag{5}$$

where A_{ij2} is the level of educational achievement for student i in school j in year 2 and ϵ_{ij2} captures unobserved random factors. Consider estimation of the following contemporaneous equation in period two

$$A_{i2} = \gamma'_S S_{i2} + v_i + \upsilon_j + w_{ij2} \tag{6}$$

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

It is not possible to separately identify β_{S2} and β_{S1} by estimating equation (5) since $S_{i2} = S_{i1}$ (perfectly colinear). With annual estimates of equation (6), one could examine the evolution of the cumulative effect, γ_S. With the exception of the initial year of randomization one could not estimate the effect of being in a small class in that particular year without invoking extra assumptions.[16] If compliance was not perfect then individual achievement outcomes in period 2 would take one of four possible sequences $[(S_{i2} = 1, S_{i1} = 1), (S_{i2} = 1, S_{i1} = 0), (S_{i2} = 0, S_{i1} = 1), (S_{i2} = 0, S_{i1} = 0)]$. In this case an individual's outcome at the conclusion of the second period can be expressed as

$$A_{i2} = S_{i1}S_{i2}A_i^{11} + (1 - S_{i1})S_{i2}A_i^{01} + S_{i1}(1 - S_{i2})A_i^{10} + (1 - S_{i1})(1 - S_{i2})A_i^{00} \qquad (7)$$

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. While imperfect compliance may break up the collinearity problem, unbiased estimates of equation (5) require that individuals switch class type exogenously. If these transitions were due to observed past test performance, individual characteristics (observed or unobserved), unobserved parental education tastes, corresponding econometric solutions are required to address these selection issues.

2.3 Attrition Bias

While an intent-to-treat analysis is robust to the problem of students changing class types in single and multi-period experiments, 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 choice based. Unlike noncompliance with treatment assignment, the presence of non-random attrition in single and multi-period intervention does not allow researchers who engage in IV or ITT analyses to recover LATE and ITT parameters, since one must account for the additional bias introduced by selective nonresponse.

Define $L_{t+1} = 1$ to indicate that a subject leaves a STAR school and attends a school elsewhere after completing grade t, if she remains in the sample next period $L_{t+1} = 0$. 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. This holds even when the mechanism that leads to attrition is not affected by the assignment and receipt of treatments because the (observed and unobserved) characteristics of respondents are in general systematically different from those of nonrespondents. In addition to bias, an inefficiency problem arises from the information loss due to the exclusion of some observations from the analysis.

Attrition may be due to selection on observables and / or selection on unobservables and econometric solutions require one to determine the factors leading to non-random attrition.[17] 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 $g(*)$.

Fewer than half of the kindergarten students participated in all four years of the experiment (3085 out of 6325 students). The participation rate varied significantly by class type across schools.[18] To test for selection on observables, we follow Becketti, Gould, Lillard and Welch (1988) and examine whether individuals who subsequently leave the

STAR experiment are systematically different from those who remain in terms of initial behavioral relationships. The following equation is estimated

$$A_{ij1} = \beta' X_{ij1} + \beta'_L L_{ij} X_{ij1} + \upsilon_j + \varepsilon_{ij1} \tag{8}$$

where A_{ij1} is the level of educational achievement for student i in school j in the first year, X_{ij1} is a vector of initial school, individual and family characteristics, L_{ij} is an indicator for *subsequent* attrition ($L_{ij} = L_{it+s}\ for\ s = 1...T-1$), υ_j is included to capture unobserved school specific attributes and ϵ_{ij1} 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 individual's characteristics and educational inputs at that point of the survey.

3 Empirical Model

In this section, we provide a simple model that guides our estimation and discuss the assumptions necessary to nonparametrically identify the structural parameters and dynamic treatment effects. Following Ben-Porath (1967) and Boardman and Murnane (1979) we view the production of education outcomes (or cognitive achievement) as a cumulative process that depends upon the potential interactions between the full history of family and school inputs as well as the child's innate characteristics. Formally, conditional on the selection of school j by child i's parents (who maximize household indirect utility), the complete history of inputs and class size treatments $[(X_{iT}...X_{i0}), (S_{ijT}...S_{ijo})]$, and independent random shocks $(\epsilon_{iT}...\epsilon_{i0})$, the child gains knowledge as measured by a score on an achievement test at period T:

$$A_{ijT} = h_T(X_{iT}...X_{i0}, S_{jT_T}...S_{jT_o}, v_i, \epsilon_{iT}...\epsilon_{i0}) \tag{9}$$

where h_T is an unknown twicely differentiable function. Note X_{ijt} is a vector of school, individual and family characteristics in year t and v_i is included to capture unobserved time invariant individual attributes.

Assuming that the unobserved factors enter additively in the two-period case, we can express achievement in each period as

$$\begin{aligned} A_{i1} &= h_1(X_{i1}, S_{i1}) + \nu_i + \varepsilon_{i1} \qquad (10) \\ A_{i2} &= h_2(X_{i2}, X_{i1}, S_{i2}, S_{i1}) + t_2 + \nu_i + \varepsilon_{i2} \qquad (11) \end{aligned}$$

where h_1 and h_2 are unknown functions. To identify the structural parameters from this model we must assume that i) the unobserved components ν_i, ε_{i1} are independent of S_{i1} ; ii) $(\varepsilon_{i1}, \varepsilon_{i2})$ is independent of $(X_{i1}, S_{i1}, X_{i2}, S_{i2})$ and iii) t_2 is a constant. Under these assumptions, the structural parameters are functionals of conditional expectations of all observed variables.[19]

To estimate dynamic treatment effects our approach builds on Miquel (2003), who demonstrates that a conditional difference-in-differences approach of the achievement equations can nonparametrically identify the causal effects of sequences of interventions.[20] The full sequence of causal effects are estimated under simple dynamic variants of 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, we consider a sequential difference in difference estimator and use data between periods of the interventions to remove common trends between individuals on alternative sequences, which permits us to recover the full sequence of dynamic average treatment effect for the treated parameters.

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

$$A_{i1} = v_i + \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \qquad (12)$$

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

$$A_{i2} = v_i + \alpha_2' X_{i2} + \alpha_1' X_{i1} + \alpha_{S2}' S_{i2} + \alpha_{S1}' S_{i1} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + \varepsilon_{i2} \tag{13}$$

and t_2 reflects period two common shock effects. Since nearly all of the explanatory variables in equations (12) and (13) are discrete dummy variables the only restrictive assumption by linearization is the additive separability of the error term. Notice, we allow the effect of being in a small class in the first year (S_{i1}) on second period achievement (A_{ij2}) to interact in unknown ways with second year class assignment (S_{i2}). 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)) and one could imagine the effect of the current class size treatment to potentially differ due to past learning experiences as well as incoming knowledge or foundation. First differencing the achievement equations generates the following system of two equations

$$\begin{aligned} A_{i2} - A_{i1} &= \alpha_2' X_{i2} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + t_2 + (\alpha_1 - \beta_1)' X_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + \varepsilon_{i2}^* \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1}^* \end{aligned} \tag{14}$$

where $\varepsilon_{i2}^* = \varepsilon_{i2} - \varepsilon_{i1}$ and $\varepsilon_{i1}^* = v_i + \varepsilon_{i1}$. Consistent estimates of the structural parameters of the education production function in equations (12) and (13) 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. As this system is triangular, parameter estimates from full information maximum likelihood are equivalent to equation by equation OLS which does not impose any assumptions on the distribution of the residuals.[22] Consistent and unbiased structural estimates of β_{S1} and of the teacher characteristics in the X_{i1} matrix can still be obtained with STAR data without information on pre-kindergarten inputs since subjects and teachers were both randomized between class types in kindergarten and to the best of our knowledge compliance issues did not arise until the following year.[23] Since not all elements in the education production function in kindergarten are assigned randomly it remains possible that a subset of the structural parameter estimates for the X_{i1} matrix may not be unbiased since they may be correlated with ε^*_{i1}.

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.[24]

The full sequence of dynamic effects can be estimated as follows

$$\begin{aligned}
\tau^{(1,1)(0,0)}(1,1) &= \alpha_{S1} + \alpha_{S2} + \alpha_{S12} \qquad (15)\\
\tau^{(1,1)(1,0)}(1,1) &= \alpha_{S2} + \alpha_{S12}\\
\tau^{(0,1)(0,0)}(0,1) &= \alpha_{S2}
\end{aligned}$$

where, $\tau^{(1,1)(0,0)}(1,1)$ provides an estimate of the average cumulative 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. These parameters presented in equation (15) are of policy interest. It is straightforward to extend the above two period regression example to T periods.

While concerns regarding substitution bias and dropout bias are addressed by assuming the effects of individual unobserved heterogeneities which include factors such as parental concern over their child's development are fixed over short time periods, attrition bias may still contaminate the results. As discussed in the preceding section it is possible to reweight the data to account for attrition due to selection on observables. We consider estimating the following attrition logit

$$Pr(L_{it+1} = 0|A_{it}, S_{it}, X_{it}) = 1\{\alpha' Z_{it} + w_{it} \geq 0) \tag{16}$$

where t is the period being studied and Z_{it} is a matrix of predetermined variables that are observed conditional on $L_t = 0$ and also include lagged dependent variables (A_{t-s}) as well as past test scores in all other subject areas. The predicted probability of staying in the sample $(\overset{\curlywedge}{p}_{it})$ are then constructed

$$\overset{\curlywedge}{p}_{it} = F_w(\hat{\alpha}' Z_{it}) \tag{17}$$

where F_w is the logistic cumulative distribution function. This method of controlling for attrition is robust with respect to any treatment experience.

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. Following Wooldridge (2002) we calculate the probability of remaining in the sample for period two $\overset{\curlywedge}{p}_{i1}$and use it to reweight observations in estimating equation (14) as follows

$$\begin{aligned} \frac{A_{i2} - A_{i1}}{\overset{\curlywedge}{p}_{i1}} &= \frac{\alpha_2' X_{i2} + (\alpha_1 - \beta_1)' X_{i1} + \alpha_{S2}' S_{i2} + \alpha_{S12}' S_{i2} S_{i1} + (\alpha_{S1} - \beta_{S1})' S_{i1} + t_2 + \varepsilon_{i2}^*}{\overset{\curlywedge}{p}_{i1}} \qquad (18) \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \end{aligned}$$

Estimates from this system of equations are $\sqrt{N}$ consistent and asymptotically normal. However, the asymptotic variance is conservative since it ignores the fact that we are weighting on the estimated and not the actual $\overset{\curlywedge}{p}_{i1}$.[25]

We estimate equation (17) for grade one as well as corresponding versions for grade two and grade three with the kindergarten sample.[26] Attrition is an absorbing state and the weights used in estimation for grades two and three $(\overset{\curlywedge}{r}_i^2$ and $\overset{\curlywedge}{r}_i^3)$ are simply the product of all past estimated probabilities

$$\begin{aligned} \overset{\curlywedge}{r}_i^2 &= \overset{\curlywedge}{p}_{i2} * \overset{\curlywedge}{p}_{i1} \\ \overset{\curlywedge}{r}_i^3 &= \overset{\curlywedge}{p}_{i3} * \overset{\curlywedge}{p}_{i2} * \overset{\curlywedge}{p}_{i1} \end{aligned} \tag{19}$$

where $\overset{\curlywedge}{p}_{is}$ are estimated probabilities for staying in the sample for period s from a logit regression using all subjects in the sample at $s-1$.[27] 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.

4 Project STAR Data

Project STAR was a large scale experiment that initially randomized assigned over 7,000 students in 79 schools into one of the three intervention groups: small class (13 to 17 students per teacher), regular class (22 to 25 students per teacher), and regular-with-aide class (22 to 25 students with a full-time teacher's aide) as the students entered kindergarten.[28] Teachers were also randomly assigned to the classes they would teach. The experiment continued until the students were in grade three. The public access data on Project STAR contains information on teaching experience, the education level and race of the teacher, the gender, race and free lunch status of the student. In addition, during each year of the experiment academic performance measures were collected from the Reading, Mathematics and Word Recognition sections of the Stanford Achievement test.[29] In our analysis, 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.

The STAR data set contains the class types that students actually were enrolled in each year. While the possibility exists that some students were switched from their randomly assigned class to another class before kindergarten started, Krueger (1999) examined actual enrollment sheets that were compiled in the summer prior to the start of kindergarten for 1581 students from 18 participating STAR schools and found that only one single student in this sample who was assigned a regular or regular/aide class enrolled in a small class.

Summary statistics on the full Project STAR kindergarten sample are provided in Appendix Table 1. In kindergarten, nearly half of the sample is on free lunch status. There are very few Hispanic or Asian students and the sample is approximately $\frac{2}{3}$ Caucasian and $\frac{1}{3}$ African American. There are nearly twice as many students attending schools located in rural areas than either suburban or inner city areas. There are very few students in the sample (9.0%) attending schools located in urban areas. Regression analysis and specification tests found no evidence of any systematic differences between small and regular classes in any student or teacher characteristics in kindergarten, suggesting that randomization was indeed successful. However, among black students those on free lunch status were more likely to be assigned to regular classes than small classes (33.67% vs. 27.69%, $\Pr(T > t) = 0.0091$, one sided test).

Following the completion of Kindergarten there were significant non-random movements between control and treatment groups as well as in and out of the sample which complicates any analysis. As a result Appendix Table 1 indicates that the sample which completed tests each year is increasingly likely to be currently attending a small class, white or Asian, female and not on free lunch status. The full set of transitions for the cohort of students who participated in Project STAR in kindergarten is shown in Figure 1. This graph displays the multitude of transitions that were outlined in the preceding section. 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.

In our empirical 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.[30] Second, subsequent incoming groups are conditionally randomly assigned (based on seat availability/capacity constraint) within each school. The second claim can be examined through simple regressions of the random assignment indicator (M_{ijT}) on individual characteristics and school indicators as follows

$$M_{ijT} = \gamma' X_{ijT} + \upsilon_j + e_{ijT} \tag{20}$$

for each group of students entering the experiment after kindergarten. If students are assigned randomly there should be no evidence of a systematic differences in baseline characteristics (as well as unknown confounders) between the treatment and control group.

OLS estimates of equation (20) are presented in the top panel of Table 1. The results clearly demonstrate that incoming students were not conditionally randomly assigned in grades one and three. The incoming students to the experiment that were on free lunch status in grades one and three were more likely to be assigned to the control group. Estimates of equation (20) that use the full sample of students are presented in the bottom panel of Table 1. They further demonstrate significant differences in student characteristics between small and regular classes in each year following Kindergarten.

5 Empirical Results

5.1 Attrition and Bounds

Past research on Project STAR has treated the data assuming attrition is ignorable and considered ITT and IV analyses. As discussed in the introduction, estimates from these analyses are distorted from a causal effect if there is evidence of non-random attrition. Further even if data were missing at random, estimates from the ITT or IV analyses can be misleading for policymakers in that they are either at the very upper end of the bounds or sometimes even lie entirely outside the bounds for the ATE. In this sub-section we present strong evidence of non-random attrition. The results demonstrate that both the ITT and IV estimates tend to lie at the very upper end of the bounds and in some grades and subject areas the IV estimates are not contained within the bounds for the ATE.

To determine whether there is evidence of non-random attrition we present estimates of equation (8) in Table 2. Wald tests indicate that the β_L coefficient vector is significantly different for attritors and non-attritors in all subject areas.[31] The attrition indicator is significantly negatively related to test score performance in all subject areas indicating that the levels of performance for subsequent attritors is significantly lower in kindergarten. The joint effect of attrition on all student characteristics and class type is significantly different from zero in all subject areas. Students on free lunch status that left scored significantly lower than free lunch students who remained in the sample in mathematics. Interestingly, female attritors out performed female non-attritors in kindergarten in all subject areas but the magnitude is small. 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 selective attrition 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. As there is no evidence that attrition patterns differed between schools in Tennessee that participated and did not participate in the STAR experiment concerns regarding selection on unobservables are reduced.[32]

To assess the degree of bias introduced by noncompliance and selective attrition due to observables as well as selective attrition due to unobservabes, we consider two different techniques to construct bounds for the ATE in the presence of missing outcome data. The first approach follows Horowitz and Manski (2000) and uses information about the support of the outcome variable to construct "worst-case" bounds for the treatment effect parameter when outcome and covariate data are non-randomly missing. Horowitz and Manski (2000) construct bounds by imputing for those with missing information either the largest or smallest possible values to compute the largest and smallest possible treatment effects. These bounds are nonparametric and hold regardless of the process that led to the missing data.[33] The second approach introduced in Lee (2005) imposes some additional assumptions to gain tighter bounds on the ATE.[34] Lee (2005) trims the test score distribution using the differential probability of remaining in the sample between the treatment and control groups. Intuitively Lee's method calculates the upper (lower) bound of the ATE by taking the conditional expectation of the outcome variable where the bottom (top) trimming factor percentage of the data is dropped for the treatment group. Those observations that exit or violate the experimental protocol are assumed to exhibit an extreme form of sample selection by achieving outcomes in either the upper or lower tail of the test score distribution. In our implementation we use the differential probability of both remaining in the sample and continuing to follow the experimental protocol as the trimming factor for the Lee (2005) bounds.

Table 3 presents estimates of the ITT and LATE that account for attrition using inverse probability weighting as well as bounds on the average treatment effect. The ITT and LATE parameters presented in the first two rows are obtained from equations based on the empirical model where we include the history of school and home inputs but not the full history of treatment. To recover the ITT we include initial class type (M_i) and consider OLS estimation of the following equation

$$\frac{A_{ijT}}{\overset{\curlywedge}{p}_{iT}} = \frac{\beta'_T X_{ijT} + \beta'_{ITT} M_i + \sum_{t=1}^{T-1} \beta'_t X_{ijt} + \upsilon_j + \psi_{ijT}}{\overset{\curlywedge}{p}_{iT}} \tag{21}$$

where ψ_{ijT} is a composite error term. The regression accounts for attrition due to observables as $\overset{\curlywedge}{p}_{iT}$ are weights that are calculated using equations. The weights are calculated using estimates from equations (17) and (19). Similarly, estimates of the LATE are obtained from 2SLS estimation of

$$\frac{A_{ijT}}{\overset{\curlywedge}{p}_{iT}} = \frac{\beta'_T X_{ijT} + \beta'_{LATE} S_{it} + \sum_{t=1}^{T-1} \beta'_t X_{ijt} + \rho_j + \psi_{ijT}}{\overset{\curlywedge}{p}_{iT}} \tag{22}$$

where $E(S_{it}, \psi_{ijT}) \neq 0$ due to noncompliance with the treatment assignment. To recover the LATE parameter M_i is used as an instrumental variable for S_{it}.[35]

ITT and LATE estimates are presented in the first two rows of Table 3 and are positively and significantly related to academic achievement in all subject areas and grades. The LATE parameter presented in the second row is greater than the estimated ITT effect in the first row since the denominator of the LATE expression (equation (4)) in our data lies strictly between 0 and 1. Since the compliance rate decreases in higher grades, the denominator of this expression decreases and the LATE parameter becomes larger in magnitude relative to the ITT.[36] The third and fourth row present results from a series of DuMouchel and Duncan (1983) statistical tests that evaluate whether weighting for attrition is necessary. In the absence of sample selection bias due to attrition, unweighted estimates are preferred since they are more efficient than the weighted estimates. Test results suggest that accounting for attrition is necessary in all grades and subject areas.[37]

The fifth and sixth rows of Table 3 present Horowitz and Manski (2000) bounds. Since the support of the outcome variable is wide so too are the treatment effect bounds. The interval is almost as consistent with extremely large negative effects as it is with extremely large positive effects. In contrast Lee (2005) bounds presented in the seventh row of Table 3 are substantially tighter. These bounds suggest that the average treatment effect is clearly positive in grade one math but might be negative in higher grades and in other subjects. The ITT and LATE estimates presented in the first two rows generally lie near the upper part of the Lee bound in each subject and grade level. In grade 3 reading the point estimate of the LATE that accounts for attrition is greater than the Lee (2005) upper bound for the ATE. Past research using Project STAR data has either reported unweigthed ITT or LATE which are presented in the two bottom rows of the Table. Notice that the point estimate of the unweighted LATE exceeds the estimate of Lee (2005) upper bound for the ATE four out of the nine cases; which respectively are in grade 2 mathematics, grade 3 word recognition, and both grades 1 and 2 reading tests. This demonstrates that estimates based on intent-to-treat analysis and instrumental variables tend to lie near the extremes and occasionally outside the theoretical bounds for the ATE.

While the estimates presented in the first seven rows of Table 3 are robust to non-compliance with the treatment assignment and non-random attrition, they ignore the multi-period nature of the STAR experiment. We next consider direct estimation of our empirical model to shed light on the effectiveness of alternative sequences of class size reductions during the experiment.

5.2 Dynamic Treatment Effect Estimates

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 average 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. After grade one, no significantly positive effect of small class exists ($P(t) \leq 10\%$) except 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 achievement in all subjects.

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 15. A maximum of 1, 6, 28 and 120 effects can be calculated for kindergarten, grades 1, 2 and 3 respectively. However, due to lack of support of some treatment paths only 78 effects can be calculated for grade 3. We present evidence that compares 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 richer picture of the structure and source of the gains in small class reductions. A significant impact from smaller classes appears in kindergarten. Following kindergarten, the positive effects of smaller classes in grade one appear only for those students who made a transition between class types. Students who substituted into small classes and dropped out of small classes both scored significantly lower than their grade one classmates in each kindergarten subject and received a significantly larger improvement in grade one achievement compared to their grade one classmates.[38] It is possible that teachers were targeting the weaker students in the class. Further, these growth rates were significantly larger than those achieved by their kindergarten classmates who did not switch in grade one. Since scaled scores are developmental and can be used to measure growth across grades within the same test subject area we can conduct these comparisons.[39]

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.[40] 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.[41]

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 $\tau^{(0,1)(0,0)}(0,1)$.

Similar to Krueger (1999) we find that students received large benefits the first year they spent in a small class in all subject areas in grade one and in grade two mathematics. In contrast to Krueger (1999), we find in all that students who first entered small classes in grade three achieved significant losses from attending a small class ($\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$) in all subject areas. Finally, students who first switched in to small class in grade two did not have statistically significant gains on reading and word recognition ($\tau^{(0,0,1)(0,0,0)}(0,0,1)$).

5.3 Discussion

The changes in the sign and significance of 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. Using classroom level regressions 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 in grades one and two, students who joined small classes scored significantly lower than their new classmates with the exception of reading for those who substituted in grade two. Only in mathematics did these students receive significantly greater growth in performance between grades for each period. Students who achieved benefits from attending small classes for the first time had significantly lower past performance in math.

Coleman (1992) suggests that the focus of US education is on the bottom of the distribution and it is much easier for teachers to identify weaker students in mathematics than other subject areas. The major challenge in investigating this claim is separating the amount of test score gains from teachers' behavior from a statistical tendency called "regression to the mean," which is created by nonrandom error in the test scores. This error leads students to score poorly at one point in time and subsequently receive scores that come closer to the average for the whole population. To investigate this issue we classified the five students in each grade one classroom that had the lowest scores on kindergarten tests in each as being a "weak" student. We included an indicator variable for being one of these "weak" students in the classroom in regression equations to explain growth in performance controlling for the full history of teacher, family and student characteristics. Using multiple regression we separately examined whether being a "weak" student in mathematics or reading or word recognition led to larger gains in test performance in all subject areas.

Consistent with the regression to the mean argument students who were weak in mathematics and word recognition received larger gains in performance relative to classmates in these subject areas. In contrast, being a "weak" student in reading significantly reduced gains in reading performance in grade 1. Supporting Coleman's hypothesis we found that students who achieved the largest gains in the classroom in reading and word recognition in higher grades were defined as "weak" students in mathematics.[42] Further, we find among "weak" students in mathematics those who substituted from regular classes to small classes received larger gains in performance in all subject areas relative to their former "weak" classmates who remained in regular classes. In general, individuals who substituted in grade one were "weak" in mathematics whereas those students who substituted in grade three were not "weak" in mathematics which may explain why $(\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1))$ is negative. Individuals who substituted in grade three did not differ significantly from their grade two classmates on their grade two performance in all subject areas.

Students who were classified as "weak" that substituted in to small classes increased the variation of background subject knowledge within small classrooms in higher grades. In higher grades, small classes had significantly more variation in past performance in mathematics and reading than regular classes.[43] Faced with relatively less variation in the incoming knowledge of students, simple 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.[44] As regular classes became less heterogeneous in knowledge the dynamic benefits of small class attendance vanished. Between class types in grades two and three there was neither any significant differences in the variation of prior performance nor significant differences in gains in performance on the word recognition examinations. While the patterns exhibited in higher grade might be explained by the existence of a trade-off between variation in incoming student performance and class type, more investigation is needed to directly test this hypothesis.[45]

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 conducted a more detailed examination of the effects of environmental stability on small classes in grade one.[46] In each grade one small class, we identified members of the largest subgroup of students who were taught by the same teacher in kindergarten. OLS regressions examining the impact of this variable on gains in performance that controlled for school indicators and the full history of student and teacher characteristics were undertaken. The results found that members of this largest subgroup had significantly smaller gains relative to their classmates in mathematics (coeff.=-6.129, s.e. 2.714) and word recognition (coeff.=-4.524, s.e. 3.008) and no significant differences in readings. Multiple regressions using the number of current classmates who were also taught with the same kindergarten teacher (instead of a simple indicator variable) also find significantly smaller gains in mathematics (coeff.=-1.797, s.e. 0.572) and word recognition (coeff.=-1.179, s.e. 0.572) for each additional former classmate. These results do not support arguments for environmental stability. Neither do they directly contradict the stability hypothesis since peer groups (classmates) were no longer exogenously formed after kindergarten.

5.4 Specification Tests

This study differs from past research on Project STAR not solely through the focus of treating the experiment as a multi-period intervention but also in accounting for both attrition due to observables and the possibility that other forms of noncompliance are due to unobservables. DuMouchel and Duncan (1983) test were used to examine the importance of accounting for attrition due to observables when estimating equation (14). Test results presented in Table 6 demonstrate that accounting for attrition due to observables is preferred in all subject areas and grade levels at conventional levels in reading and mathematics and below the 20% level in word recognition.

Assuming there does not exist selection on unobservables permits direct estimation of the structural equations (12) and (13).[47] 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 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 (14).

5.5 Robustness Checks

To check the robustness of our results we consider three strategies that increase the statistical power of the structural parameter and dynamic treatment effect estimates. Specifically we i) ignore potential nonlinear impacts of the treatments in equation (18), ii) relax the identification assumptions for the attrition model allowing us to use a larger sample, and iii) present ITT estimates for the subset of subjects who complied with their assignment throughout the study.

Since there are limited number of people on several treatment paths in Figure 1 we reestimate equation (18) removing all the non-linear impacts of class size treatment. Specifically, in period one and two we estimate

$$\begin{aligned} \frac{A_{i2} - A_{i1}}{\overset{\curlywedge}{p}_{i1}} &= \frac{\alpha_2' X_{i2} + (\alpha_1 - \beta_1)' X_{i1} + \alpha_{S2}' S_{i2} + (\alpha_{S1} - \beta_{S1})' S_{i1} + t_2 + \varepsilon_{i2}^*}{\overset{\curlywedge}{p}_{i1}} \qquad (23) \\ A_{i1} &= \beta_1' X_{i1} + \beta_{S1}' S_{i1} + \varepsilon_{i1} \end{aligned}$$

with the identical sample as in Tables 4 and 5. This model is less flexible than equation (18), and implicitly places several equality restrictions on several dynamic treatment effect paths. For example, in grade two $\tau^{(0,1,1)(0,0,1)}(0,1,1) = \tau^{(0,1,0)(0,0,0)}(0,1,0)$. Yet, it is worth stating explicitly that no additional assumptions are imposed on the underlying data to construct dynamic treatment effects when equation (23) is used in place of equation (18). The underlying empirical model differs and it is possible to construct F tests on the joint significance of the non-linear interactions of treatment receipt. The results support the use of the restricted model (equation (23)) in grade one, but the unrestricted model is preferred in four of the six specifications in higher grades.

Structural parameter and dynamic treatment effect estimates from equation (23) are jointly presented in Table 8. The qualitative picture that emerges from these results is fairly similar to that which emerged in Tables 4 and 5. In grade one, the impacts of S_{i1} and S_{ik} are smaller in magnitude as they now capture part of the negative impact of the omitted $S_{iK}S_{i1}$. An additional year of treatment appears to boost achievement in math and word recognition whereas having attended small classes in both years relative to having only been in a small class in grade one increases achievement in reading only. In grade two, we continue to find that new switchers only have achievement gains in mathematics. Similarly, in grade three, we continue to find that first time small class attendance is not positively related to achievement in any subject area and is now statistically insignificant. As before, we find that nearly every path of multiple receipt of treatment in the higher grades is not significantly related to achievement in any subject area. The results suggest that there may be lasting effects from attending a small class in kindergarten alone in reading and word recognition. For mathematics, the results appear to suggest that small class attendance in both kindergarten and grade two may have some lasting impacts.

Our second robustness check involved estimating equation (18) where the weights are constructed from an attrition model that imposes weaker data requirements using only the most recent lagged test score in that subject area for identification. In each attrition model, the lagged dependent variable was significantly associated with remaining in the sample which continues to indicate that selection on observables is not ignorable. We present weighted structural parameter estimates in Table 9.[48]

There are a few minor differences between the samples in the structural parameters. For example, in grade one, the combined effect of being in small classes both years is significantly negative in both mathematics and word recognition. This weakens evidence on a positive impact of small class attendance in grade one. 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 into 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 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. In higher grades, kindergarten small class attendance (S_{iK}) is positively related to performance in grade two reading and grade three reading and word recognition examinations. Whereas, attendance in small classes in grade one (S_{i1}) is either negatively related or unrelated to performance in both grades two and three.

Third, we consider a naive approach and assume that all noncompliance is random. We consider ITT estimation of equation (21) using only the subsample of the data that complied with their initial assignment in all prior years of the experiment. Estimates that account for non-random attrition using weights are presented in Table 10. We would anticipate that these estimates would be biased upwards since those individuals that would stay in the study are those who are receiving the largest benefits. Notice that the weighted estimates are statistically insignificant in all subject areas in both grades one and two. When accounting for the probability of remaining in the sample, having always been in small classes throughout the study leads to a positive and significant estimate in grade three math and word recognition. Overall, the quantitative results in Table 10 are similar to the dynamic average treatment effect for the treated estimates that compares the path of always being in a small class versus never attending a small class. These results increase our confidence that receiving multiple doses of small class treatment does not yield substantial impacts relative to having never attended a small class.

The results clearly suggest that the differences in our findings from earlier work are unlikely due to statistical power. Overall, these results suggest that the benefits of attending a small class early are of small magnitude and a single dose in kindergarten yields most of the benefit. The results do not provide strong evidence supporting long-term large scale class size reductions. The substantial heterogeneity in the treatment effects makes it important to understand the reason why small classes work when they are effective, and similarly understand the explanations for their failures.[49] Comparing the alternative treatment sequences sheds light on some of these circumstances but clearly more research is needed. For example, more understanding of the nature of class size and relationship with teaching practices is needed. To summarize the results suggest that small classes do not work consistently and unconditionally.

6 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 nor non-experimental estimators recover parameters of interest to policy makers, particularly if there is non-ignorable selective attrition. We introduce a strategy to estimate treatment effects in this setting and apply it 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 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 noncompliance as random whereas we find strong evidence for selection due to individual unobserved heterogeneity. We demonstrate that even if one accounts for attrition, ITT and IV estimates recover treatment effects that are outside of the bounds for the average treatment effect. Finally, estimates of conditional random assignment demonstrates 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 has some lasting impacts but there are no lasting benefits from either receiving multiple doses of treatment or receiving the treatment beyond grade one. 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 new evidence to one of the hotly debated education policy areas several questions remain. For example, 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, teacher expectations and family involvement have been collected from process evaluations as well as surveys completed by the teachers. However, the original STAR researchers have not made this information available to the general research community. Answers to this and other questions present an agenda for future research.

Notes

[1] The benefits of randomized field experiments have been widely known since the publication in 1935 of Fisher's seminal book, The Design of Experiments.

[2] The term "broken" refers to violations to the randomization protocol. Comprehensive surveys of recent developments in the economics, biostatistics and statistics literature can be found in Heckman, LaLonde and Smith (2001), Yau and Little (2001) and Frangakaris and Rubin (2002) respectively.

[3] 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.

[4] 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 aiming to reduce K-3 class sizes by roughly ten students per class and has spent more than $10 billion on this categorical program between 1996 and 2003. Brewer et al. (1999) estimate that the annual cost of these reductions ranges from 2 to 11 billion dollars in the US. The reported positive results have influenced education policies in other countries such as Canada where the government of Ontario provided school boards with $1.2 billion over three years beginning in 1997 to reduce class sizes.

[5] Krueger (1999) presents instrumental variable estimates to correct for biases related to deviations from assigned class type.

[6] It obtains this causal interpretation provided a series of assumptions detailed in Angrist, Imbens and Rubin (1996a) and listed in footnote 14 are satisfied.

[7] Todd and Wolpin (2003) present a clear discussion of why estimates from these empirical approaches should differ.

[8] While marginal structural models are the predominant method used to analyze multiperiod experiments in epidemiology (e.g. Robins (1999) or Robins et al. (2000)), estimation of causal effects from these models require that selection is only due to observables.

[9] 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.

[10] Students were added to the sample in later years because either kindergarten was not mandatory and they first entered school in grade 1, or had previously failed their grade and had to repeat it, or switched from a private school or recently moved to the school district that contained a participating school.

[11] Included among these individuals are 68 students who were moved to regular classes in grade 1 after being termed incompatible (Finn and Achilles (1990)) with their classmates in Kindergarten. Eighteen of these students returned to small classes after grade 1.

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

[13] We compare small classes versus regular or regular with aide classes. This follows Finn et al. (2001) who create a single control group and report that there are no significant differences in outcomes between regular class with and without teacher aides. Note, as many schools contained multiple classes of the same class type, there is even more transitions between classes of the same class type as well as switches between regular classes with and without teacher aides.

[14] The assumptions include 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 treatment 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 a causal effect.

[15] In other words, these individuals were induced to switch classes by the instrument (complied with initial assignment).

[16] 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.

[17] Fitzgerald, Gottschalk and Moffitt (1998) provide a framework for the analysis of attrition bias and describe specification tests to detect and methods to adjust estimates in its presence.

[18] For the full kindergarten sample, a linear probability model regression of subsequent attrition on initial class assignment yields a statistically significant impact of class type. The pattern of attrition differed substantially across schools. For example, students initially assigned to small classes in kindergarten were significantly less likely to leave the sample if their small classes outperformed the regular classes in all three subjects areas.

[19] Chesher (2003) demonstrates that a local insensitivity assumption is all that is needed to achieve local identification of the partial derivatives of structural functions in such a triangular system of equations.

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

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

treatment on outcomes at any point in time prior to actual participation. The extension to multi-period is not complex as described in Miquel (2003).

[22] 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. For example, efficiency gains are possible using the GMM procedure proposed in Hausman, Newey and Taylor (1987).

[23] The importance of randomization and the fact that compliance was near perfect in kindergarten (this evidence is discussed in the next section) is crucial to our identification strategy.

[24] We tested and found support for restricting the effect of individual unobserved heterogeneity to equal one between periods in grades two and three using a instrumental variables procedure developed in Ding and Lehrer (2004). Note our assumption is not only supported by the data but places weaker restrictions on the effects of unobserved inputs as compared to treating them as permanent unobserved heterogeneity

[25] The asymptotic variance matrix that adjusts for first stages estimates is smaller. See Wooldridge (2002) for details and a discussion of alternative estimation strategies. Note, the full set of results where the asymptotic covariance matrix of the second step estimator is computed using the results of Newey (1984) to take into account that generated regressors are used is available by request.

[26] To be specific on identification of $\hat{p}_{i1}$in equation (18) consider the example of grade one mathematics. Kindergarten reading and word recognition test scores are the sole kindergrten variables in the attrition equation that are not included in the achievement equation. For grade two mathematics all kindergarten test scores as well as grade one reading and word recognition test scores are included in the equations used to estimate weights and excluded form the achievement equation.

[27] 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.

[28] Students were randomly assigned using a random school specific starting value based on their last name using a centrally prepared algorithm that assigned every kth student from an alphabetical list to a class type.

[29] 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. 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. This allows us to use differences in scales scores as measures to track development betwen grades. As robustness checks we replicated the analysis with both percentile and standard scores (which do not possess this property) and the basic patterns reported in Section 5 also emerge.

[30] In Section 5, we present evidence of a selective attrition pattern which shows the data does not support this claim.

[31] Fitzgerald et al. (1998) demonstrate that this test is simply the inverse of examining whether past academic performance significantly affects the probability of remaining in the study in higher grades from logistic estimates of equation (16). The results presented in Appendix Table 2 suggest that students who scored higher on their most recent mathematics examination are more likely to remain in the sample at each grade level, further demonstrating that attrition due to observables is not ignorable.

[32] STAR researchers collected information on students in similar schools that did not participate in the experiment. This data has not been made available to the authors or the general research community.

[33] A larger literature has emerged that constructs bounds that allows for imperfect compliance of the treatment (Balke and Pearl 1997; Heckman and Vytlacil 1999, 2000 and Manski 1990, 1995) but assumes data is missing at random. Note Scharfstein, Manski and Anthony (2004) discuss an alternative method to construct bounds under specific

assumptions of the missing data mechanism. Some of these assumptions are similar to the monotonicity assumption that underlies Lee (2005) bounds.

[34] The critical additional assumption of Lee (2005) is monotonicity and is equivalent to that which underlies LATE estimates.

[35] It is worth noting that the use of initial assignment from randomization as an instrument for subsequent class size violates the exclusion restriction in the cumulative education model since implicitly earlier class size is an input to the production process of current academic performance. Angrist Imbens and Rubin (1996b pg. 468) note that randomization alone does not make a candidate instrument valid because randomization alone does not make the exclusion restriction more plausible.

[36] The decreasing compliance rate is the primary reason for the divergence of findings debated in Krueger (1999) and Hanushek (1999) and would exist irrespective of the method used to measure test scores (Finn and Achilles, 1999).

[37] For completeness, the unweighted ITT and LATE parameters are reported in the bottom two rows of Table 3. In general, the unweighted estimates exceed the weighted estimates as expected since selective attrition should reduce observed differences by test score performance leading to upward-biased estimates. However, the differences between the two sets of estimates appear minor.

[38] These findings are obtained from within classroom regressions that control for kindergarten and grade 1 student, family and teacher characteristics.

[39] 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. Other test score measures such as percentile scores, grade equivalent scores or standard scores are difficult to interpret since they are not comparable between grades even in the same subject area.

[40] 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 attendees gained more than always attendees 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.

[41] The regressions include school indicators as well as student and teacher characteristics. The regressor of 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 histories 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 attending a small class in reading and word recognition which explains the dynamic benefits at that time.

[42] Our results are robust to several alternative definitions of being a "weak" student. We also defined being a "weak" student as having the lowest or one of the three or four lowest scores in the classroom.

[43] 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 = 0.04) and three (P < F = 0.11) and in grade two reading (P < F= 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.

[44] 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 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.

[45] 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. Attempts at peer effect estimation with this data can be found in Boozer and Cacciola (2001) and Graham (2005) who each find evidence of large impacts. Note our findings are consistent with evidence on elementary school students presented in Hoxby (2000a) and Hoxby (2000b) who exploited natural variation in age cohorts in the population and found evidence that class size does not affect student achievement in Connecticut and peer group composition affects achievement in Texas respectively. Further, international evidence from the TIMMS study finds that Korea (where students are ability streamed in the classroom) was the only country to significantly outperform the US in both grade 4 science and mathematics.

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

[47] 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. DuMouchel and Duncan (1983) tests confirm that weighted estimates are preferred for these direct estimates of the structural equations.

[48] DuMouchel and Duncan (1983) tests suggest that weighting is also preferred with this sample.

[49] Krueger (1999) presents evidence of heterogeneous treatment impacts across schools. We find that only 25% of schools achieved significantly positive results from small classes in kindergarten.

References

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

[2] Angrist, Joshua D.; Imbens, Guido W. and Rubin, Donald B. "Rejoinder to Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association,* 1996b, 91(434), pp. 468-72.

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

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

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

[6] Boozer, Michael A. and Cacciola, Stephen. "Inside the 'Black Box' of Project STAR: Estimation of Peer Effects Using Experimental Data." Working Paper, Yale University, 2001.

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

[8] Brewer, Dominic J.; Krop, Cathy; Gill, Brian and Reichardt, Robert. "Estimating the Cost of National Class Size Reductions Under Different Policy Alternatives." *Educational Evaluation and Policy Analysis,* 1999, 21(2), pp. 179-192.

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

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

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

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

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

[14] Finn, Jeremy D. and Achilles, Charles M. “Answers about Questions about Class Size: A Statewide Experiment.” *American Educational Research Journal*, 1990, 27, pp. 557-77.

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

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

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

[18] Graham, Bryan S. “Identifying Social Interactions Through Excess Variance Contrasts.” Working Paper, University of California-Berkeley, 2005.

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

[20] Hausman, Jerry A.; Newey, Whitney K. and Taylor, William E. “Efficient Estimation and Identification of Simultaneous Equation Models with Covariance Restrictions." *Econometrica*, 1987, 55(4), pp. 849-74.

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

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

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

[24] Heckman, James J. and Vytlacil, Edward J. "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects." *Proceedings of the National Academy of Sciences,* 1999, 96, pp. 4730-4734.

[25] Heckman, James J. and Vytlacil, Edward J. "Local Instrumental Variables." *NBER Working Paper No. T0252,* 2000.

[26] Horowitz, Joel and Manski, Charles F. "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data." *Journal of the American Statistical Association,* 2000, 95, pp. 77-84.

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

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

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

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

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

[32] Lee, David S. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *NBER Working Paper No. 11721*, 2005.

[33] Manski, Charles F. "Nonparametric Bounds on Treatment Effects." *American Economic Review,* 1990, 80(2), pp. 319-323.

[34] Manski, Charles F. "Identification Problems in the Social Sciences." Cambridge, MA: Harvard University Press, 1995.

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

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

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

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

[39] Robins James M., “Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference.” *Statistical Models in Epidemiology: The Environment and Clinical Trials,* :ed by E. Halloran and D. Berry, pp. 95–134, New York: Springer Verlag, 1999.

[40] Robins, James M.; Greenland, Sander and Fu-Chang, Hu, “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*, 1999, 94(447), pp. 687-700.

[41] Robins, James M. “A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods - Application to Control of the Healthy Worker Survivor Effect,” *Mathematical Modelling*, 1986, 7, pp. 1393-1512, with 1987 Errata, *Computers and Mathematics with Applications,* 14, pp. 917-21; 1987 Addendum, *Computers and Mathematics with Applications*, 14, pp. 923-45; and 1987 Errata to Addendum, *Computers and Mathematics with Applications*, 18, pp. 477.

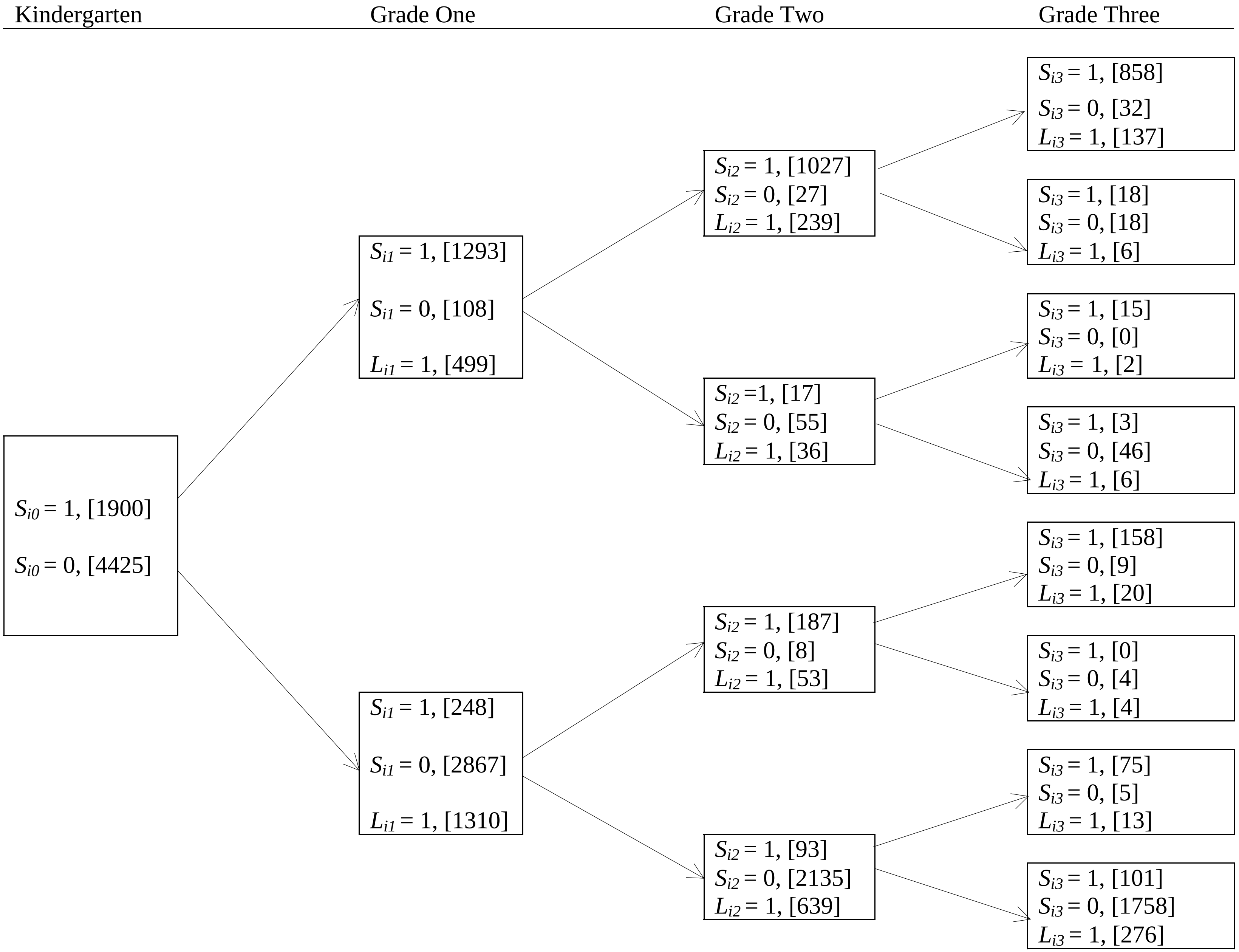
[42] Scharfstein, Daniel O.; Manski, Charles F. and Anthony, James C. “On the Construction of Bounds in Prospective Studies with Missing Ordinal Outcomes: Application to the Good Behavior Game Trial.” *Biometrics*, 2004, 60, pp. 154-164.

[43] Todd, Petra E. and Wolpin, Kenneth I. “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *Economic Journal*, 2003, 113, F3-F33.

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

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

Figure 1: Transitions During Project Star for Kindergarten Cohort



Note: Number or individuals are in [ ] parentheses.

Table 1: Testing Randomization of Student Characteristics across Class Types

	Kindergarten	Grade One	Grade Two	Grade Three
INCOMING STUDENTS				
White or Asian Student	2.35*10E-4 (0.012)	-0.275* (0.193)	-0.061* (0.041)	7.63*10E-4 (0.063)
Female Student	0.012 (0.019)	0.199* (0.126)	-0.020 (0.021)	-0.017 (0.028)
Student on Free lunch	-8.74*10E-3 (0.017)	-0.262* (0.167)	0.013 (0.022)	-0.057* (0.037)
Joint Test of Student Characteristics	0.29 [0.831]	1.83* [0.150]	1.24 [0.301]	1.01 [0.392]
Number of Observations	6300	2211	1511	1181
R Squared	0.318	0.360	0.248	0.411
FULL SAMPLE				
White or Asian Student	2.35*10E-4 (0.012)	-0.003 (0.021)	-0.008 (0.025)	-0.021 (0.027)
Female Student	0.012 (0.019)	0.007 (0.009)	0.004 (0.009)	0.008 (0.009)
Student on Free lunch	-8.74*10E-3 (0.017)	-0.038*** (0.016)	-0.030** (0.016)	-0.044*** (0.016)
Joint Test of Student Characteristics	0.29 [0.831]	2.05* [0.114]	1.38 [0.255]	2.98*** [0.037]
Number of Observations	6300	6623	6415	6500
R Squared	0.318	0.305	0.328	0359

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

Table 2: Are Attritors Different from Non-attritors?

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

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

Table 3: Traditional Single Period Causal Estimates of The Impacts of Reduced Class Size

Method	Mathematics			Reading			Word Comprehension		
	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3	Grade 1	Grade 2	Grade 3
ITT Analysis	9.023 (1.946)	5.389 (2.48)	4.329 (2.048)	10.409 (2.288)	4.698 (2.316)	7.189 (1.917)	9.036 (2.377)	4.622 (2.314)	8.127 (2.45)
IV Analysis (LATE)	10.353 (2.231)	6.417 (2.953)	5.781 (2.035)	11.948 (2.625)	5.602 (2.536)	9.584 (2.633)	10.296 (2.696)	5.513 (2.748)	10.862 (3.351)
Attrition Test for ITT	61.60 [0.00]	79.89 [0.00]	86.12 [0.00]	62.21 [0.00]	89.73 [0.00]	78.22 [0.00]	36.55 [0.00]	73.2 1 [0.00]	63.19 [0.00]
Attrition Test for IV	26.93 [0.00]	22.43 [0.00]	22.11 [0.00]	22.21 [0.00]	25.69 [0.00]	18.90 [0.00]	14.22 [0.00]	18.93 [0.00]	13.41 [0.00]
Horowitz-Manski Bounds	{-73.174, 86.647}	{-111.487, 117.794	{-144.049, 149.045}	{-67.039, 81.720}	{-114.397, 120.315}	{-130.950, 135.940}	{-102.133, 119.564}	{-97.005, 105.765}	{-112.028, 118.986}
Horowitz-Manski Bounds with school covariates	{-93.935, 94.805}	{-87.347 87.861}	{-97.246, 98.414}	{-82.452, 83.447}	{-89.447, 90.532}	{-87.716, 88.713}	{-98.434, 98.632}	{-78.441, 78.359}	{-77.924, 79.558}
Lee Bounds with no covariates	{5.844, 11.126}	{-3.010 6.502}	{-3.296, 6.005}	{5.443, 12.152}	{-2.575, 6.387}	{-0.834, 7.627}	{5.471, 11.407}	{-1.097, 7.110}	{-2.523, 8.547}
ITT Ignoring Selective Attrition	9.297 (1.894)	5.554 (2.070)	4.034 (1.637)	10.659 (2.141)	5.488 (1.86)	5.437 (1.545)	9.737 (2.283)	5.649 (2.004)	6.436 (1.922)
IV Ignoring Selective Attrition	10.852 (1.449)	6.707 (1.854)	5.642 (2.035)	12.45 (1.87)	6.783 (1.837)	7.56 (1.959)	11.323 (1.951)	6.918 (2.081)	9.004 (2.352)

Note: All of the ITT and IV estimates are statistically significant at the 5% level. The IV and ITT analyses include the full history of teacher inputs, free lunch status, race, gender and school indicators. Standard errors corrected at the classroom level are in parentheses. For the specification tests the Probability that the Null is rejected is contained in [] brackets. We present {lower bound, upper bound} in {} brackets for both the Horowitz and Manski bounds on the ATE as well as the Lee bounds analysis.

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. ***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

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%, 10% and 20% 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 Parameters and Dynamic Treatment Effect Estimates From an Education Production Functions That Ignores Non-Linear Treatment Effects

Subject Area	Mathematics	Reading	Word Recognition
Grade One			
S_{iK}	4.174 (2.829)*	9.351 (2.805)***	5.434 (3.250)**
S_{i1}	6.608 (2.488)***	2.779 (2.582)	6.415 (3.016)***
Grade Two			
S_{iK}	6.191 (4.034)*	10.479 (4.34) ***	6.035 (4.659)*
S_{i1}	-8.916 (5.191)**	-6.529 (5.949)	0.742 (5.784)
S_{i2}	12.805 (4.152)***	5.730 (4.659)	4.114 (4.138)
Grade Three			
S_{iK}	0.131 (5.286)	8.885 (5.088)**	12.057 (5.940)***
S_{i1}	-1.168 (7.588)	-0.057 (7.500)	-5.097 (8.118)
S_{i2}	11.747 (7.162)*	3.152 (6.784)	11.079 (7.655)*
S_{i3}	-2.596 (3.717)	-1.370 (3.244)	-6.679 (4.691)*
DYNAMIC TREATMENT EFFECTS			
Grade One			
$\tau^{(1,1)(0,0)}(1,1)$	10.782 (3.767)***	11.933 (3.81)***	11.849 (4.434)***
$\tau^{(1,1)(1,0)}(1,1)$	6.608 (2.488)***	2.779 (2.582)	6.415 (3.016)***
$\tau^{(1,1)(0,1)}(1,1)$	4.174 (2.829)*	9.351 (2.805)***	5.434 (3.250)**
$\tau^{(0,1)(1,0)}(0,1)$	2.434 (3.767)	-6.572 (3.813)**	0.981 (4.434)
Grade Two			
$\tau^{(0,0,1)(0,0,0)}(0,0,1)$	12.805 (4.152)***	5.730 (4.659)	4.114 (4.138)
$\tau^{(1,0,0)(0,0,0)}(1,0,0)$	6.191 (4.034)*	10.479 (4.34) ***	6.035 (4.659)*
$\tau^{(1,0,0)(0,1,0)}(1,0,0)$	15.107 (6.574)**	8.942 (7.364)	17.154 (7.427)***
$\tau^{(1,1,1)(0,0,0)}(1,1,1)$	10.080 (7.776)*	9.680 (8.714)	10.891 (8.502)*
$\tau^{(1,1,1)(1,0,0)}(1,1,1)$	3.889 (6.647)	-0.799 (7.556)	3.372 (7.112)
$\tau^{(0,0,1)(0,1,0)}(0,0,1)$	21.721 (6.647)**	12.259 (7.556)*	4.856 (7.112)
$\tau^{(0,0,1)(1,0,0)}(0,0,1)$	6.614 (5.789)	-4.749 (6.367)	-1.921 (6.231)
Grade Three			
$\tau^{(0,0,0,1)(0,0,0,0)}(0,0,0,1)$	-2.596 (3.717)	-1.370 (3.244)	-6.679 (4.691)*
$\tau^{(1,0,0,0)(0,0,0,0)}(1,0,0,0)$	0.131 (5.286)	8.885 (5.088)**	12.057 (5.940)***
$\tau^{(1,1,1,1)(0,0,0,0)}(1,1,1,1)$	8.114 (12.273)	10.580 (11.776)	11.360 (13.483)
$\tau^{(1,1,1,1)(1,1,0,0)}(1,1,1,1)$	9.151 (8.069)	1.782 (7.520)	4.400 (8.798)
$\tau^{(1,1,1,1)(1,1,1,0)}(1,1,1,1)$	-2.596 (3.717)	-1.370 (3.244)	-6.679 (4.691)*
$\tau^{(1,1,1,1)(0,1,1,1)}(1,1,1,1)$	0.131 (5.286)	8.885 (5.088)**	12.057 (5.940)***
$\tau^{(0,1,1,1)(0,0,0,0)}(0,1,1,1)$	7.983 (11.076)	1.695 (10.621)	-0.697 (12.104)
$\tau^{(1,1,1,1)(1,0,0,0)}(1,1,1,1)$	7.983 (11.076)	1.695 (10.621)	-0.697 (12.104)
$\tau^{(0,0,1,0)(,0,0,0,0)}(0,0,1,0)$	11.747 (7.162)*	3.152 (6.784)	11.079 (7.655)*
$\tau^{(1,0,1,0)(0,0,0,0)}(1,0,1,0)$	11.878 (6.426)**	12.037 (6.03)***	23.129 (7.570)***
$\tau^{(1,0,0,0)(0,0,1,0)}(1,0,0,0)$	-11.616 (6.426)**	5.733 (6.034)	0.971 (7.570)

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

Table 9: 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. ***,**,* indicate statistical significance at the 5%, 10% and 20% level respectively.

Table 10: OLS Estimates of the Cumulative Impact of Treatment for Subjects who Always Complied with Their Kindergarten Assignment

	Reading	Math	Word
Weighted Estimates			
Kindergarten	5.950***	8.595***	6.342***
	(1.280	(2.025)	(1.415)
Grade One	-11.642	-4.655	-16.51
	(14.768)	(11.692)	(16.263)
Grade Two	2.131	10.200*	-1.095
	(6.533)	(6.844)	(7.552)
Grade Three	15.231***	13.134*	5.714
	(7.322)	(7.882)	(8.913)
Unweighted Estimates			
Grade One	-0.044	-3.576*	-1.770
	(2.97)	(2.369)	(3.426)
Grade Two	6.032***	7.498***	5.969***
	(2.258)	(2.465)	(2.497)
Grade Three	4.626**	2.438	5.868**
	(2.624)	(2.902)	(3.366)

Note: Each coefficient is from a regression that includes school effects, the full history of demographic and teacher characteristics. Standard errors corrected at the classroom level are in parentheses. ***,**,* indicate statistical significance at the 5%, 10%, and 20% level respectively.

Appendix Table 1: Descriptive Statistics of the Sample that Participated in Project STAR during Kindergarten

	Kindergarten	Grade One	Grade Two	Grade 3
Class Size	20.299 (3.959)	20.386 (3.994)	20.279 (4.194)	20.400 (4.441)
Currently Receiving Small Class Treatment	0.300 (0.458)	0.347 (0.476)	0.371 (0.483)	0.396 (0.489)
Math Test Score	485.610 (47.732)	536.544 (43.929)	590.571 (44.822)	627.977 (40.181)
Reading Test Score	436.734 (31.731)	529.073 (56.694)	594.846 (45.240)	625.634 (37.125)
Word Recognition Test Score	434.375 (36.799)	521.050 (53.027)	595.653 (49.374)	622.771 (43.932)
Free Lunch Status	0.483 (0.499)	0.444 (0.496)	0.388 (0.487)	0.353 (0.478)
Student is White of Asian	0.677 (0.468)	0.702 (0.457)	0.726 (0.446)	0.753 (0.432)
Student is Female	0.486 (0.500)	0.500 (0.500)	0.514 (0.500)	0.518 (0.500)
Teacher Race is Non-White	0.159 (0.366)	0.160 (0.367)	0.188 (0.390)	0.165 (0.372)
Teacher has a Masters Degree	0.353 (0.478)	0.345 (0.475)	0.357 (0.479)	0.443 (0.497)
Teacher Years of Experience	9.624 (5.497)	11.838 (8.795)	14.053 (8.567)	13.547 (8.471)
School is In Inner City Area	0.224 (0.417)	0.191 (0.393)	0.170 (0.375)	0.146 (0.353)
School is in Suburban Area	0.216 (0.412)	0.192 (0.394)	0.197 (0.398)	0.188 (0.390)
School is in Rural Location	0.470 (0.499)	0.529 (0.499)	0.565 (0.496)	0.595 (0.492)
School is in Urban Location	0.089 (0.286)	0.088 (0.283)	0.068 (0.251)	0.072 (0.258)

Note: Each cell reports the mean and standard deviations are presented in parentheses. The sample presented in this table wrote all three exams in the current and all preceding years of the experiment.

Appendix Table 2: Logit Estimates of the Probability of Remaining in the Sample

Grade remaining in the sample	Grade One	Grade Two	Grade Three
Kindergarten Reading	0.00720*** (0.00322)	0.00230 (0.00494)	0.00041 (0.00597)
Kindergarten Mathematics	0.00865*** (0.00116)	-0.00152 (0.00189)	0.00126 (0.00252)
Kindergarten Word	-0.00035 (0.00242)	-0.00061 (0.00369)	-0.00546 (0.00464)
Grade One Reading	Not included	.00189 (0.00293)	0.00053 (0.00397)
Grade One Mathematics	Not included	.01262*** (0.00222)	-0.00494* (0.00307)
Grade One Word	Not included	.00834*** (0.00260)	0.00834*** (0.00258)
Grade Two Reading	Not included	Not included	0.00868*** (0.00404)
Grade Two Mathematics	Not included	Not included	0.00728*** (0.00289)
Grade Two Word	Not included	Not included	-0.00195 (0.00292)
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, class size treatment, school indicators, child gender and child race. Standard errors corrected at the teacher level in parentheses. ***,**,* indicate statistical significance at the 5%, 10%, and 20% level respectively.