Estimating Context-Independent Treatment Effects in Education Experiments

Weili Ding
Queen's University and NYU-SH

Steven F. Lehrer*
Queen's University, NYU-SH and NBER

Comments Welcome

Abstract

In this study, we first document that the magnitude of the estimated treatment effect in Project STAR is substantially larger in schools where fewer students were assigned to small classes. The differences in student performance across schools cannot be explained by failure in randomization, other observed school level characteristics or differences in selective test taking. Further, we show that these achievement gains are driven by students in small classes from schools where fewer students were in small classes. The results are suggestive that there was a proportionate change in motivation or effort by teachers who teach small classes but not in regular classes. Second, we introduce empirical strategies for experimental studies that aims at disentangling the pure educational effect from a specific treatment from that which is attributable to the interaction between the treatment and the social context in

*We would like to thank Giacomo De Giorgi, Caroline Hoxby, Hidehiko Ichimura, Derek Neal, Whitney Newey, Aloyius Siow and Gerard van den Berg and seminar participants at the Yale-Fudan TED International Conference, Annual CLSRN conference, IOMS Conference, NBER Economics of Education Fall Meeting, Carleton University, McMaster University, Queen's University, University of Manitoba, University of Mannheim, Syracuse University, Tilburg University, University of Oregon and the Vrije University of Amsterdam for helpful comments and suggestions. Lehrer wish to thank SSHRC for independent research support. Parts of this paper were written while the Ding and Lehrer were visitors to the Center for Labor Economics, University of California - Berkeley, and we would like to thank this department for its hospitality. We are responsible for all errors.

which the experiment takes place. Using minimal structural assumptions we disentangle the estimated treatment effect into components that are context specific and context independent. Our results indicate that between 50-70% of the estimated treatment effect in Project STAR is context specific.

1 Introduction

Randomized experiments are increasingly being used in many disciplines to test the efficacy or effectiveness of alterative policies, services and technologies. For example, the number of randomized experiments being carried out by education researchers has grown exponentially.[1] In this paper, we focus on a common concern with experiments that relates to the generalizability of any findings. Specifically would the program that is being evaluated, have the same effect if it was implemented in a different context (i.e. not in an experiment)?[2]

In many randomized education experiments, a specific treatment is randomly provided to participants at multiple schools (sites), where there is substantial unplanned variation in the fraction of individuals offered treatment across these locations. In these experiments, individuals/subjects often can easily verify their treatment status and that of other participants in the study, so they do not have expectations or beliefs on what is

[1] However, some researchers (i.e. Cook and Payne (2002)) remain skeptical that these experiments provide much value added to research and education policy debates. See also Deaton (2010) for a general critique of randomized experiments and Imbens (2010) provides a response.

[2] Substantial research in experimental economics demonstrate that qualitative factors that arise in the laboratory setting, which range from the context in which the experiment occurred and how the instructions are framed affect the outcomes in these studies (e.g. see Shogren (1993) for evidence from bargaining experiments). The notion that psychological cues and the social context have implications to economic decisions has been formally modeled in Becker (1991), Mailath and Postlewaite (2002) Laibson (2001) and Bernheim and Rangel (2004), among others. Similarly, within the field of marketing, the context in which a product is displayed has been shown to influence how consumers evaluate whether to purchase these goods. Last, qualitative contextual factors such as "bedside manner" have been found to affect health outcomes in medical studies.

the likelihood that they are receiving treatment. Further, these individuals/subjects may have preferences over the outcome of the experiment and can influence the outcome to some extent. This idea is well discussed in Hoxby (2000), who notes, that not only are the actors in an education experiment aware of the study, but that the experiment alters the incentive conditions which could result in the analysis finding that policies appear to be effective when they would not have if fully enacted.

This study has a natural relationship to research that investigate the many channels through which being part of an experimental study (and being monitored) influences its participants. The treatment literature is keenly aware of the positive psychological effect of the mere action of putting people into a study.[3] This effect is viewed by researchers as a serious confounding factor that needs controlled for in order to derive the "real" effect of undergoing a treatment, essential for human subjects. For example, in the medical literature there has been a move to performing experiments using a double-blind protocol, to lessen the potential influences of prejudices and unintentional physical cues leading to potential behavioral responses by the study participants, that may bias the estimated treatment effects. In these settings, the context specific component of the treatment effect is commonly referred to as placebo effect and it needs to be strictly distinguished from the pharmaceutical effect.[4]

[3] The notion that in an experiment selection effects could be generated that would not arise in nonexperimental settings was pointed out in Heckman (1992) who termed these selection effects as randomization bias.

[4] Malani (2006) introduces a clever strategy that exploits variation in the probability of assignment to treatment across medical trials for the same drug, to identify the importance of placebo effects. His strategy relies on the assumption that placebo effects exist if patients in higher-probability trials state better health outcomes simply because they have higher expectations about the value of the treatment from their belief that they are more likely to receive treatment, all other things being equal.

Experiments in the field of education differ from clinical medicine in two important ways. First, not only are the sample sizes in education generally much larger but these studies are implemented in many more locations and it is difficult to ensure that the experiment is implemented in the same manner across these locations. Second, many of these studies are zero blind experiments, in that all the participants can easily identify if they are receiving the "treatment" being studied, so they do not have expectations or beliefs on what is the likelihood that they are receiving treatment. The context specific effects we subsequently identify are those which we argue arise in response to psychological stimuli within schools that alters the incentives of the participants. Our empirical strategy to disentangle the pure educational effect from a specific treatment from that which is attributable to the interaction between the treatment and the social context in which the experiment takes place, exploits the random unplanned variation in the extent to which the treatment is offered across multiple sites. As the treatment are being offered in each of the participating schools, it is inevitable that the independence of the treatment and control groups would be compromised through the normal social and collegial interaction processes of the school communities. We propose that if this variation in treatment receipt is random, it may provide insights into whether some participants had a behavioral response to participation in the study itself.

We use data from Tennessee's highly influential class size experiment, Project STAR to illustrate our empirical strategy. This experiment was conducted for a cohort of students 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). Earlier analyses of Project STAR data present strong evidence

that smaller classes has a positive impact on student achievement, particularly in the first year of the experiment.[5] A salient feature from the implementation of this experiment is that teachers and principals in the participating schools were informed of the experiment and many researchers have speculated that behavioral responses occurred in this study.[6] We document that the percentage of students receiving treatment ranged from 16.129% to 44.26% across participating schools. We first demonstrate that the magnitude of the estimated treatment effect in Project STAR is substantially larger in schools where fewer students were assigned to small classes (the treatment). Assuming that given the size of the class a student is in, the percentage of students in his grade that are in small classes should not affect the small class size effect he enjoys. We are able to rule out many competing explanations for this systematic pattern of treatment effect heterogeneity and hypothesize that a context specific effect may occur, in the sense that when the treatment becomes more exclusive or rare, the positive feeling of being put on treatment becomes stronger, either because teachers and/or students derive utility from being the "selected few" thus incentivized to work harder or they feel compelled to work harder because they are the "only ones" that can show positive effect of treatment or both. Whatever the explanation, it is clear that this finding is not consistent with alternative ways small classes per se may contribute to student performance and may have important implications

[5] See Finn and Achilles (1990), Finn et al. (2001), Krueger (1999), Hanushek (1999) and Ding and Lehrer (2010), who each use different empirical strategies to estimate causal impacts with this data.

[6] Goldstein and Blatchford (1998) and Hanushek (2003) discuss such weaknesses with reference to Project STAR. In essence, researchers either speculate that some of the positive effect may be a form of Hawthorne effect, that students and teachers assigned to small classes may feel more incentivized or compelled to teach or study harder compared to their regular class counterparts. Alternatively, others have hypothesized that the opposite could be true too, as often referred to as a John Henry Effect. But the threat to the identification of positive small class effect comes solely from a Hawthorne effect since the existence of a John Henry Effect would only strengthen the results.

for external validity of the estimated effect.

The second contribution of this study is to define context specific and context independent treatment effects in the standard evaluation framework and discuss the implications for empirical design. Using an education production function framework and minimal structural assumptions, we introduce an empirical strategy that can disentangle the estimated treatment effect into these two components that differ on their external validity. Our results indicate that between 50-80% of the estimated treatment effect in Project STAR is context specific.

This paper is organized as follows. In Section 2, we provide a brief discussion of how Project STAR was designed and implemented as well as introduce the data we use for our analysis. This discussion highlights that the variation in the percentage of students receiving treatment across schools is not driven by budgetary issues and appears as random. Our reduced form empirical strategy and empirical regression results are motivated and presented in Section 3. We find that the magnitude of the estimated treatment effect is substantially larger in schools where fewer students were assigned to small classes (the treatment). These differences in student performance that exist in multiple subject areas across schools cannot be explained by failure in randomization, other observed school level characteristics or differences in selective test taking. Further, we show that these achievement gains are exclusively driven by students in small classes from schools where fewer students were in small classes. The results are suggestive that there was a proportionate change in motivation or effort by teachers who teach small classes but not in regular classes. We relate our findings to evidence within the education literature on teachers' responses to perceived principal favoritism or peer pressure. Structural analyses that attempts to document the magnitude of context specific and context independent effects are presented and discussed in Section 4. This analysis is important to identify the true effect

of smaller classes as unless any policy or program aims to divide students into treatment (small class) and control (regular class), which clearly lacks ethical or political appeal, we can no longer count on the positive feeling of being treated to yield beneficial small class effect. Our analysis suggests that context specific effects are important in practice and are consistent with recent non-experimental research that has examined large scale class size reductions in California and Ontario.[7] A concluding section summarizes our findings.

2 Experimental Design and Data

During the mid-1980s, legislation was forged and subsequently unanimously approved by the Tennessee legislature, that authorized a demonstration project called Project STAR (Student/Teacher Achievement Ratio).[8] The STAR legislation focused on specific educational goals, should assess the effects of class size in different school locations (i.e. urban vs. rural schools) and carefully defined the treatment under study. The legislation specified that the project should include schools located in the inner city, suburban, urban and rural areas. The Commissioner of Education subsequently invited all Tennessee school systems to take part and sent guidelines for participation to each local system. These guidelines indicated that the state would cover additional costs for project teachers and teacher aides, but that local systems would furnish any additional classroom space needed.

[7] In our analysis, we are holding teacher quality constant. Yet evidence presented in Jepsen and Rivkin (2009) from California's recent large scale class size reduction indicates that the teachers hired to teach the additional classrooms had limited teaching experience and lacked full certification, resulting in a dampening of the potential benefits from smaller classes. Additionally these lower quality teachers were disproportionately placed in low income neighborhoods increasing the heterogeneity in student performance between neighborhoods.

[8] See Ritter and Boruch (1999) for a comprehensive discussion of the political and institutional origins of Project STAR.

In Appendix Table 1, we reproduce the plan for distribution of students and classes in this study. This table indicates that in the kindergarten year, the number of small and regular classes within a school is determined by the enrollment. The project schools would not receive any special considerations other than class size. The students in these schools would continue to use the regular district or school curriculum, texts, etc. All participating teachers had to be certified for the grade level they were teaching. The schools were informed that within these schools, there should be no major changes in processes or organization, other than class sizes. To participate, the schools had to agree to the random assignment of teachers and students to the different class conditions.[9] To implement the experimental design, researchers at Universities developed a protocol for the random assignment of students to class type.[10]

Thus, 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). Additionally in each year of the experiment teachers were also randomly assigned to the classrooms in which they would teach. Random assignment overcomes selection bias that arises not solely by decisions made by parents themselves but also by school principals. Noncompliance with treatment assignment appears to be a limited concern in the first year of the study since Krueger

[9] See the Project STAR technical report (Word (1990)) and the public-use data manual (Finn, Boyd-Zaharias, Fish, and Gerber (2007) for more details.

[10] Sojourner (2008) provides a comprehensive survey based on his own independent research in which he discusses whether random assignment was to actual classrooms or class types. He concludes that while the documentation indicates that assignment was to class type, interviews and statistical tests do indeed provide strong but not absolute evidence of random assignment of students to class rooms.

(1999) reports following a careful examination of actual enrollment sheets compiled in the summer prior to the start of kindergarten could only identify a single student out of or 1581 students from 18 participating STAR schools that was assigned to a regular or regular/aide class but actually enrolled in a small class.

Initially, 180 schools in about 50 of the state's 141 school systems expressed interest in participating. Of these, approximately 100 schools had enough students in kindergarten to meet the size criterion for participation. The final selection of schools was based on a) including at least one school from each district that had volunteered and b) including enough schools from all four school types and all three regions of the state to permit comparisons between school types, as specified in the legislation. In total, 79 schools in 42 systems were participants in the first year.

There was substantial variation in the number of students assigned to small classes (the treatment) across these schools. Figure 1 presents a histogram of the percentage of students receiving treatment across schools in kindergarten. As seen in Figure 1 the between school variance in proportion of treated is considerable. The average school has slightly over 30% of the student body attending small classes. Schools range from 16.129% of students in treatment classes to 44.26%. The variation in proportion treated students stem from the size of the student body, but also potentially from the number of available classrooms. In our initial analysis we will estimate specifications on different quartiles of the percent of students in treatment across schools distribution.

In order to minimize issues related to the documented violations to the experimental protocol,[11] we only analyze data from the first year of the experiment. At the end of

[11] Such violations were prevaent as by grade three over 50% of the subjects who participated in kindergarten left the STAR sample and approximately 10% of the remaining subjects switch class type annually. Additionally, Ding and Lehrer (2008) present evidence of selective attrition and demonstrated that the conditional random assignment of the newly entering students failed in the second year of the experiment

each school year the majority of the students completed multiple exams to measure their performance in different dimensions. In this paper, our outcome measures (A_{iT}) are total scaled scores from the Reading, Mathematics, Word Recognition and Listening Skills sections of the Stanford Achievement test.[12]

Summary statistics for this sample are provided in 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).

The methods that we describe in the subsequent sections rely crucially on the fact that participants could infer their treatment status. Prior to conducting the analysis, it is important to state explicitly that the documentation from Project STAR themselves indicates that orientation sessions were conducted for teachers at 20 schools entering the project in kindergarten. The person conducting the orientation described the project, its

as among this group of students those on free lunch were significantly more likely to be assigned to regular (larger) classes.It should also be noted that attendance of kindergarten was not mandatory in Tennessee and students who entered school in grade 1 may differ in unobservables to those started in kindergarten which would add further statistical complications to recover the causal parameter.

[12]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 representative sample of students from across the nation.

purposes and processes, and answered questions. As these experiments were done within schools, it was near obvious for the teachers to compare workload and determine whether they were assigned to a control or treatment classroom. Also process evaluations were underway during the year but this data remains publicly unavailable.

3 An Empirical "Puzzle"

As in Krueger (1999), our analysis begins by estimating a contemporaneous specification of an education production function that relates education inputs to achievement as measured by a test score (A_{ij}) of child i in school j

$$A_{ij} = \beta_1' X_{i1} + \beta_t' T_{i1} + v_j + \varepsilon_{i1} \tag{1}$$

where X_{i1} is a vector of student and teacher characteristics, T_{i1} is an indicator if student i was assigned to a small class, v_j is a school specific fixed effect and ε_{i1} captures random unobserved factors.[13] Controlling for school effects is necessary since randomization was done within schools. By randomly assigning class type and teachers to students, T_{i1} is uncorrelated with unobserved factors such as the impact of pre-kindergarten inputs, family and community background variables, etc., permitting unbiased estimates of β_t with only data from kindergarten. The estimates of β_t can be interpreted as either an intent to treat parameter or average treatment treatment effect.[14] Inference is based on

[13] These variables are exactly the same as those used in the base specifications in Krueger (1999). For robustness we replicated the entire analysis with two alternative specifications that allowed teacher experience to have nonlinear effects. The first approach allowed different impacts in each of the first three years and the second approach included experience up to a cubic. All of the results discussed in the paper are robust to these alternative treatments of teacher experience.

[14] There were few violations to the experimental protocol in Kindergaten so the $ITT \approx ATE$. It is also possible to use non-parametric bounds strategy to calculate the ATE where once can account for issues

robust standard errors allowing for clustering at the classroom level.

An empirical "puzzle" arises when equation (1) is estimated only for subsets of students that are attending schools located in different quartiles of Figure 1.[15] Table 2 presents estimates of β_t that result from this exercise. The first row contains estimates of equation (1) for the full sample and as earlier research has documented there is a positive and significant impact of small class in all subject areas. The remaining rows of Table 2 present estimates using the different subsamples of schools defined by quartiles of the percentage of students assigned to treatment distribution. On the mathematics test the magnitude of the small class treatment effect (β_t) is largest in magnitude when equation (1) is estimated only with students from schools which had few students in treatment. Students in these schools receive nearly triple the benefit from small classes as compared to the remaining participating in the experiment. In the subject areas of reading, word recognition and listening skills we also find that the estimated impact of small classes is approximately double in magnitude in the two lower quartiles of Figure 1. Surprisingly, in all subject areas, the small class treatment is statistically insignificant if we estimate equation (1) only using students from the subsample of schools where small classes are most common (e.g. the fourth quartile of Figure 1). These results indicate that there is substantial treatment effect heterogeneity that relates to the percentage of students in treatment classes within a school. Appendix Table 2 presents estimates that the impact of small class treatment is the only explanatory variable in equation (1) whose impact varies in such a systematic manner across subject areas based on quartiles of Figure 1.[16]

related to selective test taking. This analysis is available upon request.

[15]Note that, Lechner (2002) also studies whether the treatment participation probability is a source of heterogeneity in the treatment effct, but this link is not discussed in light of the context specific effects or a violation of the Stable Unit Treatment Value Assumption in randomized experiments.

[16]This pattern of results in table 2 and Appendix table 2 is robust to other sized slices of school in the

To shed more light on what is driving this puzzle, we investigated whether the extent at which treatment if offered within schools impacted the performance of students across schools in the both the control and treatment arms of the experiment. To accomplish this, we estimate the following equation

$$A_{ij} = \beta_1' X_{i1} + \beta_p' Percent\ in\ Treatment_j + \varepsilon_{i2} \tag{2}$$

where $Percent\ in\ Treatment_j$ is the proportion of students assigned to small classrooms in school j.[17] Given the random assignment of treatment within schools and the fact that variation in the proportion of students receiving treatment across schools should be uncorrelated with unobserved school and neighborhood inputs, a significant estimate of β_p identifies whether there is a significant impact of the treatment intensity on students achievement.[18] We estimate equation (2) separately for both the control and treatment classes. For the treatment classes, on average we expect that the students would receive a similar treatment but what may vary is the potential pressure their teachers face from the colleagues. If treatment classes are rare, these teachers or students may be incentivized

data (i. e. dividing the number of schools by 3 or 5 or 6). We did not consider smaller slices (e.g. slivers) since there would be a large loss in efficiency by forming a large number of strata with this data. Our initial choice of four equal sized bins (by number of schools) was not informed by theory and it is unlikely to be the optimal subclassification for estimating the treatment effect if we consider the minimum MSE among all partitions.

[17] We consider several alternative methods to control for $Percent\ in\ Treatment_j$ including it i) by itself, assuming a linear relationship, ii) up to a quartik to allow for potential non-linearities, and iii) allowing for unsystematic patterns using a series of indicators based on which quartile of Figure 1 a given school lies within. We report results from specification iii) but are our full set are robust to these alternative treatments of this variable.

[18] This specification is similar in spirit to the manner in which Hesselius at al. (2009) identify spill-over effects in worker absences among employees in Sweeden and in control classes how Duflo and Saez (2003) identify social interaction effects.

to increase their effort. We also estimate equation (2) for the control classes. Ex ante, there is no to reason to exhibit that the proportionate response of psychological should be symmetric across class types in response to the level of treatment intensity within the school. Further, we have no ex-ante prediction as to the sign of β_p since potential "discouragement" and "John Henry" effects would push in opposite directions.

Estimates of equation (2) for both treatment and control classes are presented in Table 3. The first four columns present evidence of the impacts of $Percent\ in\ Treatment_j$ on achievement where we include indicators for which quartile of Figure 1 a given school reside in. The excluded category is the lowest quartile of Figure 1, where it is harder to find students attending small classes. In all subject areas with the exception of listening skills we find that by examining small classes alone that the larger the fraction of students in small classes within their school, the poorer these students do. The effects are driven by the top quartile in Figure 1 in all three cognitive subjective areas. Notice that the regression results for control classes presented in the last four columns of Table 3 indicate that there are no statistically significant differences in performance of students in control class rooms as treatment intensity varies. The results of Table 3 are suggestive that treatment intensity only affects student performance in all cognitive subject areas in small class rooms only.

There are many other potential explanations for the heterogeneity in treatment effects across schools based on the proportion of students that are treated. We directly investigate three candidates, i) differential failures in the randomization of student and teachers in those schools, ii) there were differences in other school level characteristics, and iii) there was more or less selective test taking within some of these schools. However, as we discuss below using the data and standard econometric methods, we can essentially rule out all of the above potential explanations.

Since random assignment of students and teachers to classes implies that there should not be systematic differences in characteristics across classrooms within school. We first conducted simple χ^2 test of difference between class types within schools and graph the p-values of these tests.[19] Second and as shown in table 4, we estimated the following equation

$$T_{ij} = \alpha_1' X_{i1} + v_j + \varepsilon_{i1}, \tag{3}$$

to provide a further check that the characteristics of students and teachers did not differ in significant ways across the schools located in different quartiles of Figure 1. With the exception of two teacher characteristics in quartile 4, there are no individually significant effects. In this quartile, teachers with advanced degrees and that are Caucasian are less likely to be assigned to small classes. Yet as Appendix Table 1 indicates these characteristics do not influence achievement levels on their own (and even in specifications with the full set of interactions so it is unlikely what can explain the puzzle. Joint tests also indicate that randomization of students and teachers was successful in all four quartiles.

We also conducted similar tests of difference between schools located in different quartiles of Figure 1 by school location, school size and aggregate race, gender and free-lunch status of the kindergarten student body. For each of these variables we cannot reject

[19] Under the null-hypothesis of random-assignment within all schools the p-values for the school test statistics are distributed U[0,1]. Panels of histograms of the schools' p-values discretized into 10 bins, informally appear close to uniform. Formally, an overall χ^2 test statistic for each variable assuming independence across schools can be computed. The p-values of these overall test statistics for each quartile appear that we cannot reject the null of random assignment with regards to the distribution of these observed characteristics. Note that this approach was also used in Sojourner (2009), Graham (2008) and Ammermueller and Pischke (2006) who each note that administrators intentionally created balance of these characteristics across classrooms, we would observe a right skewed distribution full of high p-values. Analagously a deliberate stratification along these characteristics would result in a left-skewed distribution with a larger number of low p-values.

the Null of no differences between school type.[20] We further investigated how the class sizes varied across the student body. This is illustrated in Appendix Figure 1 and note it appears that there were very few schools in which the space constraint may have applied. Schools of very different sizes were equally likely to appear in each quartile of Figure 1. Further the raw Pearson correlation between the size of kindergarten class relative to the percent of students in treated classrooms is -0.0279. Thus, we are confident that we are not simply finding a school size effect. the strongest evidence comes from estimating equation (1) based on quartiles of the distribution of kindergarten cohort size. The results of this analysis are presented in appendix table 3 and notice that there is not a systematic pattern based on cohort size which can also be viewed as a proxy for school size.

Not all students in the Project STAR experiment completed all four tests and there it is possible that administrators in specific school ensured certain students were absent from particular tests. We conducted statistical tests to see whether there were significant differences in both the rates of test completion by subject area, by class type, observed students characteristics of students who did not write specific tests and as demonstrated in appendix table 4, in no case did we find any evidence that there were differences across schools based on which quartile of Figure 1 their school resides. Additional candidate explanations for why small classes could be effective include the nature of classroom experiences, teaching methods and attention to peer effects in smaller classrooms. While we do not have access to data from the process evaluations to determine if these explanations varied across the quartiles, a survey of the existing literature on Project STAR casts doubt that any of these factors can explain the pattern reported in Table 2.[21] Thus,

[20] Project STAR researchers colleted information on school principals but this is not publically available.

[21] Existing evidence suggests that none of these factors differed significantly between regular and small classes. For instance, Evertson and Folger (1989) report no statistically significant differences between small and regular classes for the percentage of student-initiated questions and comments, percentage of

we are quite confident the puzzle is real and requires alternative explanation.

Last, it is worth noting that there are several interesting and statistically significant differences in ex-post implementation failures across the quartiles. This is demonstrated in Table 5, where we estimate a slightly modified version of equation (1) for subsets of students that are attending schools located in different quartiles of Figure 1, where the modification is using either an indicator noncompliance with treatment assignment or attrition following kindergarten. Students from small classes in schools located in the fourth quartile of Figure 1 were both significantly more likely to attrit from the study and move into regular classes. Correspondingly, there was significantly greater movement of students from regular classes to small classes as well as attrition from regular classes in those schools located in the first quartile of Figure 1. These results are suggestive that parents had a behavioral response to the treatment effect heterogeneity witnessed in Project STAR.

In summary the analysis in this section suggests that previously reported treatment effects from Project STAR in kindergarten are driven by performance in the small classes from those schools where these classrooms are rare. Since random assignment appears to have worked. As we discuss, in the next subsection the notion that teacher's motivation and effort change in this manner is consistent with evidence in several branches of both the education and economics of education literature.

students off-task or time waiting for help in reading or mathematics. Similarly, they suggest that the small difference in the amount of "disruptive behavior" (1.6% vs. 2.0%) and "inappropriate behavior" (1.6% vs. 2.1%) is too small to explain the benefits of small classes. Finn & Achilles (1999) conclude that there was no change in fundamental teaching strategies when given a small class.

3.1 Discussion on the Mechanism Underlying the "Puzzle"

The mechanism can neither be explained by a story of positive social interaction effect or some general equilibrium effect. If positive social interactions were the explanation, we would expect to find larger treatment effects in quartile 4 than quartile 1, since more individuals are exposed to treatment in the fourth quartile and there being positive spillovers. The channel through which this heterogeneity arises is unlikely through prices or incomes and we argue that it operates through responses of the teachers. The other actors who may exhibit behavioral changes are students and their parents. Given there age we do not believe that the students are unlikely to change their effort to please their instructors. In the case of parents themselves, similar to the discussion in Todd and Wolpin (2003) we believe that it is ex-ante indeterminate in which direction this may head. After all, it seems equally plausible that parents of regular class students would purchase compensating inputs and parents of small class students may either elect to purchase or not purchase inputs that could reinforce any benefits from the treatment.

To be clear, there is likely many more potential stories that can explain how these context effects arise in the Project STAR study. Our two preferred candidates are proportionate incentives and peer pressure. Regarding proportionate incentives, if only one class in a school is small, the small class teacher knows that she is solely responsible for demonstrating the effectiveness of small class, thus most incentivized to work hard. If many classes are small, the incentive for each of the small class teachers is less intense. That is, since random assignment appears to have worked, it could be the case that teacher's motivation and effort proportionately change with the fraction in treatment. Such a mechanism is consistent with evidence in the education literature on teacher responses to perceived favoritism. It is well established that teachers respond to special claims of status (Finley (1984), Kurz (1987), Becker (1952, 1955)) and principals (or

department heads) often use perceived "good assignments" to reward teachers (Lortie (1975)). Not surprisingly, school principals are known to influence teacher placements (Carey and Farris 1994), Ingersol (2003) find that 75% of principals in a nationally representative survey do this to a "great extent". Related, Spear et al (2000) survey the education literature and conclude that teachers believe their own morale is largely determined by their quality of life within the school, and claimed that the working conditions have positive effects on both morale and job satisfaction among teachers.

Teacher motivation and effort may also change with peer pressure. Kandel and Lazear (1992) present a theoretical model that suggests that if there are fewer teachers and students in small classes then the chance that they directly suffer embarrassment over direct performance comparison makes it harder to free ride.[22] Consistent with such a prediction, Mas and Moretti (2009) and Bandiera, Barankay and Rasul (2009) present evidence from fields experiments in the workplace that demonstrates that peer pressure directly affects individual productivity. In sum, we concur with Hoxby (1999) that argues not only the teachers and principals in an education experiment are likely aware of the study, but that the experiment alters the incentive conditions (or if you prefer, changes the context) for these subjects.[23] This could result in the experiment concluding that

[22] Jackson and Bruegmann (2009) present evidence that a elementary school teacher's own performance is affected by the quality of her peers in the same school who teach students in the same grade. As teachers were provided information on the experiment and the legislation that it is reasonable to postulate that if teachers believe that working conditions for them improve if class sizes are reduced. We hypothesize that a teacher's peers can also affect her classroom performance by changing her own teaching effort via peer pressure. In schools, where the presence of small classrooms are rare, this may make it harder for a teacher to may motivate their colleagues through contagious enthusiasm or through embarrassment over the unfavorable direct performance comparison.

[23] Teachers may prefer smaller class size than students or parents either because wages do not reflect working conditions fully or because teachers as a group can raise the demand for their services by lowering

certain policy intervention is effective but when fully enacted it no longer appears to be the case.

Appendix Table 6 presents some suggestive evidence consistent with these hypotheses. In this table, we estimate equation (1) on samples defined by the number of classrooms and small classes kindergarten. We report results for the four combinations which have at least 6 schools. These 4 combinations account for 69 of the 79 schools in the experiment and include (class room, small class) pairs (3,1), (4,1), (4,2) and (5,2). Recall from Appendix Table 1 these constitute schools with study designs 1 through 4. Notice that in each subject area the impact of small class on achievement is larger in magnitude and generally statistically significant in the 34 schools that only have 1 small class. In the 20 schools with 4 classrooms of which 2 are small classes, the treatment does not have a significant impact on academic performance in any subject area.

Formally testing between these alternative mechanisms is difficult but one could imagine comparing treatment effect estimates between schools with similar percentages of children in treatment classes but that have different numbers of small class teachers. For instance, we can compare treatment effects between schools with a single teacher in a small classes to those with multiple small classes that are similar in terms of the percentage of students assigned to small class. The results from this exercise are indeterminate. Alternatively, one could construct structural models to see if they can explain the pattern of results we observe. In the next draft, we will discuss estimates from a simple story of

class size. Last, in year-end interviews, Project STAR teachers were asked whether they would prefer to have a small class as their regular teaching condition. Not surprisingly, there was an overall preference for small classes, with eighty percent of the teachers who had a small class preferred a small class and 56 percent of the teachers who had an aide- would have preferred a small class. The teachers who had a regular dass chose a small class 71 percent of the time.

selective assignment of high quality teachers[24] and a model of gift exchange to see if they can account for the results. In the next section, we consider how accounting for these potential mechanisms can be accommodated within the standard strategy social scientists use to evaluate experimental data.

4 Context Specific and Context Independent Effects

In this section, we formally define context specific and context independent treatment effects in the standard evaluation framework and discuss the implications for empirical design. Using an education production function framework, we introduce two empirical strategies that can disentangle the estimated treatment effect into these two components that differ on their external validity. We discuss the assumptions required to identify the context specific and context independent effects under each strategy. Finally, we illustrate this approach with data from Project STAR. Our results indicate that between 50-80% of the estimated treatment effect in Project STAR is context specific.

4.1 Methodology and Causal Framework

Consider an education experiment that will be implemented in S schools. Within each school, we refer to being in small classes as receiving treatment and following Finn et al. (2001) attending either regular or regular with aide classes as being in the control group. We use $T_t = 1$ to denote actually being assigned to a small class in grade t and $T_t = 0$ as being assigned to a regular class. At the completion of each grade t, each

[24] This exploration is motivated by the likely substantial within school differences in teacher quality and we will examine if systematic variations in teacher quality account for the heterogeneity. The results appear in Table XI and XII using the methodology of Graham (2006).

student takes exams and scores A_t (potential outcomes; A_{1t} if attending a small class and A_{0t} if attending a regular class). With a single dose of treatment, the standard evaluation problem occurs since we cannot observe A_{1t} and A_{0t} for the same individual. In a single period experiment, without context specific effects 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.

If we allow for context specific effects, we must first make an assumption about at which level these effects arise. In our analysis, we assume that each student and teacher can only attend (be employed) in a single school in a given year, thereby ruling out spillovers across schools so the context specific effects are school specific. Under this assumption, we now define the test score a student obtains from the exam as A_{ts} where the potential outcomes now depend on not only the class type treatment operating through T_t but also a context variable that C_s that operates at the school level. Context specific effects can arise from social interactions (Hoxby (2000), Graham (2008), Ding and Lehrer (2007) among others), in our setting may result as a response to the incentive conditions that the experiment induced.

In our empirical application, we will proxy C_s using the proportion of students attending a small class in a given school.[25] As before, we continue to face a standard missing observation problem but allowing for context specific effects implies that this problem now has two dimensions: First, we do not observe the outcome of an individual with $T_t = 1$ had she received $T_t = 0$. Second, we do not know what would have happened in a school s with a context C_{s1} had this school instead been assigned to a treatment policy regime resulting in a different context C_{s2}. Allowing for context effects implies that there

[25] We verify the robustness of our results to alternative measures including the percentage of teachers in the school.

is more than two potential outcomes and represents a violation of the Stable Unit Treatment Value Assumption (SUTVA).[26] This assumption removes the context dimension of the missing data problem, allowing researchers to calculate an ATE by taking differences in means between the treatment and control group. That is, we now define $Y_i(T_t, C_s)$ as the potential outcome for individual i that applies if assigned to treatment value T_t and school context C_s. As in the literature on dynamic treatment effects (Lechner (2004), Ding and Lehrer (2010)) we formally define $\tau^{(x,y)(v,w)}$ as the average treatment effect parameter that measures the average difference in outcomes between the treatment-context pair (x, y) with another treatment-context pair (v, w) where we must hold one dimension of this problem fixed. That is either $x = v$ or $y = w$ in the definition above. For example, $\tau^{(1,0.25)(0,0.25)}$ is an estimate of the average treatment effect in schools with context 0.25 and $\tau^{(1,0)(0,0)}$ is the classical ATE, which we also define as context independent treatment effect since it is calculated when $C_s = 0$. From a policy perspective, $\tau^{(1,1)(0,1)}$ may be of prime interest for a policy that would be mandated across the board. It is worth stating that the ATE $\tau^{(1,1)(0,1)}$ can be smaller or larger than $\tau^{(1,0)(0,0)}$. Lastly, we can calculate a

[26]Implicit in many causal estimators, this assumption requires that an individual's potential outcome is not affected by the treatment status of others. Formally, SUTVA consists of two sub-assumptions. First, SUTVA rules out that there is any interference between units, such that the assignment of an individual to the treatment group should have no effect on outcomes for other individuals. That is, there is a single value for each of the potential random outcome variables for a given student regardless of the randomization assignment or mediation behavior of any other student. Second, there is a single value for each of the potential outcome random variables for a given student regardless of the method of administration of the randomized intervention. In other words, the treatments for all students are comparable. The validity of SUTVA will depend upon the specific context of the experiment. If ignored, violations of SUTVA have the possibility of adding bias to estimated treatment effects, and it is possible that these biases can go in either positive or negative direction. Rubin (1986) provides more details on the SUTVA assumption.

host of parameters such as $\tau^{(1,0.45)(1,0.2)}$, which is an estimate of the expected outcome of treated individuals when going from context 0.45 to 0.2.

4.2 Identification

We propose two strategies to identify context specific strategies. In the first strategy we can nonparametrically identify context specific effects if the context variable is randomly assigned. That is, if C_s is assigned randomly across schools and T_t is assigned randomly within schools, then one can use a difference in difference estimator under the familiar assumptions of i) common trend, ii) common support and iii) no anticipation effects to non-parametrically identify context specific effects.[27] That is, consider the following three possible comparisons of performance in figure 1: quantile 1 vs quantile 4, quantile 2 vs. quantile 4 and quantile 1 vs quantile 3. In each case there is no reason ex ante, to anticipate that the context independent treatment effect should differ. If we focus on quantiles 1 and quantiles 4 then $E[A|T_t = 1 \text{ and } Q = 1\ ] - E[A|T_t = 1 \text{ and } Q = 4] - \{E[A|T_t = 0\ and\ Q = 1] - E[A|T_t = 0 \text{ and } Q = 4]\}$ would provide this estimate. Adding covariates would lead to the well know linear difference in difference estimator given by

$$A_{ij} = \beta_1' X_{i1} + \beta_2' T_{i1} + \beta_3' I[Quatile = 1] + \beta_4' \{T_{i1} * I[Quatile = 1]\} + v_j + \varepsilon_{i1} \quad (4)$$

where β_4 provides an unbiased estimate of the context specific effect between the two quantiles under consideration. Note, that while one cannot identify β_3 with school fixed effects, this does not affect the interpretation of β_2 as the context independent treatment effect or β_4 as the context specific effect and is analogous to difference in difference applications where there are a subset of individuals in multiple states are followed over time

[27] We are extremely grateful to Hidehiko Ichimura who generously provided this idea during a seminar presentation.

and there is only a change in the treatment affecting participants from one group of states and state fixed effects are being controlled for. In this case, the regular classes are used to capture the pre-period and small classes capture the post-period.

A disadvantage of the above strategy is that it does not utilize all of the data nor allow for extrapolation to other potential contexts that may not be in the support of the experiment. In that setting, with SUTVA violated, structural assumptions are needed to identify causal parameters.[28] We consider two additional empirical strategies that draw on an underlying economic model of human capital production (Ben-Porath (1967)), allowing us to have a mapping between the causal estimates and the structural parameters. We continue to make use of both designed randomization in the STAR experiment and the unplanned random variation in the percentage of students in treatment classes to achieve identification of these parameters.

Following Ben-Porath (1967) and Boardman and Murnane (1979), we view the production of education outcomes as a cumulative process that depends upon the potential interactions between the full history of individual, family and school inputs (captured in a vector X_{ijt} in year t), class size treatments, school level context effects, school effects and independent random shocks ($\epsilon_{iT}...\epsilon_{i0}$). Formally, child i in school s gains knowledge

[28] For instance Ferracci et al. (2014) use a two-step method that includes a matching approach with non-experimental data to identify context specific effects that arise specifically from social spillovers in labor markets. Other recent work that examines causal inference in a similar two-dimensional evaluation approach and where potential outcomes depend on both dimensions include Hudgens and Halloran (2008) and Manski (2009). Also we need to discuss isues Giacomo raised around how the use of the SUTVA assumption precludes consideration of certain types of questions that social scientists/economists ask since those causal questions have answers that are dependent on distributional properties (i.e. rationing) of the population.

as measured by a test score at period T :

$$A_{isT} = h_T(X_{iT}...X_{i0}, T_{sT}...T_{soT_o}, C_{sT_T}...C_{sT_o}v_s, \epsilon_{iT}...\epsilon_{i0}) \tag{5}$$

where h_T is an unknown function. Note v_s is included to capture unobserved student invariant school attributes and as discussed in the preceding subsection, we proxy for C_{sT_T} with $Percent\ in\ Treatment_{sT_T}$.

We first linearize the production function at each time period. An individual's achievement outcome in period one is expressed as

$$A_{is1} = v_s + \beta_1' X_{i1} + \beta_{T1}' T_{i1} + \beta_c' Percent\ in\ Treatment_{s1} + \varepsilon_{i1} \tag{6}$$

where v_s is an individual school effect. Since nearly all of the explanatory variables in equation (6) are discrete dummy variables, the only restrictive assumption imposed by linearization is the additive separability of the error term. However, to identify the structural parameters we do not need to linearize the education production function. Assuming that the unobserved factors and the school level context effects enter additively, and that i) the unobserved components ν_j, ε_{i1} are independent of T_{i1} and $Percent\ in\ Treatment_{s1}$, and ii) X_{i1} is an exogenous vector that is independent of T_{i1} and $Percent\ in\ Treatment_{s1}$; the structural parameters of equation (6) are nonparametrically identified.

Based on the above assumptions, we suggest a two-step estimation method that will allow us to disentangle the context specific treatment effect from the context independent treatment effect. For the first step, we make use of the actual randomization that was carried out within schools and nonparametrically estimate treatment effects at the school level. In the second step, we exploit the fact that the $Percent\ in\ Treatment_{j1}$ was randomly assigned across schools and run a nonparametric regression that links the school treatment effect to the school level context effect. The predicted values from this regression allows us to capture what portion of the school level treatment effect arises due to the

school context effect and the remaining variation reflects the portion of the school level treatment effect that is independent of this context. Efficient estimates would use the kindergarten cohort size by school to reweight the second stage estimation and bootstrap methods are used at both stages to conduct inference. This is required account for estimation error that may arise in the first step. Specifically, we first bootstrap the treatment effect estimates in stage one and in stage two we use those estimates in the nonparametric regressions. These stage two regressions are at the school level and we also bootstrap the nonparametric curve to conduct inference. Last, the nonparametric regressions are weighted by the number of students within the school who wrote the subject exam used in the first stage estimation.

Our final empirical strategy to identify and disentangle context specific from context independent treatment effects relaxes the additive separability assumption between school inputs and contexts (particularly between T_{i1} and $Percent\ in\ Treatment_{j1}$) and draws on a recent literature that has developed econometric methods to optimally reallocate inputs across groups both in the absence and in the presence of social spillovers.[29] Rather than using nonparametric copula functions in our two step estimation approach as in Graham et al. (2009) we do the following. In the first step we essentially nonparametrically estimate the production function in equation by introducing a specification that fully saturates all of the interactions between the school inputs. That being said, we do treat the school fixed effects as being additively separable.[30] As all of the inputs with the exception

[29]For example, Bhattacharya (2009) and Graham et al. (2009, 2010) independently develop econometric methods for evaluating the effects of reallocating inputs in various scenarios regarding social spillovers. They each frame the problem as a decision by a social planner who wish to maximize some objective function (e. g. average student performance) subjects to technology and resource constraints. Maximization of the mean is analogous to maximizing productive efficiency.

[30]Many of the nonparametric strategies developed in the econometric literature that we build upon

of $Percent\ in\ Treatment_{j1}$ are discrete we do not run into a curse of dimensionality. As before, we exploit randomization of all inputs conditional on each other and school characteristics to identify the structural parameters. In the second step we average the estimated production function over potential distributions of $Percent\ in\ Treatment_{j1}$. As in Graham et al. (2009, 2010) this amounts to conducting policy simulations where we examine how achievement would vary as the context changes in both small and regular sized classrooms. In particular, by conducting policy experiments where we can extrapolate to situations where $Percent\ in\ Treatment_{j1} = 0$ allows us to calculate the context independent treatment effect $\tau^{(1,0)(0,0)}$. Bootstrap based methods will be used to conduct inference in this procedure that only involves a single stage of estimation.

4.3 Results

We present estimates and analyses of the plausibility of each of the identification strategies. The results of using the first empirical strategy to disentangle this exercise are presented in Table 6. In the top panel, the context specific effects are presented by quantile. Notice that the behavioral responses are larger in lower quartiles. Comparing these estimates to what appeared in Table 2, we observe that on average the context specific component

can only allow for a limited set of explanatory variables in practice. For instance, Graham et al. (2009, 2010) illustrate their approaches with project STAR data but do not incorporate school fixed effects. Last, we consider two specifications of inputs. In the first we treat $Percent\ in\ Treatment_{j1}$ as linear and interactions with all the inputs. In the second specification, we include $Percent\ in\ Treatment_{j1}$ on its own as well as a quadratic term and all of the interactions between these variables and all of the other inputs. Naturally, the more higher order terms of $Percent\ in\ Treatment_{j1}$ are included in the specification the closer we can correctly approximate the nonparametric relationship. That is there is a conflict between the higher order required by the effiency of the estimator and the number of observations and the size of the vector of school inputs.

accounts for between 50-80% of the estimated treatment effect. These estimates are calculated using the nonparametric estimates presented in Appendix Figure 2. Estimates of the context independent portion of the treatment effect appear in the bottom panel of Table 6. On average, these are small in magnitude and are not statistically significant at conventional levels. Last, we conducted policy experiments where we extrapolated out of sample to calculate $\tau^{(1,0)(0,0)}$ and $\tau^{(1,1)(0,1)}$. In both cases we found small impacts and using bootstrapped samples we computed standard errors that indicate that class size did not significantly affect achievement.

Results on the linear difference in difference appear in Table X. Briefly and this will be rewritten is that the main story echoes above when quartile 4 is used, quartile 3 does suggest roughly 40-50% of the effect is context independent and statistically significant. Other big point is context specific in two of the panels in math is huge. Present in an appendix t-tests of common support LPT4.doc file and discuss why common trend and no anticipation effects have to hold in this setting.

4.4 Discussion

Since context specific effects may arise in educational experiments, researchers may wish to ex-ante consider strategies to design experiments in a manner to minimize these complications. For example, researcher can now randomize treatment both across and within schools. In typical experimental studies, the randomization is often done within the multiple sites (e.g. classsrooms, school or school districts) where the study is being conducted and the data is subsequently pooled across sites for the analysis. Randomizing within a site (often called blocking in the experimental design literature) makes one more certain that unobserved characteristics (i.e block factors) are balanced between the treatment and control groups. Randomizing across sites would ensure that in some schools everyone gets

treatment and in other schools no one receives the treatment. If one assumes that at the extremes, context specific effects are likely minimized then a matching metric could be used to measure similarity between the schools. However, it is worth noting that randomizing across schools may reopen other critiques as one may be concerned that unobserved characteristics are not balanced across the matched schools.[31]

5 Conclusion

Many social experiments occur at multiple sites with variable treatment intensity, in which participants have preference for a particular outcome, they not only know they are being studied but can easily correctly verify their treatment status. In this paper, we exploit the random variation in the fraction of individuals offered treatment across locations (schools in Project STAR) to explain treatment effect heterogeneity witnessed in the study. This variation in the experiment was not planned by the experimenters. We argue that if class size effect varies in a systematic manner with this uncontrolled difference in the fraction in treatment, it has important implications for external validity of the estimated effect. Intuitively, in a medical trial one would want to know how much of a causal effect is a pharmacological effect and how much a placebo effect. Here we want to understand how much of the class size treatment effect in Project STAR is context specific and context independent, where the context is the uncontrolled fraction of individuals offered treatment. We postulate that context specific effects have less external validity than context independent effects. Decomposing the amount of context-independent effects is clearly of importance not just for the immediate cost-benefit exercise but also for policy

[31] That being said, it would be simple to apply Rosenbaum (2002) bounds to demonstrate the sensitivity of the ATE to the presence of such a hidden bias, if such a bias were to exist, while remaining agnostic about the presence of this bias.

discussion of large-scale implementation. After all, the effects from different sources may have different external validity.

With data from Project STAR, we present evidence of treatment effect heterogeneity being driven by the performance of students in treatment classes within schools with low treatment intensity. Assuming that class size effects are additively separable from context specific effects we find that the context specific effects account for between 50-70% of the estimated treatment effect in Project STAR. We believe the method introduced in this paper have wide applicability and can be included not only in an evaluator's but also an empirical economist's tool box.

There are several other points that deserve some emphasis. First, the effects of being observed by others is likely very different than being observed by an individual who is implementing an experiment. Only in zero blind experiments does this concern arise. A growing body of research in behavioral economics indicates that social image concerns are a motivator of both pro-social behavior and contributions to public goods.[32] Yet, the extent to which these social image concerns are important is likely school specific and research is needed to understand how educators use social norms to create focal points that stigmatize certain behaviors more than others within schools. That is, to the extent that a teacher's social image can be manipulated to increase her effort, the

[32]For example, Neckermann and Frey (2007), in an experiment within a corporate setting, find that awards given to workers who contribute to a public good are more effective — in terms of the expressed intention of the subjects to contribute to the public good — when the awardees are made public. See Frey and Oberholzer-Gee (1997), Gneezy and Rustichini (2000) and Harbaugh (1998a, 1998b) for additional evidence from the laboratory. Benabou and Tirole (2006) have formalized these effects in a model where individuals perform altruistic activities to increase their social reputation and self-respec tand the provision of public goods. This line of research on the role of social image is consistent with claims in other disciplines (e.g. Goode 1978, Wedekind 1998, Nowak and Sigmund 2000, Price 2003).

level of success is likely sensitive to the details of the school environment in which she is employed. Second, this pathway differs substantially from the impact of an implementer. A common critique of experiments is that subjects may try to please experimenters by adjusting their behavior.[33] While double-blind experiments reduce the likelihood that implementer effects occur, the use of process evaluations to gain a qualitative component of the pathways through which treatments are effective, ensure that these issues also exist in zero-blind experiments. Third, standard empirical methods to estimate causal effects assume that the assumption of SUTVA is satisfied. Yet, if either context effects or social interactions are present, alternative methods such as that which is developed in this paper are required to identify causal parameters.

In this study, we only examined data from the first year of the STAR experiment. Future research is needed to develop strategies to identify context independent effects in general multi-period experiments as well as with Project STAR. There is an additional feature of the Project STAR public data that has not been exploited. During the experiment, STAR researchers also collected data on students from 23 comparison schools in which 0% of the students received treatment. Using this data, one could estimate educate production function and then use these structural parameters to predict outcomes in the schools where the experiment took place. The forecast errors could then capture the amount of additional effects arising from the experiment and this could be regressed on the percent treated in the school as in the third step of our empirical strategy. Alternatively, this data could be used in a validation study of the method developed in this paper. However, there are serious challenges in using this data as information on the kindergarten performance of the students in these schools remains unavailable. Data

[33] That being said teachers may interpret the experimenter's monitoring of their behavior as distrust and reduce their effort.

for students from these schools is currently only available for the 1st grade, which then leads to concerns on how one should deal with issues related to noncompliance in treatment assignment, selective attrition and the non-random assignment of students in the refreshment sample.[34] Developing strategies to overcome these methodological challenges presents an agenda for future research.

[34]See Ding and Lehrer (2010) for a discussion of estimating treatment effects in contaminated multi-period experiments.

References

[1] Akerlof, G. (1982), "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics* 97(4), 543–569.

[2] Bandiera, O., I. Barankay and I. Rasul (2005), "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *Quarterly Journal of Economics* 120(3), 917-962.

[3] Ben-Porath, Y. (1967), "The Production of Human Capital and the Life-Cycle of Earnings." *Journal of Political Economy* 75(4), 352-365.

[4] Boardman, A. E. and R. J. Murnane (1979), "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(1), 113-121.

[5] Cook, T. D. and M. R. Payne (2002), "Objecting to the Objections to Using Random Assignment in Educational Research," in *Evidence Matters: Randomized Trials in Education Research* edited by F. Mosteller and R. F. Boruch. Washington, D.C.: Brookings Institute Press.

[6] Deaton,A., (2010), "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature* 48(2), 424–455.

[7] Ding, W. and S. F. Lehrer (2010), "Estimating Treatment Effects from Contaminated Multi-Period Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics* 92(1), 31-42.

[8] Duflo, E. and E. Saez (2003), "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence From a Randomized Experiment," *Quarterly Journal of Economics* 118(3), 815-842.

[9] Encinosa III, W. E., M. Gaynor, and J. B. Rebitzer (2007), "The Sociology of Groups and the Economics of Incentives: Theory and Evidence on Compensation Systems," *Journal of Economic Behavior & Organization* 62(2), 187-214.

[10] Frey, B.S. and Oberholzer-Gee, F., 1997: "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," American Economic Review, 87, 4, 746-755.

[11] Harbaugh, W. T. (1998a), "What do Donations Buy? A Model of Philanthropy Based on Prestige and Warm Glow," Journal of Public Economics, 67, 269-284. H

[12] Harbaugh, W. T. (1998b), "The Prestige Motive for Charitable Transfers," American Economic Review, 88, 2, 277-282.

[13] Ferracci, M., G. Jolivet, and G. J. van den Berg (2014), "Treatment Evaluation in the Case of Interactions within Markets," *Review of Economics and Statistics* 96(5), 812–823.

[14] Finn, J. D., and C. M. Achilles (1990), "Answers about Questions about Class Size: A Statewide Experiment," *American Educational Research Journal* 27(3), 557 - 577.

[15] Finn, J. D.; Gerber, S. B., Achilles, C. M. and J. Boyd-Zaharias (2001), "The Enduring Effects of Small Classes," *Teachers College Record* 103(2), 145-83.

[16] Finn, J. D., J. Boyd-Zaharias, R. M. Fish, and S. B. Gerber (2007), Project STAR and Beyond: Database Users Guide, Heros, Inc., available online.

[17] Gneezy, U., and A. Rustichini, 2000: "Pay Enough or Don't Pay At All," Quarterly Journal of Economics, 115, 3, 791-810.

[18] Goldstein, H. and P. Blatchford (1998), "Class Size and Educational Achievement: A Review of Methodology with Particular Reference to Study Design," *British Educational Research Journal* 24(3), 255-268.

[19] Graham, B. S., G. W. Imbens, and G. Ridder (2006), "Complementarity and Aggregate Implications of Assortative Matching," *NBER Working Paper No. W14860.*

[20] Graham, B. S., G. W. Imbens, and G. Ridder (2009), "Measuring the Average Outcome and Inequality Effects of Segregation in the Presence of Social Spillovers," *mimeo*, New York University.

[21] Hanushek, E. A. (2003), "The Failure of Input Based Schooling Policies" *The Economic Journal* 113(483), 64-98.

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

[23] Hattie, J. (2005), "The Paradox of Reducing Class Size and Improving Learning Outcomes," *International Journal of Educational Research* 43(2), 387–425

[24] Heckman, J. J. (1992), "Randomization and social policy evaluation," in Evaluating Welfare and Training Programs, editors Charles Manski and I. Garfinkel. Cambridge, MA: Harvard University Press..

[25] Hesselius, P., P Johansson and J. P. Nilsson (2009) "Sick of Your Colleagues' Absence?," *Journal of the European Economic Association* 7(2-3), 583-594.

[26] Hoxby, C. M. (2000), "The Effects of Class Size on Student Achievement: New Evidence from Population Variation," *Quarterly Journal of Economics* 115(4), 1239-1285.

[27] Hudgens, M. G., and M. E. Halloran (2008), "Toward Causal Inference With Interference," *Journal of the American Statistical Association* 103(482), 832-842.

[28] Ichino, Andrea, and Giovanni Maggi (2000)."Work Environment And Individual Background: Explaining Regional Shirking Differentials In A Large Italian Firm", *Quarterly Journal of Economics* 115, 1057-1090.

[29] Jackson, C. K. and E. Bruegmann (2009), "Teaching Students and Teaching Each Other: the Importance of Peer Learning for Teachers," forthcoming in the *American Economic Journal: Applied Economics.*

[30] Jepsen, C., and S. Rivkin (2009), "Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size," *Journal of Human Resources* 44(1), 223–250.

[31] Kandel, E., and E. Lazear (1992), "Peer Pressure and Partnerships," *Journal of Political Economy* 100(4), 801-17.

[32] Koedel, C. (2008), "An Empirical Analysis of Teacher Spillover Effects in Secondary School,", forthcoming in the *Economics of Education Review.*

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

[34] Lavy, V. (2002), "Evaluating the Effects of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy* 110(6), 1286-1317.

[35] Lechner, M. (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," *Review of Economics and Statistics*, 84(2), 205-220.

[36] Lechner, M. (2004) "Sequential Matching Estimation of Dynamic Causal Models." Working Paper, University of St. Gallen.

[37] Manski, C. F. (2009), "Identification Of Treatment Response With Social Interactions," *mimeo,* Northwestern University.

[38] Mas, A. and E. Moretti (2009), "Peers at Work", *American Economic Review* 99(1), 112-145.

[39] Nowak, M.A. and Sigmund, K., (2000), "Shrewd Investments," Science, 288, 5467, 819-820.

[40] Price, M. K. (2003), "Pro-Community Altruism and Social Status in a Shuar Village," Human Nature, 14, 2, 191-195.

[41] Ritter, G. W. and R. F. Boruch (1999), "The Political and Institutional Origins of a Randomized Controlled Trial on Elementary School Class Size: Tennessee's Project STAR," *Educational Evaluation and Policy Analysis* 21(2), 111-125.

[42] Rosenbaum, P. R., (2002), "*Observational Studies*," Second Edition, New York: Springer.

[43] Rotemberg, Julio (1994). "Human relations in the workplace", Journal of Political Economy 102, 684—718.

[44] Sojourner, A. (2008), "Inference on Peer Effects with Missing Peer Data: Evidence from Project STAR," *mimeo,* Northwestern University.

[45] van der Berg, G. (2007), "An Economic Analysis of Exclusion Restrictions for Instrumental Variable Estimation," *IZA Discussion Papers #2585.*

[46] Wedekind, K. (1998), "Give and Ye Shall Be Recognized," Science, 280, 5372, 2070-2071.

[47] Word, E. e. a. (1990), "The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990," *Discussion Paper*, Tennessee State Department of Education.

Figure 1: Histogram and Kernel Density Estimate of the Number of Schools Based on Fractions of Students in Treatment Classes in Kindergarten.

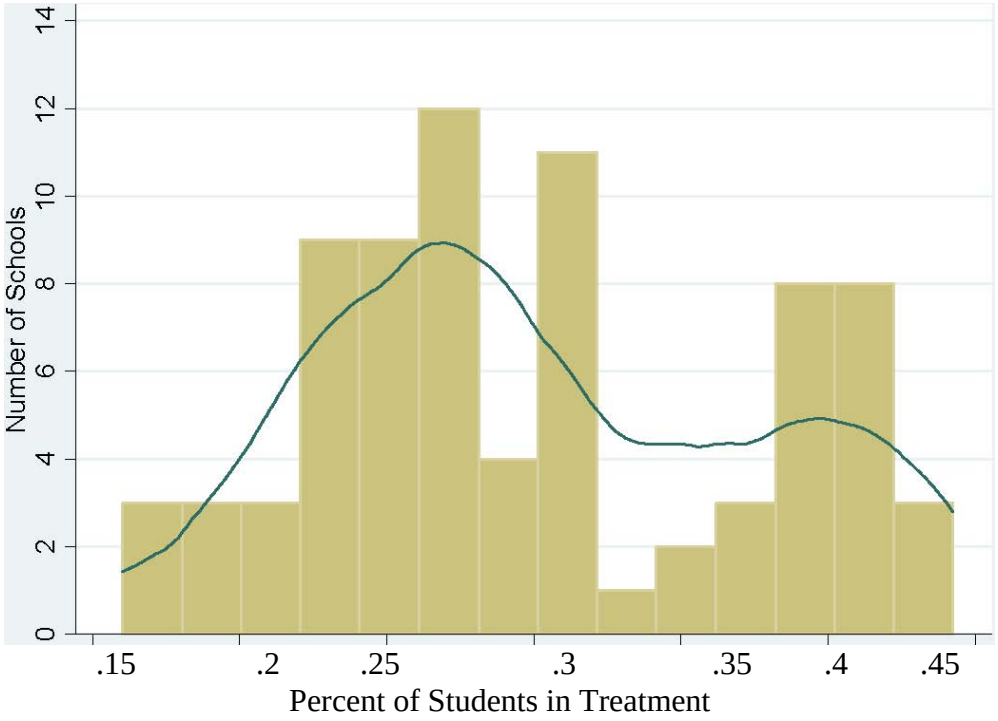


Table 1: Summary Statistics

Variable	Full sample	Small Classes	Regular Classes
Small Class Treatment	.3004 (.4585)	1 (0)	0 (0)
Student is on Free Lunch	.4846 (.4996)	.4705 (.4992)	.4907 (.4997)
Female Student	.4862 (.4998)	.4858 (.4999)	.4863 (.4999)
Teaching Experience	9.2683 (5.8017)	8.9195 (5.8129)	9.4181 (5.7912)
Teacher has a Master's Degree	.3471 (.4752)	.3137 (.4641)	.3615 (.4792)
Teacher is White	.1638 (.3701)	.1389 (.346)	.1745 (.3796)
Percentage of Kids Within the School Receiving Treatment	.3004 (.0729)	.3181 (.0722)	.2928 (.0718)
Kindergarten Cohort Size	88.5311 (28.6616)	88.0005 (28.1191)	88.7589 (28.8915)
Number of Class Rooms	4.4895 (1.3314)	4.5047 (1.308)	4.4829 (1.3415)
Number Small Class Rooms	1.7502 (.647)	1.8384 (.6512)	1.7123 (.6415)
Math Test Scores	485.3771 (47.6979)	490.9313 (49.5101)	482.9954 (46.7035)
Reading Test Scores	436.7253 (31.7063)	440.5474 (32.4974)	435.0842 (31.2212)
Word Recognition Test Scores	434.1793 (36.7588)	438.1362 (37.4366)	432.4839 (36.3374)
Listening Skills Test Scores	537.4746 (33.1397)	539.8568 (33.1593)	536.452 (33.0828)
White Student	.6695 (.4704)	.6811 (.4662)	.6645 (.4721)
African American Student	.3258 (.4687)	.3126 (.4637)	.3314 (.4707)
Hispanic Student	.0008 (.0281)	.0021 (.0458)	.0002 (.015)
Asian Student	.0022 (.047)	.0016 (.0397)	.0025 (.0498)
Student has another race	.0014 (.0377)	.0021 (.0458)	.0011 (.0336)
Student is Native American	.0003 (.0178)	.0005 (.0229)	.0002 (.015)
Class size kindergarten	20.3382 (3.9806)	15.1168 (1.4981)	22.5801 (2.2251)
School is in urban district	.0898 (.2859)	.0958 (.2944)	.0872 (.2822)
School is in the inner city	.2258 (.4181)	.2105 (.4078)	.2323 (.4224)
School is in rural disctrict	.4612 (.4985)	.4537 (.498)	.4644 (.4988)
School is in suburban district	.2232 (.4165)	.24 (.4272)	.216 (.4116)
Teacher is new and has no prior experience	.0477 (.2132)	.0316 (.1749)	.0547 (.2274)
Student Attrits from the Study at Some Point	.5126 (.4999)	.4874 (.5)	.5234 (.4995)
Student does not Comply with Treatment Assignment	.1051 (.3064)	.1037 (.3049)	.1058 (.3076)
Student Attrits from study After Kindergarten	.2862 (.452)	.2632 (.4405)	.296 (.4566)
Student does not Comply with Treatment Assignment	.0563 (.2309)	.0568 (.2316)	.056 (.23)
Observations	6325	1900	4425

Note: Standard deviations in parentheses

Table 2: Estimates of the Average Impact of Assignment to Small Classes in Kindergarten

	Mathematics	Reading	Word Recognition	Listening Skills
Full Sample	8.690 (2.014)***	5.966 (1.274)***	6.335 (1.411)***	3.570 (1.207)***
Treatment Effect Estimates by Quartile of Figure 1				
Quartile 1 (16.1-24.1%)	19.475 (5.368)***	10.472 (3.488)***	8.171 (3.245)**	6.889 (2.767)**
Quartile 2 (24.8-28.6%)	8.240 (3.718)**	10.033 (2.567)***	11.880 (2.961)***	4.534 (273)*
Quartile 3 (29.5-38.1%)	8.439 (3.387)**	3.876 (2.114)*	5.109 (2.497)**	2.496 (1.864)
Quartile 4 (38.2-44.3%)	4.930 (4.070)	3.361 (2.550)	3.035 (2.878)	2.720 (2.501)

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications include school fixed effects, student characteristics and teacher demographics. Estimates of other education inputs are presented in Appendix Table 1.

Table 3: How does the percentage of treatment within a school influence outcomes in small and regular classrooms?

Subject Area	Mathematics	Reading	Word Recognition	Listening Skills
Small Class Students Only				
Quartile 2 of Figure 1	-9.057 (8.868)	1.164 (5.405)	2.221 (5.568)	-1.091 (4.696)
Quartile 3 of Figure 1	-14.935 (7.984)*	-4.939 (5.183)	-2.020 (5.510)	-4.872 (4.109)
Quartile 4 of Figure 1	-17.051 (8.099)**	-9.330 (5.111)*	-9.914 (5.383)*	-3.886 (4.154)
White or Asian Student	10.223 (4.776)**	4.323 (2.888)	3.472 (3.219)	19.390 (3.072)***
Student is on Free Lunch	-20.957 (2.698)***	-15.534 (1.863)***	-17.871 (2.095)***	-13.982 (1.601)***
Female Student	3.009 (2.200)	4.787 (1.505)***	3.144 (1.728)*	2.798 (1.411)**
Teaching Experience	-0.083 (0.381)	0.023 (0.252)	0.111 (0.297)	-0.199 (0.206)
Teacher has a Master's Degree	-0.147 (4.714)	-1.880 (3.147)	0.186 (3.718)	-0.046 (2.667)
Teacher is White	6.642 (7.082)	1.700 (4.560)	-1.610 (4.919)	5.668 (4.525)
Constant	504.284 (9.798)***	447.010 (5.979)***	445.154 (6.335)***	535.885 (5.213)***
Observations	1762	1739	1755	1753
R-squared	0.08	0.09	0.08	0.15
Regular Class Students Only				
Quartile 2 of Figure 1	3.637 (5.537)	3.129 (3.495)	0.085 (3.866)	1.903 (2.856)
Quartile 3 of Figure 1	-5.881 (4.634)	0.555 (2.953)	-0.135 (3.258)	-2.508 (2.649)
Quartile 4 of Figure 1	-0.980 (4.320)	-1.656 (2.756)	-4.409 (3.379)	0.231 (2.603)
White or Asian Student	11.021 (4.664)**	5.396 (2.627)**	6.023 (2.798)**	17.860 (2.266)***
Student is on Free Lunch	-19.561 (2.149)***	-15.277 (1.296)***	-17.431 (1.471)***	-15.438 (1.377)***
Female Student	9.460 (1.508)***	6.630 (0.986)***	6.704 (1.203)***	3.208 (0.934)***
Teaching Experience	1.070 (0.392)***	0.736 (0.244)***	0.673 (0.264)**	0.789 (0.247)***
Teacher has a Master's Degree	-4.983 (3.549)	-1.542 (2.449)	-1.680 (2.681)	-2.261 (1.943)
Teacher is White	7.236 (6.438)	4.024 (3.641)	3.786 (3.903)	3.528 (3.404)
Constant	472.057 (5.867)***	428.015 (3.546)***	428.199 (3.962)***	523.394 (3.041)***
Observations	4109	4050	4096	4084
R-squared	0.10	0.11	0.10	0.19

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects.

Table 4: Checking Randomization

	Full sample	Schools in Quartile 1 of Figure 1	Schools in Quartile 2 of Figure 1	Schools in Quartile 3 of Figure 1	Schools in Quartile 4 of Figure 1
White or Asian Student	0.003 (0.020)	0.048 (0.051)	0.017 (0.031)	0.001 (0.032)	-0.043 (0.042)
Student is on Free Lunch	-0.011 (0.014)	0.008 (0.027)	-0.023 (0.025)	-0.022 (0.022)	0.004 (0.030)
Female Student	0.001 (0.010)	-0.012 (0.018)	-0.012 (0.019)	0.027 (0.021)	-0.031 (0.019)
Teaching Experience	-0.003 (0.005)	-0.001 (0.009)	-0.007 (0.010)	-0.002 (0.009)	0.002 (0.011)
Teacher has a Master's Degree	-0.061 (0.068)	0.038 (0.138)	0.001 (0.143)	0.003 (0.127)	-0.327 (0.120)***
Teacher is White	-0.063 (0.102)	-0.199 (0.202)	0.149 (0.218)	0.076 (0.174)	-0.496 (0.230)**
Constant	0.361 (0.056)***	0.233 (0.128)*	0.314 (0.113)***	0.325 (0.103)***	0.617 (0.117)***
Observations	6325	1433	1499	1907	1486
R-squared	0.03	0.02	0.01	0.01	0.10

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects.

Table 5: Linear Probability Model Estimates of How Do Ex-Post Implementation Failures Vary Across Schools Based on Kindergarten Treatment Assignment?

	Full sample	Schools in Quartile 1 of Figure 1	Schools in Quartile 2 of Figure 1	Schools in Quartile 3 of Figure 1	Schools in Quartile 4 of Figure 1
Does Not Complying with Treatment Assignment Immediately After Kindergarten					
Small Class Treatment	0.002 (0.009)	-0.052 (0.025)**	-0.047 (0.012)***	-0.013 (0.009)	0.084 (0.020)***
Attrits from the Experiment Immediately After Kindergarten					
Small Class Treatment	-0.037 (0.013)***	-0.112 (0.025)***	-0.115 (0.026)***	-0.058 (0.022)**	0.085 (0.027)***
Does Not Complying with Treatment Assignment at Some Point After Kindergarten					
Small Class Treatment	0.000 (0.011)	-0.044 (0.032)	-0.042 (0.019)**	-0.004 (0.015)	0.053 (0.022)**
Attrits from the Experiment at Some Point After Kindergarten					
Small Class Treatment	-0.053 (0.013)***	-0.077 (0.027)***	-0.110 (0.027)***	-0.066 (0.023)***	0.017 (0.024)
Observations	6325	1433	1499	1907	1486

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include specifications include school fixed effects, student characteristics and teacher demographics.

Table 6A: Weighted Estimates of the Context Specific Treatment Effects by Quartile of Figure 1

	Math	Reading	Word Recognition	Listening skills
Quartile 1	12.010 (8.262)	8.428 (4.768)	6.973 (6.881)	3.945 (3.812)
Quartile 2	9.688 (2.775)	10.200 (1.444)	11.638 (1.753)	6.162 (1.251)
Quartile 3	6.101 (2.656)	4.403 (2.591)	5.914 (3.234)	3.858 (1.856)
Quartile 4	4.147 (3.708)	3.211 (4.030)	2.926 (4.515)	1.516 (2.259)

Table 6B: Weighted Estimates of the Context Independent Treatment Effects by Quartile of Figure 1

	Math	Reading	Word Recognition	Listening skills
Quartile 1	7.735 (9.853)	1.955 (5.923)	1.063 (7.618)	2.722 (4.734)
Quartile 2	-2.529 (3.977)	-0.242 (2.996)	0.264 (3.453)	-1.779 (2.773)
Quartile 3	2.457 (4.317)	-0.439 (3.349)	0.697 (4.090)	-1.231 (2.640)
Quartile 4	0.896 (4.558)	0.159 (4.776)	0.100 (5.360)	1.312 (3.390)

Appendix Table 1: Plan for Distribution of Students and Classes in Within-School Design: Project STAR (1985-1986)

Design Type	Enrollment (ADM)	Classes (N)	Extra room Needed
One	57-67	(3) S,R,R/A	No
Two	68-78	(4) S,S,R,R/AA	Yes
Three	79-92	(4) S,R,R/A,R/A or S,R,R,R/A	No
Four	93-109	(5) S, S,R,R,R/A or S,S,R,R/A,R/A	Yes
Five	110-134	(6) S,S,R,R,R/A,R/A or	Yes
Six	135+	Individually designed	Yes

S=Small Class (1 :13-17);R=Regular Class (1:22-25);
R/A=Regular Class with a Full-time Teacher Aide (1:22-25)

Note: This is a simply a reproduction of table II-1 from the Project STAR Technical Report

Appendix Table 2: Estimates of the Full Education Production

	Full Sample	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Mathematics					
Small Class Treatment	8.690 (2.014)***	19.475 (5.368)***	8.240 (3.718)**	8.439 (3.387)**	4.930 (4.070)
White or Asian Student	16.965 (2.395)***	15.872 (5.372)***	20.001 (5.584)***	14.321 (3.140)***	18.038 (5.670)***
Student is on Free Lunch	-20.039 (1.328)***	-19.294 (2.896)***	-20.711 (2.817)***	-18.473 (2.496)***	-22.860 (2.626)***
Female Student	6.416 (1.124)***	9.120 (2.697)***	5.913 (2.144)***	6.593 (1.994)***	4.284 (2.262)*
Teaching Experience	0.430 (0.202)**	0.301 (0.446)	0.683 (0.352)*	0.197 (0.342)	0.623 (0.427)
Teacher has a Master's Degree	-2.089 (2.071)	-2.434 (4.094)	-2.495 (4.487)	-3.256 (3.661)	0.127 (4.715)
Teacher is White	0.931 (3.828)	9.609 (8.086)	-18.253 (5.620)***	0.035 (5.985)	15.394 (8.725)*
Constant	474.450 (2.821)***	474.715 (6.274)***	476.386 (5.258)***	474.063 (4.995)***	470.654 (6.034)***
Observations	5871	1326	1386	1787	1372
R-squared	0.27	0.25	0.35	0.27	0.19
Reading					
Small Class Treatment	5.966 (1.274)***	10.472 (3.488)***	10.033 (2.567)***	3.876 (2.114)*	3.361 (2.550)
White or Asian Student	7.929 (1.613)***	7.893 (2.776)***	10.998 (3.600)***	5.371 (2.675)**	7.320 (3.381)**
Student is on Free Lunch	-14.669 (0.904)***	-13.935 (1.988)***	-14.760 (1.994)***	-14.948 (1.653)***	-15.443 (1.597)***
Female Student	5.406 (0.780)***	7.211 (1.708)***	5.948 (1.476)***	4.927 (1.396)***	3.927 (1.720)**
Teaching Experience	0.303 (0.126)**	0.543 (0.262)**	0.397 (0.261)	0.301 (0.218)	0.036 (0.233)
Teacher has a Master's Degree	-0.689 (1.254)	-1.753 (2.186)	-0.528 (3.105)	0.341 (2.376)	-0.451 (2.632)
Teacher is White	0.403 (2.726)	8.814 (5.407)	-14.379 (6.808)**	-0.178 (3.938)	7.146 (4.782)
Constant	431.381 (1.843)***	426.009 (3.125)***	431.289 (3.875)***	433.659 (3.489)***	431.805 (3.799)***
Observations	5789	1323	1359	1761	1346
R-squared	0.27	0.27	0.36	0.26	0.19
Word Recognition					
Small Class Treatment	6.335 (1.411)***	8.171 (3.245)**	11.880 (2.961)***	5.109 (2.497)**	3.035 (2.878)

White or Asian Student	7.172 (1.915)***	5.751 (3.568)	12.452 (4.204)***	3.611 (3.342)	6.660 (3.628)*
Student is on Free Lunch	-15.904 (1.067)***	-16.531 (2.347)***	-15.428 (2.286)***	-15.922 (2.002)***	-16.043 (1.929)***
Female Student	5.027 (0.937)***	7.637 (1.959)***	5.608 (1.888)***	4.335 (1.729)**	2.883 (1.973)
Teaching Experience	0.310 (0.139)**	0.668 (0.264)**	0.286 (0.317)	0.364 (0.239)	0.031 (0.266)
Teacher has a Master's Degree	0.321 (1.478)	-0.360 (3.178)	2.499 (3.900)	-0.033 (2.542)	-0.041 (3.107)
Teacher is White	-0.652 (3.148)	4.838 (5.753)	-14.638 (9.057)	-1.047 (4.609)	6.859 (5.573)
Constant	429.786 (2.215)***	426.592 (4.478)***	426.571 (4.597)***	433.566 (4.220)***	429.291 (4.326)***
Observations	5851	1317	1379	1781	1374
R-squared	0.23	0.25	0.28	0.24	0.18
Listening Skills					
Small Class Treatment	3.570 (1.207)***	6.889 (2.767)**	4.534 (2.473)*	2.496 (1.864)	2.720 (2.501)
White or Asian Student	18.011 (1.702)***	17.496 (4.138)***	19.626 (3.776)***	16.554 (2.562)***	17.846 (3.713)***
Student is on Free Lunch	-15.147 (0.903)***	-13.240 (1.934)***	-15.284 (1.896)***	-15.524 (1.743)***	-16.763 (1.695)***
Female Student	2.680 (0.738)***	4.693 (1.879)**	1.318 (1.379)	2.017 (1.252)	2.989 (1.451)**
Teaching Experience	0.243 (0.150)	0.487 (0.471)	0.532 (0.215)**	-0.027 (0.205)	0.238 (0.251)
Teacher has a Master's Degree	0.753 (1.238)	7.233 (2.702)***	0.083 (2.874)	-0.002 (1.924)	-1.221 (2.447)
Teacher is White	3.796 (2.613)	16.455 (5.788)***	-13.340 (4.721)***	4.293 (3.637)	6.500 (4.056)
Constant	527.113 (1.915)***	517.124 (5.561)***	527.683 (3.746)***	529.932 (2.760)***	528.295 (3.698)***
Observations	5837	1316	1374	1776	1371
R-squared	0.26	0.22	0.29	0.26	0.29

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects.

Appendix table 3: The Impact of Small Class Across Schools Defined by Quartiles of the Size of the Kindergarten Student Body

	Mathematics	Reading	Word Recognition	Listening Skills
Full Sample	8.690 (2.014)***	5.966 (1.274)***	6.335 (1.411)***	3.570 (1.207)***
Quartile 1	11.495 (4.296)***	5.127 (2.658)*	6.699 (3.197)**	5.623 (3.166)*
Quartile 2	3.513 (3.881)	5.408 (2.720)*	7.773 (3.098)**	2.743 (2.369)
Quartile 3	8.473 (4.459)*	8.771 (2.635)***	6.980 (3.049)**	4.636 (2.258)**
Quartile 4	12.157 (3.493)***	5.882 (2.270)**	6.302 (2.453)**	3.410 (2.139)

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects as well as all the student and teacher controls used in Appendix table 1.

Appendix table 4: Does Attending a Small Class make you more likely not to write an examination in kindergarten?

	Full sample	Quartile 1 of Figure 1	Quartile 2 of Figure 1	Quartile 3 of Figure 1	Quartile 4 of Figure 1
Math	0.002	0.024	-0.003	-0.015	0.015
	(0.007)	(0.015)	(0.015)	(0.010)	(0.012)
Reading	-0.001	0.022	-0.020	-0.025	0.032
	(0.008)	(0.016)	(0.017)	(0.012)**	(0.020)
Word Recognition	0.003	0.020	-0.004	-0.014	0.022
	(0.007)	(0.017)	(0.016)	(0.010)	(0.013)*
Listening	0.001	0.019	-0.000	-0.018	0.017
	(0.007)	(0.017)	(0.015)	(0.011)*	(0.012)

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects as well as all the student and teacher controls used in Appendix table 1.

Appendix table 5: School Characteristics by Quartile of Figure 1

Variable	Quartile 1 of Figure 1	Quartile 2 of Figure 1	Quartile 3 of Figure 1	Quartile 4 of Figure 1
Total Class Rooms for Kindergarten	3.9444 (1.3492)	3.619 (1.1609)	4.800 (1.0563)	4.100 (1.0208)
Kindergarten cohort size	79.6111 (29.7245)	71.381 (24.9028)	95.35 (23.2385)	74.300 (21.704)
School has a grade range from K-6	0.5556 (0.5113)	0.4762 (0.5118)	0.4444 (0.5113)	0.550 (0.5104)
Total school enrollment kindergarten	636.1667 (242.1497)	529.6667 (133.8639)	680.6 (216.707)	529.6 (119.8865)
Percent students that are bussed kindergarten	50.8333 (43.6514)	53.5714 (37.894)	59.7 (35.683)	52.55 (34.2121)
Percent students that are white in kindergarten	76.75 (35.7445)	84.1111 (24.4538)	84.75 (23.0145)	77.8333 (33.9658)
Average number of students in small class	14.2509 (1.2406)	15.1523 (1.5745)	15.6284 (1.3211)	14.6545 (1.4639)
Average number of students in regular class	22.6198 (1.8269)	22.0378 (2.286)	22.7235 (1.9414)	21.4909 (2.201)
School is located in the inner city	.2222 (.4278)	0.2381 (0.4364)	0.200 (0.4104)	0.150 (0.3663)
School is located in urban district	.0556 (.2357)	0.0476 (0.2182)	0.150 (0.3663)	0.100 (0.3078)
School is located in suburban district	.2222 (.4278)	0.1429 (0.3586)	0.200 (0.4104)	0.350 (0.4894)
School is located in rural district	0.500 (0.5145)	0.5714 (0.5071)	0.450 (0.5104)	0.400 (0.5026)
Percent of students receiving free/reduced price lunch kindergarten	0.4351 (.2692)	0.5324 (0.2526)	0.4873 (0.291)	0.4301 (0.2824)
Percent of students that are African American	0.3072 (0.4255)	0.2746 (0.379)	0.3219 (0.4059)	0.2948 (0.4053)
Percent of students that are white or Asian	0.6909 (0.424)	0.7232 (0.3776)	0.6761 (0.4047)	0.7008 (0.4032)
Percent of students that are female	0.4952 (0.0799)	0.4874 (0.0468)	0.475 (0.0568)	0.4899 (0.0642)
Percent of teachers that are new to the profession	0.0213 (0.0626)	0.0696 (0.1352)	0.0296 (0.0729)	0.0638 (0.1157)
Average years of teaching experience	9.5706 (2.1194)	8.7422 (2.8779)	10.0478 (2.9871)	8.8037 (2.6499)
Percentage of teachers with master's degree	0.4248 (0.3359)	0.3387 (0.2913)	0.3227 (0.2421)	0.3602 (0.2327)

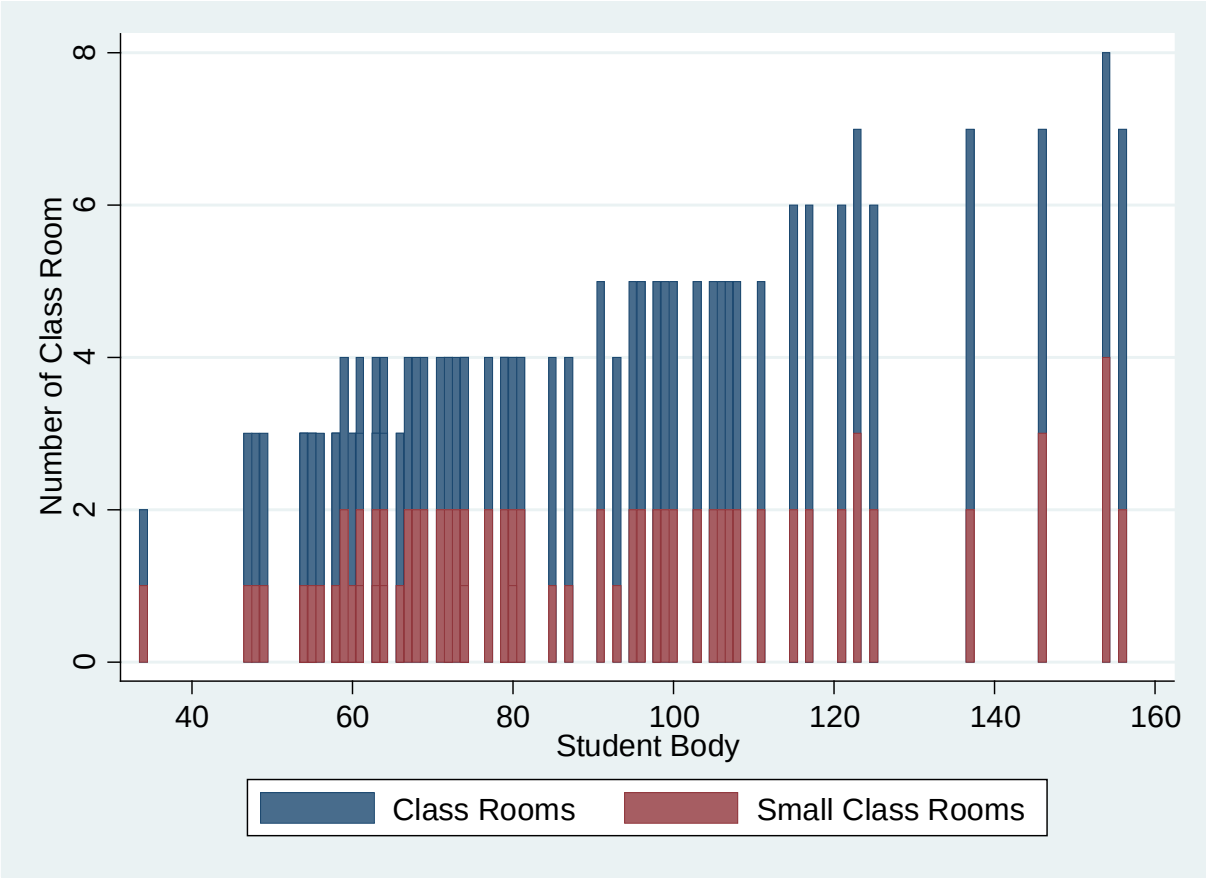
Note: Standard deviation in parentheses.

Appendix table 6: The Impact of Small Class Across Samples of Schools Defined by Number of Classrooms and Small Classes in Kindergarten

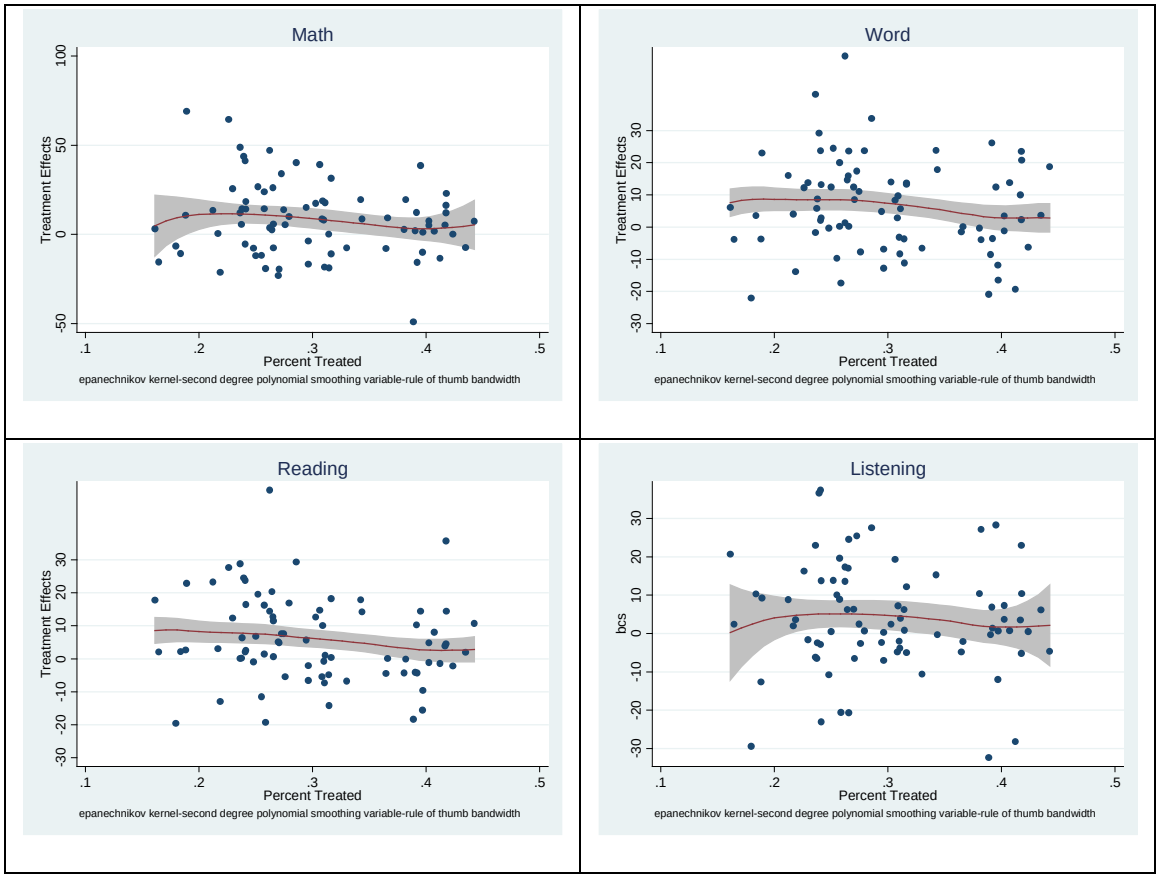
	3 Classrooms 1 Small Class 28 Schools	4 Classrooms 1 Small Class 6 Schools	4 Classrooms 2 Small Classes 20 Schools	5 Classrooms 2 Small Classes 15 Schools
Math				
Small Class Treatment	9.399 (3.501)***	12.581 (9.241)	3.189 (4.142)	8.417 (3.621)**
Observations	1508	456	1337	1436
R-squared	0.35	0.22	0.18	0.30
Reading				
Small Class Treatment	7.073 (2.461)***	11.660 (2.958)***	3.073 (2.567)	3.618 (2.317)
Observations	1494	456	1310	1401
R-squared	0.36	0.18	0.20	0.24
Word Recognition				
Small Class Treatment	8.185 (2.812)***	6.952 (3.311)**	3.022 (2.817)	5.014 (2.933)*
Observations	1502	454	1334	1425
R-squared	0.29	0.13	0.20	0.20
Listening				
Small Class Treatment	5.182 (2.436)**	4.515 (2.551)*	0.999 (2.438)	1.065 (1.966)
Observations	1499	454	1332	1420
R-squared	0.24	0.27	0.23	0.29

Note: Robust standard errors clustered at the classroom level in parentheses. ***, ** and * respectively denote statistical significance at the 1%, 5% and 10% level. The specifications also include school fixed effects as well as all the student and teacher controls used in Appendix table 1.

Appendix Figure 1: Histogram of the Classroom Breakdown by the Size of the Kindergarten Cohort Across Schools



Appendix Figure 2: Local Linear Regression Estimates of the Structural Model of the Total Treatment Effect



Note: This graph presents estimates from a local linear regression of school treatment on the percent of students receiving treatment that is weighted by the size of the student body who completed the subject examination in kindergarten. Both the context specific and context independent treatment effects as well as the standard errors are calculated from these estimates on the bootstrapped samples. Recall, bootstrapping is conducted at both stages of the estimation procedure.

Table X: Linear Difference in Differences Estimates of the Context Independent (Common small class size) and Context Specific Treatment Effects

Quartile 1 vs 4	Math	Reading	Word	Listen	Math	Reading	Word	Listen
Common Small Class Size Effect	3.101 (3.785)	2.746 (2.690)	2.367 (3.033)	2.062 (3.197)	4.428 (3.647)	3.466 (2.651)	3.136 (3.060)	3.944 (3.162)
Extra Small Class Effect Context	15.47 (8.167)	6.978 (4.908)	5.295 (5.135)	3.899 (5.881)	15.15* (7.409)	6.949 (4.155)	5.092 (4.709)	2.797 (4.920)
White or Asian Student					17.04** (4.536)	7.772** (2.734)	6.523* (2.874)	18.10** (3.199)
Free Lunch					-21.12** (2.114)	-14.74** (1.419)	-16.28** (1.747)	-15.04** (1.406)
Female					6.600** (1.644)	5.494** (1.074)	5.197** (1.204)	3.943** (1.262)
Teacher experience					0.486 (0.350)	0.270 (0.199)	0.315 (0.249)	0.305 (0.264)
Teacher has Masters					-0.766 (3.085)	-1.261 (2.003)	-0.619 (2.659)	2.780 (2.122)
Non-White Teacher					11.82* (5.286)	7.847* (3.203)	5.554 (3.496)	12.46** (3.824)
Constant	483.9** (1.069)	434.0** (0.686)	431.6** (0.750)	536.8** (0.825)	472.3** (4.333)	429.3** (2.899)	428.3** (3.482)	523.1** (3.133)
Observations	2,698	2,669	2,691	2,687	2,698	2,669	2,691	2,687
R-squared	0.015	0.011	0.005	0.004	0.081	0.083	0.062	0.092

Quartile 1 vs 3	Math	Reading	Word	Listen	Math	Reading	Word	Listen
Common Small Class Size Effect	9.034* (3.630)	4.296 (2.372)	5.466* (2.622)	3.025 (2.060)	8.306* (3.358)	3.736 (2.233)	5.015 (2.597)	2.444 (1.848)
Extra Small Class Effect Context	9.538 (8.096)	5.428 (4.741)	2.196 (4.904)	2.936 (5.349)	10.73 (7.813)	6.303 (4.109)	2.877 (4.513)	4.031 (4.368)
White or Asian Student					14.83** (2.554)	6.258** (1.777)	4.346 (2.174)	17.11** (1.811)
Free Lunch					-18.86** (2.185)	-14.53** (1.670)	-16.23** (1.873)	-14.64** (1.478)
Female					7.609** (1.581)	5.813** (1.079)	5.706** (1.247)	3.123** (1.144)

Teacher experience					0.227	0.395	0.473*	0.135
					(0.297)	(0.198)	(0.207)	(0.209)
Teacher has Masters					-2.917	-0.526	-0.240	2.537
					(2.985)	(1.810)	(2.103)	(1.623)
Non-White Teacher					3.467	3.047	0.939	8.643*
					(5.511)	(3.610)	(3.185)	(3.718)
Constant	481.5**	435.4**	433.9**	535.8**	475.1**	431.1**	431.2**	525.9**
	(0.923)	(0.570)	(0.606)	(0.578)	(4.363)	(2.940)	(3.057)	(3.244)
Observations	3,113	3,084	3,098	3,092	3,113	3,084	3,098	3,092
R-squared	0.019	0.011	0.007	0.004	0.077	0.080	0.065	0.084

Quartile 2 vs 4	Math	Reading	Word	Listen	Math	Reading	Word	Listen
Common Small Class Size Effect	3.101	2.746	2.367	2.062	3.197	2.450	2.401	1.975
	(3.781)	(2.688)	(3.030)	(3.193)	(3.659)	(2.795)	(3.249)	(3.018)
Extra Small Class Effect Context	4.429	6.807	9.030*	1.946	4.484	7.067	8.940	2.112
	(5.337)	(3.953)	(4.340)	(4.292)	(5.082)	(3.966)	(4.490)	(4.074)
White or Asian Student					19.06**	9.357**	9.768**	18.74**
					(4.724)	(3.203)	(3.333)	(3.324)
Free Lunch					-21.39**	-14.78**	-15.42**	-15.73**
					(2.257)	(1.276)	(1.602)	(1.483)
Female					5.062**	4.925**	4.280**	2.138*
					(1.321)	(0.931)	(1.242)	(1.032)
Teacher experience					0.670	0.205	0.131	0.370
					(0.368)	(0.205)	(0.238)	(0.217)
Teacher has Masters					-2.421	-1.589	0.211	-1.684
					(3.220)	(1.962)	(2.342)	(1.962)
Non-White Teacher					-2.828	-4.218	-4.206	-3.873
					(7.120)	(5.671)	(7.241)	(4.886)
Constant	484.4**	434.6**	430.6**	537.2**	474.6**	432.3**	428.7**	528.9**
	(0.905)	(0.658)	(0.730)	(0.742)	(4.193)	(2.947)	(3.569)	(2.668)
Observations	2,758	2,705	2,753	2,745	2,758	2,705	2,753	2,745
R-squared	0.004	0.012	0.012	0.002	0.076	0.084	0.065	0.094

Note Robust standard errors in parentheses clustered at the classroom level in parentheses.
Specifications include school fixed effects. ** p<0.01, * p<0.05

Table XI: Experiments with Simulated Teacher Quality

Q1-Uni						Q2-Uni				
	Math	Read	Word	List			Math	Read	Word	List
SC Mean	19.516	10.564	8.454	6.754		SC Mean	7.947	9.745	11.485	4.567
	(4.752)	(2.645)	(3.303)	(3.584)			(5.152)	(3.549)	(4.358)	(3.100)
SC SE	6.671	3.941	4.183	4.011		SC SE	5.556	3.801	4.378	3.542
	(0.691)	(0.587)	(0.406)	(0.637)			(0.649)	(0.483)	(0.552)	(0.400)
SC T-Stat	2.929	2.704	2.041	1.739		SC T-Stat	1.449	2.575	2.631	1.327
	(0.678)	(0.669)	(0.837)	(0.922)			(0.967)	(0.919)	(0.964)	(0.926)
TQ Mean	-0.101	-0.205	-0.670	0.313		TQ Mean	0.699	0.699	0.929	-0.091
	(11.978)	(6.649)	(8.275)	(8.877)			(7.007)	(8.931)	(10.982)	(7.881)
TQ SE	10.560	5.697	7.014	7.532		TQ SE	10.535	7.007	8.212	6.458
	(1.639)	(1.108)	(1.024)	(1.609)			(1.620)	(1.135)	(1.380)	(1.068)
TQ T-Stat	-0.008	-0.045	-0.090	0.039		TQ T-Stat	0.083	0.110	0.126	-0.004
	(1.150)	(1.180)	(1.221)	(1.145)			(1.296)	(1.294)	(1.353)	(1.244)
Obs	500	500	500	500		Obs	500	500	500	500

Q1-Norm						Q2-Norm				
	Math	Read	Word	List			Math	Read	Word	List
SC Mean	1.956	4.248	5.436	6.490		SC Mean	-5.12257	13.43032	15.45953	-6.624605
	(12.017)	(9.012)	(11.017)	(8.050)			11.26143	9.579055	10.24608	10.47923
SC SE	13.247	9.721	12.498	9.574		SC SE	13.28488	9.289468	10.8038	11.72162
	(1.453)	(1.217)	(1.418)	(1.114)			1.797091	1.428456	1.285409	1.879799
SC T-Stat	0.145	0.471	0.454	0.679		SC T-Stat	-0.400854 7	1.490673	1.453607	-0.540777 1
	(0.924)	(0.977)	(0.920)	(0.861)			0.891523 9	1.056671	0.966454 2	0.900820 9
TQ Mean	0.278	-0.013	-0.365	0.340		TQ Mean	1.350549	1.056017	0.983506 1	0.955829
	(8.918)	(6.713)	(8.184)	(5.992)			8.480564	7.166854	7.585485	7.87773
TQ SE	7.065	5.012	6.191	4.825		TQ SE	6.841961	5.121547	5.531753	6.084678
	(1.248)	(0.923)	(1.104)	(0.890)			1.412725	1.312146	1.237605	1.419621
TQ T-Stat	0.041	-0.010	-0.060	0.070		TQ T-Stat	0.183078 7	0.180597 7	0.164519 8	0.180076 2
	(1.329)	(1.385)	(1.370)	(1.341)			1.296053	1.417101	1.420643	1.322397
Obs	500	500	500	500		Obs	500	500	500	500

Q3-Uni						Q4-Uni				
	Math	Read	Word	List			Math	Read	Word	List
SC Mean	8.204	3.870	5.091	2.568		SC Mean	4.849	3.379	2.861	2.800
	(4.805)	(3.193)	(3.607)	(2.685)			(6.371)	(3.473)	(3.763)	(3.464)
SC SE	5.263	3.402	3.947	2.968		SC SE	6.536	3.905	4.359	3.935
	(0.532)	(0.354)	(0.417)	(0.395)			(0.858)	(0.532)	(0.537)	(0.506)
SC T-Stat	1.578	1.164	1.308	0.901		SC T-Stat	0.712	0.843	0.628	0.712
	(0.943)	(0.956)	(0.918)	(0.931)			(1.002)	(0.889)	(0.865)	(0.907)
TQ Mean	0.538	-0.003	0.032	-0.190		TQ Mean	0.241	-0.085	0.411	-0.228
	(11.529)	(7.694)	(8.674)	(6.435)			(15.939)	(8.640)	(9.387)	(8.672)
TQ SE	9.772	6.445	7.390	5.571		TQ SE	12.901	7.431	8.300	7.520
	(1.208)	(0.864)	(1.014)	(0.906)			(1.752)	(1.084)	(1.154)	(1.015)
TQ T-Stat	0.046	-0.005	-0.009	-0.037		TQ T-Stat	0.020	-0.010	0.053	-0.026
	(1.201)	(1.183)	(1.157)	(1.142)			(1.309)	(1.190)	(1.171)	(1.197)
Obs	500	500	500	500		Obs	500	500	500	500
Q3-Norm						Q4-Norm				
	Math	Read	Word	List			Math	Read	Word	List
SC Mean	10.50809	-0.9415233	-5.922727	3.440819		SC Mean	1.271197	-3.635062	-3.053818	-4.083434
	10.54665	7.039421	8.230388	9.12161			11.95849	8.712887	10.204	8.388023

SC SE	12.95026	7.90901	9.202165	9.145289		SC SE	11.79878	8.600587	10.67994	8.304938
	1.27855	0.836320 3	0.994365	1.092998			1.381435	1.08434	1.251288	0.977168 3
SC T-Stat	0.824965	-0.117431	-0.651566 3	0.405612 4		SC T-Stat	0.104430 3	-0.399760 5	-0.279978 2	-0.499180 3
	0.837657 3	0.905540 7	0.920800 9	1.010182			1.04183	1.030261	0.974438 8	1.035106
TQ Mean	0.729038 3	0.451293 7	0.472821 3	0.292005 4		TQ Mean	0.257901 7	0.075308 2	0.360359 7	-0.059590 5
	7.740682	5.165954	6.039573	6.705986			8.919268	6.575082	7.686274	6.27036
TQ SE	6.861655	4.419918	5.26052	5.300945		TQ SE	6.758767	4.94891	6.103601	4.669639
	1.09122	0.779736 4	0.882545 2	0.953094 2			1.109863	0.891104 6	1.078602	0.837530 9
TQ T-Stat	0.112560 7	0.103691 4	0.093507 6	0.065451 3		TQ T-Stat	0.042106 3	0.047667 5	0.080750 9	-0.015222 6
	1.147091	1.188086	1.181933	1.258336			1.381708	1.351363	1.29892	1.422135
Obs	500	500	500	500		Obs	500	500	500	500

Table XII: Estimates of Teacher Quality Effects of Graham (2006) Assuming no Peer Effects

Math								
Quartile	1	2	3	4	1	2	3	4
γ^2	5.97 (2.20)	1.74 (2.19)	1.27 (1.16)	4.58 (2.81)	4.39 (3.26)	1.36 (1.81)	4.82 (1.70)	0.11 .
Regular-with-aide	-58.39 (116.53)	74.48 (87.45)	-53.72 (80.70)	80.76 (124.31)	-16.37 (102.55)	59.85 (80.15)	-282.66 (42.48)	126.77 (133.09)
Large-school*small	- -	- -	- -	- -	182.77 (357.59)	41.95 (218.97)	-211.98 (128.32)	235.27 (131.97)
Large-school*regular-with-aide	- -	- -	- -	- -	-57.10 (197.34)	35.29 (215.25)	208.79 (112.01)	- -
p-value H0: γ^2 =1	0.03	0.74	0.82	0.21	0.30	0.84	0.03	.
F 1st-Stage	5.49	12.19	24.78	15.17	5.53	8.81	6.63	0.00
Number of classrooms	71	76	96	82	71	76	96	82
School fixed effects	y	y	Y	y	y	y	y	y
Reading								
Quartile	1	2	3	4	1	2	3	4
γ^2	8.03 5.06	2.12 3.10	-0.53 2.71	18.10 26.64	1.66 1.09	1.56 3.87	13.03 4.74	0.08 0.00
Regular-with-aide	-44.94 70.45	70.23 56.68	-17.42 55.25	81.94 210.14	-18.07 21.13	32.74 75.15	77.57 79.05	30.66 41.29
Large-school*small	- -	- -	- -	- -	325.73 260.96	24.93 64.35	-245.94 149.26	118.75 37.94
Large-school*regular-with-aide	-	-	-	-	-4.26	95.92	-20.25	-
p-value H0: γ^2 =1	-	-	-	-	84.85	96.86	138.38	-
F 1st-Stage	0.17	0.72	0.57	0.52	0.55	0.89	0.01	.
Number of classrooms	12.97	2.61	3.50	0.46	3.06	1.68	5.09	0.00
School fixed effects	71	76	96	82	71	76	96	82
School fixed effects	y	y	Y	y	y	y	y	y
Word Recognition								

Quartile	1	2	3	4	1	2	3	4
γ^2	5.10 3.42	1.64 2.58	0.02 3.20	12.68 9.84	1.34 1.58	0.05 3.28	10.13 5.56	0.14 0.00
Regular-with-aide	-22.14 55.87	71.90 66.81	-36.05 82.69	79.25 129.41	-16.90 38.27	37.68 98.74	100.75 106.70	47.41 46.89
Large-school*small	- -	- -	- -	- -	217.37 189.86	118.75 113.73	-219.86 161.47	176.05 54.88
Large-school*regular-with-aide	-	-	-	-	17.30	43.75	-26.40	-
p-value H0: γ^2 =1	-	-	-	-	81.12	103.00	170.86	-
F 1st-Stage	0.24	0.80	0.76	0.2399	0.83	0.77	0.11	.
Number of classrooms	10.42	4.66	3.96	1.97	3.47	2.90	2.11	0.00
School fixed effects	71	76	96	82	71	76	96	82
School fixed effects	y	y	Y	y	y	y	y	y
Listening Skills								
Quartile	1	2	3	4	1	2	3	4
γ^2	6.81 4.65	1.43 2.44	1.89 1.81	3.75 2.14	2.66 1.78	3.21 3.34	3.69 1.16	1.20 .
Regular-with-aide	101.52 108.66	-8.61 53.56	42.62 48.36	13.13 50.35	-37.75 34.36	48.15 59.82	3.98 18.28	16.49 50.18
Large-school*small	- -	- -	- -	- -	170.33 178.67	-93.25 118.73	-31.68 39.76	66.39 51.36
Large-school*regular-with-aide	- -	- -	- -	- -	206.32 172.45	-132.15 109.49	33.79 58.57	- -
p-value H0: γ^2 =1	0.22	0.86	0.62	0.2045	0.36	0.51	0.02	.
F 1st-Stage	10.99	7.57	11.46	21.59	8.83	3.33	8.75	0.00
Number of classrooms	71	76	96	82	71	76	96	82
School fixed effects	y	y	y	y	y	y	y	y