Estimating Context-Independent Treatment Effects in Education Experiments

Weili Ding
Queen's University

Steven F. Lehrer*
Queen's University and NBER

September 2010

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 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. Second, we introduce an empirical strategy 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 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.

Preliminary - Please Do Not Quote

*We would like to thank Aloyius Siow and Gerard van den Berg and seminar participants at the 2009 Yale-Fudan TED International Conference, 2010 Annual CLSRN conference, 2010 IOMS Conference, Queen's University, University of Manitoba, University of Mannheim, Syracuse University, Tilburg University and the Vrije University of Amsterdam for helpful comments and suggestions. Ding and Lehrer wish to thank SSHRC for independent research support. We are responsible for all errors.

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. However, some researchers remain skeptical that these experiments provide much value added to research and policy debates.[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]For example, Cook and Payne (2002) discuss and provide objections to a variety of arguments that appear in the education literature.

[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. 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

[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.

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 will attempt to 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.[4] 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.[5] 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

[4]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.

[5]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.

classes per se may contribute to student performance and may have important implications 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 50-70% 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.[6] 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).[7] 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 participate and sent guidelines for participation to each local system. These guidelines indicated that the state would cover additional costs for project teachers

[6] 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.

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

and teacher aides, but that local systems would furnish any additional classroom space needed. Thus, the number of small and regular classes within a school is determined by the size of the student body subject to potential space constraints. 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 process, organization, etc. other than class sizes. To participate, the schools had to agree to the random assignment of teachers and students to the different class conditions.[8] To implement the experimental design, researchers at Universities developed a protocol for the random assignment of students to class type.[9]

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 as Krueger (1999) reports that an examination of

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

[9] 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.

actual enrollment sheets that were compiled in the summer prior to the start of kindergarten for 1581 students from 18 participating STAR schools could only identify a single student in this sample who 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,[10] we only analyze data from the first year of the experiment. At the end of

[10]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.[11]

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.

[11] 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.[12] 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.[13] Inference is based on

[12] 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.

[13] 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.[14] 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 1 presents estimates that the impact of small class treatment is the only explanatory variable in equation (1) whose impact varies across samples based on quartiles of Figure 1.[15]

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

[14]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.

[15]This pattern of results in table 2 and Appendix table 1 is robust to other sized slices of school in the

There are many 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 conduct simple χ^2 test of difference between class types within schools and graph the p-values of these tests.[16] 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 quatile of Figure 1 appear that we cannot reject the null of random assignment with regards to the distribution of these observed characteristics.

We also conducted similar tests of difference between schools located in different quar-

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.

[16] This approach was 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.

tiles 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 the Null of. no differences between school type.[17] We further investigated how the class sizes varied across the student body. This is illustrated in Figure 2 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. Appendix Figure 1 presents a plot of the size of kindergarten class relative to the percent of students in treated classrooms and the raw Pearson correlation between these factors is -0.0279. Thus, we are confident that we are not simply finding a school size effect.

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 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.[18] Thus, we are quite confident the puzzle is

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

[18] 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 students off-task or time waiting for help in reading or mathematics. Similarly, they suggest that the

real and requires alternative explanation.

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.[19] 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.[20] 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 to increase their effort. We also estimate equation (2) for the control classes. Ex ante,

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.

[19] 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.

[20] 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.

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. Thus, the 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.

Last, it is worth noting that there are several interesting and statistically significant differences in ex-post implementation failures across the quartiles. 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.

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".

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.[21] 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

[21] 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.

study, but that the experiment alters the incentive conditions (or if you prefer, changes the context) for these subjects. This could result in the experiment concluding that certain policy intervention is effective but when fully enacted it no longer appears to be the case.

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.

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 50-70% 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 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.[22] 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

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

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 is more than two potential outcomes and represents a violation of the Stable Unit Treatment Value Assumption (SUTVA).[23] 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 $C_s = 0$. Lastly, we can calculate a 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.

With SUTVA violated, structural assumptions are needed to identify causal parameters.[24] We consider two empirical strategies (introduced and the assumptions required

[23] 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 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. 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.

[24] For instance Ferracci et al. (2010) use a two-step method that includes a matching approach with

for identification are discussed in the next section) 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 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.

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 makes one more certain that unobserved characteristics are balanced between the treatment and control groups. Randomizing across similar sites would ensure that only in some schools everyone gets treatment and in other school no one receives 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.[25]

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).

[25] 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.

4.2 Empirical Strategies to Identify Context Specific and Context Independent Treatment Effects

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 as measured by a test score at period T :

$$A_{isT} = h_T(X_{iT}...X_{i0}, T_{sT}...T_{soT_o}, C_{sT_T}...C_{sT_o}v_s, \epsilon_{iT}...\epsilon_{i0}) \tag{3}$$

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 $Percent\ in\ Treatment_{sT_T}$.

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

$$A_{is1} = v_s + \beta_1' X_{i1} + \beta_{T1}' T_{i1} + \beta_c' Percent\ in\ Treatment_{s1} + \varepsilon_{i1} \tag{4}$$

where v_s is an individual school effect. Since nearly all of the explanatory variables in equation (4) 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 (4) 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 to reweight the second stage estimation and bootstrap methods are used to conduct inference.

Our second 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.[26] 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,

[26] 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.

we do treat the school fixed effects as being additively separable.[27] As all of the inputs with the exception 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, conduction simulations where we set $Percent\ in\ Treatment_{j1} = 0$, we can calculate the context independent treatment effect and use bootstrap based methods to conduct inference.

4.3 Results

The results of using the first empirical strategy to disentangle this exercise are presented in Table 4. In each cell, the context specific effect are presented in red and the corresponding total treatment effect as well as its standard error that correspond to Table 2 are presented in black. Notice that the behavioral responses are larger in lower quartiles. On average

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

the entries in red appear to account for between 50-70% of the estimated treatment effect.

Note, we need to work on getting the standard errors correctly. The main challenge relates to the notion that estimation error may arise in step 1 of the procedure and we need to develop a bootstrap based procedure to account for this feature. This has beeen completed and in the next draft we will present estimates and standard errors for both the context specific and context independent portion of the treatment effect.

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 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.[28] 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 level of success is likely sensitive to the details of the school environment in which she is

[28]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).

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.[29] 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 for students from these schools is currently only available for the 1st grade, which then

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

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.[30] Developing strategies to overcome these methodological challenges presents an agenda for future research.

[30] 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] Deaton, A., (2009), "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development," *NBER Working Paper #14690.*

[6] 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.

[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] 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.

[10] 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

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

[12] 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.

[13] Ferracci, M., G. Jolivet, and G. J. van den Berg (2010), "Treatment Evaluation in the Case of Interactions within Markets," *IZA Discussion Paper No. 4700.*

[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.

Estimating Context-Independent Treatment Effects in Education Experiments

Weili Ding
Queen's University

Steven F. Lehrer
Queen's University and NBER

Summer Workshop in Industrial Organization
and Management Strategy
Shanghai

August 2010

Some Preliminaries

Multi-Billion Dollar Class Size Reduction Initiatives throughout North America over the Past Decade; Incredible Attention Paid to Project STAR in Policy Debates and Decisions.

United States Department of Education (1998):

"In sum, due to the magnitude of the Project STAR longitudinal experiment, the design, and the care with which it was executed, the results are clear: This research leaves no doubt that small classes have an advantage over larger classes in student performance in the early primary grades"

Remain unexplained is substantial heterogeneity in small class size effect across schools in Tennessee's Project STAR experiment asking what makes small class work?

We will present a "puzzle" that could explain the heterogeneity experienced in different schools.

Project STAR

The Student Teacher Achievement Ratio (STAR) project was a large-scale, four-year, experimental study of class size reduction.

Over 6,000 students in 79 schools upon entering kindergarten were randomly assigned into one of three interventions: small class, regular class and regular-with-aide class (no difference with aide).
Classroom teachers were also randomly assigned to classes.

The interventions were initiated as the students entered school in kindergarten and continued through the third grade.

Random assignment is designed to circumvent selection in treatment.

A variety of implementation problems came up in those four years.

Our focus here will be on the initial implementation.

Data

Only kindergarten year; scaled test scores for norm referenced test; examine four subject areas separately.

Summary Statistics

Variable	Mean	Std. Dev.
Math Test Score	485.377	47.698
Reading Test Score	436.725	31.706
Word Recognition Test Score	434.179	36.759
Listening Skills Test Score	537.475	33.1397
African American Student	0.365	0.481
Hispanic Student	0.0018	0.0428
Student is on free lunch	0.484	0.4998
Student is female	0.471	0.499
Teacher year of experience	9.259	5.809
Teacher has masters degree	0.347	0.476
Teacher is not white	0.165	0.371

Multiple Sites and Implementation

Following unanimous approval by the legislature, the Commissioner of Education invited all Tennessee school systems to participate in Project STAR. School principals had to first respond to a request from the researchers. If the school was large enough, they were essentially admitted.

The guidelines indicated that the state would cover additional costs for STAR project teaching staff, but that local systems would furnish any additional classroom space needed. The number of small and regular classes within a school is determined by the size of the student body subject to potential space constraints.

The project schools would not receive “any special considerations other than class size”. To participate, the schools had to agree to random assignment of teachers and students to different classes.

Figure 1: Variation across schools on Fractions of Students in Treatment

Schools range from 16.129% of students in small classes to 44.26%.

Mean = 0.3010607

Std. Dev. =0.0738637

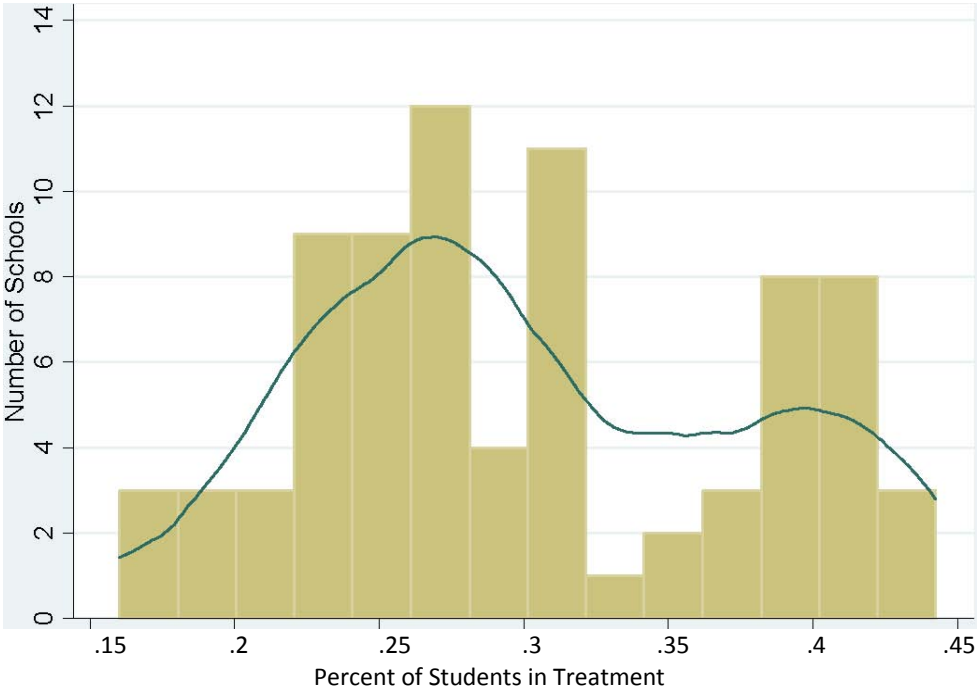
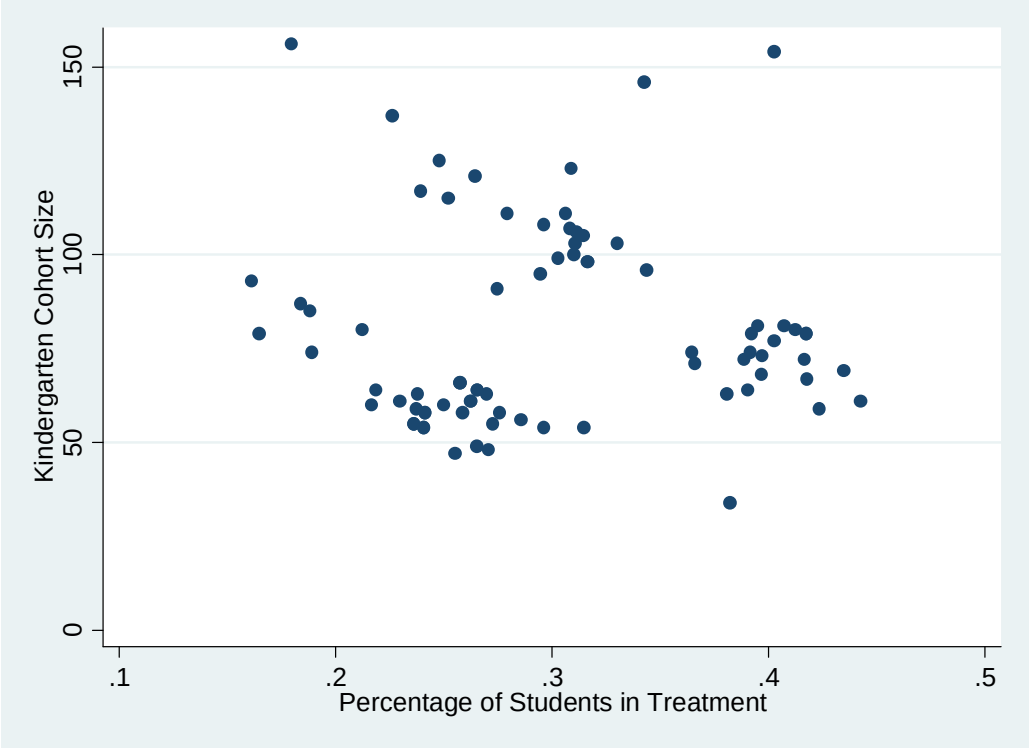


Figure 2: No Correlation between Cohort Size and Fraction in Treatment



Raw Pearson Correlation is -0.0279

A “Puzzle”

Basic specification to estimate causal impacts (ITT or ATE)

$$A_{ik} = \beta_1 X_{ik} + \beta_2 S_{ik} + v_j + \varepsilon_{ijk}$$

School effects included since randomization is done within schools.
Nothing new: We find the same causal impacts as everyone else.

When we decide to estimate the above equation based on where the schools are located in Figure 1, we get the following pattern:

Puzzle Part 1: Small Class Effect (β_2) by Quartile of Fractions in Treatment

Subject	Math	Reading	Word Recognition	Listening skills
Full sample	**8.613 (2.025)*** **	**5.950 (1.280) *** **	**6.331 (1.415) *** **	**3.522 (1.213) *** **
Quartile 1 (16.1-24.8%)	**19.409 (5.414) *** **	**10.383 (3.518)*** **	**8.036 (3.269)*** **	**6.667 (2.807) ** **
Quartile 2 (25.0-29.6%)	**7.671 (3.705)** **	**9.958 (2.591)*** **	**11.902 (2.975)*** **	**4.383 (2.475)* **
Quartile 3 (30.3-36.5%)	**8.558 (3.403)** **	**3.964 (2.122)* **	**5.217 (2.504)** **	2.627 (1.877)
Quartile 4 (36.6-44.3%)	5.043 (4.104)	3.370 (2.566)	3.026 (2.889)	2.828 (2.527)

Striking difference between Q1 and Q4. Only small class effect differs across quartiles, not impact of other inputs such as free lunch, race and location.

The pattern is robust to other sized slices of school in the data (1/5, 1/3, 1/6). Note we used slices not slivers.

Puzzle Part 2A: Comparing Small Class Student Performance across Schools

We estimate the following equation:

$$A_{ik} = \gamma_1 X_{ik} + \gamma_2 1\{\text{Quartile of Percent in Small Classes}_j\} + \xi_{ijk}$$

	MATH	**Reading**	**Word Recognition**	**Listening skills**
Quartile 2	-9.018 (8.873)	1.181 (5.407)	2.238 (5.567)	-1.011 (4.702)
Quartile 3	**-14.728** **(7.994)***	-4.843 (5.190)	-1.906 (5.512)	-4.739 (4.120)
Quartile 4	**-16.995** **(8.113)****	**-9.319** **(5.115)***	**-9.944** **(5.378)***	-3.805 (4.164)

On average, small class student performance drops as the fraction of students receiving small class treatment in their school increases.

Puzzle Part 2B: Comparing Regular Class Student Performance across Schools

	MATH	**Reading**	**Word Recognition**	**Listening skills**
Quartile 2	3.625 (5.624)	2.838 (3.529)	-0.325 (3.904)	1.493 (2.863)
Quartile 3	-6.439 (4.679)	-0.024 (2.948)	-0.795 (3.247)	-3.271 (2.635)
Quartile 4	-1.595 (4.352)	-2.139 (2.746)	-4.968 (3.372)	-0.466 (2.577)

For students in control, much smaller and not significant difference in average performance by quartile.

Results robust to other specifications: linear, quartic and fraction of teachers in treatment.

Summing up

The puzzle suggests that class size treatment effects could in part be driven by excellent small class performance in schools where smaller fractions of students got treatment. Are we sure? If yes, Why?

First, what cannot explain these striking differences across quartiles:

Anything funny going on across quartiles in STAR data? NO

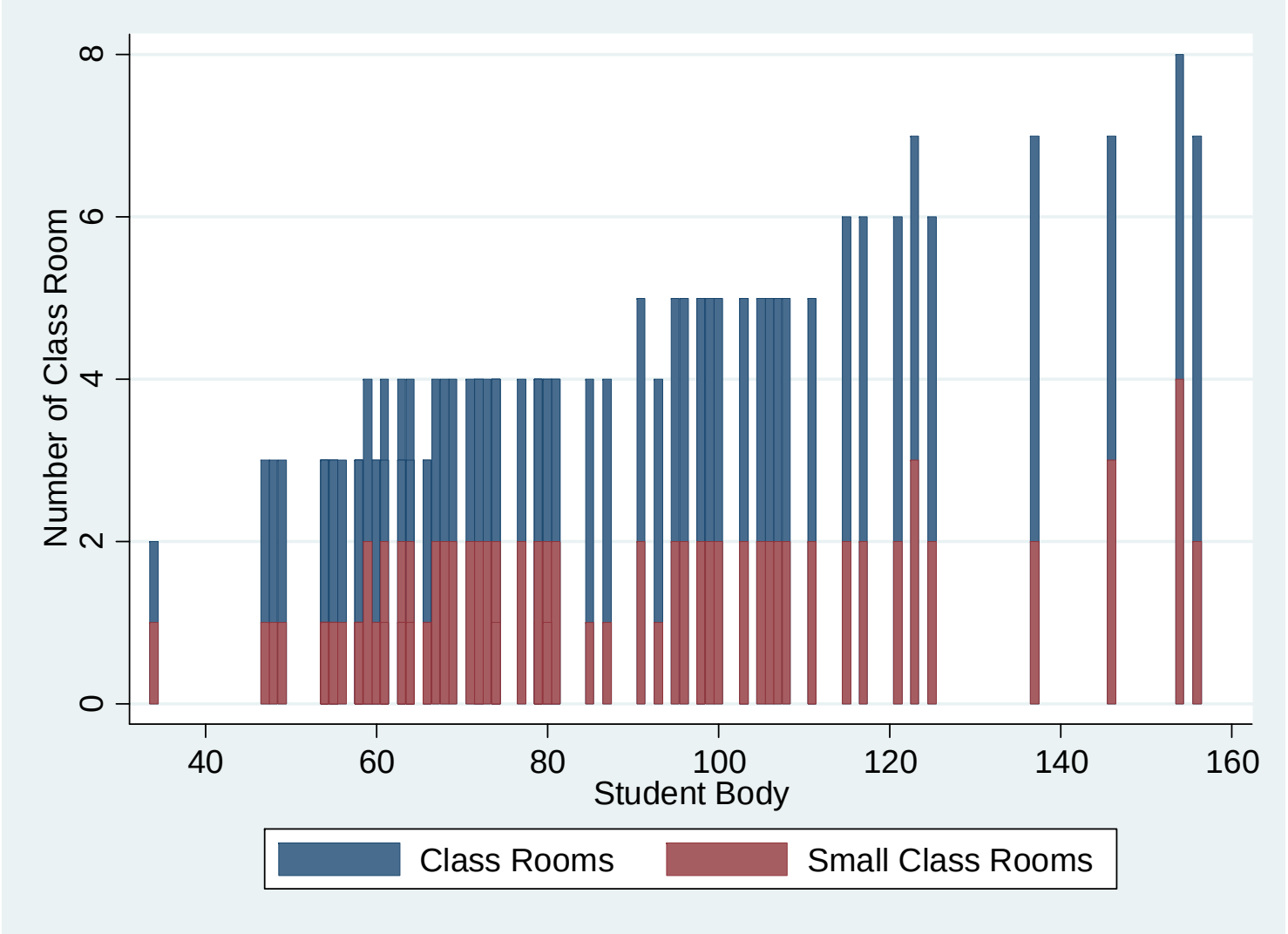
-Randomization of student and teachers in schools.

-Number of students in small and regular sized classrooms.

-Other school level characteristics (location, **size**, student race, free lunch).

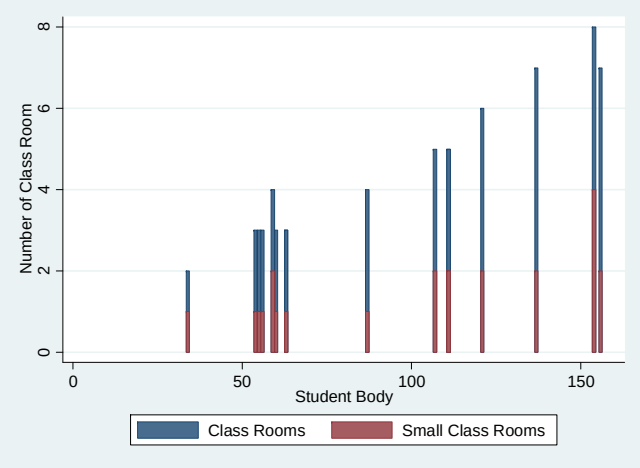
-Selective test taking and the characteristics of these students.

Kindergarten Cohort Size and Classroom Breakdown

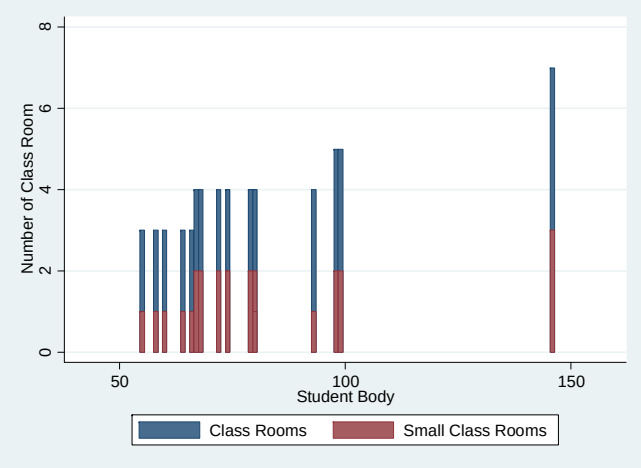


By School Location

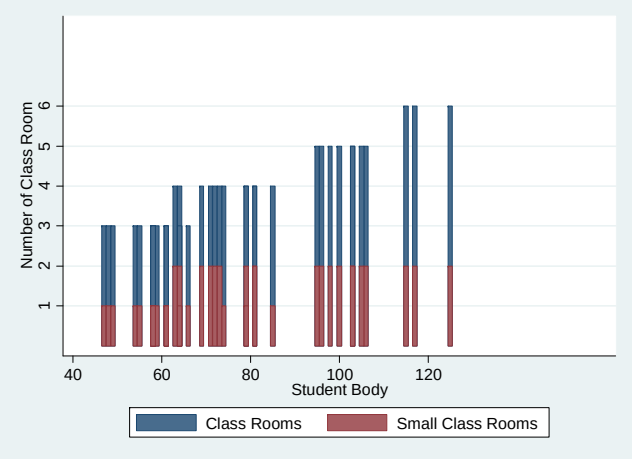
Inner City



Suburban



Rural



Urban

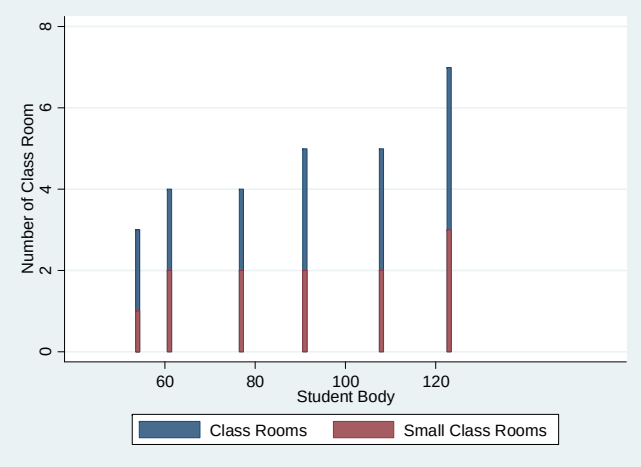
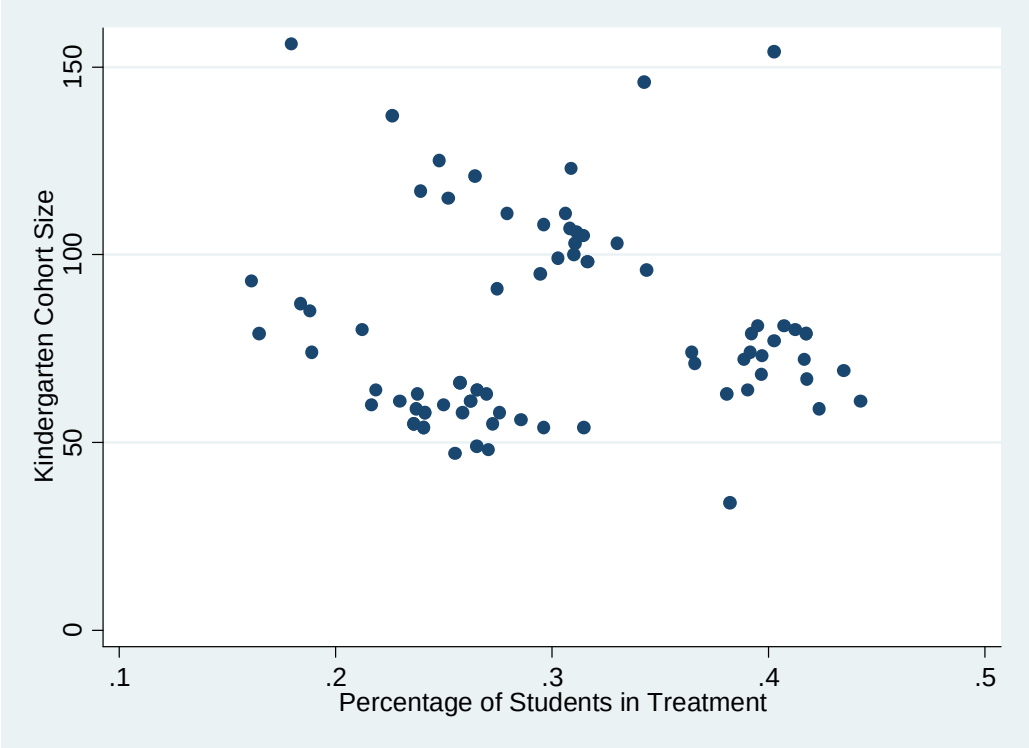


Figure 2: No Correlation between Cohort Size and Fraction in Treatment



Raw Pearson Correlation is -0.0279

Candidate Explanations for Why Small Classes Could be Effective:

- the nature of classroom experiences
- teaching methods
- attention to peer effects in smaller classrooms.

A survey of the existing literature casts serious doubt that they can explain the pattern discovered:

Evertson and Folger (1989) report no statistically significant differences between small and regular classes for % student-initiated questions and comments, % students off-task, or time waiting for help in reading or mathematics. Difference in the amount of "disruptive behavior'' (1.6% vs. 2.0%) and "inappropriate behavior'' (1.6% vs. 2.1%) too small to explain the benefits of small classes.

Finn & Achilles (1999) find no change in fundamental teaching strategies when given a small class.

Finn et al. (2001) and Achilles et al., (1994) find teachers in STAR small classes spent less time on managerial/organizational tasks.

Thus, we are quite confident the puzzle is real and requires alternative explanation. Moreover, our confidence comes from the discovery process.

How Did We Come Across this Pattern? Not Random Luck!

Many experiments are done at multiple sites but the implementation across sites is rarely identical. Our strategy is to exploit the random variation in the fraction of individuals offered treatment across locations (schools in Project STAR) to explain treatment heterogeneity. This variation in the experiment was not induced 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.

In an education experiment:

- Individuals/subjects often can verify their treatment status within the experiment, as well as the experimenters. (zero blind)
- Individuals/subjects may have preference over the outcome of the experiment (teachers like smaller classes, students/families like more resources, Hoxby(1999)) and can influence such outcome to some extent.
-

Thus these individuals/subjects may be incentivized to respond differently to their treatment assignment (i.e. Hawthorne effects, John Henry effects).

Note this is neither a story of positive social interaction effect (Q4 > Q1) or some general equilibrium effect (not through price).

Where might context effects arise in STAR?

-Students (Possible but unlikely given their age)

-Parents (Direction?)

-Teachers (Large stakes in results)

Why do we think psychological effects via teachers are plausible?

The experimenters themselves went to 20 of the 79 schools prior to implementation and met all the teachers to explain the "philosophy behind the experiment". All school principals aware of the study.

Even if a teacher still did not realize they were being studied, it would be obvious for them to compare workload.

Documentation indicates that process evaluations underway.

Relevance for other applications depends on whether you believe subjects could verify their treatment status.

STORY #1: 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.

Consistent with the education literature on teacher responses to perceived favouritism: 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)).

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".

STORY #2: peer pressure

Literature in Economics also consistent with this form of psychological effects:

Teacher motivation and effort may change with peer pressure

- Kandel and Lazear (1992), model, embarrassment over direct performance comparison and harder to free ride.
- Mas and Moretti (2009) and Bandiera, Barankay and Rasul (2009), experiments, peer pressure affects individual productivity

There are likely many more potential stories that can explain how these context effects arise in the Project STAR study…

As an aside in case you were still worried it is not a school size effect

Subject	Math	Reading	Word Recognition	Listening skills
Full sample	**8.613 (2.025)*** **	**5.950 (1.280) *** **	**6.331 (1.415) *** **	**3.522 (1.213) *** **
Quartiles Based on the Size of Kindergarten Student Body				
Quartile 1 (34-59)	**11.340 (4.346)** **	**4.903 (2.679)* **	**6.461 (3.209)** **	5.347 (3.214)
Quartile 2 (60-72)	3.628 (3.918)	**5.497 (2.726)** **	**7.874 (3.088)** **	2.949 (2.379)
Quartile 3 (73-96)	**7.765 (4.547)* **	**8.342 (2.716)*** **	**6.474 (3.152)** **	**4.238 (2.237)* **
Quartile 4 (98-156)	**12.263 (3.498)*** **	**5.975 (2.271)*** **	**6.422 (2.455)** **	3.485 (2.140)

Disentangling Context Effects from Treatment Effects

In the standard evaluation framework we define

$S_t = 1$ Attending *small class* in period t

$S_t = 0$ Attending regular class in period t

A_t is Achievement at the end of period *t* and there are 2 potential outcomes.

$$A_t = S_t A_{1t} + (1-S_t)A_{0t}$$

If we allow for context specific effects, potential outcomes A_{ts} now depend on not only the class type treatment operating through S_t but also a context variable C_s that operates at the school level.

$$A_{its} = A_i(S_t, C_s)$$

Context effects can arise from many phenomena and there are now more than two potential outcomes so multiple treatment effects can be defined.

Allowing for context effects implies there is a violation of the Stable Unit Treatment Value Assumption. This assumption is implicit in many causal estimators and requires that an individual's potential outcome is not affected by the treatment status of other.

We Need to Make Some Assumptions

Note, earlier research with Project STAR either assumes zero context effects or for policy, if non-zero it would get fully translated into in a different setting.

We suggest that the education production function in each subject area is given by $A_{ij}=f(X_{ij},S_{ij}, \text{Percent in Treatment}_j, v_j, \varepsilon_{ij})$

where f(*) is an unknown function to be estimated.

Randomization helps as we do not need historical inputs. But no within variation of the Percent in Treatment$_j$ variable if linearized is that if v_j is treated as fixed.

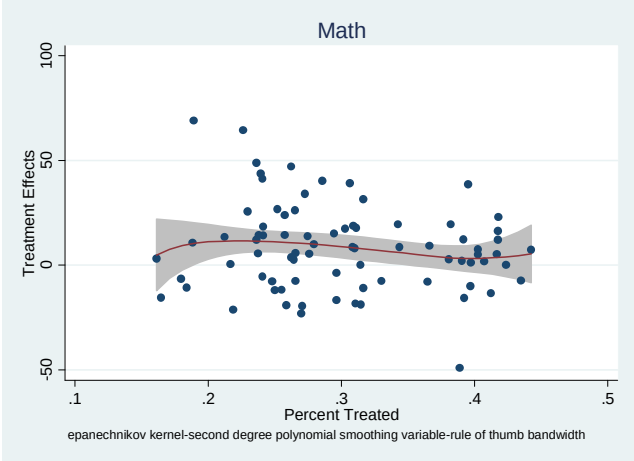
First crack: Assume the impact of psychological effects and mechanical effects are additively separable.

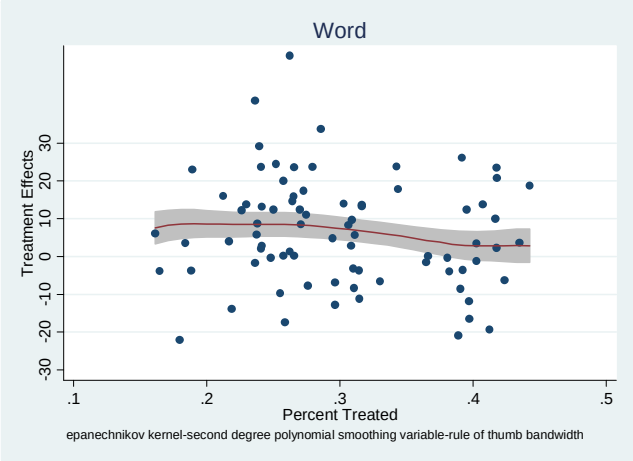
Step 1 Get nonparametric estimate of treatment effects by school (79).

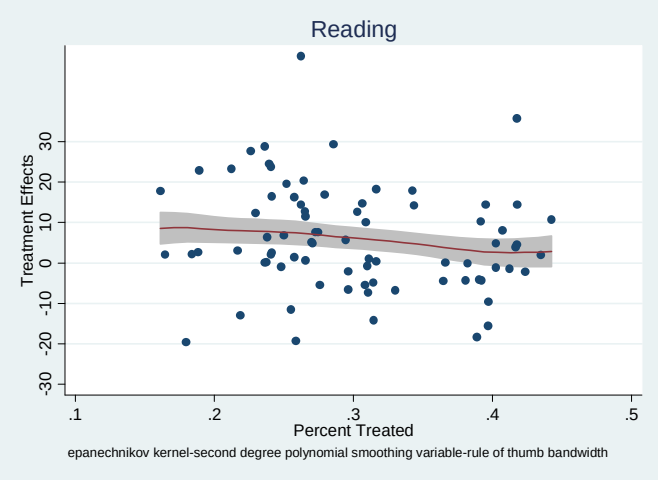
Step 2 Use nonparametric regression relating school treatment effect to percent treated within school (Results robust to different weighting schemes).

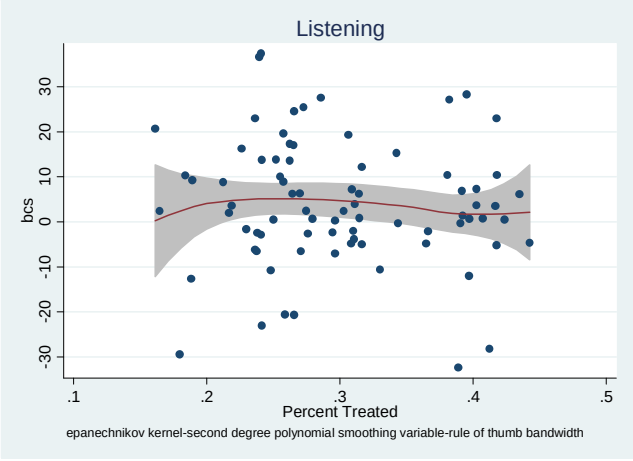
Step 3 Use the mean of the estimated curve within quartiles to calculate the predicted impact of percent treated on the school treatment effect.

Results from First Crack -Decomposing the Treatment Effect









Numerically

	Math	Reading	Word Recognition	Listening skills
Quartile 1	9.967 19.409 (5.414)	7.253 10.383 (3.518)	5.339 8.036 (3.269)	3.436 6.667 (2.807)
Quartile 2	10.193 7.671 (3.705)	7.008 9.958 (2.591)	8.164 11.902 (2.975)	5.119 4.383 (2.475)
Quartile 3	6.718 8.558 (3.403)	4.205 3.964 (2.122)	6.228 5.217 (2.504)	4.035 2.627 (1.877)
Quartile 4	3.693 5.043 (4.104)	2.817 3.370 (2.566)	3.037 3.026 (2.889)	1.899 2.828 (2.527)

Behavioral responses are larger in lower quartiles.

Note, we need to work on getting the standard errors correctly.

Briefly (unless time permits)

We are thinking two other ways to decompose the existing treatment effect into context specific effects and context independent treatment effects:

1. Use 23 comparison schools with 0% in treatment to estimate epf, predict outcome for experiment schools, regress forecast errors on percent treated…

    -Unfortunately data on performance from comparison school (useful to serve as a validation) is not available for kindergarten. Maybe 1st grade? How to deal with noncompliance?

Third crack: Structural model of education production function and then extrapolate to calculate treatment effect when $C_s = 0$.

Some benefits: can relax additive separability assumption.

Con #1: Difficult to account for school effects
Con #2: Results are arguably implausible

Additional Implications of Considering Context Effects

Many experiments are done at multiple sites but the implementation across sites is rarely identical.

If you believe that context specific effects are likely, then there are some simple ways to modify one's experimental design:

-Randomization across and within schools. (In laboratory studies individuals generally randomize treatments across sessions but there are a number of studies that randomize within individuals) But randomizing across schools reopens other critiques

Conclusions

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. Should an individual's outcome depend on their own as well as other individuals' treatment status?

Decomposing the amount of context-independent effects is clearly of importance not just for the immediate cost-benefit exercise but also for policy discussion of large-scale implementation. 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.

We believe the method discussed can be in an evaluator’s as well as an empirical economist’s tool box.