IZA DP No. 4566

How Consistent Are Class Size Effects?

Spyros Konstantopoulos

November 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

How Consistent Are Class Size Effects?

Spyros Konstantopoulos
Michigan State University
and IZA

Discussion Paper No. 4566
November 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4566
November 2009

ABSTRACT

How Consistent Are Class Size Effects?

Evidence from Project STAR has suggested that on average small classes increase student achievement. However, thus far researchers have focused on computing mean differences in student achievement between smaller and larger classes. In this study I focus on the distribution of the small class effects at the school level and compute the inconsistency of the treatment effects across schools. I use data from Project STAR and estimated small class effects for each school on mathematics and reading scores from kindergarten through third grade. The results revealed that school-specific small class effects are both positive and negative and that although students benefit considerably from being in small classes in some schools, in other schools being in small classes is a disadvantage. Small class effects were inconsistent and varied significantly across schools. Full time teacher aide effects were also inconsistent across schools and in some schools students benefit considerably from being in regular classes with a full time aide, while in other schools being in these classes is a disadvantage.

JEL Classification: I20

Keywords: small classes, treatment variability, meta-analysis

Corresponding author:

Spyros Konstantopoulos
College of Education
Michigan State University
Erickson Hall
East Lansing, MI 48824
USA
E-mail: spyros@msu.edu

Class size reduction has been identified by some researchers as a promising school mechanism that can increase student achievement (Finn & Achilles, 1990; Krueger, 1999; Konstantopoulos, 2008a). The effects of class size on student achievement have been of great interest to educational researchers and policy makers the last two decades. As a result, many states have introduced class-size reduction programs. California, for example, introduced a class size reduction program that provided financial incentives to schools that reduce class size in the early grades to twenty or fewer students per classroom. Wisconsin adopted a program that reduced class size to fifteen students per classroom in early grades in schools with high percentages of students from disadvantaged backgrounds. Perhaps the best evidence about class size effects has been produced from Project STAR, a large-scale randomized experiment designed to investigate class size effects. The results of independent analyses have indicated that on average smaller classes had positive effects on students' achievement in early grades (Finn & Achilles, 1990; Hanishek, 1999; Krueger, 1999; Nye et al., 2000).

Large scale experiments such as Project STAR allow researchers to compute an overall average effect for class size and determine its statistical significance. Although typically the main interest in empirical studies lies in computing an average treatment effect, it is also important to compute the inconsistency of the treatment effect. In particular, because randomization took place within each school in Project STAR class size effects can be estimated for each school separately and therefore the researcher can determine whether and how school context interacts with the treatment effect. The idea is that class size effects may differ across the sample of schools mainly because of differences in school context. Thus far, researchers have provided adequate documentation about the average effects of class size on student achievement. Although such school interventions are typically designed to be consistent across settings it is possible that class size effects across schools. The inconsistency of class size effects across schools has not been well discussed thus far, and hence we have not gained a good understanding about why and how these effects are consistent or vary across schools.

Project STAR is a randomized block design where randomization of students to classrooms of different sizes took place within schools (the blocks). Because of the nature of the design one can compute school specific treatment effects and their variability of across schools in the sample. Large variability of class size effects across schools would indicate large differences in class size effects between schools that are mainly due to differences in school context. In contrast, small variability, that is not statistically different from zero, of class size effects would indicate that class size effects are consistent and do not interact with school context. The computation of class size effects for each school results in the creation of a distribution of effects that can be used to identify the schools where the treatment was more (or less) successful and determine the school characteristics and context that contributed to the varying degrees of success. This process may facilitate our understanding of how class size effects are shaped by school context and may identify the optimal conditions under which the treatment becomes more effective. Ultimately such knowledge can be useful in understanding the mechanism of the intervention, in rethinking and redesigning the treatment as well as optimizing its implementation in order to ensure high levels of effectiveness (see Konstantopoulos, 2008b; Turpin & Sinacore, 1991).

In this study I examined the consistency of the small class effects across schools using data from project STAR. In particular, I computed the small class effect in each school and then used meta-analytic methods to compute the overall average effect as well as the variability of the school-specific class size effects across all schools. The analysis was conducted for each grade (i.e., kindergarten, first, second, and third) separately. Because Project STAR intended to gauge the effects of having a full time teacher aide in the classroom on student achievement, which represents the pupil teacher ratio in the classroom, I also examined the consistency of the effect of having a full-time aide in a regular classroom across all schools. Although the average effect of having a full time aide in the classroom has been shown to be small and non-significant (Nye et al., 2000), it is critical to examine whether the full time aide effect varies between schools and interacts with school context. For instance, it is possible that the full time aide effect is more (or less) pronounced in some schools than in others and it would be useful to identify schools where the full time aide effect is beneficial to students.

The Consistency of Treatment Effects

The consistency of class size effects is closely related to the notion of the generalizability of the treatment effects and the concept of external validity, which is concerned with the degree to which the causal relationship holds across schools (see Shadish, Cook, & Campbell, 2002). Even though external validity and generalization have typically been expressed in qualitative terms, Shadish et al. (2002) argue that there is a conceptual similarity between generalizability of treatment effects and interactions between treatments such as small classes and school context. Evidence of an interaction between school context and class size effects would indicate low external validity and low generality of the effects across different settings. One way to evaluate the generalizability of class size effects is to examine the inconsistency or variability of the treatment effects across schools. Since the class size reduction intervention was designed to assign randomly students (and teachers) in different types of classrooms within schools, it is possible that class size-school interactions took place and produced differential treatment effects across schools. Because schools may differ in leadership, organization, climate, and commitment to the intervention it is plausible that the effectiveness of class size reduction programs will vary across schools. That is, in some schools the class size effect may be more beneficial to students than in other schools.

The consistency of treatment effects is also related to the notion of scale up (Schneider & McDonald, 2006). It is noteworthy that some research programs are dedicated to understand how treatment effects vary across contexts. For example, the Interagency Educational Research Initiative (IERI), funded jointly by the National Science Foundation, the National Institute of Child Health and Human Development, and the Institute of Educational Sciences is a major program of research devoted to the problem of determining which educational interventions produce consistent effects across classrooms or schools.

Threats to Validity of Project STAR

Randomization

A key advantage of randomized experiments such as Project STAR is that successful randomization ensures that students in different types of classrooms within schools have on average similar observed and unobserved characteristics. Since the same individuals can’t be assigned to different conditions the idea is to create equivalent groups of individuals, on average, across conditions. Hence, when randomization holds, differences in characteristics across treatment types are only due to chance and are not systematic. Randomization is a crucial aspect of the internal validity of any experiment and therefore of Project STAR. The important question is whether random assignment effectively eliminated preexisting differences between students assigned to different types of classrooms. The fact that the randomization of students to different types of classrooms was carried out by the consortium of researchers who carried out the experiment, enhances its credibility. However it is good practice to check whether there were any differences on preexisting observed characteristics of students. Note that examining differences in preexisting observed characteristics does not prove that randomization worked well.

Simply, this procedure can only provide evidence about whether randomization was not successful for observed variables. Unfortunately, no pretest scores were collected in Project STAR so it is not possible to examine differences in pre kindergarten achievement. However, one could check randomization using student variables such as age, race, gender, and SES. Kreuger (1999) examined the effectiveness of the randomization among the three treatment groups, small, regular, and regular classes with a full time aide, and found for three observed variables such as SES, minority group status, and age there were no significant differences between classroom types across all schools. Krueger pooled data from all schools and classrooms in the sample to conduct this analysis.

Nonetheless, since random assignment of students to small and regular classes was conducted within schools, each school represents a small-scale experiment study. It is reasonable then, to examine whether randomization was successful within schools. Thus, in the present study I used data from each school and conducted F- and chi-squared tests to examine whether randomization worked well. For continuous variables such as age I used the typical ANOVA F-test, and for categorical variables such a race, gender, and SES I used chi-squared tests of independence.

In kindergarten, I found that there were significant differences among classroom types with respect to age in 5 out of 79 schools (6%), with respect to gender in 4 out of 79 schools (5%), with respect to race in 1 out of 79 schools (1%), and with respect to SES in 6 out of 79 schools (7-8%). In first grade, I found that there were significant differences among classroom types with respect to age in 11 out of 76 schools (14%), with respect to gender in 0 out of 76 schools (0%), with respect to race in 4 out of 76 schools (5%), and with respect to SES in 10 out of 76 schools (14%). In second grade, I found that there were significant differences among classroom types with respect to age in 8 out of 75 schools (10-11%), with respect to gender in 1 out of 75 schools (1%), with respect to race in 3 out of 75 schools (4%), and with respect to SES in 5 out of 75 schools (6-7%). Finally, in third grade I found that there were significant differences among classroom types with respect to age in 7 out of 75 schools (9%), with respect to gender in 1 out of 75 schools (1%), with respect to race in 2 out of 75 schools (2-3%), and with respect to SES in 10 out of 75 schools (14%). Overall, these results do not suggest systematic differences for gender and race. That is, it appears that the observed gender and race differences occurred by chance, and this result is consistent with what one would expect if randomization were successful. However, for age and SES the observed significant differences were greater than 5 percent and in some grades greater than 10 percent. These percentages are larger than the typical 5 percent chance that social science researchers universally accept as random chance. Hence, for these two variables the evidence is not so consistent with what one would expect had randomization worked.

Attrition

Large scale longitudinal studies such as Project STAR are likely to experience attrition. Some of the students who participated in Project STAR one year were not part of the experiment the following year. Approximately 28 percent of the students who participated in Project STAR in kindergarten were not part of the study in the first grade. The attrition rate from first to second grade was nearly 25 percent. Twenty percent of the students dropped out of the study after the second grade and thus they were not present in the third grade. Overall, about 50 percent of students were part of Project STAR all four years.

Attrition can potentially affect the class size estimates if within small or regular size classes the students who drop out of the study are systematically different than those who remain in the study. This mechanism would introduce selection bias in the estimates of class size. Systematic differences among groups are typically examined for outcomes of interest. In Project STAR such outcomes were mathematics and reading achievement. For instance, suppose that the students who dropped out from small classes in one year have significantly lower achievement than students who dropped out from regular size classes. This suggests that students who are in small classes and remain in the study may have higher achievement than those in regular classes who stayed in the study because of differential attrition. In this example the class size effect will likely be overestimated. In contrast, if students who dropped out from small classes have higher achievement than those who dropped out from regular classes, then small class effects may be underestimated. In any case if such selection mechanisms take place the class size effects will be biased either upwards or downwards.

Previous analyses that examined the effects of differential attrition on class size estimates with Project STAR data conducted analyses pooling data across all schools (Krueger, 1999; Nye et al., 2000). In this study I reexamined the effects of differential attrition on class size estimates conducting analysis within each school since in Project STAR a small-scale experiment was conducted within each school and in principle attrition in one school is independent of attrition in other schools. Specifically, I examined mean differences in mathematics and reading scores between students who stayed in the experiment and where in small classes (or in a regular class with a full time aide) and those who stayed in the experiment and where in regular classes. The analysis was repeated for each grade (kindergarten, first, and second grade). For example, I used t-tests to determine whether the students who went from kindergarten to first grade and where in small classes (or in regular classes with a full time aide) in kindergarten had on average different kindergarten achievement than students in regular classes that year.

In kindergarten, I found significant differences in mathematics or reading scores between stayers in small classes (or regular classes with a full time aide) and regular classes in 15 percent of the participating schools. The results indicated that stayers in small classes (or regular classes with a full time aide) had higher average achievement than those in regular classes. In first grade, achievement differences between stayers in different types of classrooms were detected in more than 20 percent of the participating schools. Again, stayers in small classes (or regular classes with a full time aide) had higher average achievement than those in regular classes. In second grade, achievement differences between stayers in different classrooms were detected in nearly 10 percent of the participating schools and the stayers in small classes (or regular classes with a full time aide) had overall higher average achievement than those in regular classes. Overall, these percentages are larger than the typical 5 percent chance that social science researchers universally accept as random chance. Hence, one could argue that such differences may be systematic. If so, the results produced by the within school analysis suggest that some positive selection may have taken place and therefore the small class advantage may have been overestimated. In addition, it is not impossible that differential attrition may have created differences among students with respect to other observed and unobserved characteristics.

Taken together the results of the analysis that checked randomization and attrition by school provide some support to previous work that has expressed some concerns about the randomization in Project STAR and has argued that the small class effect may be biased upwards (Hanushek, 1999).

Method

Data

Project STAR was a four-year large scale field experiment that involved students in seventy-nine elementary schools in forty-two districts in Tennessee. During the first year of the study, within each school, kindergarten students were assigned randomly to classrooms in one of three treatment conditions: smaller classes (with thirteen to seventeen students), larger classes (with twenty-two to twenty-six students), or larger classes with a full-time classroom aide. Teachers were also assigned randomly to classes of different types. Some students entered the study in the first grade or subsequent grades, and were assigned randomly to different types of classes at that time. Teachers at each subsequent grade level were also assigned randomly to classes as the experimental cohort passed through the grades. Districts had to agree to participate for four years and allow school visits for verification of class sizes, interviewing, and data collection, including extra student testing. They also had to allow research staff to assign pupils and teachers randomly to class types and to maintain the assignment of students to class types from kindergarten through grade three. Overall, more than 11,000 students in 79 schools participated in the experiment over the four-year period.

Project STAR has high internal validity because, within each school, students and teachers were assigned randomly to classes of different sizes. In addition, because Project STAR is a large-scale randomized experiment that includes a broad range of schools and districts (urban, rural, wealthy, and poor), it has higher external validity than smaller-scale studies. Moreover, the study was part of the everyday operation of the schools that participated and hence there is a lower likelihood that novelty effects affected the class size estimates.

Data Analysis

Because random assignment was conducted within schools in Project STAR it is natural to compute class size effects within each school and then pool all estimates across schools to calculate an overall treatment effect. Conceptually Project STAR is a series of experiments that took place in each school throughout the State of Tennessee and therefore

Project STAR data resemble meta-analytic data where each school contributes one class size effect. Therefore it is appropriate to use univariate meta-analysis to analyze the data. Treatment effects can be computed for each school separately (for mathematics or reading), but since each school specific estimate of class size effects is measured with different precision a weighted scheme is necessary to combine estimates together in order to calculate one overall treatment effect across schools.

The computation of class size effects within each school is crucial because it adjusts for possible school effects or differences in achievement between schools (Krueger, 1999; Konstantopoulos, 2008a). To compute class size effects within each school I used linear regression and regressed standardized mathematics or reading scores separately on two dummies that represent small class or regular class with a full-tile aide (regular class being the omitted category) and controlled for gender, race, and SES effects. Note that I computed intention to treat effects and not effects of class size as implemented or received because intention to treat effects are unbiased by design (see Friedman, 2006). In contrast, modeling the received treatment could produce a biased coefficient since that estimate could be affected by unobserved factors related to principals, parents, and teachers. The mean differences I computed for each school were in standard deviation units and indicated the standardized mean difference in achievement between small and regular classes or between regular with full time aide and regular classes. Once the effect sizes for the class size effects were computed for each school I used mixed or random effects meta-analytic regression to combine the estimates (see Konstantopoulos & Hedges, 2004). I used the inverse of the variance of each school-specific effect size as a weight in the weighted regression and I treated the school-specific estimates as random between schools.

I employed the SAS procedure proc mixed to analyze the data. The first model included only the intercept and therefore I computed the weighted mean across schools and the variance of the class size effects between schools. In subsequent models I used school characteristics as predictors to determine their predictive power in explaining variance in class size effects between schools. In particular, I included in the regression equation school composition such as percent of minority and disadvantaged students in a school, percent of students who are present in a school in a year, percent of teachers with graduate degrees and average teacher experience in each school, school urbanization such as urban, rural, or suburban school, school size per grade and number of classrooms per grade in each school. Finally, I also included in the model district fixed effects since it is plausible that districts may have contributed to the between school variability of the class size effects. District fixed effects were modeled as binary indicators.

Results

Small Class Effects

School interventions such as class size reduction programs aim to positively affect, increase student achievement. However, the intention of the intervention does not always match the empirical estimates of the treatment effects. Specifically, in Project STAR the computation of small class effect sizes for each school resulted in an array of estimates that were both positive and negative. Table 1 summarizes the percent of small class estimates that were positive or negative by grade in columns one to four. Column five represents the total number of school estimates in each grade. All percentages were computed using the total number of estimates in each grade as the denominator. For example, in kindergarten mathematics 33 percent of the estimates were negative and four percent of the estimates were negative and significant. The remaining 67 percent of the estimates were positive and 24 percent of the estimates were positive and significant. The percentages were similar in grades 1 through 3. The results for reading were comparable, only the percentage of significant negative estimates in second and third grade was smaller than in mathematics. Overall these results suggest that nearly two-thirds of the small class estimates in each grade were positive and one-fourth were significant. In contrast, one-third of the small class estimates in each grade were negative and a small proportion of the estimates were significant.

In kindergarten mathematics the schools with positive and significant small class estimates were mainly inner city and rural schools. In reading the schools with positive and significant small class estimates were inner city, rural, and suburban. The same pattern was detected in first grade where the schools with positive and significant small class estimates where inner city, rural, and some suburban schools. In second grade the schools with positive and significant small class estimates where inner city, rural, and suburban schools both for mathematics and reading. Finally, the same pattern was observed in the third grade.

The range of small class effects across schools for each grade is presented in Table 2. The minimum and maximum values are expressed in standard deviation units. In kindergarten the range was greater than 3 standard deviations in mathematics and nearly 3 standard deviations in reading. It is noteworthy that in mathematics the maximum value is positive and slightly greater than 1.5 standard deviations, whilst the minimum value is negative and nearly 1.5 standard deviations. In reading the results were similar. This suggests that the average student in a school that benefits the most from small classes is nearly two grades ahead than the average student who is in a school that benefits the least from small classes (Hill, Bloom, Black, & Lipsey, 2008). Hill et al. estimated that the annual mathematics gain in kindergarten is nearly 1.3 standard deviations and the range of small class effects is more than twice as large. According to Hill et al., in reading the estimated annual gain in kindergarten is nearly 1.5 standard deviations and the range of small class effects is approximately twice as large. The range of small class effects in first grade was greater than 2.5 standard deviations in mathematics and 2 standard deviations in reading. In second grade the range of small class effects was greater than 2 standard deviations in mathematics and in reading. Finally, in the third grade the range of small class effects was nearly 2 standard deviations in reading and smaller than 2 standard deviations in mathematics. It appears that the range became smaller over time as students moved through grades.

These results suggest that students in schools that benefit the most from small classes are at least 2 grades ahead in achievement than their peers in schools that benefit the least from small classes. This is not a trivial difference especially in early grades. In addition, these results show that treatment effects vary considerably across different school context and that although it is plausible to hypothesize that class size reduction would affect student achievement positively, in practice some of the school-specific effects are negative and substantial. That is, although the intervention was designed to affect student achievement positively, in reality students in some schools will be at a disadvantage when being in small classes compared to students in other classes.

The results from the unconditional mixed effects meta-regression are reported in Table 3. The term unconditional means that no predictors are included in the model and that a weighted average (the estimate of the intercept) is computed across schools. In kindergarten the average small class benefit in mathematics was 0.19 standard deviations and significant, which suggests that across all schools students in small classes in kindergarten scored about one-fifth of a standard deviation higher than their peers in regular classes. The variance of the small class effects across schools was 0.21 and statistically significant. That is, in some schools the benefits of small class membership is more pronounced than in other schools which is consistent with the results in Table 2. The average small class advantage in reading was slightly larger, 0.24 standard deviations, and significant which suggests that across all schools students in small classes in kindergarten scored about one-fourth of a standard deviation higher than their peers in regular classes. The variability of the small class effect across schools was 0.21 and statistically significant which indicates a significant interaction between small classes and school context. In first grade the average small class benefit in mathematics was 0.28 standard deviations and significant. The variance of the small class effects across schools was 0.17 and statistically significant. The average small class advantage in reading was 0.25 standard deviations and significant and the variability of the small class effect across schools was 0.13 and statistically significant. In second grade the average small class benefit in mathematics was nearly 0.20 standard deviations and significant. The variance of the small class effects across schools was 0.18 and statistically significant. The average small class advantage in reading was 0.24 standard deviations and significant and the variability of the small class effect across schools was 0.11 and statistically significant. Finally, the average small class benefit in third grade mathematics was 0.16 standard deviations and significant. The variance of the small class effects across schools was 0.08 and statistically significant. The average small class advantage in reading was 0.23 standard deviations and significant and the variability of the small class effect across schools was 0.08 and statistically significant. Overall, the average small class advantage across grades was one-fifth of a standard deviation or larger. The variance of small class effects was significant across schools in all grades; however the variance estimates became smaller over time.

In order to identify the kinds of school characteristics that may be responsible for the inconsistency of the treatment effects I also used a mixed effects meta-analytic regression that included several observed school characteristics as predictors. The results suggested that in kindergarten mathematics school characteristics and district effects explained 13 percent of the between-school variance of the small class effect. District effects were responsible for 10 of the 13 percent of the variance explained. In kindergarten reading however, the school characteristics and district effects did not explain any between-school variance in the small class effect. In grades 1 through 3 school characteristics and district effects did not explain any between-school variance in mathematics or reading. These results indicate that the class size effect is more school dependent in mathematics than in reading, but only in kindergarten. In other grades observed school characteristics did not explain any between school variance. Nonetheless, the remaining inconsistency of the class size effect was still significant at the .05 level. Most of the variability in the effects is unexplained and therefore it seems that both in reading and mathematics it is the unobserved school characteristics that are responsible for the inconsistency of the class size effects.

Regular Class with Full Time Aide Effects

The percent of the regular class with a full time teacher aide estimates that were positive or negative by grade are reported in columns one to four in Table 4. In kindergarten mathematics 41 percent of the estimates were negative and 17 percent of the estimates were negative and significant. The remaining 59 percent of the estimates were positive and 18 percent of the estimates were positive and significant. The proportion of significant estimates dropped considerably in grades 1 through 3. In grades 2 and 3 the percentage of negative estimates was slightly higher than that of the positive estimates. The results for reading were comparable. Overall these results showed higher percentages of positive full time aide estimates across grades. The full time aide estimates however seemed more effective in kindergarten that in other grades.

The range of regular class with full time aide effects across schools is presented in Table 5. Again, the estimates are expressed in standard deviation units. In kindergarten the range was nearly 2.5 standard deviations in mathematics and reading. This means that the average student in a school that benefits the most from regular classes with a full time aide is up to two grades ahead than the average student in a school that benefits the least from full time aide in a regular classroom (see Hill, Bloom, Black, & Lipsey, 2008). The range of regular class with full time aide effects in first grade was smaller than 1.5 standard deviations in mathematics and larger than 1.5 standard deviations in reading. That is, the average student in a school that benefits the most from regular classes with a full time aide is at least one grade ahead than the average student in a school that benefits the least from regular classes with a full time aide. In second grade the range of regular class with full time aide effects was greater than 1.5 standard deviations in mathematics and smaller than

1.5 standard deviations in reading. This suggests that the average student in a school that benefits the most from regular classes with a full time aide is nearly 2 grades ahead in reading and one grade in mathematics than the average student in a school that benefits the least from regular classes with a full time aide. Finally, in the third grade the range of regular class with full time aide effects was greater than 1.5 standard deviations in reading and mathematics. Again using the empirical benchmark by Hill et al. it appears that the average student in a school that benefits the most from regular classes with a full time aide is nearly 3 grades ahead in mathematics and 4 to 5 grades ahead in reading and than the average student in a school that benefits the least from regular classes with a full time aide. Overall, these differences are not trivial and show considerable variation of full time aide effects.

The results from the unconditional mixed effects meta-regression are reported in Table 6. Across all grades the average full time aide effect was close to zero and statistically insignificant. That is, on average, reducing pupil teacher ratio in a classroom does not increase student achievement significantly or meaningfully. The estimates of the variance of the regular class with full time aide effects across schools were statistically significant in kindergarten and second grade only. In these two grades there was significant interaction between full time aide effects and school context. In other grades the between-school variance was not significantly different than zero. Still, the full time aide effects were positive and significant in some schools and small and insignificant or negative in other schools.

In order to identify the kinds of school characteristics that may be responsible for the inconsistency of the treatment effects I also used a meta-analytic regression that included several observed school characteristics. The results suggested that in kindergarten mathematics school characteristics and district effects explained 42 percent of the full time aide effect across schools and district effects were responsible for 22 percent of the 42 percent. In kindergarten reading however, the school characteristics and district effects did not explain any between-school variance of the regular class with full time aide effect. In grades 1 through 3 school characteristics and district effects did not explain any variance in mathematics or reading. These results indicate that the full time aide effect is more school dependent in mathematics than in reading, but only in kindergarten. Nonetheless, the remaining inconsistency of the full time aide effect was still significant at the .05 level in kindergarten. Most of the variability in the effects is unexplained and therefore it seems that both in reading and mathematics it is the unobserved school characteristics that are responsible for the inconsistency of the full time aide effects.

Conclusion

This study examined the consistency of class size effects from kindergarten through third grade using data from Project STAR. Analyses were conducted within each school and then estimates were combined across all schools using meta-analytic methods. The main objective of the study was to compute the between school variance of the school distribution of class size effects. First, the findings provide additional support to the notion that the average small class effect is significant, positive, and important in early grades. Across grades the small class effect in mathematics was nearly one-fifth of a standard deviation. In reading the effect was slightly larger, especially in kindergarten and first grade where the effect was closer to one-fourth of a standard deviation. Second, the small class effects vary significantly across schools in all grades for both mathematics and reading. The inconsistency of the effect is larger in kindergarten and becomes smaller as students transition through grades. In the first and second grade the between-school variance of the small class effect was larger in mathematics than in reading indicating that perhaps mathematics is more likely to be affected by school context. Overall, the significant inconsistency of the small class effect strongly suggested that school context interacts with small class effects. In addition, the significant variation of small class effects across schools indicates that the treatment has low external validity or generality and does not scale up across the schools in the sample. The small class effect is positive and significant in some schools and negative and significant in others. District fixed effects and observed school characteristics explained a small proportion of the between school variance of the small class effect only in kindergarten mathematics.

The average regular class full time aide effect was small and non-significant across grades showing that on average decreasing pupil teacher ratio in the classroom does not effect student achievement positively. However, the full time aide effects vary significantly across schools in kindergarten and in second grade in mathematics and reading. The variance estimates were larger in mathematics than in reading which suggested that mathematics may be a more school dependent subject matter. As with small class, these results indicate that school context interacts with full time aide effects and that reducing pupil teacher ratio is beneficial in some schools, but not in others. District fixed effects and observed school characteristics explained a good proportion of the between school variance of the full time aide effect only in kindergarten mathematics.

These findings indicate that in large scale studies such as Project STAR computing the average treatment effect does not provide a complete picture of the effect of the intervention. The variance of the class size effects across schools provides additional important information. When random assignment is conducted within schools computing treatment by school interactions is essential. Such interactions show whether the treatment effects are inconsistent and whether school context influences class size effects. The results of the study demonstrated that school context matters and shapes class size effects. It appears that some schools know how to make use of small classes or full time teacher aides than other schools since the effects are more pronounced in some schools and less pronounced or negative in others.

The schools that benefit most from class size reduction give overall a substantial advantage to their students compared to students in schools that benefit the least from small classes. In some cases the small class benefit is as large as or larger than a two-grade achievement gain in early grades (Hill et al., 2008). It was noteworthy that a good proportion of schools with positive and significant small class effects were inner city or rural schools which are schools that perhaps need to most boost from such a school intervention. The schools that benefit most from a full time teacher aide in a regular classroom also give a substantial advantage to their students compared to students in schools that benefit the least from a full time aide. However, the benefit is not as considerable as the small class benefit and typically less than a two-grade achievement gain in early grades (Hill et al., 2008).

The within school analysis that addressed threats to the validity of the experiment demonstrated differences among classroom types for student characteristics such as age and SES that seem to be somewhat systematic, not entirely due to chance. It is unclear then that randomization was entirely successful for these two variables and some caution about randomization may be warranted. In addition, within school analysis to address the effects of differential attrition produced results that suggest some selection bias from grade to grade. This selection seems to be positive and therefore it appears that the class size effects may be overestimated. These findings provide some support to previous studies that had questioned that class size effects were unbiased (Hanushek, 1999).

Nonetheless, Project STAR is one the best education experiments ever conducted (Mosteller, Light, & Sacks, 1996) and the findings of the present study do not invalidate the important of Project STAR. Simply the results of the analyses reported here suggest that it is best practice for researchers to examine threats of the validity of any experimental study using different methods. Still, Project STAR data have most likely provided the best evidence about class size effects and may have provided the best case scenario for class size reduction programs.

Finally, one could identify the schools that benefit the most or the least for either small classes or from full time teacher aides in regular classes in Project STAR. Ideally, the next step would be to study these schools and determine the specific factors that helped maximize the benefit. In the same vein one could study the schools that benefited the least from the class size effects and identify the factors that hindered the success of the treatment. This micro process would help with reevaluating the nature of the intervention and modifying its implementation so that it is most effective. Eventually such useful information would inform future studies and would most likely maximize the advantage of class size reduction efforts. Unfortunately such school data are not available in Project

STAR and such micro analysis is not permitted. A new large-scale experiment would give us the opportunity to study such schools and understand how the class size mechanism is enacted.

References

Finn, J D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27,* 557-577.

Freedman, D. A. (2006). Statistical models for causation. What inferential leverage do They provide? *Evaluation Review, 30,* 691-713.

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,* 143-163.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2,* 172-177..

Konstantopoulos, S. (2008a). Do small classes reduce the achievement gap between low and high achievers? Evidence from Project STAR. *Elementary School Journal, 108,* 275-291.

Konstantopoulos, S. (2008b). Computing power of tests for the variability of treatment effects in designs with two levels of nesting. *Multivariate Behavioral Research, 43,* 327-352.

Konstantopoulos, S., & Hedges, L.V. (2004). Meta-Analysis. In D. Kaplan (Ed.), *Handbook of Quantitative Methodology for the Social Sciences* (pp. 281-297). New York: Sage.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics, 114,* 497-532.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons learned from skill grouping and class size. *Harvard Educational Review, 66,* 797-842.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (2000). Effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal, 37,* 123-151.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Schneider, B., & McDonald, S.K. (2006). *Scale up in education: Ideas in principle.* Lanham, MD: Rowman & Littlefield.

Turpin, R. S., & Sinacore, J. M. (1991). *Multisite evaluations*. San Francisco, CA:

Jossey-Bass.Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and Health surveys. *International Statistical Review*, *64*, 265-294.

Table 1. Percentages of School Estimates of Small Class by Grade

	Positive		Negative		N
Mathematics	Overall	Significant	Overall	Significant	
Kindergarten	67%	24%	33%	4%	79
First Grade	67%	30%	33%	3%	76
Second Grade	70%	24%	30%	8%	74
Third Grade	65%	17%	35%	5%	75
Reading					
Kindergarten	68%	25%	32%	5%	79
First Grade	72%	24%	28%	7%	75
Second Grade	66%	22%	34%	1%	74
Third Grade	70%	18%	30%	1%	74

Table 2. Range of Small Class Effects

	Minimum	Maximum
Mathematics		
Kindergarten	-1.52	1.66
First Grade	-1.31	1.45
Second Grade	-0.97	1.26
Third Grade	-0.84	0.94
Reading		
Kindergarten	-1.17	1.83
First Grade	-0.97	1.27
Second Grade	-0.99	1.23
Third Grade	-0.70	1.39

Table 3. Estimates of Average Small Class Effect and its Variability Across Schools

	Coefficient	SE	Variance	SE
Mathematics				
Smalll Class in Kindergarten	0.190*	0.061	0.207*	0.045
Smalll Class in First Grade	0.280*	0.056	0.166*	0.039
Smalll Class in Second Grade	0.195*	0.060	0.181*	0.044
Smalll Class in Third Grade	0.158*	0.047	0.084*	0.028
Reading				
Smalll Class in Kindergarten	0.241*	0.061	0.209*	0.046
Smalll Class in First Grade	0.247*	0.053	0.134*	0.033
Smalll Class in Second Grade	0.235*	0.051	0.113*	0.032
Smalll Class in Third Grade	0.227*	0.045	0.079*	0.027

* $p < 0.05$

Table 4. Percentages of School Estimates of Full Time Aide by Grade

	Positive		Negative		N
Mathematics	Overall	Significant	Overall	Significant	
Kindergarten	59%	18%	41%	17%	79
First Grade	59%	7%	41%	3%	76
Second Grade	49%	8%	51%	4%	74
Third Grade	48%	1%	52%	4%	75
Reading					
Kindergarten	55%	13%	45%	10%	79
First Grade	61%	4%	39%	3%	75
Second Grade	54%	9%	46%	3%	74
Third Grade	50%	3%	50%	0%	74

Table 5. Range of Regular Class Full Time Aide Effects

	Minimum	Maximum
Mathematics		
Kindergarten	-1.47	1.09
First Grade	-0.56	0.74
Second Grade	-0.97	0.88
Third Grade	-0.84	0.80
Reading		
Kindergarten	-0.99	1.64
First Grade	-0.63	1.03
Second Grade	-0.59	0.70
Third Grade	-0.54	1.10

Table 6. Estimates of Average Regular Class Full Time Aide Effect and its Variability Across Schools

	Coefficient	SE	Variance	SE
Mathematics				
Regular Aide Class in Kindergarten	0.022	0.061	0.214*	0.046
Regular Aide Class in First Grade	0.053	0.031	0.011	0.011
Regular Aide Class in Second Grade	0.031	0.039	0.045*	0.019
Regular Aide Class in Third Grade	-0.019	0.035	0.024	0.015
Reading				
Regular Aide Class in Kindergarten	0.055	0.055	0.160*	0.037
Regular Aide Class in First Grade	0.056	0.033	0.021	0.013
Regular Aide Class in Second Grade	0.045	0.036	0.031*	0.015
Regular Aide Class in Third Grade	0.026	0.031	0.005	0.012

* $p < 0.05$