Non-Robustness of the Carry-Over Effects of Small Classes in Project STAR

손기태 (대외경제정책연구원 부연구위원)
Class size reduction (CSR) is an enduring school reform in an effort to improve academic achievement and has been widely encouraged or required in the US. Supporters of CSR often rely on the positive contemporaneous and carry-over effects of Project STAR, which is a state-wide, long-term randomized experiment. This paper checks robustness of the carry-over effects of small classes in STAR to find that the effects are mostly driven by students of small classes in a small number of STAR schools. In the process of the check, it turns out that, in contrast to the protocol of randomization, observable student characteristics in these schools are non-randomly distributed between small classes and regular classes in such a way as to increase the academic achievement of students of small classes and lower that of students of regular classes.

Keywords: Carry-over effects, Robustness, Project STAR

I. Introduction

Class size reduction (CSR) is an enduring school reform that attracts not only researchers but also the broader public in the US. According to a compilation of
Education Commission of the States (2005), CSR has been encouraged or required in 24 states since 1977. The Student Achievement Guarantee in Education (SAGE) program in Wisconsin is well-publicized for its success in raising academic achievement (e.g., Molnar, et al, 1999) whereas a CSR program in California is widely criticized for its failure (e.g., Stecher and Bohrnstedt, 2002). At present, a CSR program in Florida garners wide attention because of its timing and size. Not just individual states but the federal government also encourages CSR directly or indirectly. A federal CSR program was separately authorized in Fiscal Year 1999 and then folded into Title II in 2001.

Although many surveys and experiments exist on CSR, Project STAR (Student Teacher Achievement Ratio) stands out as a large-scale, long-term randomized experiment. Mosteller, Light, and Sachs (1996, p. 814) extol that STAR is “one of the great experiments in education in U.S. history,” and Finn and Achilles (1999, p. 97) even claim that the experiment “came to eclipse all of the research that preceded it.” Hence, it would not be surprising to find that the arguably positive effects of CSR in STAR played a non-negligible role in justifying the implementation of CSR across the states.

Whether CSR improved academic achievement in STAR during the experiment is still open to debate. Researchers engaged in the experiment and some independent researchers argue that students of small classes outperformed students of regular classes with or without full-time teacher aids (Word et al., 1990; Finn and Achilles, 1990; Krueger, 1999). On the other hand, other researchers call into question robustness of the results to non-random attrition and noncompliance with the treatment, and they point out that no statistically significant treatment effects exist in a majority of schools (Hanushek, 1999; Ding and Lehrer 2005, 2010; Sohn, 2010).

In contrast to the heated debate on the effects of CSR in STAR, relatively few objections have been raised on the carry-over effects of CSR beyond the experiment period. The former group argues that students of small classes achieved higher scores on tests of cognitive skills even after the experiment ended (Finn, Fulton, Zaharias, and Nye, 1989; Nye, Hedges, and Konstantopoulos, 1999; Krueger and Whitmore, 2001). Implicitly, the doubt on robustness of the earlier effects could make the claim of the carry-over effects weak. And yet, no explicit arguments have
been forwarded on robustness of the carry-over effects.

This paper fills this void in the literature. Because STAR is such as an influential experiment in both research and policy, it is necessary to check robustness of the effects of CSR not only during but also beyond the experiment period. This paper finds that the carry-over effects disappear once a small number of STAR schools are excluded, which indicates that the carry-over effects are not robust.

Two points need to be emphasized in the beginning. First, this paper does not discuss the effects of CSR per se. The goal of this paper is focused on the effects of CSR in one, but important experiment, namely STAR. However, this paper does not concern whether attending small classes raises academic achievement during the experiment. Much research has been done on this topic although clear conclusion is yet to be drawn. Instead, this paper tests robustness of the carry-over effects of small classes after the experiment.

Second, this paper limits its attention to academic achievement. Finn, Fulton, Zaharias, and Nye (1989) argue that students of small classes in third grade show higher noncognitive skills such as effort, initiative, and participatory behavior in fourth grade than students of regular classes in third grade. When Dee and West (2008) apply more sophisticated statistical methods, however, they find that the positive effects of CSR do not persist throughout eighth grade at least in STAR. This paper is not interested in effects of CSR in STAR that do not persist.

This paper is structured as follows. In Section 2, STAR and its literature is briefly introduced. The data and estimation procedures are explained in Sections 5 and 6, respectively. The results are discussed in Section 5, and Section 6 concludes.

II. Project STAR

STAR is a large-scale randomized experiment performed in Tennessee in 1985 through 1989, costing about $12 million. 6,000-7,000 students in 75-79 schools participated in the experiment each year, totaling 12,000 students in the whole period. Although the experiment was costly, it was justified on the ground that research on CSR had not established a definite answer to its effects.
For example, reviewing 76 studies published after 1954, Ryan and Greenfield (1975) find that observable and unobservable variables are not adequately controlled for to appreciate the effects of CSR clearly. Instead of relying on a qualitative review, Glass and Smith (1978, 1979) apply a then-new statistical procedure called meta-analysis to summarize quantitatively a total of 725 comparisons in 77 studies on CSR. This meta-analysis yields that CSR has substantial, positive effects on academic achievement, which are robust to “source of data, subject taught, duration of instruction, pupil IQ and type of achievement measure” (Glass and Smith, 1979, p. 12). And yet, Educational Research Service (1980) points out that, in contrast to the claim that the positive effects of CSR are robust, the results are mostly driven by only 14 experimental studies. Even in these 14 studies, only six studies are relevant to class size situations that are typical of elementary and secondary schools. In a similar line, Slavin (1984, p. 10) dismisses the results of the meta-analysis, demonstrating that the positive effects are “entirely due to studies of tutoring, not of class size as it is usually understood.” Hence, when the implementation of STAR was considered, the effects of CSR were still open to debate.

If STAR had been ideally implemented, it would have proceeded as follows. Schools in Tennessee are randomly selected for the experiment, and under the mandatory enrollment in kindergarten schools, kindergarten students in the selected schools are randomly assigned to small classes with 13-17 students (henceforth, S), regular classes with 22-25 students without teacher aids (henceforth, R), and regular classes with 22-25 students with teacher aids (henceforth, RA). Throughout the experiment period, no students leave or enter the schools, and all of the students remain in their initially assigned class types. Teachers are also randomly assigned to each class type every year and take no differential treatments. Finally, all the students take tests on cognitive skills before and after the experiment starts to check whether random assignment is properly implemented.

Unfortunately, the experiment was not ideally implemented. Schools volunteered for the experiment, and only schools large enough to accommodate at least 57 students were selected. Also, because kindergarten education was not mandatory at that time, kindergarten students enrolled in STAR schools were likely to be a selected

1) The results of these papers are expanded and published in a book by Glass, Cahen, Smith, and Filby (1982).
group. Moreover, students constantly left and entered STAR schools, so the attrition rate was almost 50 percent. Because students left and entered each year, class sizes of S, R, or RA were not contained within its designated range. Hence, it was not rare to find some S’s had more students than other R’s or RA’s. Teachers were randomly assigned to each class type, but 54 out of 340 second grade teachers from 15 STAR schools were provided with a three-day training course. Finally, because students did not take tests before the experiment, it is uncertain whether the assignment was randomly made.

Some problems could be minor; For example, although schools were not randomly selected, this problem could be addressed by random assignment of students within schools. If actual class sizes went beyond their designated ranges, the variable of class size rather than class type can be used to estimate the effects of CSR. Also, because teachers were randomly assigned every year and the number of trainees was small, the three-day training would not have large differential effects on the academic achievement of students of different class types.

A more serious controversy arises on the size of effects of selective attrition of students and noncompliance with the treatment. One side argues that the size is negligible and the positive effects are robust (e.g., Finn and Achilles, 1999; Krueger, 1999) whereas the other side asserts that the size could be substantial and the positive effects are not robust (e.g., Hanushek, 1999; Ding & Lehrer, 2010). Recently, Ding and Lehrer (2010) demonstrate that, once selective attrition and noncompliance with the treatment are accounted for, positive effects of CSR in STAR are absent in second and third grade.

So far, controversy has been limited to the positive effects of CSR in STAR during the experimental period, namely kindergarten through third grade. And yet, it is argued that the positive effects persist beyond the experiment period, Finn, Fulton, Zaharias, and Nye (1989) estimate that students of S in third grade score higher than students in R and RA (henceforth, RR) in third grade by effect sizes of 0.11-0.16 in reading and math when they become fourth graders. Nye, Hedges, and Konstantopoulos (1999) extend the study period to eighth grade to find that students of S in third grade outperform their counterparts even in eighth grade by effect sizes of 0.133 and 0.158 in reading and math, respectively. Effect sizes are similar
in fourth and sixth grades, Finn, Gerber, Achilles, and Boyd-Zaharias (2001) not only confirm these results covering Grades K-8 but also emphasize that staying longer in small classes improves academic achievement more. Looking beyond eighth grade, Krueger and Whitmore (2001) focus on academic achievement in high school. They estimate that students initially assigned to S are 2.7 percentage points more likely to take the ACT or SAT and score 0.12-0.13 standard deviation (SD) higher in the ACT than those in RR.

In contrast to the heated debates on the effects of CSR in Grade K-3, to the best of our knowledge, few arguments have been exchanged for the effects of CSR beyond third grade. This paper initiates this exchange by demonstrating that the effects are not robust to the exclusion of a small number of schools. In this process, one understands why the effects of CSR persist for students who initially attended this small number of schools but not for others. It turns out that CSR has nothing to do with these differential effects. The main reason is that students are non-randomly assigned into S and RR in the small number of schools.

**III. Data**

This paper uses the follow-up data of STAR students. At the end of the experiment in the spring of 1989, when most of the STAR students had completed third grade, all the STAR students returned to regular classes, but data on their academic achievement were collected throughout high school. For this study, academic achievement is measured by test scores on math and reading, whether the student graduated high school, whether the student took the ACT or SAT, and ACT equivalent scores. All the variables are typically used in the literature of STAR.

In this paper, students of R and students of RA are aggregated as usually treated in the literature. The reason is that half of the students of R and RA were randomly reassigned to the other class type in the second year and stayed there until the end of the experiment. In fact, a part-time teacher aid was available 25-33 percent of the time on the average even in R because it was required to provide services which had been typically available. Partly for the above reasons, academic achievement
was statistically insignificantly different between students of R and RA. Hence, there are essentially two class types in this paper, namely, S on the one hand and R and RA combined (henceforth, RR) on the other hand.

This paper starts with distinguishing effective schools from ineffective schools in each grade. Following Sohn (2010), effective schools are defined as schools where the test scores of students of S are statistically significantly higher at the five percent level than students of RR in both math and reading. By contrast, ineffective schools are defined as schools where test scores are statistically significantly higher at the five percent level in neither math nor reading.

Because randomization took place within schools at each entry grade and non-random attrition and noncompliance with the treatment are suspected, the initially assigned school is used to indicate whether the student was assigned to an effective or ineffective school.

### IV. Estimations

The main estimation strategy is similar to Krueger and Whitmore (2001) and Dee and West (2008). More specifically,

$$y_{icsw} = \beta_0 + \beta_1 (\text{SMALL}_{csw}) + \beta_2 x_{icsw} + \alpha_{sw} + \epsilon_{icsw} \quad (1)$$

where $y_{icsw}$ is a dependent variable for student $i$ in classroom $c$ of school $s$ in entry wave $w$. Depending on specifications, $y_{icsw}$ is a variable of total math scale score CTBS (Comprehensive Tests of Basic Skills)$^{2}$, total reading scale score CTBS, probability of graduating high school, probability of taking the ACT or SAT college-admissions tests by senior year of high school, or ACT-converted scores. All the test scores are standard normalized to facilitate interpretation, $\text{SMALL}_{csw}$ indicates that the student is initially assigned to a small class, $x_{icsw}$ includes student characteristics, namely, race, gender, eligibility for free-lunch and days since birthday and its square. Free lunch status indicates that the student was eligible for free lunch at

---

2) CTBS is a norm-referenced test battery,
least in one grade during the experiment. Although it is not necessary to control for \( X_{tsw} \) if randomization was perfectly made, to do so would improve precision of the estimations. \( \alpha_{tsw} \) represents school-by-entry dummies. Before Krueger and Whitmore (2001), school fixed effects are usually entered to estimate the effects of CSR in STAR because randomization took place within schools. However, Krueger and Whitmore (2001) explains that because students were randomized not only within schools but also at their entry levels, controlling for school-by-entry dummies reflects the nature of randomization more clearly. Dee and West (2008) follow this rationale. Finally, \( \varepsilon_{tsw} \) is adjusted to reflect heteroscedasticity clustered at the school-by-entry-wave level.

In the second specification, \( SMALL_{tsw} \) in (1) is replaced by four dummies, which indicate number of years in small classes with a dummy of zero year being omitted. This specification is of interest because Finn and Achilles (1999) and Finn, Gerber, Achilles, and Boyd-Zaharias (2001) argue that as students stay longer in small classes, his test scores increase even further. It is possible to include the variable of the number of years in small classes rather than dummies of years in small classes, but the former imposes a linear restriction, which forces staying two years to have effects twice those of staying one year, and so on. By using the dummies, this arbitrary restriction is relaxed. In fact, it turns out that staying in S for one or two years does not have positive effects on test scores. Also, it is noteworthy that the dummies could be endogenous because student attrition is not random.

In the third specification, \( SMALL_{tsw} \) is replaced by class size instrumented by the status of S. Class size rather than class type is introduced because the ranges of the number of students are overlapped between S and RR and attrition took place nonrandomly. In addition, the variable of class size provides more precise effects of CSR of one student than the dichotomous variable , namely \( SMALL_{tsw} \). The rationale behind the instrument strategy is that class type is randomly assigned so S is correlated with a small number of students in a classroom but it is uncorrelated with \( Y_{tsw} \). Krueger (1999) and Krueger and Whitmore (2001) use this method,
V. Results

1. Descriptive Statistics

Table 1 reports descriptive statistics of independent and dependent variables. Descriptive statistics are listed only for effective and ineffective schools because statistics for schools other than effective schools are in between these two cases. By scrutinizing the statistics, one can understand differences between effective and ineffective schools and speculate about robustness of the effects of CSR in STAR.

Five points stand out. First, the number of ineffective schools is always, mostly more than twice, higher than that of effective schools. When the total number of STAR schools is concerned, effective schools account for only about one fourth of the total number. This is one of the main reasons that Hanushek (1999) and Ding and Lehrer (2005) raise doubts on the positive effects of S in STAR.

Second, the difference of the class sizes of RR and S is virtually zero between effective and ineffective schools. More specifically, the difference is 4.6 students and 4.3 students for effective and ineffective schools, respectively, and both numbers are statistically significantly different at the 5 percent significance level. The difference of 0.3 student is minuscule, however.

Third, all of the independent variables differ between RR and S in effective schools whereas it is not the case for ineffective schools. Of course, the treatment variable, namely class size, differs between RR and S in both types of schools. And yet, there are 5.6 percentage points more girls of S than RR in effective schools whereas no difference of it is found in ineffective schools. Also, 9.3 percentage points more White/Asian students are initially assigned to S than to RR in effective schools, but the difference for ineffective schools is only 2.3 percentage points. Moreover, the difference of the latter is only weakly significant. Similarly, 9.1 percentage points fewer students are eligible for free lunch in S than RR in effective schools, but there is no statistically significant difference at conventional levels in ineffective schools. Finally, students of S are 50.7 days younger than those of RR in effective schools whereas the difference is only 11.4 days for ineffective schools,
〈Table 1〉 Descriptive Statistics: Comparison by Initially Assigned Class Type within Each Type of Schools

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Effective School</th>
<th>~</th>
<th>Ineffective School</th>
<th>~</th>
<th>N</th>
<th>Effective School</th>
<th>~</th>
<th>Ineffective School</th>
<th>~</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Size</td>
<td>22.1 17.5</td>
<td></td>
<td>4.6 (0.061)***</td>
<td></td>
<td>2778</td>
<td>22.2 17.9</td>
<td></td>
<td>4.3 (0.042)***</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>Girls</td>
<td>0.453 0.509</td>
<td></td>
<td>-0.056 (0.023)***</td>
<td></td>
<td>2778</td>
<td>0.477 0.468</td>
<td></td>
<td>0.009 (0.014)</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>White</td>
<td>0.472 0.566</td>
<td></td>
<td>-0.093 (0.023)***</td>
<td></td>
<td>2778</td>
<td>0.706 0.729</td>
<td></td>
<td>-0.023 (0.013)*</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>Free Lunch</td>
<td>0.646 0.556</td>
<td></td>
<td>0.091 (0.022)***</td>
<td></td>
<td>2778</td>
<td>0.569 0.568</td>
<td></td>
<td>0.000 (0.014)</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>Days between birthday and 1977.1.1</td>
<td>1077.5 1128.2</td>
<td></td>
<td>-50.7 (9.8)***</td>
<td></td>
<td>2778</td>
<td>1110.7 1122.1</td>
<td></td>
<td>-11.4 (5.6)**</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>Dependent Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math Scores in Fourth Grade</td>
<td>-0.193 0.178</td>
<td></td>
<td>-0.371 (0.070)***</td>
<td></td>
<td>1003</td>
<td>0.089 0.074</td>
<td></td>
<td>0.016 (0.044)</td>
<td></td>
<td>2217</td>
</tr>
<tr>
<td>Math Scores in Sixth Grade</td>
<td>-0.123 0.222</td>
<td></td>
<td>-0.345 (0.057)***</td>
<td></td>
<td>1498</td>
<td>0.051 0.058</td>
<td></td>
<td>-0.006 (0.038)</td>
<td></td>
<td>3284</td>
</tr>
<tr>
<td>Math Scores in Eighth Grade</td>
<td>-0.149 0.242</td>
<td></td>
<td>-0.392 (0.062)***</td>
<td></td>
<td>1446</td>
<td>0.062 0.063</td>
<td></td>
<td>-0.000 (0.038)</td>
<td></td>
<td>3146</td>
</tr>
<tr>
<td>Reading Scores in Fourth Grade</td>
<td>-0.174 0.351</td>
<td></td>
<td>-0.525 (0.059)***</td>
<td></td>
<td>1429</td>
<td>0.039 0.114</td>
<td></td>
<td>-0.075 (0.041)*</td>
<td></td>
<td>2908</td>
</tr>
<tr>
<td>Reading Scores in Sixth Grade</td>
<td>-0.161 0.196</td>
<td></td>
<td>-0.357 (0.056)***</td>
<td></td>
<td>1406</td>
<td>0.054 0.083</td>
<td></td>
<td>-0.029 (0.039)</td>
<td></td>
<td>3290</td>
</tr>
<tr>
<td>Reading Scores in Eighth Grade</td>
<td>-0.209 0.171</td>
<td></td>
<td>-0.380 (0.063)***</td>
<td></td>
<td>1445</td>
<td>0.076 0.124</td>
<td></td>
<td>-0.048 (0.038)</td>
<td></td>
<td>3152</td>
</tr>
<tr>
<td>High School Graduate</td>
<td>0.752 0.793</td>
<td></td>
<td>-0.041 (0.030)</td>
<td></td>
<td>1106</td>
<td>0.780 0.801</td>
<td></td>
<td>-0.021 (0.018)</td>
<td></td>
<td>2589</td>
</tr>
<tr>
<td>Taking ACT or SAT</td>
<td>0.292 0.426</td>
<td></td>
<td>-0.134 (0.021)***</td>
<td></td>
<td>2778</td>
<td>0.343 0.341</td>
<td></td>
<td>0.003 (0.013)</td>
<td></td>
<td>5882</td>
</tr>
<tr>
<td>ACT or ACT-Equivalent Scores</td>
<td>18.52 19.48</td>
<td></td>
<td>-0.97 (0.33)***</td>
<td></td>
<td>892 19.56 19.56</td>
<td></td>
<td>-0.000 (0.219)</td>
<td></td>
<td>2009</td>
<td></td>
</tr>
</tbody>
</table>

Note: Standard errors are reported in parentheses. ***: p value < 0.01, **: p value < 0.05, *: p value < 0.10.
Overall, students of S in effective schools have characteristics that have positive effects on academic achievement. For example, Krueger (1999, Column 8 in Panel A of Table 5) finds that girls, White/Asian students, and students ineligible for free lunch outperform their counterparts by 4.39, 8.44, 13.07 percentile ranks, respectively.

Fourth, dependent variables are statistically insignificantly different between RR and S in ineffective schools at the five percent significance level whereas differences always exit for effective schools except the probability of graduating high school.

Fifth, academic achievement for ineffective schools is in between those of S and RR in effective schools regardless of measures used, not surprisingly, except the probability of graduating high school. For example, math scores in fourth grade are \(-0.193\) SD for students in RR and \(0.178\) SD for students of S in effective schools. On the other hand, the scores range \(0.074\) SD\(-0.089\) SD for students initially assigned to ineffective schools, and the difference is statistically insignificant. This pattern appears for all, except one, of the dependent variables.

The overall pattern is as follows. The difference of class sizes of RR and S are almost the same both in effective and ineffective schools, but almost all measures of academic achievement indicate that students of S are superior to students of RR in effective schools but not for ineffective schools. In other words, the treatment is the same, but the outcomes differ between effective and ineffective schools. Moreover, the academic achievement of students both in RR and S in ineffective schools is in between those of students in RR and R in effective schools, which indicates that some variables other than the difference of class sizes increase academic achievement of students of S and decrease that of students in RR in effective schools. In fact, observable student characteristics are distributed, intentional or not, in such a way as to yield the results, which makes one suspect that unobservable student characteristics are also unevenly distributed. Furthermore, the proportion of effective school is rather small relative to the entire STAR schools.

Sohn (2010) demonstrates that uneven distributions of observable characteristics in effective schools yield the same results for kindergarten through third grade, and this section sketchily demonstrates that the pattern persist throughout eighth grade. In fact, the persistence may be unsurprising because some important characteristics are fixed such as sex, race, free lunch status, and birthday. It could be argued that free
lunch status is not fixed, but from the point of view of students, this characteristic is given.

In examining descriptive statistics, one can surmise that the effects of small classes on academic achievement in STAR are driven by a small number of STAR schools and understand why effective schools are effective. Below, robustness is checked more systematically by testing whether small classes are effective to raise academic achievement even in schools other than effective schools.

2. Math Achievement in Grades 4, 6, and 8

Table 2 show effects associated with S on math in fourth, sixth, and eighth grades. Although not shown, similar results are obtained for fifth and seventh grades. As can be seen, positive effects are mostly driven by students in effective schools, the proportion of which is about a quarter of the samples. Specifically, in fourth grade, when all the samples are considered, students who were initially assigned to S scored 0.091 SD higher in math than students who were initially assigned to RR. And yet, when the estimation is made only for students in effective schools, the size of the positive effects becomes much larger, namely 0.247 SD. When students in schools other than effective schools are considered, the positive effects disappear. Not surprisingly, students in ineffective schools did not get any positive effects of S. Similar patterns are observed for sixth and eighth grades. Among students in effective schools, students of S in effective schools score 0.155 SD and 0.161 SD higher than students of RR in sixth and eighth grade, respectively. However, students of S in schools other than effective schools or ineffective schools do not show any positive effects.

The case of cumulative effects of S does not differ. For all the samples in fourth grade, one or two years in S do not persist. Three years in S yield 0.130 SD positive but weakly significant effects. Only students who stayed for four years in S show statistically significantly positive effects of 0.110 SD. Note that staying in S is likely to be endogenous as explained above. Hence, these positive effects capture effects of some factors (e.g., stable families) and cannot be entirely attributed to staying for four years in S.
### Table 2: Robustness of Effects of Class Size Reduction: Math

<table>
<thead>
<tr>
<th></th>
<th>Fourth grade</th>
<th></th>
<th>Sixth grade</th>
<th></th>
<th>Eighth grade</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Schools</td>
<td>Effective Schools</td>
<td>Other than Effective Schools</td>
<td>All Schools</td>
<td>Effective Schools</td>
<td>Other than Effective Schools</td>
</tr>
<tr>
<td>Small (OLS)</td>
<td>0.091 (0.033)**</td>
<td>0.247 (0.053)**</td>
<td>-0.021 (0.043)</td>
<td>0.028 (0.029)</td>
<td>0.155 (0.056)**</td>
<td>-0.011 (0.034)</td>
</tr>
<tr>
<td>1 Year</td>
<td>0.033 (0.072)</td>
<td>0.147 (0.092)</td>
<td>-0.003 (0.086)</td>
<td>-0.144 (0.095)</td>
<td>0.008 (0.054)</td>
<td>0.080 (0.106)</td>
</tr>
<tr>
<td>2 Years</td>
<td>0.043 (0.094)</td>
<td>0.206 (0.211)</td>
<td>-0.016 (0.105)</td>
<td>0.012 (0.131)</td>
<td>0.067 (0.058)</td>
<td>0.359 (0.116)**</td>
</tr>
<tr>
<td>3 Years</td>
<td>0.130 (0.076)*</td>
<td>0.161 (0.204)</td>
<td>0.121 (0.079)</td>
<td>0.009 (0.084)</td>
<td>0.024 (0.075)</td>
<td>0.012 (0.156)</td>
</tr>
<tr>
<td>4 Years</td>
<td>0.110 (0.044)**</td>
<td>0.292 (0.069)**</td>
<td>0.045 (0.052)</td>
<td>0.006 (0.056)</td>
<td>0.029 (0.039)</td>
<td>0.174 (0.067)**</td>
</tr>
<tr>
<td>Class Size (2SLS)</td>
<td>-0.021 (0.006)**</td>
<td>-0.042 (0.011)**</td>
<td>-0.012 (0.007)*</td>
<td>-0.004 (0.007)</td>
<td>-0.008 (0.006)</td>
<td>-0.030 (0.014)**</td>
</tr>
<tr>
<td>N</td>
<td>3974</td>
<td>1003</td>
<td>2971</td>
<td>2217</td>
<td>5942</td>
<td>1498</td>
</tr>
</tbody>
</table>

Note: Included in all the specifications, but not reported in this table are variables of race, gender, eligibility for free-lunch, days between birthday and January 1, 1977, its square term, school-by-entry wave fixed effects. Standard errors, adjusted for school-by-entry wave, are reported in parentheses. ***: p value < 0.01, **: p value < 0.05, *: p value < 0.10.
When effects of class size rather than class type are estimated, positive effects are observed mostly for students of S only in effective schools. For example, for all the samples in fourth grade, a class with one student smaller raises 0.021 SD, but these positive effects are largely due to effective schools. Similar patterns appear for sixth and eighth grades.

3. Reading Achievement in Grades 4, 6, and 8

As described in Table 3, positive effects of S are larger on reading although the pattern is similar to the case of math. For example, the average score of reading in fourth grade is 0.163 SD higher for the entire samples in fourth grade, but the size of the positive effects is much larger for students in effective students than students in schools other than effective schools. Students in ineffective schools did not benefit from being assigned to S across the grades.

When statistical significance is ignored, the number of years in S is positively associated with higher scores of reading throughout eighth grade, which is consistent with Finn and Achilles’s (1999) arguments. When statistical significance is considered, however, only staying in S for four years yield statistically significant effects except fourth grade. Also, as grades advance, the size of the positive effects of staying in S for four years decreases from 0.204 SD in fourth grade to 0.119 in eighth grade for the entire samples. The results are similar for students in effective schools although the size is larger.

When class size rather than class type is considered, even students in ineffective schools benefit from smaller sizes of classes for reading in fourth grade. And yet, the positive effects disappear beyond fourth grade at conventional levels of significance. Positive effects of smaller class sizes seem to persist for students in schools other than effective schools in fourth and eighth grade, but the size is about half of that for effective schools,
### Table 3: Robustness of Effects of Class Size Reduction: Reading

<table>
<thead>
<tr>
<th></th>
<th>Fourth grade</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>Sixth Grade</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>Eighth Grade</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>Ineffective Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (OLS)</td>
<td>0.163 (0.035)***</td>
<td>0.320 (0.093)***</td>
<td>0.113 (0.038)***</td>
<td>0.045 (0.046)</td>
<td>0.071 (0.028)***</td>
<td>0.179 (0.064)***</td>
<td>0.037 (0.032)</td>
<td>0.009 (0.035)</td>
<td>0.088 (0.029)***</td>
<td>0.169 (0.073)***</td>
<td>0.065 (0.031)***</td>
<td>0.008 (0.036)</td>
<td></td>
</tr>
<tr>
<td>1 Year</td>
<td>0.051 (0.057)</td>
<td>0.171 (0.099)*</td>
<td>0.020 (0.066)</td>
<td>-0.104 (0.075)</td>
<td>0.012 (0.054)</td>
<td>0.064 (0.110)</td>
<td>-0.001 (0.062)</td>
<td>-0.056 (0.069)</td>
<td>0.062 (0.054)</td>
<td>0.153 (0.123)</td>
<td>0.041 (0.060)</td>
<td>0.068 (0.070)</td>
<td></td>
</tr>
<tr>
<td>2 Years</td>
<td>0.154 (0.075)**</td>
<td>0.308 (0.168)*</td>
<td>0.104 (0.081)</td>
<td>0.055 (0.102)*</td>
<td>0.077 (0.066)</td>
<td>0.337 (0.146)**</td>
<td>0.008 (0.076)</td>
<td>-0.019 (0.096)</td>
<td>0.060 (0.062)</td>
<td>0.237 (0.166)</td>
<td>0.016 (0.068)</td>
<td>-0.022 (0.082)</td>
<td></td>
</tr>
<tr>
<td>3 Years</td>
<td>0.188 (0.076)**</td>
<td>0.302 (0.223)</td>
<td>0.151 (0.073)**</td>
<td>0.101 (0.081)</td>
<td>0.076 (0.074)</td>
<td>0.087 (0.183)</td>
<td>0.072 (0.078)</td>
<td>0.014 (0.092)</td>
<td>0.078 (0.088)</td>
<td>-0.071 (0.275)</td>
<td>0.122 (0.079)</td>
<td>0.050 (0.093)</td>
<td></td>
</tr>
<tr>
<td>4 Years</td>
<td>0.201 (0.054)***</td>
<td>0.370 (0.121)**</td>
<td>0.147 (0.059)**</td>
<td>0.095 (0.068)</td>
<td>0.108 (0.039)**</td>
<td>0.218 (0.088)**</td>
<td>0.067 (0.042)</td>
<td>0.066 (0.048)</td>
<td>0.119 (0.040)***</td>
<td>0.219 (0.086)***</td>
<td>0.086 (0.045)*</td>
<td>0.060 (0.053)</td>
<td></td>
</tr>
<tr>
<td>Class Size (2SLS)</td>
<td>-0.040 (0.006)***</td>
<td>-0.062 (0.015)***</td>
<td>-0.033 (0.007)***</td>
<td>-0.024 (0.008)***</td>
<td>-0.015 (0.006)***</td>
<td>-0.037 (0.015)***</td>
<td>-0.008 (0.007)</td>
<td>-0.006 (0.007)</td>
<td>-0.026 (0.007)***</td>
<td>-0.043 (0.014)***</td>
<td>-0.020 (0.007)***</td>
<td>-0.012 (0.008)</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>5483</td>
<td>1429</td>
<td>4054</td>
<td>2908</td>
<td>5948</td>
<td>1496</td>
<td>4452</td>
<td>3290</td>
<td>5745</td>
<td>1445</td>
<td>4300</td>
<td>3152</td>
<td></td>
</tr>
</tbody>
</table>

Note: Included in all the specifications, but not reported in this table are variables of race, gender, eligibility for free-lunch, days between birthday and January 1, 1977, its square term, school-by-entry wave fixed effects. Standard errors, adjusted for school-by-entry wave, are reported in parentheses. ***: p value < 0.01, **: p-value < 0.05, *: p-value < 0.10.
4. High School Outcomes

Table 4 shows that, consistent with math and reading achievement in fourth through eighth grades, high school outcomes are mostly driven by effective schools. Although being initially assigned to S does not improve the probability of graduating high school, it raises the probability of taking the ACT or SAT by 2 percentage points. Although specifications slightly differ, this result is similar to Krueger and Whitmore’s (2001, Column 3 in Table 5) marginal effects, namely 1.7 percentage points. And yet, statistically significant effects at conventional levels are observed only for students attending effective schools with much improvement in the probability.

Because the probability of taking the ACT or SAT is higher for students of S, less skilled students would be more likely to take the ACT or SAT. If this is the case, students initially assigned to S do not necessarily earn higher scores in the ACT or SAT. In fact, ACT-converted scores are not statistically different between students of S and RR. Because of this endogeneity problem, Krueger and Whitmore (2001) use the Heckman selection model and linear truncation method to demonstrate that the students initially assigned to S earned higher scores. Because the assumption of normal errors for the Heckman selection model is too strong, they use the linear truncation method, which does not rely on the assumption. And yet, because both methods are imperfect, they derive bounds to check whether the results obtained from both methods are included in the bounds. We do not delve into such an extent, but it suffices for the purpose at hand to show that students of S in effective schools account for almost all the improvement in ACT-converted scores. In fact, only students of S in effective schools show higher scores of 0.449 SD although the estimate is weakly significant.

When the number of years in small classes is concerned, some interesting facts emerge. Before analyzing the results, one point needs to be pointed out. Notice that the dummies of interest are different from those of Nye, Hedges, and Konstantopoulos (1999). Our dummies indicate only one year in S, only two years in S, and so on whereas the dummies in the latter indicate one year in S, two and more years in S, and so on. Hence, our estimates are not directly comparable to

3) A logit or probit model yields similar results.
**Table 4** Robustness of Effects of Class Size Reduction: High School Outcomes

<table>
<thead>
<tr>
<th></th>
<th>Graduate</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>Take ACT or SAT</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>ACT Scores</th>
<th>All Schools</th>
<th>Effective Schools</th>
<th>Other than Effective Schools</th>
<th>Ineffective Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small (OLS)</td>
<td></td>
<td>0.008 (0.014)</td>
<td>-0.003 (0.035)</td>
<td>0.010 (0.015)</td>
<td>&lt;-0.000 (0.018)</td>
<td>0.020 (0.011)*</td>
<td>0.091 (0.020)**</td>
<td>-0.001 (0.012)</td>
<td>-0.012 (0.013)</td>
<td>0.110 (0.143)</td>
<td>0.449 (0.278)*</td>
<td>-0.003 (0.167)</td>
<td>0.001 (0.185)</td>
</tr>
<tr>
<td>1 Year</td>
<td></td>
<td>-0.022 (0.028)</td>
<td>-0.023 (0.061)</td>
<td>-0.021 (0.031)</td>
<td>-0.040 (0.036)</td>
<td>-0.085 (0.017)**</td>
<td>-0.049 (0.036)</td>
<td>-0.094 (0.019)**</td>
<td>-0.113 (0.021)**</td>
<td>0.421 (0.271)</td>
<td>0.094 (0.594)</td>
<td>0.498 (0.310)</td>
<td>0.499 (0.370)</td>
</tr>
<tr>
<td>2 Years</td>
<td></td>
<td>-0.010 (0.036)</td>
<td>-0.032 (0.091)</td>
<td>-0.006 (0.040)</td>
<td>-0.017 (0.048)</td>
<td>-0.007 (0.026)</td>
<td>0.036 (0.052)</td>
<td>-0.021 (0.030)</td>
<td>-0.016 (0.039)</td>
<td>0.311 (0.599)</td>
<td>0.984 (0.833)</td>
<td>0.137 (0.456)</td>
<td>0.357 (0.484)</td>
</tr>
<tr>
<td>3 Years</td>
<td></td>
<td>-0.011 (0.034)</td>
<td>-0.063 (0.071)</td>
<td>-0.001 (0.038)</td>
<td>0.017 (0.045)</td>
<td>0.049 (0.027)**</td>
<td>0.071 (0.059)</td>
<td>0.043 (0.031)</td>
<td>0.053 (0.036)</td>
<td>-0.016 (0.372)</td>
<td>0.083 (0.756)</td>
<td>-0.040 (0.426)</td>
<td>-0.364 (0.509)</td>
</tr>
<tr>
<td>4 Years</td>
<td></td>
<td>0.040 (0.019)**</td>
<td>0.028 (0.045)</td>
<td>0.043 (0.021)**</td>
<td>0.029 (0.021)</td>
<td>0.168 (0.021)**</td>
<td>0.272 (0.032)**</td>
<td>0.133 (0.025)**</td>
<td>0.116 (0.025)**</td>
<td>0.083 (0.186)</td>
<td>0.539 (0.293)*</td>
<td>-0.333 (0.221)</td>
<td>-0.306 (0.253)</td>
</tr>
<tr>
<td>Class Size</td>
<td>(2SLS)</td>
<td>-0.006 (0.003)**</td>
<td>-0.001 (0.006)</td>
<td>-0.008 (0.003)**</td>
<td>-0.006 (0.003)*</td>
<td>-0.007 (0.003)**</td>
<td>-0.018 (0.005)**</td>
<td>-0.003 (0.003)</td>
<td>-0.002 (0.003)</td>
<td>-0.046 (0.034)</td>
<td>-0.101 (0.060)*</td>
<td>-0.029 (0.041)</td>
<td>-0.039 (0.045)</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>4600</td>
<td>1106</td>
<td>3494</td>
<td>2589</td>
<td>10747</td>
<td>2778</td>
<td>7969</td>
<td>5882</td>
<td>3578</td>
<td>892</td>
<td>2686</td>
<td>2009</td>
</tr>
</tbody>
</table>

Note: Included in all the specifications, but not reported in this table are variables of race, gender, eligibility for free-lunch, days between birthday and January 1, 1977, its square term, school-by-entry wave fixed effects. Standard errors, adjusted for school-by-entry wave, are reported in parentheses, **: p value < 0.01, ***: p value < 0.05, *: p value < 0.10.
the dummies in the latter, and the dummies in the latter tend to yield higher estimates.

The probability of graduating high school is 4 percentage points higher for students of S in the whole samples. And yet, this improvement is driven by students of S in schools other than effective schools. Moreover, students attending S only for one year show a lower probability to take ACT or SAT, which is mostly driven by students of S in schools other than effective schools and in ineffective schools. And yet, this probability is statistically significantly higher for students who stayed in S for four years in any types of schools although the improvement is highest for students of S in effective schools. This pattern suggests that the variable could be endogenous indeed. If students, who stayed in S only for one year, belonged to unstable families and these family characteristics negatively affect academic achievement, it is possible to see the aforementioned pattern. In fact, although statistically insignificant, similar patterns can be found in math and reading scores in fourth through eighth grades, especially pronounced for math. This pattern does not seem to be observed for ACT-converted scores, but it appears for students who stayed in S for three and four years in schools other than effective schools and in ineffective schools, A possible reason is double endogeneity, namely, taking the ACT or SAT on the one hand and the number of years in S on the other hand. For example, students in unstable families may be less likely to take the ACT or SAT.

When class size is the variable of interest, students in a one student smaller class graduate high school 0.6 percentage point more. However, this effect is largely explained by students of S in schools other than effective schools and in ineffective schools, which is in stark contrast to previous results. It is difficult to come up with good explanations for this. Finally, the probability to take ACT or SAT is explained by students of S in effective schools, and the case is similar for ACT-converted scores.

VI. Conclusions

Two points need to be clear. If the positive effects of S on academic achievement are driven by effective schools, which seems to be the case as reported above, one
could argue that this result is unsurprising because effective schools are defined as schools where students benefited from S. And yet, the distinction between effective schools and ineffective schools is drawn based on test scores before fourth grade, in which sense, initial random assignment of S is exogenous to academic achievement in and beyond fourth grade. If the positive effects of S exist and persist, the effects should appear in advanced grades whether or not students attended effective schools even if the effects did not appear during the experiment. Moreover, if some effects disappear when a small number of samples are excluded from the estimation, these effects are considered “not robust,” by definition. Hence, if the carry-over effects of S disappear when a small number of schools, namely effective schools, are removed from the estimations, the carry-over effects are thought to be not robust.

Another could argue that because some of the observable student characteristics are controlled for, the uneven distributions of the characteristics should not have effects independent of the treatment on the academic achievement of students in effective schools. The issue at hand is not about controlling for observable characteristics, but about possible uneven distributions of unobservable characteristics that are relevant for academic achievement, but imperfectly controlled for. For example, the fact that more students eligible for free lunch in RR than S in effective schools is likely to indicate that more academically successful students, independent of class size, were initially assigned to S in effective schools but not other schools. It seems that S captures the effects of these unobservable characteristics.

This paper argues that the carry-over effects of small classes in STAR are not robust. In general, students initially assigned to S appear to outperform students initially assigned to RR in math and reading in and beyond fourth grade, but this happens mostly for students in a small number of STAR schools, namely effective schools. Once scrutinized, observable student characteristics are unevenly distributed in such a way as to raise the academic performance of students of S in effective schools but not for students of S in ineffective schools. And yet, the differences of class size of S and RR are almost the same in both effective and ineffective schools. Therefore, the treatment is the same in both effective and ineffective schools, but the positive effects are estimated only for effective schools during and beyond the experiment period. The natural question would be “what causes these biased pos-
itive effects?”. Considering the uneven distributions of observable characteristics, it seems probable that uneven distributions of unobservable characteristics are the culprit and that $S$ in effective schools captures these effects.

Non-robustness of small classes in STAR warns that researchers and policy-makers should be cautious when they rely on STAR to promote CSR. They often cite the positive effects of $S$ during and beyond the experiment period. However, it is still controversial whether or not the positive effects of $S$ actually exist during the experiment. Furthermore, this paper casts serious doubt on the carry-over effects.

It should be emphasized at the end that this paper does not claim that CSR is ineffective for raising academic achievement. At the outset, it is already stated that this paper focuses only the carry-over effects of $S$ on academic achievement in STAR. Although the carry-over effects are not robust, CSR might be more effective when it is combined with particular subjects, teaching methods, or student compositions. In extreme, it may be that CSR alone is not effective to raise academic achievement at all. So far, CSR seems to be studied in isolation of other factors, Finn, Pannozoo, and Achilles (2003) provide a survey which is closely related to this subject, but they presume that small classes are beneficial for academic achievement and selectively collect studies to explain this debatable presumption. The CSR in California shows that across-the-board CSR is not likely to succeed without rigorous preparations. Future research on CSR would be more fruitful to search factors complementary to CSR in order to optimize CSR.

∥REFERENCES∥


Ding, Weili and Steven F. Lehrer. 2005. Class size and student achievement: Experimental estimates of who benefits and who loses from reductions, Queen's Economics Department Working Paper 1046, Queen's University, Ontario, Canada.


