HARRIS SCHOOL WORKING PAPER
SERIES 06.06

WHAT HAVE RESEARCHERS LEARNED FROM PROJECT STAR?

Diane Whitmore Schanzenbach

What Have Researchers Learned from Project STAR?

Diane Whitmore Schanzenbach

Harris School

University of Chicago

August 2006

Thanks go to Alan Krueger, Dan Goldhaber and Diane Ravitch for helpful comments, and to Rachel Henry Currans-Sheehan for comments and research assistance. All remaining errors are my own.

Project STAR was a large-scale randomized trial of reduced class sizes in grades K-3. Because of the scope of the experiment, it has been used in many policy discussions. For example, the California state-wide Class Size Reduction was justified in part on the successes of Project STAR. Recent (failed) proposals for Federal assistance for class size reductions in the Senate were motivated by Project STAR research. Even the recent discussion of small schools often conflates the notion of small schools and smaller classrooms.

Because of the importance of Project STAR, it has been studied by many scholars looking at a wide variety of outcomes and even exploiting the randomization using its variation to understand variations in inputsother aspects of the education production function that do not directly relate to class size. This paper provides an overview of the academic literature using the Project STAR experiment.

1. What was Project STAR?

Project STAR was a randomized experiment that assigned students to a small-size class (target of 13-17 students), a regular-size class (target of 22-25 students), or a regular-size class with a full-time teacher's aide. Teachers were also randomly assigned to class types. It is important to note that randomization was done within school, so all analysis presented here is done looking at within-school differences by class size. The experiment took place in 79 Tennessee public schools for a single cohort of students in kindergarten through third grade in the years 1985-89. An eventual 11,600 students and 1,330 teachers took part in the experiment. The experiment was funded by the Tennessee State Legislature under Governor Lamar Alexander (later the Secretary of Education under President George H. W. Bush, and currently a U.S. Senator), at a total cost of approximately $12 million (Word et al., 1990, Mosteller, 1995).

In the ideal implementation of this experiment, students were to remain with the same randomly assigned class-type from kindergarten through the end of third grade. In practice, though, there were several major sources of deviation from this model. Students who entered a participating school while the cohort was in first, second or third grades were added to the experiment and randomly assigned to a class type. There were a substantial number of new entrants – 45 percent of eventual participants entered after kindergarten. An especially large group of students entered in first grade – fully one-third of first grade participants were new in grade one – in part because at the time kindergarten was not required in Tennessee. A relatively large fraction of students exited Project STAR schools (45 percent of overall participants), due to school moves, grade retention or grade skipping, which also caused deviations from the original plan (Hanushek, 1999). Students who were male, black or on free or reduced price lunch were more likely to both exit and enter Project STAR. In addition, in response to parental concerns about fairness to students, all students in regular and regular-aide classes were re-randomized in first grade.

Finally, a smaller number of students (about 10 percent of participants) were moved from one class type to another in a non-random manner. It has been reported that most of these moves were due to student misbehavior, and were not typically the result of parental requests for moves to small classes (Krueger, 1999).[1] This weakness of the experiment can be addressed through use of an "intent-to-treat" setup – that is, to use the variation caused by initial randomly assigned class type instead of the actual (possibly non-random) class type attended.

The experiment only manipulated class size, and did not provide additional teacher training, new curriculum or any other intervention. One exception is that teachers in 15 schools were offered a 3-day training seminar between years 2 and 3 of the experiment (i.e., as the students were entering second grade). The training was given to all teachers in the Project STAR grades in those schools, and occurred prior to teacher assignment to class type. Investigation of the impacts of the training has shown that

[1] Of all transitions, 25 percent were into small (more desirable) classes.

teachers receiving the additional training performed no better (or worse) than teachers who were not offered the training.

A few aspects of the sample may limit the validity of the study to be generalized to other settings. In order to be eligible to participate in the program, schools were required to have a minimum cohort size of 57 students – enough to sustain both a regular and regular/aide classroom of 22 students, and one small class of 15 students. As a result, the schools that participated are about 25 percent larger on average than other Tennessee schools (see Table 1). The implications of this requirement will be discussed in more detail below. Because of requirements imposed by the Legislature for geographic diversity, schools in inner-cities are over-represented and the mix of students included is more economically disadvantaged and more likely to be African-American than those in the state overall. Even though the percent non-white in STAR mirrors closely the percent in the United States overall (33 vs. 31 percent), this masks the fact that there were very few Hispanic and Asian students in Tennessee at the time compared to the rest of the nation. Finally, it is worth noting that average school spending in Tennessee was about three-fourths of the nationwide average, and teachers were less likely to have a master's degree. If additional resources have greater impacts when the baseline levels are already low, this might mean that schools with higher levels of spending could experience a smaller impact of class size reduction.

2. Was Randomized Assignment Maintained?

There have been volumes of research looking at the relationship between class size and student performance in non-experimental settings (summarized by Hanushek 1986, 1997), but Project STAR is the first (and only) large-scale experiment to address class size. An experiment typically offers more compelling evidence than a non-experimental study, because it allows researchers to isolate the impact of the policy they are trying to test. In the absence of an experiment, the effect of a policy may be confounded by other observable or unobservable factors that may be correlated with the policy. To solidify this idea, take the following model of student achievement as an example:

(1) $$Y_{ij} = aS_{ij} + bF_{ij} + \varepsilon_{ij}.$$

Here *Y* represents a measure of student achievement for student *i* in school *j*. *S* contains information on school-level inputs that impact achievement, *F* contains family inputs, and *ε* is an error term. Both *S* and *F* measure inputs over the child's entire lifetime, and may contain inputs that are not observable to the econometrician. These omitted factors lead to biased coefficients if the omitted variables are correlated with included variables. For example, if students are assigned to small classes or better teachers in a compensatory manner – perhaps because of low baseline test scores, or low levels of family inputs – but that information is not available to the researcher, the estimated impact of school resources will be biased. Similarly, bias will result if parents who are more involved in their children's education are more likely to push for a smaller class or better teachers, and parental involvement is not measured in the dataset.

The benefit of using a randomized experiment is that the treatment assignment is unrelated to any omitted characteristics. With a well-designed experimental assignment, a straightforward comparison of means by class type will provide an unbiased estimate of the impact of class size on achievement. In the case of (an idealized version of) Project STAR, the equation to be estimated is as follows (Krueger, 1999):

(2) $$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 AIDE_{cs} + X_{ics}\gamma + \alpha_s + \upsilon_{ics}$$

where *SMALL* and *AIDE* are indicator variables for small-class and regular/aide-class assignments, respectively, and *c* indexes class *c* in school *s*. *X* is a vector of student-level characteristics. When treatments are randomized, student-level covariates are not related to class assignment and their inclusion should not change the estimated effect on class size, but should just contribute to the overall explanatory power of the model. A school-level fixed effect, *α*, is included, so that identification of class-type effects are identified off of within-school comparisons. Finally, the error term υ contains class-level and individual-level components.

In practice, the non-random transitions and new entrants described above complicate the approach somewhat. Because of non-random transitions after initial assignment, it would be inappropriate to use current-year class type; instead, initial class-type assignment (the "intent-to-treat" measure) is used throughout all estimations in this paper.[2] That is, all impacts are measured with regard to the class that students were assigned to, and not the class that they actually attended. The intent-to-treat measure likely under-states the impact of small classes by up to 15 percent (Krueger, 1999).[3] As described above, new entrants into the program were randomly assigned to class types. So, even though new entrants in first, second, and third grades on average are more disadvantaged than the kindergarten entrants, randomization allows us to compare new entrants in each grade to other new entrants in the same school across class types. In practice, then, the school-level fixed effect in equation (2) is replaced with a fixed effect that combines school with a student's grade of entry (K, 1, 2 or 3) to the experiment.

Impacts of reduced class size are straightforward to measure as the within-school (and entry wave) difference between class types, provided the randomization was done correctly. A compelling check of randomization is to examine a pre-test to ensure that

---

[2] In some early work (for example Finn and Achilles, 1990, Word et al., 1990) current class size was used instead of initial assignment.

[3] This conservative "intent-to-treat" measure based on randomly assigned is typically considered preferable to models which measure the impact of actual class-type attended in cases in which there is non-random movement between classes. A simple example may help to illustrate this: if a child were moved from a regular class to a small class because his parents insisted on the move, it is also reasonable to assume that the parents are especially active in other aspects of the student's education, say by monitoring homework especially closely or providing other education-enhancing opportunities. The problem arises because we do not have perfect measures of the home environment. In the ideal case in which class type is randomly assigned, these home environment measures are uncorrelated with class type and their impacts are absorbed in the error term in equation 2. When the effect of actual (non-random) class attended is measured instead, some of the impacts of the active home environment also may be picked up because actual attendance may be correlated with this "home environment" component of the error term. Using the experimentally induced variation – even though not all students attend their assigned class type, and some students' test scores will "count" toward the regular class they were assigned to, even though they actually attended small classes – circumvents this problem, but provides an understatement of the true impact. Krueger (1999) provides a more detailed discussion of this matter.

there are no measurable differences in the dependent variable between class types before the program begins. Unfortunately, no baseline test measure was collected. Another way to investigate whether randomization was done properly is to compare student characteristics that are related to student achievement but cannot be manipulated in response to treatment, such as student race, gender and age. If there are no systematic differences in observable characteristics across class types, this provides support that the randomization was done properly. Table 2 presents student characteristics by entry wave and class type. The joint P-value for a test of equality across the columns is conditional on school fixed effects. The first four rows show that there are no systematic differences in background characteristics between class type along race, gender, free-lunch status and age. It is also apparent that later entrants to Project STAR are more disadvantaged, with a substantially higher fraction of later entrants on free lunch and likely to be older (which may signal that they were retained in grade).[4]

Another drawback is that initial random assignment was not recorded, rather initial *enrollment* is measured. If parents successfully lobbied for a class change in the days between class assignments and the beginning of school, it may be masked in the data. To test whether this is a serious limitation, Krueger (1999) was able to collect data on initial assignment from 18 participating schools for 1581 students. He found that only 0.3 percent of students failed to attend their initially assigned class type in kindergarten, and only one of those was moved into a small class from a regular-sized one. If rates were similar at the other schools, then this does not appear to be a serious limitation.

If families feel that they are well-served by attending smaller classes (or, are upset that their child was randomly assigned to a regular class), a response might be a differential attrition rate or better attendance rates by class type.[5] Rows 5 and 6 in Table

[4] In addition, tests of the interaction between race, gender and free lunch status and small-class assignment, and a model saturated with all the interaction terms between race, gender and free lunch status, show no statistically significant relationship between baseline characteristics and small-class assignment (p-values of 0.94 and 0.15, respectively.)

[5] Better attendance rates in small classes might also be caused by fewer classmates from whom to pick up

2 provide some evidence that this is true for the kindergarten entry cohort. Small class students were 3 to 4 percentile points more likely to stay at a Project STAR school through third grade, and on average missed a fraction of a day less during their kindergarten year.[6] If students who gain the most from small classes were the ones induced to stay, then the impact of small classes may be overstated during the experimental period. Long-term follow up data – which adds back in the early exits from Project STAR – will alleviate this problem. The differential attrition rate subsided in the entering waves after kindergarten. Row 7 confirms that there was indeed a "program" – students assigned to small classes had about 7 fewer students in their class than those assigned to non-small classes. There was no difference in class size between regular and regular/aide classrooms. Finally, row 8 previews the results described in the next section. With the exception of the second grade entry wave, by the end of the first year in Project STAR the students in small classes were statistically significantly outperforming those in non-small classes.

Finally, it is crucial that teachers were randomly assigned. If the most effective teachers were disproportionately placed with small (or regular) classes, then the class size effect would be picking up this effect as well. There is only limited data available to confirm random assignment of teachers, which is displayed in Table 3. Overall, almost all of the teachers are female, and about 80 percent of them are white. Average years experience ranges between 9 and 15, depending on the class type and year. In most cases, there is no within-school difference across teacher race, gender, experience level or highest level of education.[7] Where there is a significant difference in teacher

---

germs and illnesses.

[6] Another potential source of bias is selective withdrawal prior to kindergarten entry: Krueger (1999) reports that of initially assigned students, 10.4 percent assigned to small classes for kindergarten failed to enroll in kindergarten in the fall. Comparable figures for regular classes were 14.3 percent (difference with regard to small classes has t=1.86), and for regular with aide classes were 12.2 percent (difference between small classes t=0.86).

[7] As Hanushek (1999) points out, though, these characteristics – and most other observable characteristics – are not good predictors of teacher effectiveness. Recent work by Jacob and Lefgren (2005) indicates that principals can generally identify which teachers are at the extremes of effectiveness across their entire

characteristics across class type (for total experience in first grade, and master's degree in third grade), the "best" attributes (more experience, and higher degrees) are not more likely to be found in small classes, so there is no evidence that assignment was done based on seniority, measured qualifications, or anything else that would violate random assignment.

3. K-3 Test Score Results

Because of randomization, the impact of being assigned to a small class can be measured by comparing average test scores across class types. As described above, students were randomly assigned to small and regular classes within schools by entry wave, so all analysis controls for separate school effects for each entering cohort while estimating the treatment effect of being assigned to a small class. Results in this section report the estimated coefficient on small class treatment from a regression estimated from the following equation:

(3) $$Y_{igs} = \beta_{0g} + \beta_{1g} SMALL_{is} + \beta_{2g} AIDE_{is} + \beta_{3g} X_{is} + \alpha_{sw} + \varepsilon_{igs}.$$

Here g indexes the grade (K-8) of the test score. Both the *SMALL* and *AIDE* variables are measured as initial assignment, and not actual class attendance. The fixed effect varies by school and entry wave w. The coefficient on the control for classes with a teacher aide is not reported here. Because there is no impact of teacher aides on student performance, the small-class effect is similar whether or not aide classes are controlled.

The dependent variable is the mean math and reading score on the Stanford Achievement Test (SAT-9) for each grade. In cases that had missing test scores for one test but not both, the score for the non-missing test is used. Test scores are presented as z-scores – that is, the scores are standardized by subtracting the mean score for non-small classes and dividing by the standard deviation. The coefficient on the indicator variable

school, but are not able to distinguish between the middle 60-80 percent of teachers in terms of effectiveness. It may be reasonable to conclude that teachers were not allocated based on their potential impact if principals are not able to easily distinguish between most teachers.

for small class can be interpreted as the standard deviation impact of the treatment.

Table 4 reports the impact of initial assignment to a small class on student test scores, and Figure 1 represents the small-class impact graphically. As many researchers have found (Finn and Achilles 1990; Word et al. 1990; Word et al. 1994; Krueger 1999; Krueger and Whitmore 2001; and Nye et al. 2002), the table indicates that overall students benefit about 0.15 standard deviations from assignment to a small class. When the results are disaggregated by race, it appears that black students benefited about twice as much as whites (0.24 vs. 0.12 standard deviations) from being assigned to a small class. Krueger and Whitmore (2002) find that this result is largely driven by a larger treatment effect for all students regardless of race in predominantly black schools, suggesting that benefits from additional resources are higher in such schools. There is also a small, positive within-school interaction between small class and an indicator variable for black students, which means that black students gain a little more from small classes than their white classmates do. Both of these findings suggest that reduced class size might be an effective strategy to reduce the black-white achievement gap.

Similar – but less stark – differences appear between free lunch students (who must have a family income less than 185 percent of the poverty line to qualify) and non-free lunch students. In third grade, free-lunch students gain about 0.055 standard deviations more than non-free lunch students. Boys appear to have slightly larger small-class gains than girls, but the difference is not statistically significant (Whitmore 2005). There is also considerable heterogeneity in impact size based on teacher experience, as illustrated in Panel E. Whether or not the results are conditioned on school fixed effects, students with more experienced teachers show large, statistically significant gains. On the other hand, students who have a teacher with fewer than 5 years of experience show smaller and often not statistically significant increases from small classes. This could help explain the difference between the large impacts found in Project STAR and the disappointing results from the California state-wide class reduction (Bohrnstedt and

Stecher 2002). In California, many districts were forced to hire new, inexperienced teachers in order to reduce class size, and there is evidence in Project STAR that these inexperienced teachers are not made particularly more effective when given small classes.

One concern about the results is that they may be driven by Hawthorne Effects, meaning that students and teachers behaved differently simply because they were given special treatment. If this were the case, then benefits from small classes would dissipate if the class size reduction were made permanent and the students and teachers were no longer being studied and made to feel special. Krueger (1999) addresses this by investigating differences in achievement using the variation in regular-sized classes. There is little reason to think that Hawthorne Effects would cause some classes in the treatment group to behave differently relative to other treatment group classes. Class size in regular-sized classes ranged from 16 to 30 students, but the bulk of the distribution was between 20 and 26 students. Whether or not school effects are controlled, students in a regular class with slightly fewer members out-scored larger regular classes. The estimated magnitude of a one-student reduction in class size was consistent with the magnitude of the experimental results (which estimates the impact of a seven-student reduction).

Another concern might arise regarding generalizability of the findings to a larger population, because of the size restrictions in place for participation in the experiment. Schools were required to be large enough to support 3 classrooms per grade, and on average Project STAR schools were about 30 percent larger than schools across Tennessee or the United States (see Table 1). If larger schools are somehow more or less effective with additional resources, then the findings in Project STAR may not be generalizable to smaller school settings.[8] One way to attempt to address this using the available data is to compare the small-class advantage across the wide range of school

[8] Thanks to Dan Goldhaber for pointing this out.

sizes in Project STAR. In third grade, the largest school had over 200 students, while the smallest participating school had only 56 students. To test whether school size is related to the magnitude of the class size effect, I separated schools into enrollment quartiles and tested whether the small-class impact was related to school size. The lowest quartile had an average of just about 60 students, while the highest had more than twice that number at 130 students. There was no statistically significant relationship between school size and test performance, or between school size and the small-class impact. This suggests that the limitation of Project STAR to only relatively large school sizes is not a major drawback in terms of generalizability to other samples.

4. Follow-up Studies on Test Performance

In 4th grade, the class-size reduction experiment concluded and all students were returned to regular-sized classes. At the same time, the assessment test was changed from the SAT-9 to the Comprehensive Test of Basic Skills (CTBS). Both tests are multiple-choice standardized tests that measure reading and math achievement, and are taken by students at the end of the school year. The CTBS results are scaled in the same manner as the SAT-9, in terms of standard deviation units. One important difference in the data is that all students in public schools statewide who had ever participated in Project STAR are included in the follow-up study, even if they had been retained a grade.[9] As a result, some students took the 4th grade test in 1990, while others took it in later years or even took it more than once. In the analysis reported here, all scores from grade X – no matter what year a student was in that grade – are compared. In the event of multiple attempts at grade X's test, the first available score is used.[10] As before, all estimates are conditional on school-by-entry wave fixed effects.

---

[9] It is estimated that 20 percent of students had been retained a grade by eighth grade, but this probability did not vary with initial class assignment.

[10] Krueger and Whitmore (2001) used a different approach, and deducted a small number of points from the test scores of students who were retained. Since class-type is not correlated with the probability of grade retention, the results are robust to either approach.

Results for grades 4-8 are reported in Table 5 and illustrated in Figure 1. Overall, there is a persistent positive impact of small-class assignment that is statistically significant (or borderline significant) through 8th grade, as has been found in previous studies (Achilles et al. 1993; Nye et al. 1995; Krueger and Whitmore 2001). The magnitude of the gain is one-third to one-half the size that was observed while the students were in the experimental classes. When the results are disaggregated, though, the impact appears to remain stronger with black and free-lunch students than with more advantaged students.

Another potential measure of student achievement is whether they take either the SAT or ACT college-entrance exam, which can be used as an early proxy for college attendance.[11] In order to measure this, the ETS and ACT corporations matched Project STAR student data to their national databases of test records, as described in Krueger and Whitmore (2001, 2002). To examine whether assignment to a small class influences the college-entrance exam test-taking rate, a binary variable indicating that a college entrance exam was taken is the dependent variable in equation (3). The impact of small-class assignment on college test taking is included as the final column in Table 5. Overall, test taking rates increase by about 2 percentage points. For white students, the impact is small and not statistically significant. On the other hand, black students were 5 points more likely to take the SAT or ACT if they were assigned to a small rather than regular-size class.[12] This corresponds to a rate of 38 percent of black students assigned to small classes took at least one of the college entrance exams, compared with 33 percent in regular classes. The chance of such a large difference in test-taking rates between the small and regular class students occurring by chance is less than one in 10,000. Krueger and Whitmore (2002) interpret the magnitude of these effects as a reduction in the black-

[11] Dynarski and Schanzenbach have been collecting college performance data, so that soon a more direct measure of college behavior will be available.

[12] The numbers reported here are slightly different from those reported in Krueger and Whitmore (2001), because at that time data were only available for students who graduated on track with their kindergarten class. Updated data allow the impact on students who were retained to be studied.

white test-taking gap. In regular classes, the black-white gap in taking a college entrance exam was 12.9 percentage points, compared to 5.1 percentage points for students in small classes. Thus, assigning all students to a small class is estimated to reduce the black-white gap in the test-taking rate by an impressive 60 percent. After controlling for increased selection into the test among small-class students, the impact on test scores for blacks is 0.15 standard deviations – about the same as the test-score impact in third grade.

5. Follow-up Studies on Non-Test Outcomes

Increased investments in school quality may also affect the frequency of negative social outcomes such as crime, welfare receipt and teen pregnancy. For example, the Perry Pre-School Project (Schweinhart et al., 2005) was an intervention that increased pre-school quality, and has yielded large, persistent effects on outcomes of participants through age 40 despite disappointing impacts on standardized test scores when children were younger. To date, only limited outcomes have been studied in the Project STAR data, but as the sample ages more research should be done on outcomes such as earnings, criminal behavior and welfare utilization.

One outcome measure is criminal arrest data, which come from Tennessee State Department of Corrections records as reported in Krueger and Whitmore (2002). Also, because the match was only performed in Tennessee, any crime committed by a student who has moved to another state is not included in the data set, which likely leads to a downward-biased estimate of the difference in criminal behavior by class assignment.

Criminal convictions are in this sample are rare (but much more likely for males): only 7.5 percent of male Project STAR students were reported as having been arrested. Because the outcome is so rare, results are estimated using a probit model, where the dependent variable equals one if a student has been convicted of a crime. Using this approach, it is found that black males assigned to small classes are 1.2 percentage points less likely to be convicted of a crime than those in regular-size classes, but the impact is imprecisely measured (t-ratio = 0.77).

Sentence length is measured as the maximum sentence (in days) faced by individuals for their specific crimes. Data are not available on length of actual sentence or time served, but maximum sentence length provides a measure of the severity of the crime committed. The sentences range from one year for minor theft and drug offenses to 8-12 years for aggravated robbery and serious drug offenses. Students without convictions were assigned a zero sentence length. Regression results indicate that black males in small classes on average committed crimes that carried 12 fewer days (or 24 percent) of maximum prison time than their peers in regular-sized classes back in elementary school. This effect is also imprecisely measured (t-ratio = 0.88). Class size appears to have no effect on either crime rates or sentence length for white males.

Another outcome that has been measured is the teen birth rate. Birth records, like crime records, were matched in the State of Tennessee only. If both of a newborn child's parents were STAR students, the birth record is counted for both the mother and the father. Birth records were only available by calendar year. The analysis is restricted to births during 1997 and 1998 because most students graduated high school in 1998. Unlike all other outcomes discussed so far, the birth records were not made available to researchers on an individual basis and were aggregated up to a school-by-class type level. Small class assignment is associated with a statistically significant 1.6 percentage point (or 33 percent impact, t-ratio 2.29) lower teen birth rate for white females, but has no measured impacts on black females.

In addition to investigating straightforward impacts of the program, several researchers have exploited aspects of the randomization to answer other important questions about the education production function that are not directly related to class size. Dee (2004) uses the random assignment of teachers and students to investigate the impact of having a teacher of the same race as the student on achievement. He finds that having an own-race teacher increases a student's performance by a statistically significant 3-5 percentile rank points for both black and white students. Dee and Keys (2005) find that students perform better in math when they are randomly assigned to a teacher who is receiving merit pay. Graham (2005), Schanzenbach (2006) and Cascio and Schanzenbach (2006) all use randomization into class-types to identify impacts of peer composition on student achievement. Graham finds that being randomly assigned to a classroom with average peer test scores in the 75th percentile leads to a 1.1 standard deviation increase in student performance, while peers with average scores in the 25th percentile decrease own performance by 0.9 standard deviations. Schanzenbach finds that if a child's peers are made up of a higher-than-average fraction of girls, there is about a 2 percentile point positive impact on the student's own test score. The effect seems to be working through two channels: girls, on average, have higher test scores, but even after factoring that part out, there is a positive effect of more girls in the classroom. This suggests that girls may change the culture of a classroom – at least in the early grades – to facilitate more learning for both boys and girls. Cascio and Schanzenbach find that conditional on a student's own age, being older-than-average relative to one's classmates is associated with higher test scores.

6. Why Might Small Classes Matter?

As described previously, the Project STAR experiment only manipulated class size. There were no changes in curriculum, there was no additional teacher training at most schools, and few if any new teachers needed to be hired to implement this because of the limited scope. What, then, causes the impact on student performance?

Lazear (2001) puts forth a useful theory of educational production. In it, gains from class size reduction are driven by a decrease in the amount of time that the classroom is being disrupted. A simple summary of the model is as follows: a child is behaving in class at a given moment with probability p, and misbehaving with probability $(1-p)$. In the model, misbehaving might be disrupting class by talking or fighting, or could be as benign as asking questions that slow down the class or monopolizing the teacher's time. When there are n children in the classroom, p^n is the probability that the entire class is behaving and learning is taking place (assuming that p is independent across children). Assuming a constant disruption rate, having fewer students in the class means that in a larger fraction of time learning is taking place.

In the model, the impact of reducing class size depends not only on the size of the class, but also on the behavior of the students in it. As a result, the Lazear theory predicts that class size effects should be larger for classes with more poorly behaved students. There is some evidence that teachers report higher rates of misbehavior in predominantly black schools compared to predominantly white schools, so the theory fits in well with the observation that class size has a larger impact in black schools.

Another potential mechanism is that early interventions improve non-cognitive skills in addition to the cognitive skills measured by standardized test scores (Heckman and Krueger, 2005; Heckman 2006). During kindergarten (at least during the period studied here), a primary focus in the classroom is to build non-cognitive skills such as listening, sitting still and cooperating. To the extent that kindergarten teachers could be more effective in teaching these skills when in small classes, we might see an improvement in non-cognitive skills that will spill over to later outcomes even if cognitive gains are modest. I test two approaches to isolating impacts of small classes on non-cognitive skills. An indirect test for the impact of non-cognitive skills is to compare the observed increase on a non-test outcome to the increase in that measure that would be predicted by the improvement in test scores. For example, there was a large increase in college-test taking for black students assigned to small class, which is a signal that more small-class students are going on to college. Is that increase larger than what would be predicted by the test-score increase observed in third grade (or kindergarten)? To test this, compare the observed relationship between test-taking rates and standardized test scores in the early grades. Within regular-sized classes only, the probability that a black student takes a college entrance exam increases by 0.6 percentage points for each additional percentile rank attained on the third grade test (r-squared = 0.17). As reported in Krueger and Whitmore (2001), the percentile rank increase for black students in third grade is 7.6 points for small classes. The point estimate implies that 4.5 points of the observed 8 point increase in college-test taking for small class students can be attributed to increased test score performance in third grade alone.[13] When accounting for sampling variability, this is consistent with the entire increase being due to an increase in cognitive test scores, or a portion of the increase being due to improvements in non-cognitive skills.

A perhaps more direct way to isolate the impact on non-cognitive skills is to investigate a few measures of these skills collected in the data (such as listening, self-concept and motivation). Measures of self-concept and motivation are from the Self-Concept and Motivational Inventory (SCAMIN), which was given at the end of the school year in all grades. The measure has been found to be only moderately reliable, and is thought to be most reliable for middle-income, suburban students (Finn and Achilles 1990, Rouse and Cashin 2001).

Results are presented in Table 6. In columns (1) – (4), self-concept and motivation scores in standard deviation units are presented. As with other results, this table shows the coefficient on an indicator for initial small class assignment in a regression framework that controls for school-by-entry wave fixed effects. By the end of kindergarten, there are some apparent increases in self-concept (column 1) overall and for females, black students and students on free lunch, and an increase in motivation (column 3) for females. This impact dissipates entirely by the end of third grade (columns 2 and 4), which could be due to lack of validity in the test for older students or could reflect a

[13] Another approach is to predict college-test taking based on third grade test score and a small class indicator. Using this approach, the small class indicator does not predict college-test taking in a statistically significant manner (coefficient = 0.028, standard error = 0.027), which is consistent with the entire test-taking effect being driven by increased test scores. The approach yields similarly non-significant findings predicting test taking behavior on 8th grade scores and a small class indicator (coefficient = 0.030, standard error 0.020).

catching up of students in regular classes.

Columns (5) and (6) present scores for the listening subsection of the Basic Skills First test, a curriculum-based test given to all children in Project STAR. Like the self-concept and motivation measures, listening is measured imperfectly and may in fact reflect a combination of cognitive and non-cognitive skills. Here there are strong positive impacts of small classes on listening skills for most subgroups of students in kindergarten. By third grade, the impacts are smaller and only marginally significant (in contrast to the robust over time findings for the cognitive tests of math and reading), but this may suggest that one important mechanism for later improved outcomes is that small classes increase a student's non-cognitive skills. Overall, these approaches seem to indicate that increased non-cognitive skills caused by small classes may play a roll in the observed positive outcomes, but that the results are also consistent with the outcomes being driven by cognitive skills alone.

7. How Do Small Class Impacts Compare to Other Interventions?

As a policy intervention, is reducing class size an economically worthwhile investment? One way to measure this is to compare the long-term benefits of increased test scores to the costs of reducing class size. To do this, I update the Krueger and Whitmore (2001) cost-benefit analysis and solve for r, the internal rate of return, in the following equation:

(4) $$\sum_{t=1}^{4} C_t \Big/ (1+r)^t = \sum_{t=14}^{61} (E_t \beta \delta) \Big/ (1+r)^t ,$$

where C_t is the cost of reducing class size in year t, and E_t is annual earnings in year t. Additionally, I present net present value calculations for various assumptions about the rate of return. For this calculation, it is assumed that students begin kindergarten at age 5, begin working at age 18 and retire at age 65. The β term relates a one standard deviation gain in test scores during high school to an increase in future earnings. It would be preferable to have a parameter that relates increases in elementary school test scores to later earnings, I have found no such estimate in the literature. The δ term is the increase in test scores for students assigned to small classes.

The left hand side of the equation represents the present discounted value of the costs of reduced class size. C_t is the additional cost of reducing class size by 7 students, which requires a 47 percent increase in number of classes. Assuming that the additional cost is proportional to current average spending per pupil, the additional cost each year will be 47 percent of $10,551 – which is the average national spending based on 2001-2002 figures, inflated by the Consumer Price Index to 2005 prices. Since students on average received 2.3 years of class size reduction, those costs are reflected as full years of class size reduction in grades K and 1, and 0.3 times the additional cost in grade 2. In this calculation, no costs are borne in grade 3.

The right hand side of the equation measures benefits over a student's working life of reduced class size. The test score increase (δ) is 0.152 standard deviations, from column (4) of Table 4, and the estimate of β is a 0.20 increase in wages from a 1 standard deviation increase in test scores (from Neal and Johnson 1996). To forecast future earnings, average earnings for each age between 18 and 65 were calculated for 2005 from the 2006 March Current Population Survey, and future real earnings growth was assumed to be 0, 1 or 2 percent per year. Results are presented in Table 7.

Using these assumptions, the estimated internal rate of return from the class size effect in Project STAR ranges from 5 to 10 percent. The net present value of the investment ranges from $3,000-$50,000. Of course, exact numbers should be viewed with some caution, as the calculation is based on many assumptions that may or may not prove to be reasonable. The benefits only include future increased earnings, and ignore potential impacts on crime and other behavior since those results are on the margin of statistical significance. If future follow-ups of Project STAR find lasting impacts on other outcomes such as crime or welfare use, then those benefits should be added to the equation and the internal rate of return would increase.

Another way to think about whether the investment in smaller classes is worthwhile is to compare it to other proposed interventions. Some reforms such as improved teacher training have been shown to be ineffective (Jacob and Lefgren, 2004), and therefore do not make an appropriate comparison. The most promising interventions aside from smaller class size come from the new literature on school choice through either vouchers or charter schools. Comparing Project STAR results to the largest positive effects found in voucher experiments – the New York City results found in Howell et al. (2002) – small classes yield a 30 percent higher test score improvement, but also cost 50 percent more (Krueger and Whitmore 2002). Compared to recent work by Hoxby and Rockoff (2004) on three charter schools in Chicago that employed random assignment, small classes yield about a 35 percent higher test score improvement. It is difficult to compare the costs associated with charter schools, which are allotted approximately the same in per-pupil revenues from the school district but also generally raise substantial external funding. In order to determine which would be a better investment – school choice or reduced class size – we would need to have some measure of willingness to pay for improved test scores and compare the smaller improvements associated (so far) with choice to the larger cost associated with class size reduction.

8. Conclusion

Mosteller (1995) described Project STAR as "one of the most important educational investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools." Given the scarcity of large-scale educational experiments like Project STAR, it is important to learn as much as possible from the experiment. Researchers have combed through the experiment to learn not only about test score effects of reduced class size, but also to gain insight into classroom dynamics. Overall, Project STAR indicates that reducing class size is a reasonable economic investment: the benefits are sizeable and long-lasting, especially for black students, and the overall benefits outweigh the costs.

References

Achilles, C. M., Barbara A. Nye, Jayne B. Zaharias, and B. DeWayne Fulton (1993). "The Lasting Benefits Study (LBS) in Grades 4 and 5 (1990-1991): A Legacy from Tennessee's Four-Year (K-3) Class-Size Study (1985-1989), Project STAR." HEROS, Inc.

Bohrnstedt, G. W. & Stecher, B.M. (2002). *What we have learned about class size reduction in California.* Sacramento, CA: California Department of Education.

Cascio, Elizabeth U. and Diane Whitmore Schanzenbach (2006). "First in the Class? Academic Redshirting and Education Production," mimeo, Dartmouth College.

Dee, Thomas S. (2004). "Teachers, Race, and Student Achievement in a Randomized Experiment," *The Review of Economics and Statistics*, 86(1): 195-210.

Dee, Thomas S. and Benjamin J. Keys (2005). "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment", *Journal of Policy Analysis and Management* 23(3): 471-488.

Finn, Jeremy D., and Charles M. Achilles. (1990). "Answers and Questions About Class Size: A Statewide Experiment." *American Educational Research Journal* 27 (Fall): 557-77.

Graham, Bryan S. (2005). "Identifying Social Interactions through Excess Variance Contrasts," mimeo, University of California – Berkeley.

Hanushek, Eric A. (1986). "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 24 (September): 1141-77.

Hanushek, Eric A. (1997). "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19(2): 141-64.

Hanushek, Eric A. (1999). "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and fro Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis*, 21: 154-164.

Heckman, James J. (2006). "Skill Formation and the Economics of Investing in Disadvantaged Children," *Science*¸312(5782): 1900-1902.

Heckman, James J. and Alan B. Krueger (2005). *Inequality in America: What Role for Human Capital Policies?* Cambridge: MIT Press.

Howell, W., P. Wolf, D. Campbell, and P. Peterson. (2002). "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of*

Policy Analysis and Management. 21(2): 191-218.

Hoxby, Caroline M. and Jonah E. Rockoff. (2004). "The Impact of Charter Schools on Student Achievement," mimeo, Columbia Business School.

Jacob, Brian A. and Lars Lefgren (2004) "The Impact of Teacher Training on Student Achievement: Quasi- Experimental Evidence from School Reform Efforts in Chicago" *Journal of Human Resources*, 39(1).

Jacob, Brian A., and Lars Lefgren. (2005) "Principals as Agents: Subjective Performance Measures in Education." NBER Working Paper 11463.

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

Krueger, Alan B. and Diane M. Whitmore (2001). "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal* 111:1-28.

Krueger, Alan B. and Diane M. Whitmore (2002). "Would Smaller Classes Help Close the Black-White Achievement Gap?" in John E. Chubb and Tom Loveless, ed, *Bridging the Achievement Gap*. Washington: Brookings Institution Press.

Lazear, Edward P. (2001). "Educational Production," *Quarterly Journal of Economics*. August, 2001. pp. 777-803.

Mosteller, Frederick (1995). "The Tennessee Study of Class Size in the Early School Grades," *The Future of Children*, 5(2): 113-127.

Neal, Derek and William R. Johnson (1996). "The Role of Pre-Market Factors in Black-White Wage Differences," *Journal of Political Economy*, October, vol. 104: 869-895.

Nye, B., Fulton, B.D., Boyd-Zaharias, J., and Cain, V.A. (1995). The lasting benefits study, eighth grade technical report. Nashville, TN: Center of Excellence for Research in Basic Skills, Tennessee State University.

Nye, B., Hedges, L.V., & Konstantopoulos, S. (2002). Do low achieving students benefit more from small classes? Evidence from the Tennessee class size experiment. *Educational Evaluation and Policy Analysis*: 24, 201-217.

Rouse, Kimberly A. Gordon and Susan E. Cashin (2001). "Children's Self-Concept and Motivation Assessment: Initial Reliability and Validity," Paper presented at the Annual Conference of the American Psychological Association.

Schanzenbach, Diane Whitmore (2006). "Classroom Gender Composition and Student

Achievement: Evidence from a Randomized Experiment," mimeo, University of Chicago.

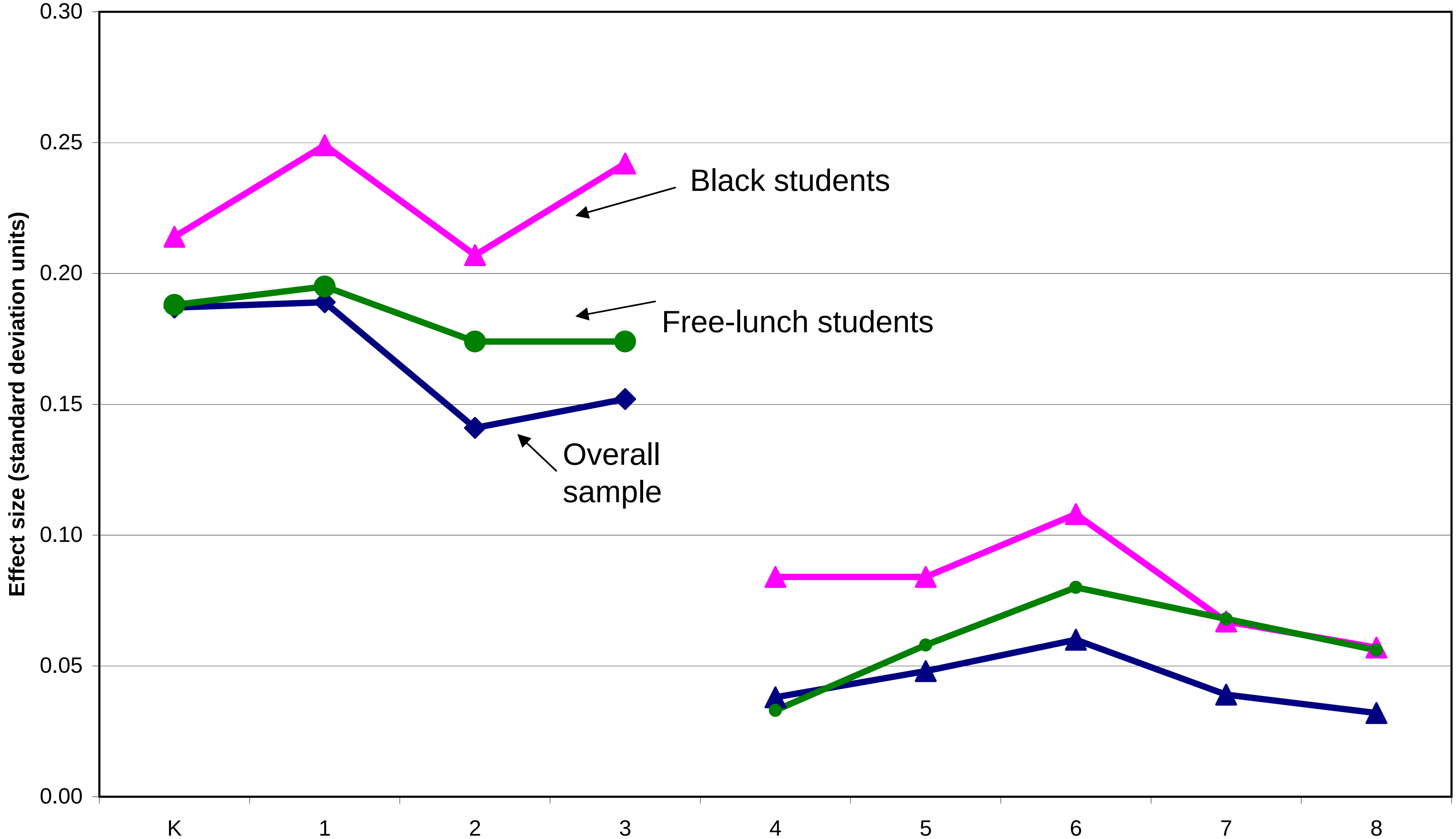
Schweinhart, Lawrence J., Jeanne Montie, Zongping Xiang, W. Steven Barnett, Clive R. Belfield, and Milagros Nores. (2005). *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40.*

Whitmore, Diane (2005). "Resource and Peer Impacts on Girls' Academic Achievement: Evidence from a Randomized Experiment," *American Economic Review*, May.

Word, Elizabeth, J. Johnston, Helen Pate Bain, et al. (1990). *Student/Teacher Achievement Ratio (STAR): Tennessee's K-3 Class Size Study.* Final summary report 1985-1990. Nashville: Tennessee State Department of Education.

Word, Elizabeth R., J. Johnston, Helen Pate Bain, et al. (1994). *The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project: Technical Report 1985-1990.* Nashville: Tennessee State Department of Education.

Figure 1: Small-Class Impacts on Test Scores



Notes: Data points represent the coefficient on a small-class indicator in a regression of grade-specific test scores on initial class type, school-by-entry wave fixed effects and student characteristics.

Table 1: Characteristics of students in STAR, Tennessee and the United States

	STAR (1)	Tennessee (2)	United States (3)
Percent minority students	33.1	23.5	31.0
Percent black students	31.7	22.6	16.1
Percent of children below poverty level	24.4	20.7	18.0
Percent of teachers with master's degree or higher	43.4	48.0	47.3
Average ACT score	19.2	19.8	21.0
Average 3rd grade enrollment across schools	89.1	69.5	67.1
Average current expenditures per student across schools	$3,423	$3,425	$4,477

Notes: With the following exceptions, data are from the 1990 Common Core of Data (CCD) from the Department of Education. For comparability, the Project STAR characteristics were calculated from the CCD. (Nevertheless, the characteristics were very similar when calculated directly from Project STAR data.) Teacher education data are for 3rd grade teachers from Project STAR data, and for 1993-94 public elementary and secondary school teachers from the Digest of Education Statistics. Race and poverty statistics for the U.S. are from the Census Bureau. ACT scores for Tennessee and U.S. are from ACT, Inc.

Table 2: Comparison of Mean Characteristics by Entry Wave

	(1)	(2)	(3)	(4)
Panel A: Students to entered STAR in Kindergarten				
	Small	Regular	Regular/Aide	P-value
1. Free lunch	0.47	0.48	0.50	0.46
2. White/Asian	0.68	0.67	0.66	0.66
3. Age in 1985	5.44	5.43	5.42	0.38
4. Female	0.49	0.49	0.48	0.87
5. Attrition rate	0.49	0.52	0.53	0.01
6. Days absent	10.0	10.5	10.9	0.01
7. Class size in kindergarten	15.1	22.4	22.8	0.00
8. Standardized test score in kindergarten	0.17	0.00	0.00	0.00
Panel B: Students to entered STAR in First Grade				
	Small	Regular	Regular/Aide	P-value
1. Free lunch	0.59	0.62	0.61	0.29
2. White/Asian	0.62	0.56	0.64	0.28
3. Age in 1985	5.78	5.86	5.88	0.12
4. Female	0.49	0.44	0.46	0.33
5. Attrition rate	0.53	0.51	0.47	0.07
6. Days absent	8.2	7.7	7.7	0.95
7. Class size in first grade	15.9	22.7	23.5	0.00
8. Standardized test score in first grade	-0.04	-0.24	-0.09	0.00
Panel C: Students to entered STAR in Second Grade				
	Small	Regular	Regular/Aide	P-value
1. Free lunch	0.66	0.63	0.66	0.58
2. White/Asian	0.53	0.54	0.44	0.15
3. Age in 1985	5.94	6.00	6.03	0.48
4. Female	0.43	0.45	0.46	0.56
5. Attrition rate	0.37	0.34	0.35	0.58
6. Days absent	NA	NA	NA	NA
7. Class size in second grade	15.5	23.7	23.6	0.00
8. Standardized test score in second grade	-0.11	-0.16	-0.27	0.40
Panel D: Students to entered STAR in Third Grade				
	Small	Regular	Regular/Aide	P-value
1. Free lunch	0.60	0.64	0.69	0.18
2. White/Asian	0.66	0.57	0.55	0.21
3. Age in 1985	5.95	5.92	5.99	0.40
4. Female	0.43	0.47	0.46	0.62
5. Attrition rate	NA	NA	NA	NA
6. Days absent	6.0	7.6	7.6	0.00
7. Class size in third grade	16.0	24.1	24.4	0.00
8. Standardized test score in third grade	-0.10	-0.20	-0.30	0.01

Notes: P-value is for F-test of equality of all three groups, conditional on school fixed effects. Sample size in Panel A ranges from 6299 to 6324, in Panel B ranges from 2240 to 2314, in panel C ranges from 1585 to 1679, and in Panel D ranges from 1202 to 1283. Free lunch pretains to the fraction receiving free lunch when they were randomly assigned. Test scores are scaled in units of the standard deviation of non-small classes, with mean non-small class score=0.

Table 3: Teacher Characteristics by Grade

	(1)	(2)	(3)	(4)
Panel A: Kindergarten Teachers				
	Small	Regular	Regular/Aide	P-value
1. White	0.866	0.780	0.838	0.327
2. Female	1.000	1.000	1.000	---
3. Master's Degree	0.323	0.364	0.374	0.653
4. Total Experience	9.0	9.1	9.7	0.404
Sample size	127	99	99	
Panel B: First-Grade Teachers				
	Small	Regular	Regular/Aide	P-value
1. White	0.813	0.835	0.820	0.462
2. Female	0.976	1.000	1.000	0.139
3. Master's Degree	0.358	0.313	0.390	0.674
4. Total Experience	12.2	10.7	12.7	0.069
Sample size	123	115	100	
Panel C: Second-Grade Teachers				
	Small	Regular	Regular/Aide	P-value
1. White	0.794	0.770	0.804	0.583
2. Female	0.992	0.980	0.972	0.291
3. Master's Degree	0.344	0.340	0.439	0.305
4. Total Experience	13.0	12.7	13.9	0.542
Sample size	131	100	107	
Panel D: Third-Grade Teachers				
	Small	Regular	Regular/Aide	P-value
1. White	0.775	0.798	0.776	0.650
2. Female	0.964	0.944	0.972	0.465
3. Master's Degree	0.377	0.427	0.523	0.077
4. Total Experience	13.3	13.6	14.8	0.368
Sample size	138	89	107	

Notes: P-values for F test of equality across the three class types, conditional on school fixed effects.

Table 4: Small Class Effects on Test Scores

	(1)	(2)	(3)	(4)
Panel A: Overall	Kindergarten	Grade 1	Grade 2	Grade 3
	0.187	0.189	0.141	0.152
	(0.039)	(0.035)	(0.034)	(0.030)
Panel B: By Race	Kindergarten	Grade 1	Grade 2	Grade 3
Black	0.214	0.249	0.207	0.242
	(0.074)	(0.063)	(0.054)	(0.060)
White	0.172	0.161	0.105	0.115
	(0.042)	(0.040)	(0.042)	(0.034)
Panel C: By Free Lunch Status	Kindergarten	Grade 1	Grade 2	Grade 3
Free Lunch	0.188	0.195	0.174	0.174
	(0.046)	(0.042)	(0.041)	(0.039)
No Free Lunch	0.177	0.194	0.126	0.118
	(0.051)	(0.047)	(0.047)	(0.041)
Panel D: Gender	Kindergarten	Grade 1	Grade 2	Grade 3
Male	0.209	0.192	0.144	0.172
	(0.041)	(0.040)	(0.039)	(0.400)
Female	0.157	0.180	0.132	0.122
	(0.049)	(0.040)	(0.042)	(0.040)
Panel E: Teacher Experience	Kindergarten	Grade 1	Grade 2	Grade 3
Low Experience (<=4 years)	0.310	0.057	0.073	0.171
	(0.121)	(0.081)	(0.064)	(0.100)
High Experience (>5 years)	0.181	0.269	0.179	0.154
	(0.044)	(0.056)	(0.037)	(0.034)

Notes: Standard errors clustered on classroom in parenthesis. Coefficients represent the impact of being randomly assigned to a small class on test scores, in terms of standard deviation units. Each cell represents a separate regression. Other covariates included in the model (where appropriate) were indicators for white, female and free lunch, and school-by-entry wave fixed effects.

Table 5: Small Class Effects on Long-Term Test Scores

	Grade 4 (z-score) (1)	Grade 5 (z-score) (2)	Grade 6 (z-score) (3)	Grade 7 (z-score) (4)	Grade 8 (z-score) (5)	Took college entrance test (1=yes) (6)
Panel A: Overall						
	0.035	0.048	0.060	0.040	0.036	0.024
	(0.025)	(0.024)	(0.025)	(0.025)	(0.025)	(0.010)
Panel B: By Race						
Black	0.078	0.080	0.105	0.066	0.063	0.050
	(0.048)	(0.043)	(0.045)	(0.042)	(0.046)	(0.018)
White	0.026	0.028	0.043	0.031	0.027	0.011
	(0.027)	(0.029)	(0.029)	(0.031)	(0.030)	(0.013)
Panel C: By Free Lunch Status						
Free Lunch	0.029	0.058	0.080	0.067	0.064	0.031
	(0.036)	(0.031)	(0.034)	(0.031)	(0.034)	(0.014)
No Free Lunch	0.048	0.036	0.027	0.003	-0.012	0.018
	(0.035)	(0.034)	(0.036)	(0.038)	(0.038)	(0.017)
Panel D: Gender						
Male	0.046	0.063	0.065	0.086	0.061	0.029
	(0.032)	(0.034)	(0.036)	(0.032)	(0.035)	(0.014)
Female	0.011	0.026	0.039	-0.017	0.002	0.015
	(0.037)	(0.033)	(0.031)	(0.035)	(0.034)	(0.016)

Notes: Standard errors clustered on initial school in parenthesis. Coefficients represent the impact of being randomly assigned to a small class on test scores, in terms of standard deviation units. Each cell represents a separate regression. Other covariates included in the model (where appropriate) were indicators for white, female and free lunch, and school-by-entry wave fixed effects.

Table 6: Small Class Impacts on Potential Measures of Non-Cognitive Skills

	Self-Concept		Motivation		Listening	
	Kindergarten	Grade 3	Kindergarten	Grade 3	Kindergarten	Grade 3
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Overall	0.135	0.024	0.031	-0.084	0.172	0.104
	(0.055)	(0.035)	(0.030)	(0.046)	(0.075)	(0.055)
Panel B: By Race						
Black	0.251	-0.032	0.080	-0.141	0.274	0.141
	(0.093)	(0.071)	(0.063)	(0.085)	(0.132)	(0.106)
White	0.083	0.052	0.008	-0.058	0.112	0.089
	(0.061)	(0.036)	(0.031)	(0.055)	(0.086)	(0.063)
Panel C: By Free Lunch Status						
Free Lunch	0.218	0.003	0.051	-0.111	0.253	0.125
	(0.076)	(0.053)	(0.048)	(0.073)	(0.098)	(0.078)
No Free Lunch	0.063	0.034	0.006	-0.002	0.090	0.064
	(0.062)	(0.045)	(0.034)	(0.068)	(0.087)	(0.062)
Panel D: Gender						
Male	0.088	0.061	-0.024	0.005	0.146	0.145
	(0.063)	(0.042)	(0.035)	(0.055)	(0.085)	(0.072)
Female	0.178	0.000	0.086	-0.153	0.204	0.067
	(0.064)	(0.050)	(0.038)	(0.066)	(0.083)	(0.068)

Notes: Standard errors clustered on classroom in parenthesis. All tests are scaled in standard deviation units. Self-concept and motivation scores are from the SCAMIN test, and listening scores are from the Basic Skills First assessment test.

Table 7: Cost-Benefit Analysis of Small Classes

Discount rate (*r*)	Net present value, assuming annual real wage growth of:		
	no growth	1% per year	2% per year
0.02	16,617	29,267	48,335
0.03	8,421	16,892	29,492
0.04	2,924	13,654	17,162
0.05	-821	3,178	8,965
Internal Rate of Return (*IRR*)	4.75%	5.82%	9.95%

Notes: All figures in 2005 dollars. Assumes 2.3 years' exposure to small classes and a test score increase of 0.152 standard deviation. A one standard deviation increase in average test scores is associated with a 0.2 increase in earnings.