NCEE 2010-4018

U.S. DEPARTMENT OF EDUCATION

Evaluation of the DC Opportunity Scholarship Program

Final Report

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Evaluation of the DC Opportunity Scholarship Program

Final Report

June 2010

Patrick Wolf, Principal Investigator, University of Arkansas
Babette Gutmann, Project Director, Westat
Michael Puma, Chesapeake Research Associates
Brian Kisida, University of Arkansas
Lou Rizzo, Westat
Nada Eissa, Georgetown University
Matthew Carr, Westat

Marsha Silverberg, Project Officer, Institute of Education Sciences

NCEE 2010-4018
U.S. Department of Education

ies NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Evaluation and Regional Assistance
John Q. Easton
Acting Commissioner

June 2010

This report was prepared for the Institute of Education Sciences under Contract No. ED-04-CO-0126. The project officer was Marsha Silverberg in the National Center for Education Evaluation and Regional Assistance.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the reports.

To order copies of this report,

- Write to ED Pubs, U.S. Department of Education, P.O. Box 22207, Alexandria, VA 22304.
- Call in your request toll free to 1-877-4ED-Pubs. If 877 service is not yet available in your area, call 800-872-5327 (800-USA-LEARN). Those who use a telecommunications device for the deaf (TDD) or a teletypewriter (TTY) should call 1-877-576-7734.
- Fax your request to 703-605-6794.
- Order online at www.edpubs.gov.

This report also is available on the IES website at http://ies.ed.gov/ncee.

Upon request, this report is available in alternate formats such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department's Alternate Format Center at 202-260-9895 or 202-205-8113.

Contents

Contents (continued)

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Figures

List of Figures (continued)

Acknowledgments

This report is the sixth of a series of annual reports mandated by Congress. We gratefully acknowledge the contributions of a significant number of individuals in its preparation and production.

Staff from the Washington Scholarship Fund provided helpful information and have always been available to answer our questions.

We are also fortunate to have the advice of an Expert Advisory Panel. Members include: Julian Betts, University of California, San Diego; Thomas Cook, Northwestern University; Jeffrey Henig, Columbia University; William Howell, University of Chicago; Guido Imbens, Harvard University; Rebecca Maynard, University of Pennsylvania; and Larry Orr, formerly of Abt Associates and now an independent consultant.

The challenging task of assembling the analysis files was capably undertaken by Yong Lee, Quinn Yang, and Yu Cao at Westat. The management and conduct of the data collection was performed by Juanita Lucas-McLean and Bonnie Ho of Westat. Expert editorial and production assistance was provided by Evarilla Cover and Saunders Freeland of Westat. Jeffery Dean of the University of Arkansas ably assisted with the intermediate outcomes analysis and the drafting of chapter 4, and appendices E and H. Daniel Bowen helped to confirm the contents of the tables and in the production of appendix F.

Disclosure of Potential Conflicts of Interests[1]

The research team for this evaluation consists of a prime contractor, Westat, and two subcontractors, Patrick Wolf (formerly at Georgetown University) and his team at the University of Arkansas Department of Education Reform and Michael Puma of Chesapeake Research Associates (CRA). None of these organizations or their key staff has financial interests that could be affected by findings from the evaluation of the DC Opportunity Scholarship Program (OSP). No one on the seven-member Expert Advisory Panel convened by the research team once a year to provide advice and guidance has financial interests that could be affected by findings from the evaluation.

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the particular tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Executive Summary

The *District of Columbia School Choice Incentive Act of 2003,* passed by Congress in January 2004, established the first federally funded, private school voucher program in the United States. As part of this legislation, Congress mandated a rigorous evaluation of the impacts of the Program, now called the DC Opportunity Scholarship Program (OSP). This final evaluation report presents the longer term effects of the Program on families who applied and were given the option to move from a public school to a participating private school of their choice.

The evaluation compares the outcomes of 2,300 eligible applicants randomly assigned to receive an offer (treatment group) or not receive an offer (control group) of an OSP scholarship through a series of lotteries. Although data on most of these outcomes—test scores, high school graduation, perceptions of school safety and satisfaction—were collected annually over four or five years, each year's estimated impacts are cumulative in that they represent students' entire educational experience between their application to the Program and the year the data were obtained. Some students offered scholarships never used them, while others used their scholarships to attend a participating private school at some point during the four- to five-year period. Based on analysis of the final, spring 2009 data we find:

- ***There is no conclusive evidence that the OSP affected student achievement.*** On average, after at least four years students who were offered (or used) scholarships had reading and math test scores that were statistically similar to those who were not offered scholarships (figure ES-2). The same pattern of results holds for students who applied from schools in need of improvement (SINI), the group Congress designated as the highest priority for the Program. Although some other subgroups of students appeared to have higher levels of reading achievement if they were offered or used a scholarship, those findings could be due to chance. They should be interpreted with caution since the results were no longer significant after applying a statistical test to account for multiple comparisons of treatment and control group members across the subgroups.

- ***The Program significantly improved students' chances of graduating from high school.*** Although students may not have raised their test scores in reading and math as a result of the OSP, they graduated at higher rates. The offer of an OSP scholarship raised students' probability of completing high school by 12 percentage points overall (figure ES-3). The graduation rate based on parent-provided information was 82 percent for the treatment group compared to 70 percent for the control group. The offer of a scholarship improved the graduation prospects by 13 percentage points for the high priority group of students from schools designated SINI in 2003-05 (79 percent graduation rate for the treatment group versus 66 percent for the control group).

- ***The OSP raised parents', but not students', ratings of school safety and satisfaction*** (figures ES-4 and ES-5). Parents were more satisfied and felt school was safer if their child was offered or used an OSP scholarship. The Program had no effect on students' reports on school conditions.

The DC Opportunity Scholarship Program

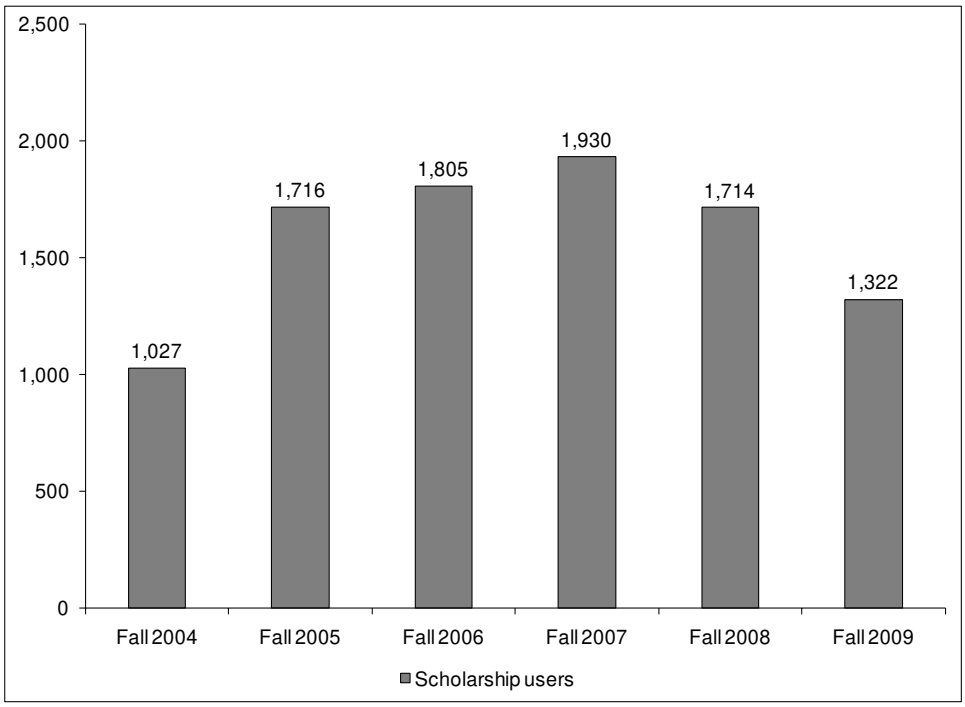
The purpose of the new scholarship program was to provide low-income residents, particularly those whose children attend schools in need of improvement (SINI) or corrective action under the *Elementary and Secondary Education Act*, with "expanded opportunities to attend higher performing schools in the District of Columbia" (Sec. 303). The scholarship, worth up to $7,500, could be used to cover the costs of tuition, school fees, and transportation to a participating private school. The statute also directed that scholarships be awarded by lottery any year in which there are more eligible applicants than available scholarships or open slots in private schools and that priority in the lotteries be given first to students attending SINI public schools.

The Program has been operated by the Washington Scholarship Fund (WSF). To date:

- 8,480 students have applied;
- 5,547 have been deemed eligible;
- 3,738 have been awarded Opportunity Scholarships; and
- 2,881 students used their scholarships within a year of receiving them.

The Program's $13-14 million annual appropriation has been sufficient to support about 1,700 scholarship students each year, if each student uses the full value of his or her scholarship. The Program enrolled 1,027 scholarship students in the fall of 2004, its initial year of partial implementation, and grew to its peak enrollment of 1,930 students in the fall of 2007 (figure ES-1). Language in a federal appropriations statute closed the Program to new applicants in the spring of 2009. In the fall of 2009, the OSP supported 1,322 continuing scholarship students attending 1 of the 52 private schools in the District participating in the Program that year.

Figure ES-1. Scholarship Users, Fall 2004-2009



SOURCE: WSF's enrollment and payment files.

The Congressionally Mandated Evaluation of the OSP

Guided by language in the statute, the evaluation of the OSP relied on lotteries of eligible applicants—random chance—to create two statistically equivalent groups who were followed over time and whose outcomes were compared to estimate Program impacts. A total of 2,308 eligible applicants in the first two years of Program implementation were entered into scholarship lotteries (492 in year one, called "cohort 1," and 1,816 in year two, called "cohort 2"). Across the cohorts, 1,387 students were randomly assigned to the impact sample's treatment group (offered a scholarship), while the remaining 921 were assigned to the control group (not offered a scholarship).

The OSP law also prescribed what types of impacts or outcomes would be assessed as part of the evaluation. These outcomes included student test-score performance in reading and math, educational attainment (in our case, parent reports of high school graduation), school safety, the success of the Program in expanding options (for which we have used "school satisfaction" as an indicator), and the effect of the OSP on District of Columbia Public Schools (DCPS) as well as private schools. To provide context for understanding the effects of the Program, the evaluation also provides a description of the patterns of school and student participation in the OSP. Data on these outcomes and issues were

collected primarily through annual surveys of parents, students in grade 4 or higher, and principals of both public and private schools in the District. Test scores were derived from evaluation-administered assessments using the SAT-9.[2]

The impacts of the Program were computed by comparing the outcomes of the treatment group with those of the control group, controlling for students baseline (pre-Program) reading and math scores and other demographic characteristics.[3] The impacts of the Program were assessed for the complete sample of eligible study participants as well as for several student subgroups, including the high priority set of students who applied from SINI public schools.[4]

Program Impacts

This final report on the OSP examines the effects of the Program on students and their parents near the end of the 2008-09 school year. The analysis is both consistent with and different from that presented in prior evaluation reports examining shorter-term impacts. It is consistent in that impacts are presented in two ways: the impact of the *offer* of an OSP scholarship, derived straight from comparing the average outcomes of the treatment and control groups, and (2) the impact of *using* an OSP scholarship, statistically adjusting for students who declined to use their scholarships. Like the earlier reports, the final estimates provide impacts on achievement, safety, and satisfaction.

Two parts of the analysis are different this year. First, in previous analyses, the two cohorts of students in the impact sample had the potential to experience the same number of years in the Program (e.g., three years after application). In spring 2009, the last year evaluation data were collected, cohort 1 students who applied in 2004 (14 percent of the sample) could have used their scholarships for five years

---

[2] By the 2008-09 school year, a total of 296 students (13 percent of the impact sample) had aged to the point where they would have completed 12th grade based on their grade upon application to the Program. The primary outcome measure used for the evaluation, the Stanford Abbreviated Achievement Test, ninth edition (referred to as the SAT-9 and published by Harcourt Educational Measurement in San Antonio, Texas), does not have a version for students beyond 12th grade, so these students effectively "graded-out" of the achievement portion of the study for purposes of this final impact report. Among the remaining 2,012 members of the impact sample, 69.5 percent of both the treatment and control groups effectively responded to test score data collection efforts in the final year of the study. The data they generated were then adjusted to account for nonrespondents before the impact analysis was conducted. For the other measures, response rates were obtained from 63 to 75 percent of eligible sample members, depending on the survey.

[3] There were no statistically significant differences in baseline measures between the treatment and control group samples overall and for those that provided data for this report's analyses.

[4] In all four years of the impact evaluation, the subgroups included students who had attended SINI 2003-05 and not SINI 2003-05 public schools at the time of application, lower baseline test-score performers and higher baseline test-score performers, and males and females. In the first three years of the evaluation, impacts were also estimated for two additional subgroup pairs: cohort 1 and cohort 2, and students entering grades K-8 and grades 9-12 at baseline. By the final year of the evaluation, all of the students in the baseline grade 9-12 subgroup and most of the students in the cohort 1 subgroup had aged to the point that they were no longer eligible for an achievement test, making the cohort and grade-level subgroups too small to analyze reliably.

while cohort 2 students who applied a year later (86 percent of the sample) could only have used their scholarship for four years. For this reason, we refer to impacts as "after at least four years" since a small portion of the sample—both treatment and control—were in the study a year longer.[5] Another important difference is that for the first time we are able to estimate the impacts of the Program on educational attainment. Most students who applied to the Program were in grades K-5. But by 2009, 22 percent of the impact sample (approximately 500 students) had aged to the point that they could have completed 12th grade and graduated from high school. This number of students was sufficient to reliably estimate impacts on this outcome; this is the first time random assignment has been used to estimate the causal relationship between a school voucher program (or private schooling) and educational attainment, thus providing a more rigorous estimate than previous studies that have addressed this issue. There are some limitations to this analysis, however: it is based on parent reports rather than school administrative records, and it represents a relatively small share of the study sample.

In examining the longer term impacts of the Program, we found:

Student Achievement

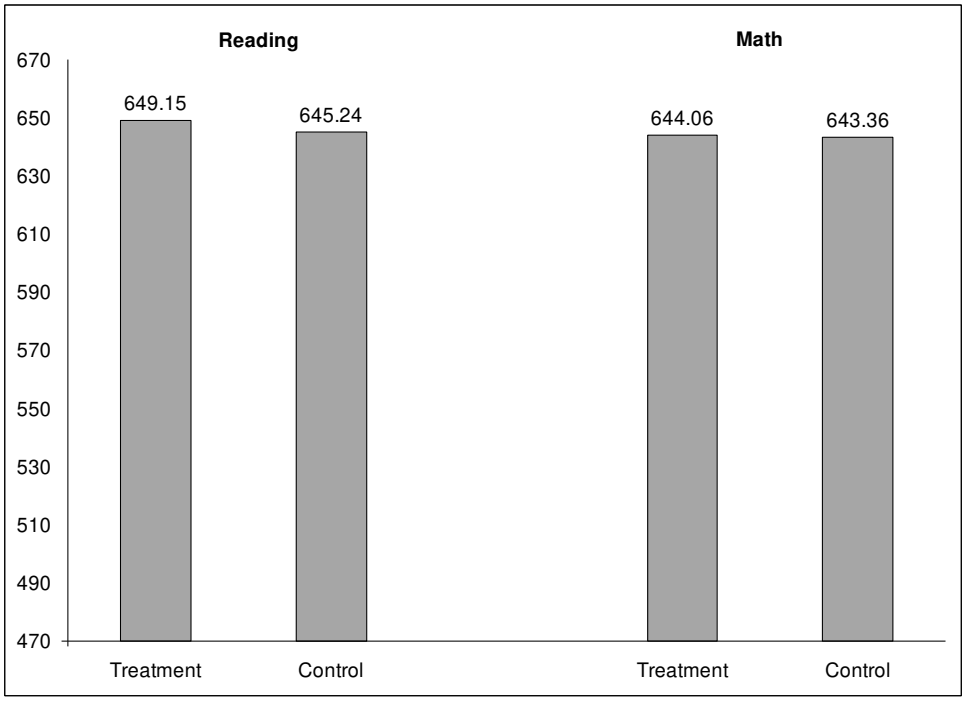
- Overall reading and math test scores were not significantly affected by the Program, based on our main analysis approach. On average over the 40-plus months of potential participation, the treatment group scored 3.90 points higher in reading and .70 points higher in math than the control group, but these differences were not statistically significant (figure ES-2).

- No significant impacts on achievement were detected for students who applied from SINI 2003-05 schools, the subgroup of students for whom the statute gave top priority, or for male students, or those who were lower performing academically when they applied.

- The Program may have improved the reading but not math achievement of the other three of six student subgroups. These include students who came from not SINI 2003-05 schools (by 5.80 scale score points), who were initially higher performing academically (by 5.18 points), or who were female (5.27 points). However, the impact estimates for these groups may be due to chance after applying a statistical test to adjust for multiple comparisons.

[5] Combining the two cohorts in this way was necessary to ensure that the sample size (number of students) for analysis was sufficient to detect impacts of a policy-relevant size and to provide results that could be applied to both cohorts. We were unable to collect data from cohort 1 in their fourth year after application because the legislative decision to extend the OSP and the evaluation came too late.

Figure ES-2. Achievement (SAT-9 Scale Score Points) After At Least Four Years



NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. The possible range of SAT-9 scale scores varies by grade level. The value at which the x-axis intersects the y-axis in this figure (470) represents the minimum average reading score possible given the grade composition of the control group sample in the final year. The minimum average math score possible for the control group sample was 502. The maximum possible reading score and math score was 835 and 832, respectively. Valid N for reading = 1,328; math = 1,330. Separate reading and math sample weights used.

High School Graduation (Educational Attainment)

- The offer of an OSP scholarship raised students' probability of completing high school by 12 percentage points overall. The graduation rate based on parent-provided information[6] was 82 percent for the treatment group compared to 70 percent for the control group (figure ES-3). There was a 21 percent difference (impact) for using a scholarship to attend a participating private school.

- The offer of a scholarship improved the graduation prospects by 13 percentage points for the high-priority group of students from schools designated SINI in 2003-05 (79 percent for the treatment group versus 66 percent for the control group) (figure ES-3). The impact of using a scholarship on this group was 20 percentage points.

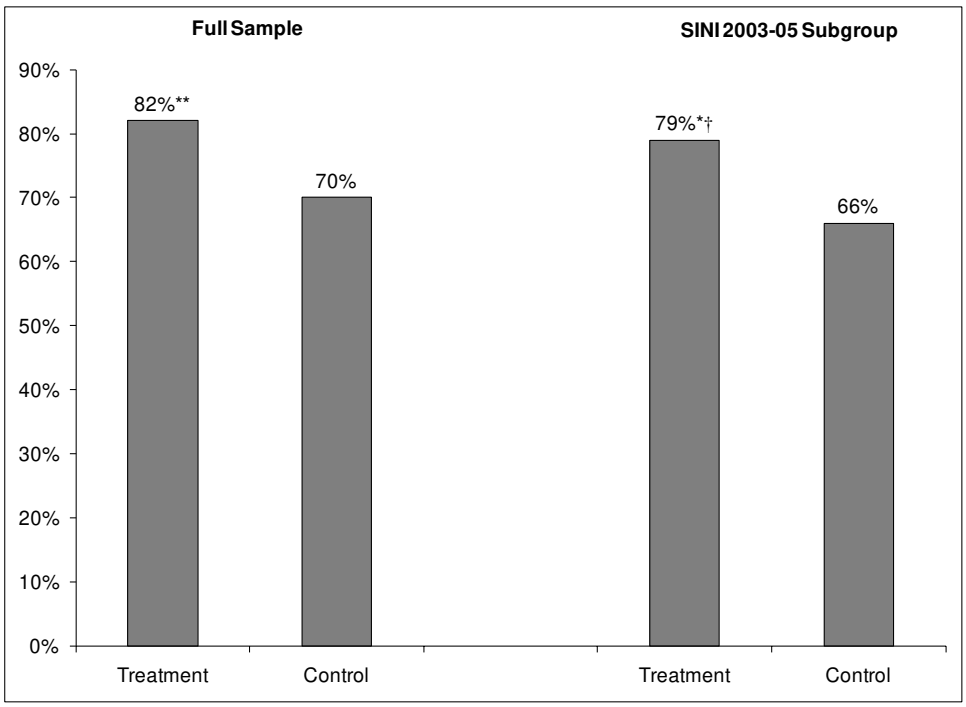
- Two other subgroups had statistically higher graduation rates as a result of the Program. Those who entered the Program with relatively higher levels of academic performance had a positive impact of 14 percentage points from the offer of a scholarship and 25 percentage points from the use of a scholarship. Female students

[6] These data were obtained through follow-up telephone surveys with parents of students in the study forecasted to have completed 12th grade by the summer of 2009. A total of 63 percent of parents in the target sample responded to this survey.

had a positive impact of 20 percentage points from the offer of a scholarship and 28 percentage points from the use of a scholarship.

- The graduation rates of students from the other subgroups were also higher if they were offered a scholarship, but these differences were not statistically significant.

Figure ES-3. High School Graduation Rates for the Overall Sample and the SINI 2003-05 Subgroup, 2008-09



*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

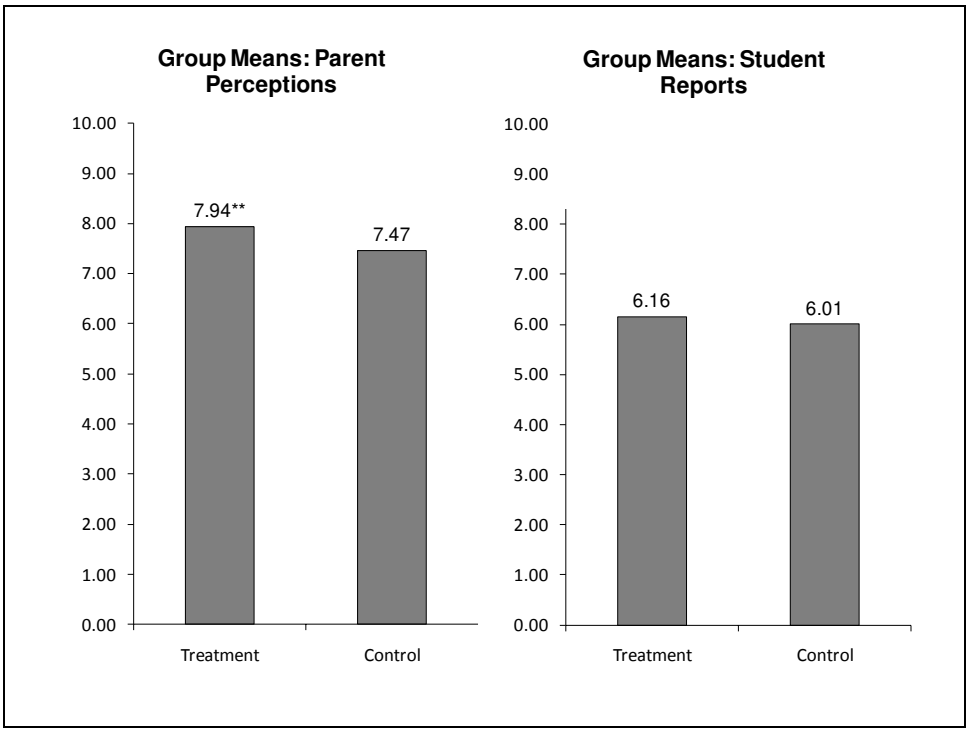
† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* = 316, including SINI 2003-05 *N* = 231, not SINI 2003-05 *N* = 85. High school graduation determined via parental self-reports.

School Safety and Satisfaction

At least four years after random assignment, the OSP had a positive impact overall on parents' ratings of school safety and satisfaction, but not on students' reports of those same outcomes (figures ES-4 and ES-5). For example, parents were 8 percentage points more likely to give their child's school a grade of A or B if offered a scholarship as compared with the control group; however, student reports of school satisfaction were comparable whether they were in the treatment or control groups.

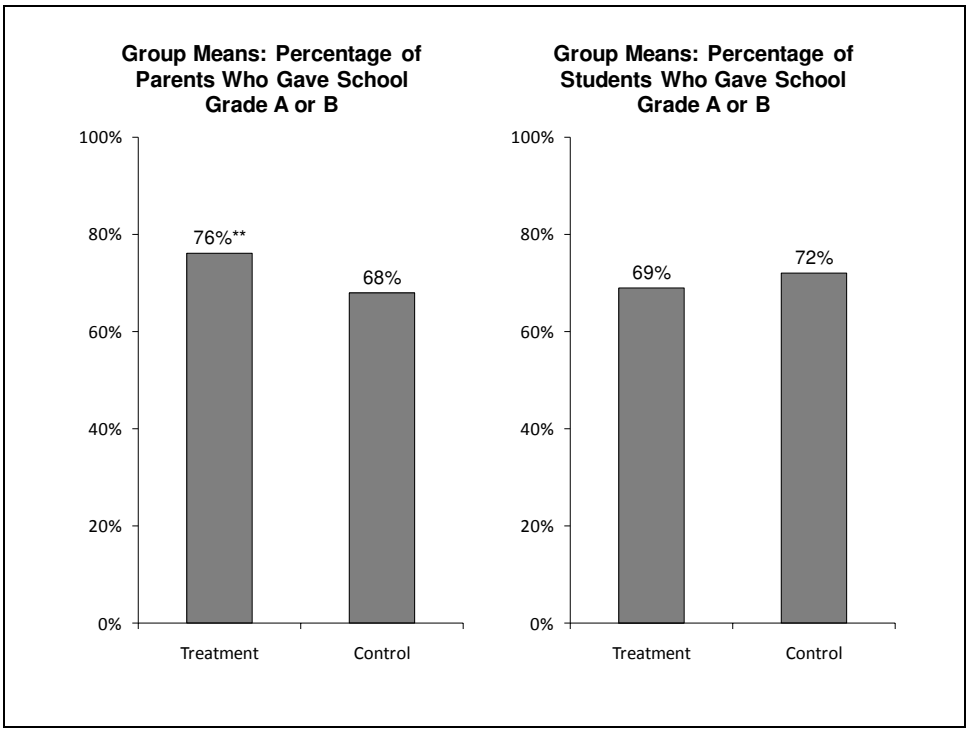
Figure ES-4. Parent Perceptions and Student Reports of Safety and an Orderly School Climate, 2008-09



**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Parent perceptions are based on a 10-point scale; student reports are based on an 8-point scale. For parent perceptions, valid N = 1,224; parent survey weights were used. For student reports, valid N = 1,054; student survey weights were used. The survey was given to students in grades 4-12. Means are regression adjusted using a consistent set of baseline covariates.

Figure ES-5. Parent and Student Reports of School Satisfaction, 2008-09



**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. For parent reports, valid *N* = 1,227; parent survey weights were used. For student reports, valid *N* = 1,001; student survey weights were used. The survey was given to students in grades 4-12. Means are regression adjusted using a consistent set of baseline covariates.

Program Context: Student and School Participation in and Response to the OSP

Understanding how and under what conditions the Program operated is important context for interpreting the impacts. For example, the degree to which students used their scholarships provides some signal of the attractiveness of the OSP and the ability of the Program and its participating schools to accommodate student needs. How the characteristics of the private schools differed from the public school options available may have influenced parent choices and students' educational experiences. Public and private schools' exposure to the OSP, through enrollment losses and gains, and any changes principals made to retain or attract students could indicate a more complete picture of the OSP and its potential for affecting the public and private schools in the area.

Students

As has been true in other school choice programs, not all students offered an OSP scholarship actually used it to enroll in a participating private school. And over the years, some students lost their eligibility for the Program. For example, by 2008-09, a total of 94 of the 1,387 members of the treatment group were no longer eligible to receive scholarships because they had "graded out" of the Program, which means that they would have moved beyond 12th grade. Looking across the remaining members of the impact sample's treatment group who had four (cohort 1) or five (cohort 2) years of potential Program participation:

- 282 out of 1,293 (22 percent) never used the OSP scholarships offered to them.
- 660 treatment students (51 percent) used their scholarships, but not consistently, during the school years after the scholarship award. Among these students are an estimated 147 who may have been forced by circumstances to stop using their scholarship. Students could become "forced-decliners" because the school they continued to attend converted from a participating Catholic school to a public charter school (confirmed for 35 treatment students),[7] their family income grew to exceed the Program's income limit (confirmed for 21 treatment students), their family moved out of DC (confirmed for 29 students), or they may have faced a lack of space for them in a participating high school when they transitioned from 8th to 9th grade (estimated for 62 treatment students).[8] Among the students who partially used their scholarship over at least four years after random assignment, 17 percent (9 percent of eligible treatment group students overall) used their OSP scholarship in 2008-09.
- The remaining 351 treatment group students (27 percent) used their scholarship during all years available to them after the scholarship lottery.

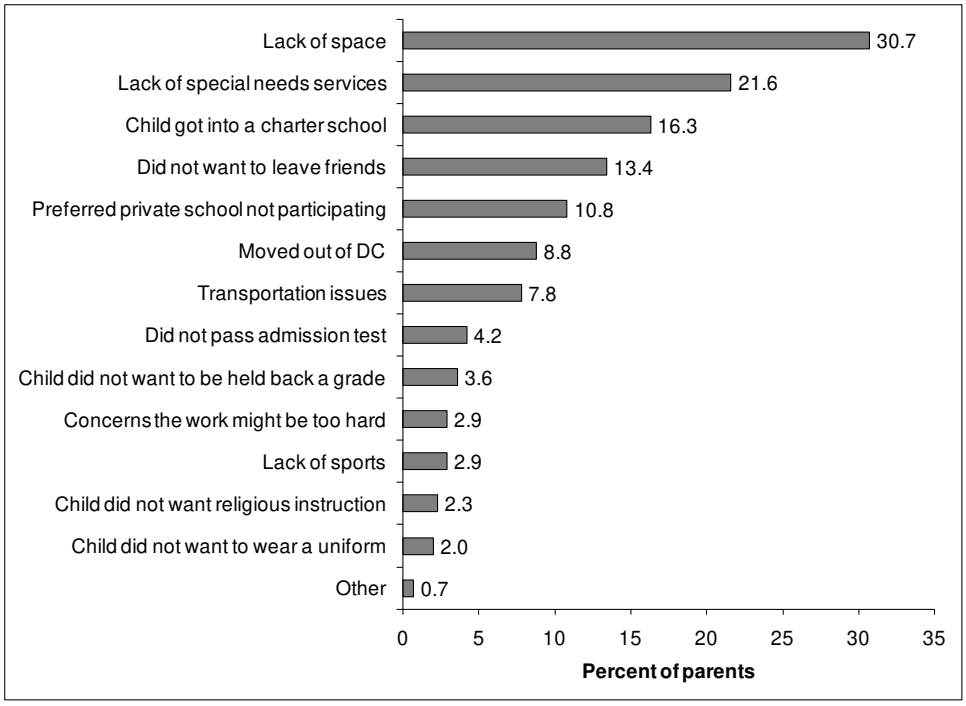
Across the years, the most common reasons given by parents for never using an OSP scholarship that was awarded to their child was a lack of space at their preferred private school (30.7 percent), the absence of special needs services (21.6 percent), and that their child was admitted to a preferred public charter school (16.3 percent) (figure ES-6).

---

[7] Based upon survey data, 35.9 percent of 97 treatment group students who used a scholarship to attend one of these Catholic schools in grades K-7 in 2007-08 continued to attend the same school when it converted to a public charter school in 2008-09.

[8] The estimate of the number of students forced to decline their scholarships due to the lack of high school slots was calculated by comparing the higher rate of scholarship continuation for 7th graders moving to 8th grade with the lower rate of scholarship continuation for 8th graders moving to 9th grade. The difference between those two continuation rates, applied to the number of OSP students moving from 8th to 9th grade, generates the estimate of forced decliners due to high school slot constraints of 62 (20 in year two plus 30 in year three plus 12 new cases in 2008-09). It is impossible to know for certain if all 62 of these students declined to use the scholarship solely or primarily because of high school slot constraints, and not for other reasons, or if some treatment students were forced to decline their scholarship at the very start due to high school slot constraints. It also is impossible to know if some students declined to even attempt to renew their scholarships because they knew their family exceeded the income limit, or how many treatment students moved out of DC and never informed the evaluators that they had "moved out" of Program eligibility. Therefore, the total estimate of 147 forced decliners for 2008-09 is simply an estimate based on the limited data available.

Figure ES-6. Reasons Given by Parents of Treatment Students for Never Using an OSP Scholarship

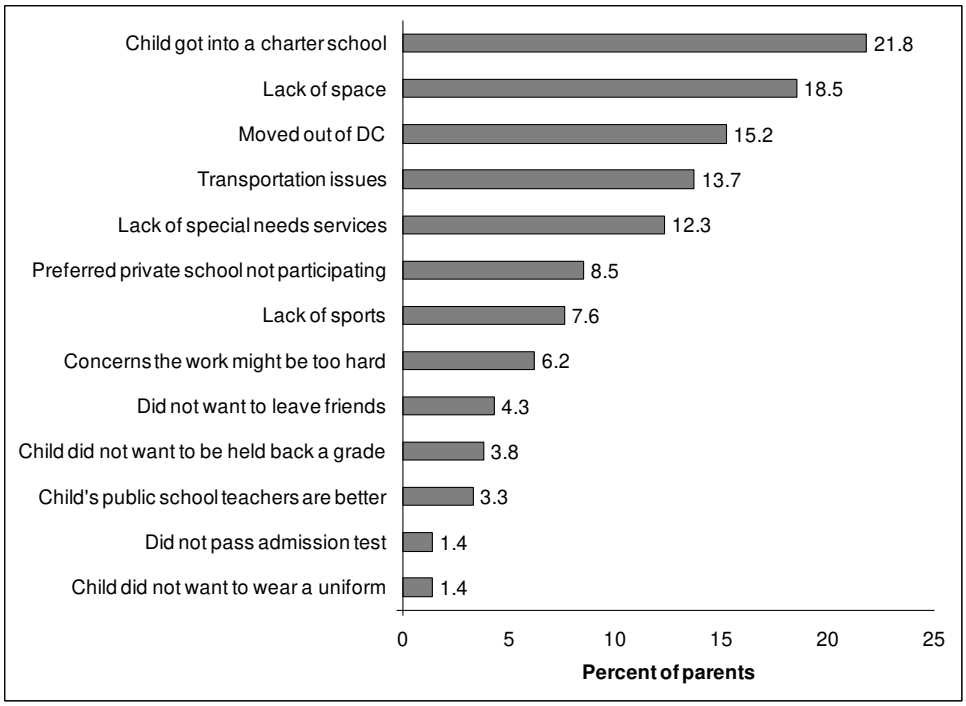


NOTES: Responses are unweighted. Respondents were able to select multiple responses each year, and some respondents participated in data collection for multiple years. Percentages given represent the sum of all responses obtained across years one through four of data collection (i.e., longitudinal responses) divided by the sum of all respondents (N = 306) across all of those same years (i.e., longitudinal respondents). As a result, this figure includes initial responses from parents of students who subsequently graded out of the Program. Categories with responses from fewer than three parents in any year are collapsed into the "Other reasons" category for confidentiality reasons.

SOURCE: Impact Evaluation Parent Surveys.

Among students who initially used a scholarship but then left the Program, the most common reasons for leaving were that the child was admitted to a preferred public charter school (21.8 percent), a lack of space at their preferred private school (18.5 percent), and that the family moved out of DC (15.2 percent) (figure ES-7).

Figure ES-7. Reasons Given by Parents of Treatment Students for Not Continuing to Use an OSP Scholarship



NOTES: Responses are unweighted. Respondents were the parents of treatment students who used a scholarship in a previous year but not in a subsequent year (N = 211). The reasons for not using were drawn from the parent responses the first year after their child stopped using a scholarship. Respondents appear in the data only one time (i.e., unique respondents), though they may have provided multiple reasons for not continuing to use a scholarship. This figure includes initial responses from parents of students who subsequently graded out of the Program.

SOURCE: Impact Evaluation Parent Surveys.

Schools

Fifty-two of 90 private schools in the District of Columbia were participating in the Program at the start of the 2008-09 school year, down from a peak of 68 schools in 2005-06.[9] Among the 22 schools that participated at some point but left the Program are seven Catholic schools that, in their last year in the Program (2007-08), enrolled 112 treatment group students; these schools converted to become public charter schools in 2008-09 and therefore no longer could be OSP voucher recipients. Overall in 2008-09, the last year of the evaluation, 14 percent of treatment group students attended a private school

[9] While, technically, 56 individual campuses were participating in the OSP from the start of the 2008-09 school year, the research team treats four of the schools with dual campuses as single entities because they have one financial office that serves both campuses, following the classification practice used by the National Center for Education Statistics in its Private School Survey. The 52 schools represent a net loss of nine schools since the prior year. Eleven schools stopped participating, while two new schools participated for the first time in 2008-09. The total number of private schools operating in DC declined from 109 in 2004-05 to 90 in 2008-09.

that charged tuition above the statutory cap of $7,500, and 80 percent attended a faith-based school, with most of them (53 percent) attending the 15 participating Catholic parochial schools (table ES-1).

Table ES-1. Features of Participating OSP Private Schools Attended by the Treatment Group in 2008-09

Characteristic	Weighted Mean
Charging over $7,500 tuition (percent of treatment students attending)	14.2%
Tuition	$7,252
Enrollment	292.1
Faith-based	81.7%
Archdiocesan Catholic	53.3%

NOTES: School *N* for tuition amounts and religious affiliations = 38; *N* for enrollment totals = 31. When a tuition range was provided, the mid-point of the range was used. The weighted mean was generated by associating each student with the characteristics of the school he/she was attending and then computing the average of these student-level characteristics.

SOURCES: OSP School Directory information, 2008-09, WSF; National Center for Education Statistics' Private School Survey, 2007-08.

The schools attended by the evaluation's treatment group (both those who used their scholarship to enroll in a participating private school and those who did not) differed in some ways from the schools attended by students who were not offered scholarships (the control group) in 2008-09:[10]

- Students in the treatment group were less likely than those in the control group to attend a school that offered special programs for students who may be academically challenged; these include programs or services for non-English speakers (32 vs. 57 percent) and for students with learning problems (75 vs. 90 percent);
- Students in the treatment group were less likely to be in schools with special programs for advanced learners (38 vs. 49 percent); and
- Students in the treatment group were less likely than those in the control group to attend a school with a cafeteria facility (76 vs. 91 percent), a nurse's office (50 vs. 82 percent), counselors (77 vs. 87 percent), and art programs (84 vs. 92 percent).

These features of the public and private schools in DC in 2008-09 could, hypothetically, reflect a response by the schools to the OSP. School choice theory suggests that a thriving private school scholarship program provides competition to the public schools and could generate improvements to the public school system, the private school system, or both (see Chubb and Moe 1990; Henig 1994). Such systemic changes could take place if significant percentages of students in the public school system, or in

[10] Differences in the characteristics of schools are noted here only if the difference was statistically significant at the .05 level or higher. In 2008-09, statistically similar proportions of the treatment and control groups were enrolled in schools that offered a computer lab (95 vs. 91 percent), separate library (77 vs. 79 percent), gyms (68 and 71 percent), individual tutors (58 vs. 63 percent), music programs (93 vs. 91 percent), and after-school programs (91 vs. 88 percent).

specific schools, apply for, receive, and use scholarships to transfer to private schools. Systemic changes also could occur in the private sector, if private schools adjust program operations to better attract or retain scholarship students.

As mandated in the statute, we examined how DC public and private schools were affected by the OSP by analyzing how these underlying components of the competitive school theory played out in DC. A maximum of 1,700 to 2,000 students—about 3 percent of those in DCPS public schools (including charter schools)—could be supported by the OSP to attend a private school in any year. We found that 3.2 percent of students in the DC public schools, cumulatively, used an OSP scholarship to transfer out between 2004 and 2009, joining continuing students in the Program. OSP-related student transfers ranged from 0 to 21 percent of enrollment across individual schools during that period. Just over one-quarter (28 percent) of public school principals reported making any changes to their operations in order to retain students who might be interested in the OSP or private schools in general (figure ES-8).[11] On average, OSP students made up 16 percent of participating private schools' student populations, with a range of 0 to 65 percent.[12] Fifty-two percent of principals at those schools indicated they had made changes to encourage OSP students to attend their schools (figure ES-9).[13]

These findings can be placed in some context. The cumulative exposure of DCPS to the Program across five cohorts of students using OSP scholarships represents less than one-fifth of the average annual mobility of students in the district (3 percent versus 20 percent).[14] Given these figures, OSP-related transfers to private schools may not have been distinguishable from the larger share of other student departures. In addition, school choice theory suggests that if any significant system-wide change in public schools is likely, the loss of students to the scholarship program should also entail a loss of funding for the public schools and school system affected by such transfers (Hoxby 2003). However, the law that established the OSP ensured that DCPS would gain, rather than lose, funds, and district officials

[11] The response rate for the public school principal survey in 2008-09 was 75 percent.

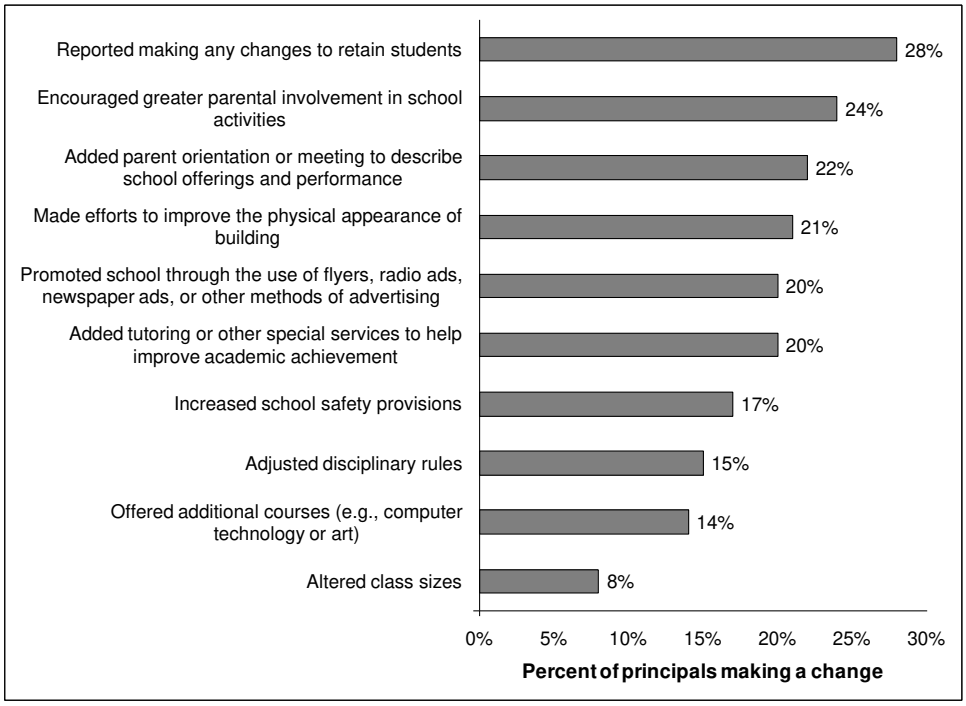
[12] Private schools were deemed by the WSF, the program operator, as participating if they agreed to take OSP vouchers even if no OSP students were admitted.

[13] The response rate for the private school principal survey in 2008-09 was 72 percent.

[14] "A student is defined as 'mobile' if the student attended a different school or was not enrolled in the snapshot from the prior month." http://www.osse.dc.gov/seo/lib/seo/dc_student_mobility_report2008_06_10.pdf. The DC Office of the State Superintendent of Education (OSSE) reported the average monthly student mobility rate in the district was about 2 percent from the fall of 2007 to the spring of 2008. Taken across a 10-month school year, that monthly rate translates into an annual average mobility rate of 20 percent.

were not given information to determine how many students left individual public schools as a result of the Program.[15]

Figure ES-8. Public School Responses to the OSP, 2008-09

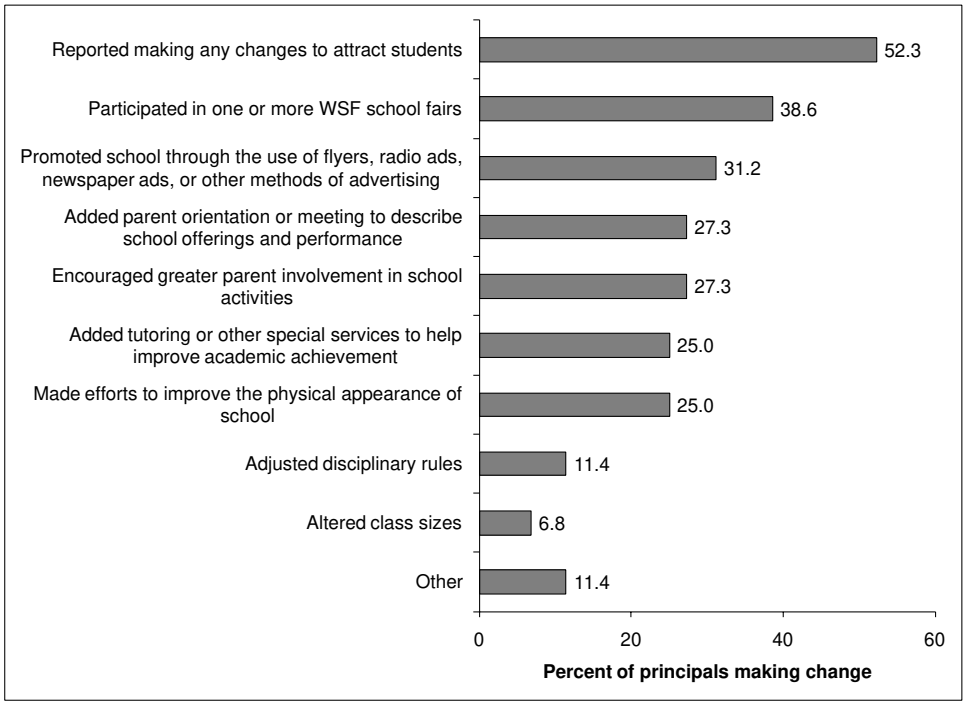


NOTES: Responses are unweighted. Respondents were able to select multiple responses. The survey question is "In the past five years or since you became principal, have you made any changes specifically to encourage students interested in private schools (or the Opportunity Scholarship Program) to remain enrolled in your school?" If the principal answered yes, then the principal was asked to indicate which (of the following) changes were made. For all percentages, the numerator is the number of principals who answered "yes" to making a change and the denominator is the total number of survey respondents (*N*=168). The response rate for the public school principal survey was 75 percent.

SOURCE: Impact Evaluation Public School Surveys, 2008-09.

[15] The appropriations law that established the OSP and each subsequent appropriations bill that funded the Program provided approximately $13 million for the OSP, approximately $13 million for DC charter schools, and approximately $13 million for the traditional public schools in DCPS. Because of the confidentiality provisions in the law, neither WSF nor IES could reveal information about which or how many students left individual DC schools.

Figure ES-9. Participating Private School Responses to the OSP, 2008-09



NOTES: Responses are unweighted. Respondents (N = 44) were able to select multiple responses. Categories with responses from fewer than three principals are collapsed into the "Other reasons" category for confidentiality purposes. The response rate for the private school principal survey was 72 percent.

SOURCE: Impact Evaluation Principal Surveys, 2008-09.

This final report on the impacts of the OSP adds to the growing body of evidence on private school voucher programs in the United States. As is the case with previous evaluations of such programs, our study had some limitations. We studied early Program applicants, and not all members of our original sample participated in data collection each year. By the final year, 13 percent of the students were no longer eligible to take the K-12 achievement assessment because they had "graded out," reducing the size of the analysis sample and the precision we had to detect effects. In addition, some of our measures, including high school graduation and school safety, are based on respondents' recall and perceptions and not on more conclusive administrative records. Finally, it is important to note that the findings in this report are a reflection of the particular Program elements that evolved from the law passed by Congress and the characteristics of the students, families, and schools, both public and private, that exist in the Nation's capital. The same program implemented in another city might yield different results, and a different scholarship program administered in Washington, DC, might also produce different outcomes.

1. Introduction

The *District of Columbia School Choice Incentive Act of 2003*,[1] passed by Congress in January 2004, established the first federally funded, private school voucher program in the United States. Since that time, more than 8,400 students have applied for what is now called the DC Opportunity Scholarship Program (OSP), and a rigorous evaluation of the Program, mandated by Congress, has been underway. This last formal report from the ongoing evaluation describes the impacts of the Program at least four years after families who applied and were given the option to move from a public school to a participating private school of their choice.[2]

1.1 DC Opportunity Scholarship Program

The purpose of the new scholarship program was to provide low-income parents, particularly those whose children attend schools identified for improvement or corrective action under the *Elementary and Secondary Education Act*, with "expanded opportunities to attend higher performing schools in the District of Columbia" (Sec. 303). According to the statute, the key components of the Program include:

- To be eligible, students entering grades K-12 must reside in the District and have a family income at or below 185 percent of the federal poverty line.
- Participating students receive scholarships of up to $7,500 to cover the costs of tuition, school fees, and transportation to a participating private school.
- Scholarships are renewable for up to five years (as funds are appropriated), so long as students remain eligible for the Program and remain in good academic standing at the private school they are attending.
- In a given year, if there are more eligible applicants than available scholarships or open slots in private schools, applicants are to be awarded scholarships by random selection (e.g., by lottery).
- In making scholarship awards, priority is given to students attending public schools designated as in need of improvement (SINI) under the *No Child Left Behind (NCLB) Act* and to families that lack the resources to take advantage of school choice options.

[1] Title III of Division C of the *Consolidated Appropriations Act*, 2004, P.L. 108-199.

[2] As described later in this chapter, the impact sample is comprised of eligible applicants from the first two years of the Program. Results in this final report are five years after random assignment for cohort 1 and four years after random assignment for cohort 2.

- Private schools participating in the Program must be located in the District of Columbia and must agree to requirements regarding nondiscrimination in admissions, fiscal accountability, and cooperation with the evaluation.

In late March 2004 following passage of the legislation, the Washington Scholarship Fund (WSF), a 501(c)3 organization in the District of Columbia, won the grant competition conducted by the U.S. Department of Education (ED) to implement the OSP under the supervision of both ED's Office of Innovation and Improvement and the Office of the Mayor of the District of Columbia. The grant was for five years, coinciding with the five-year period of Program implementation specified in the statute (Sec. 302(7)). When Congress decided to continue the Program for a sixth year, the WSF was awarded a second grant in spring 2008. Since receiving the original grant, the WSF finalized the Program design, established protocols, recruited applicants and schools, awarded scholarships, and placed and monitored scholarship awardees in participating private schools.

At the time of this report writing, the future of the OSP and WSF's role in it is uncertain. In December of 2009, Congress passed and President Obama signed a new appropriations law that extended the Program for a seventh year. The law provides funds to continue the scholarships of current participants but closes the Program to new applicants. Although a bill has been introduced in the Senate to reauthorize and expand the Program, there has been no action taken on the proposal. In the meantime, the WSF announced that it would no longer administer the Program.

The funds appropriated for the OSP between fiscal years 2004 and 2010 have been sufficient to support approximately 1,700 to 2,000 students annually, depending on the cost of the participating private schools that they attend and the proportion of the school year in which they maintain their enrollment.

To date, there have been six rounds of applicants to the OSP (table 1-1):

- Applicants in spring 2004 (cohort 1) and spring 2005 (cohort 2) represent the majority of Program applicants since the program began; the evaluation sample was drawn from these two cohorts.[3]

[3] Reports describing detailed characteristics of cohorts 1 and 2 (Wolf, Gutmann, Eissa, Puma, and Silverberg 2005; Wolf, Gutmann, Puma, and Silverberg 2006) can be found on the Institute of Education Sciences' website at: http://www.ies.ed.gov/ncee.

- A smaller number of applicants in spring 2006 (cohort 3), spring 2007 (cohort 4), and spring 2008 (cohort 5) were recruited and enrolled by WSF in order to keep the Program operating at capacity each year. [4]

- Applicants in spring 2009 (cohort 6) were recruited and their eligibility was determined. Eligible applicants were not enrolled in the Program when it was determined that federal funding would not be provided for new students for the 2009-10 school year.

Table 1-1. OSP Applicants by Program Status, Cohorts 1 Through 6, Years 2004-09

	Cohort 1 (Spring 2004)	Cohort 2 (Spring 2005)	Total Cohort 1 and Cohort 2	Cohort 3 (Spring 2006), Cohort 4 (Spring 2007), and Cohort 5 (Spring 2008)	Cohort 6 (Spring 2009)	Total, All Cohorts
Applicants	2,692	3,126	5,818	2,034	628	8,480
Eligible applicants	1,848	2,199	4,047	1,284	216	5,547
Scholarship awardees	1,366	1,088	2,454	1,284	NA	3,738
Scholarship users in initial year of receipt	1,027	797	1,824	1,057	NA	2,881
Scholarship users fall 2005	919	797	1,716	NA	NA	1,716
Scholarship users fall 2006	788	684	1,472	333	NA	1,805
Scholarship users fall 2007	678	581	1,259	671	NA	1,930
Scholarship users fall 2008	496	411	909	807	NA	1,714
Scholarship users fall 2009	386	319	705	617	NA	1,322

NOTES: NA means "not applicable" because scholarships could not be awarded or used in a given year. Because most participating private schools closed their enrollments by mid-spring, applicants generally had their eligibility determined based on income and residency, and the lotteries were held prior to the administration of baseline tests. Therefore, baseline testing was not a condition of eligibility for most applicants. The exception was applicants entering the highly oversubscribed grades 6-12 in cohort 2. Those who did not participate in baseline testing were deemed ineligible for the lottery and were not included in the eligible applicant figure presented above, though they were counted in the applicant total. In other words, the cohort 2 applicants in grades 6-12 had to satisfy income, residency, and baseline testing requirements before they were designated eligible applicants and entered in the lottery.

The initial year of scholarship use was fall 2004 for cohort 1, fall 2005 for cohort 2, fall 2006 for cohort 3, fall 2007 for cohort 4, and fall 2008 for cohort 5.

SOURCES: OSP applications and WSF's enrollment and payment files.

[4] Because the influx of cohort 2 participants essentially filled the Program, the WSF recruited and enrolled a much smaller number of students in each succeeding year, primarily to replace OSP students who left the Program between the second and fifth year of implementation. WSF limited applications to students entering grades K-6 for cohort 3 and grades K-7 for cohorts 4 and 5 because there were few slots available in participating high schools, as large numbers of students from cohorts 1 and 2 advanced to those grades. Applications also were limited to students previously attending public schools or rising kindergarteners, since public school students are a higher service priority of the Program than are otherwise eligible private school students. See chapter 2 for more detail on the cohort 1 and 2 exits from the Program that enabled WSF to accommodate cohorts 3 through 5.

Among the applicants, those determined eligible for the Program represent just over 10 percent of all children in Washington, DC, who meet the OSP's eligibility criteria, according to 2000 Census figures.[5] During fall 2009, a total of 1,322 students were using Opportunity Scholarships to attend participating private schools.

1.2 Mandated Evaluation of the OSP

In addition to establishing the OSP, Congress mandated that an independent evaluation of it be conducted, with annual reports on the progress of the study. The legislation indicated that the evaluation should analyze the effects of the Program on various academic and nonacademic outcomes of concern to policymakers and use ". . . the strongest possible research design for determining the effectiveness" of the Program.[6]

The evaluation was developed to be responsive to these requirements. In particular, the foundation of the evaluation is a randomized controlled trial (RCT) that compares outcomes of eligible applicants (students and their parents) randomly assigned to receive or not receive a scholarship.[7] This decision was based on the mandate to use rigorous evaluation methods, the expectation that there would be more applicants than funds and private school spaces available, and the statute's requirement that random selection be the vehicle for determining who receives a scholarship. An RCT design is widely viewed as the best method for identifying the independent effect of programs on subsequent outcomes (e.g., Boruch, de Moya, and Snyder 2002, p. 74; Cook and Campbell 1979, p. 56). Random assignment has been used by researchers conducting impact evaluations of other scholarship programs in Charlotte, NC; New York City; Dayton, OH; and Washington, DC (Greene 2001; Howell, Wolf, Campbell, and Peterson 2002; Mayer et al. 2002).

---

[5] See previous evaluation reports, including Wolf, Gutmann, Puma, Rizzo, Eissa, and Silverberg 2007, p. 8.

[6] *District of Columbia School Choice Incentive Act of 2003,* Section 309 (a)(2)(A).

[7] The law clearly specified that such a comparison in outcomes be made (see Section 309 (a)(4)(A)(ii)).

Key Research Questions

The research priorities for the evaluation were shaped largely by the primary topics of interest specified in the statute.[8] This legislative mandate led the evaluators to focus on the following research questions:

1. *What is the impact of the Program on student academic achievement?* Does the award of a scholarship improve a student's academic achievement in the core subjects of reading and mathematics? Does the use of a scholarship improve student achievement?

2. *What is the impact of the Program on other student measures?* Does the award of a scholarship or the use of a scholarship improve other important aspects of a student's education that are related to school success such as high school graduation?

3. *What effect does the Program have on school safety and satisfaction?* Does the award of a scholarship or the use of a scholarship increase student and/or parent perceptions of safety in schools? Does receiving or using a scholarship increase student and/or parent satisfaction with schools?

4. *What is the effect of attending private versus public schools?* Because some students offered scholarships will choose not to use them, and some members of the control group will attend private schools, the study will also examine the results associated with private school attendance with or without a scholarship.[9]

5. *To what extent is the Program influencing public schools and expanding choice options for parents in Washington, DC?* That is, to what extent has the scholarship program had a broader effect on public and private schools in DC, such as instructional changes by public schools to respond to the new competition from private schools.

These research questions are consistent with the topics that scholars and policymakers have identified as important questions of interest surrounding private school scholarship programs. For broad summaries of

---

[8] Specifically, "The issues to be evaluated include the following: (A) A comparison of the academic achievement of participating eligible students … to the achievement of … the eligible students in the same grades … who sought to participate in the scholarship program but were not selected. (B) The success of the programs in expanding choice options for parents. (C) The reasons parents choose for their children to participate in the programs. (D) A comparison of retention rates, dropout rates, and (if appropriate) graduation and college admission rates…. (E) The impact of the program on students, and public elementary schools and secondary schools, in the District of Columbia. (F) A comparison of the safety of the schools attended by students who participate in the programs and the schools attended by students who do not participate in the programs. (G) Such other issues as the Secretary considers appropriate for inclusion in the evaluation." (Section 309 (4)). The statute also says that, "(A) the academic achievement of students participating in the program; (B) the graduation and college admission rates of students who participate in the program, where appropriate; and (C) parental satisfaction with the program" should be examined in the reports delivered to the Congress. (Section 310 (b)(1)).

[9] The statute requests comparisons between "program participants" and nonparticipants. Since the central purpose of the Program is to provide students with the option of attending a private school, the evaluation team has understood this provision as consistent with the examination of the effects of actual attendance at a private school. Previous experimental evaluations of scholarship programs have examined the effects of actual private school attendance on study participants (Howell et al. 2006, pp. 144-167; Greene 2001; Rouse 1998). The analysis of the effect of attending a private versus a public school is included in an appendix and not in the main text of this report. Private school attendance could be achieved with or without an OSP scholarship and therefore the analysis does not address Congress' primary interest in measuring the Program's effect.

the research literature on school choice, see, for example, Coulson 2009, Gill et al. 2007, Rouse and Barrow 2009, and Wolf 2008.

In addition, the evaluation is exploring the mechanisms by which the Program may or may not have an effect on the key outcomes, by examining the Program's impact on a set of intermediate outcomes that include home educational supports, student motivation and engagement, instructional characteristics, and school environment factors. Finally, the pattern of impact results raises a number of questions regarding why impacts have been observed regarding some outcomes but not others, or for some subgroups of participating students but not for other subgroups. A variety of factors, such as the length of scholarship use, that could potentially or partially explain the pattern of results that we report here will be explored in a later, second volume to this final evaluation report.

Student Recruitment, Random Assignment, and the Creation of the Impact Analysis Sample

The recruitment, application, and lottery process conducted by WSF with guidance from the evaluation team created the foundation for the evaluation's randomized trial and determined the group of students for whom impacts of the Program are analyzed. Because the goal of the evaluation was to assess both the short-term and longer term impacts of the Program, it was necessary to focus the study on early applicants to the Program (cohorts 1 and 2) whose outcomes could be tracked over at least four years during the evaluation period. During the first two years of recruitment, WSF received applications from 5,818 students. Of these, approximately 70 percent (4,047 of 5,818) were eligible for the Program (table 1-1).

Once students applied and were verified eligible for the Program, the next step was to determine whether they would receive a scholarship. The statute specifies that lotteries be conducted to award scholarships when the Program is "oversubscribed," that is, when the number of eligible applicants exceeds the number of available slots in participating private schools.[10] Further, the statute specifies that certain groups of applicants be given priority in any such lotteries, which led to the following rank ordering:

1. Applicants attending a public school in need of improvement (SINI) under *No Child Left Behind (NCLB)* (highest priority);

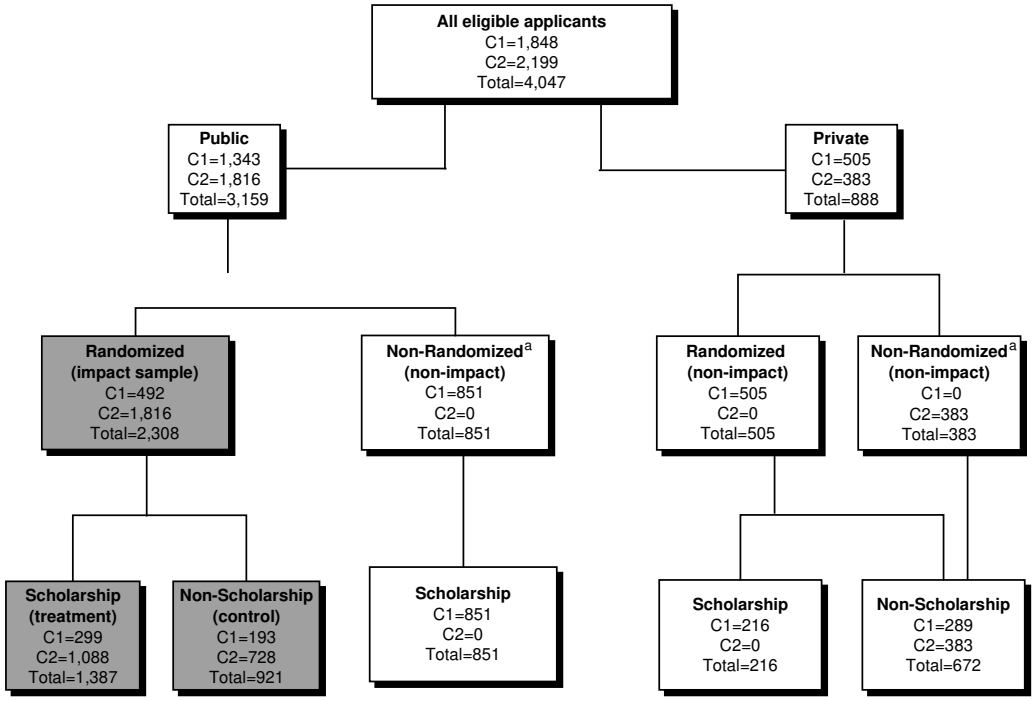
2. Non-SINI public school applicants (middle priority); and

[10] However, because the extent of oversubscription varied significantly by grade, in practice the determination of whether to hold a lottery was considered within grade bands: those applying for grades K-5, those applying for grades 6-8, and those applying for grades 9-12.

3. Applicants already attending private schools (lowest priority).

Not all applicants faced the conditions that necessitated scholarship award by lottery (Wolf et al. 2006).[11] In addition, some applicants who were eligible for a lottery (in oversubscribed grades) could not be included in the impact analysis sample. For example, because the evaluation was intended to measure the effects of providing access to private school, the impact analysis focuses on the population of applicants for whom private schooling represented a new opportunity. Thus, the impact sample for the evaluation comprised all eligible applicants who were previously attending public schools (or were rising kindergarteners) AND were subject to a lottery to determine whether they would receive an Opportunity Scholarship (figure 1-1, shaded area).

Figure 1-1. Construction of the Impact Sample From the Applicant Pool, Cohorts 1 and 2



NOTES: C1 = Cohort 1 (applicants in spring 2004)
C2 = Cohort 2 (applicants in spring 2005)
Total = C1 and C2

[a]The group of applicants who were not randomly assigned includes: in cohort 1, public school applicants from SINI schools or who were entering grades K-5 (all received a scholarship) and in cohort 2, private school applicants, the lowest priority group (none received a scholarship because it was clear the Program would be filled with higher priority public school applicants).

[11] In the first year of Program implementation (spring 2004 applicants, or cohort 1), for example, there were more slots in participating schools than there were applicants for grades K-5; therefore, all eligible K-5 applicants from SINI and non-SINI public schools automatically received scholarships, and no lotteries were conducted at that level. In contrast, there were more eligible public school applicants in cohort 2 (spring 2005) than there were available slots at all grade levels, so that all of those applicants were subject to a lottery to determine scholarship awards. One other difference is that, because there were sufficient funds available in school year 2004-05, applicants seeking an OSP scholarship but who were already attending a private school were entered into a lottery the first year. In cohort 2, there was sufficient demand from public school applicants that lotteries were conducted only for them; applicants who were already attending a private school (the lowest priority group) were not entered into a lottery and did not receive scholarships (figure 1-1).

The total pool of eligible applicants comprised 1,848 applicants in cohort 1 (spring 2004) and 2,199 applicants in cohort 2 (spring 2005). Of those eligible applicants, 492 in cohort 1 and 1,816 in cohort 2 met the criteria to be randomly assigned by lottery to the evaluation's treatment and control groups. In cohort 1, a total of 299 students were randomized into the treatment condition and 193 into the control condition. In cohort 2, some 1,088 students were randomized into the treatment condition and 728 into the control condition.[12] The evidence from cohort 1 and cohort 2 students was consolidated for the evaluation of overall Program impacts in order to increase the statistical power of the analysis and to provide results that could be applied to both first and second year applicants to the Program. The impact sample comprising these groups totals 2,308 students: 1,387 students in the treatment condition and 921 in the control condition.[13] The more than 2,300 students in the impact sample is a large group relative to the impact samples of 803 to 1,960 students used in other evaluations of private school scholarship programs (Howell et al. 2002).

Data Collection

The evaluation gathered information annually from students and families in the study, as well as from their schools, in order to address the key research questions. These data include:

- **Student assessments.** Measures of student achievement in reading and math for public school applicants came from the Stanford Achievement Test-version 9 (SAT-9)[14] administered by either the District of Columbia Public Schools (DCPS) (cohort 1 baseline) or the evaluation team (cohort 2 baseline and all follow-up data collection). The evaluation testing took place primarily on Saturdays, during the spring, in locations throughout DC arranged by the evaluators. The testing conditions were similar for members of the treatment and control groups.

---

[12] Although the scholarship award probabilities differed somewhat across cohort 1 and cohort 2, the factors that went into the design of the probabilities each year were the same. Each cohort of applicants was divided into three grade classifications: K-5, 6-8, and 9-12. The number of eligible applicants was compared with the number of available slots in participating schools as determined by the WSF based on school reports. If the number of slots exceeded the number of applicants in a given "grade band," then all students in that band automatically received scholarship awards. If the number of applicants exceeded the number of slots, then customized scholarship award probabilities were set for the SINI students (i.e. higher probability of award) and the non-SINI students (i.e. lower probability of award) in the band. These differences in the probability of assignment to treatment and control based on cohort, grade band, and SINI status were factored into the sample weights as described in appendix section A.7.

[13] As part of the control group follow-up lottery to reward control group members who cooperate with the evaluation's testing requirements, five members of the control group (cohort 1) were awarded scholarships by lottery in the summer of 2005, seven members of the control group (cohorts 1 and 2) were awarded scholarships by lottery in the summer of 2006, seven members of the control group (cohorts 1 and 2) were awarded scholarships by lottery in the summer of 2007, and four members of the control group (cohort 2) were awarded scholarships by lottery in the summer of 2008. Control group students who win a follow-up incentive lottery remain in the analysis as control group members, even though they have been awarded scholarships, to preserve the integrity of the original random assignment. Whether or not they use their scholarship to attend a private school, they are treated as control group members for purposes of the intent-to-treat (ITT) and Bloom adjusted impact-on-treated (IOT) analyses.

[14] *Stanford Abbreviated Achievement Test (Form S)*, Ninth Edition. San Antonio, TX: Harcourt Educational Measurement, Harcourt Assessment, Inc., 1997.

- **Parent surveys.** The OSP application included baseline surveys for parents applying to the Program. These surveys were appended to the OSP application form and therefore were completed at the time of application to the Program. Each spring after the baseline year, surveys of parents of all applicants were conducted at the Saturday testing events while parents were waiting for their children to complete their outcome testing. The parent surveys provided the self-reported outcome measures for parental satisfaction and safety.

- **Student surveys.** Each spring after the baseline year, surveys of students in grades 4 and above were conducted at the outcome testing events. The student surveys provided the self-reported outcome measures for student satisfaction and safety.

- **Principal surveys.** Each spring, surveys of principals of all public and private schools operating in the District of Columbia were conducted. Topics included self-reports of school organization, safety, and climate; principals' awareness of and response to the OSP; and, for private school principals, why they were or were not OSP participants.

- **Parent Follow Up surveys.** In the summer of 2009, parents of students who had turned 16 by June 30, 2009, were contacted (by telephone) to determine if their child was still enrolled in high school or had graduated from high school. The parent follow-up surveys provided the self-reported outcome measure for educational attainment (that is, high school graduation) analyzed for the first time in this report.

Several methods were used to encourage high levels of response to the final year of data collection in spring 2009 (year 5 for cohort 1 and year 4 for cohort 2). Study participants were invited to at least five different data collection events if they were a member of the treatment group and at least six different data collection events if they were a member of the control group. Impact sample members received payment for their time and transportation costs if they attended a data collection event. The events were held on Saturdays except for one session that was staged on a weeknight. Multiple sites throughout DC were used for these events, and participants were invited to the location closest to their residence. When the address or telephone number of a participant was inaccurate, such cases were submitted to the tracing office at Westat and subject to intensive efforts to update and correct the contact information.

After these initial data collection activities were completed, the test score response rate[15] for 2009 was 63.5 percent. The treatment group response rate was 63.9 percent, and the control group response rate was 62.7 percent, a response rate differential of 1.2 percentage points lower for the control group compared to the treatment group. Although that differential was not statistically significant, to reduce the likelihood of nonresponse bias and increase the generalizability of the study results, a random

[15] A total of 296 students initially in the impact sample (202 in cohort 1 and 94 in cohort 2) were forecasted to have graduated from high school before the spring of 2009, based on their grade upon Program application at least four years after random assignment. The Stanford Achievement Test that is mandated as the evaluation test does not have a version for students beyond 12th grade. As a result, these "grade outs" were not invited to data collection events, and therefore are not counted in the main set of response rate calculations presented in this report.

subsample of half of the nonrespondents in both the treatment and control groups was drawn and subjected to intensive efforts at nonrespondent conversion (see appendix section A.7). Since these initial nonrespondents were selected at random, each one that was successfully converted to a respondent counts double in the analysis, as he or she "stands in" for an approximately similar initial nonrespondent that was not subsampled (see Kling, Ludwig, and Katz 2005; Sanbonmatsu, Kling, Duncan, and Brooks-Gunn 2006). The "effective" response rate after subsample conversion is the number of actual respondents prior to the subsample plus two times the number of subsampled respondents, all divided by the total number of students in the impact sample.

As a result of the subsample conversion process, the final effective test score response rate for 2009 data collection was 69.5 percent, and the differential rate of response between the treatment and control groups was reduced to 0.1 percentage points higher for the treatment group.[16] The effective parent survey response rate was 66.1 percent.[17] The effective student survey response rate was 66.5 percent.[18] The public school principal survey response rate was 75 percent, and the private school principal survey response rate was 72 percent. The response rate for the parent follow-up survey was 63.2 percent.

Missing outcome data create the potential for nonresponse bias in a longitudinal evaluation such as this one, if the nonrespondent portions of the sample are different between the treatment and control groups. Response rates for the various data collection instruments all differed by less than 2.5 percent between the treatment and control groups. Study respondents in the final year of data collection were statistically similar to the overall impact sample on 13 of 14 baseline (pre-Program) characteristics before nonresponse weights were applied—a difference predicted by chance—and indistinguishable from

[16] Specifically the overall effective response rates were 69.5 percent for the treatment group and 69.4 percent for the control group. Prior to drawing the subsample, response rates for the control group were 50.6 percent (cohort 1) and 64.3 percent (cohort 2). Response rates (after drawing the subsample) for the control group were 53.1 percent (cohort 1) and 67.6 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 54.6 percent (cohort 1) and 71.2 percent (cohort 2). Prior to drawing the subsample, response rates for the treatment group were 52.9 percent (cohort 1) and 66.6 percent (cohort 2). Response rates (after drawing the subsample) for the treatment group were 59.6 percent (cohort 1) and 68.5 percent (cohort 2). After subsample weights were applied, the effective response rates for the treatment group were 65.5 percent (cohort 1) and 70.5 percent (cohort 2). There were 296 impact sample students awarded scholarships who were no longer eligible for data collection by spring of 2009; these students are excluded from the main response rate calculations. See appendix A, tables A-6 and A-8 for a detailed breakdown of the response rates.

[17] Specifically, the overall effective response rates were 65.6 percent for the treatment group and 67.0 percent for the control group. Response rates (after drawing the subsample) for the control group were 49.4 percent (cohort 1) and 65.0 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 50.9 percent (cohort 1) and 69.1 percent (cohort 2). Response rates (after drawing the subsample) for the treatment group were 53.7 percent (cohort 1) and 65.0 percent (cohort 2). After subsample weights were applied, the effective response rates for the treatment group were 59.3 percent (cohort 1) and 67.1 percent (cohort 2). See appendix A, table A-10 for a detailed breakdown of the response rates.

[18] Specifically, the overall effective response rates were 67.3 percent for the treatment group and 65.1 percent for the control group. Response rates (after drawing the subsample) for the control group were 50.6 percent (cohort 1) and 64.5 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 52.2 percent (cohort 1) and 67.2 percent (cohort 2). Response rates (after drawing the subsample) for the treatment group were 59.2 percent (cohort 1) and 66.3 percent (cohort 2). After subsample weights were applied, the effective response rates for the treatment group were 65.1 percent (cohort 1) and 68.0 percent (cohort 2). See appendix A, table A-11 for a detailed breakdown of the response rates.

each other on all these features after applying weights to adjust for nonresponse. More important, treatment group respondents did not differ significantly from control group respondents on any of the 14 baseline covariates examined, either before or after the application of the nonresponse weights. Thus, the impact results presented here do not appear affected by participant nonresponse and do appear to be generalizable to the original sample of students in the study. Sections A.3 and A.7 of appendix A provide additional details about data sources, collection methods, response rates, subsampling for nonresponse conversion, and final nonresponse sample weights.

The test score response rate of 69.5 percent for this report's analysis of the OSP at least four years after random assignment is higher than the response rates obtained in any of the three previous experimental evaluations of privately-funded K-12 scholarship programs three years after random assignment. The previous evaluations of such programs in New York City and Washington, DC, reported year three test score response rates of 67 percent and 60 percent, respectively (Howell et al. 2006, p. 47). A previous experimental evaluation of the publicly funded Milwaukee Parental Choice Program reported test score response rates in year four of 25 percent for the treatment group and 15 percent for the control group (Rouse 1998, p. 598).

Research Methodology

The evaluation of the OSP was designed as an RCT or experiment. Experimental evaluations take advantage of a randomization process that divides a group of potential participants into two statistically similar groups—a treatment group that receives the offer of admission to the intervention or program and a control group that does not receive the offer of admission, with the control group's subsequent experiences indicating what likely would have happened to the members of the treatment group in the absence of the intervention (Fisher 1935). Most analyses of experimental data use covariates measured at baseline in statistical models to improve the precision of the impact estimates. The results—comparing the experiences of the two groups—can then be interpreted in relatively straightforward ways as revealing the actual impact of the Program on outcomes of policy interest.

Certain specific features of this experimental evaluation are important to convey. A power analysis based on actual respondent counts in 2009 indicated that the analysis of impacts in year four or five is likely to be sufficiently powered to detect achievement impacts of .13 standard deviations for the entire study sample and .15 to .23 standard deviations for the subgroups of interest. The same power analysis, applied to the outcome of educational attainment, indicated that the evaluation is likely to be sufficiently powered to detect programmatic impacts on the rate of high school graduation of .26 standard

deviations for the entire study sample and .31 to .39 standard deviations for the subgroups of interest (see appendix A, section A.2).[19] Observations were weighted after data collection, using baseline characteristics associated with study nonresponse, to re-establish the equivalence of the treatment and control groups in the face of differential rates of nonresponse (see appendix section A.7). A consistent set of 15 baseline student characteristics related to student achievement was included in the regression models that generated the estimates of Program impact (see appendix section A.8). In cases where impacts were estimated for subgroups of participants, or a large set of intermediate outcomes of the Program were estimated, the Benjamini-Hochberg method of adjusting standard errors was used to reduce the risk of false discoveries due to multiple comparisons (see appendix B). Finally, sensitivity tests were conducted to determine the robustness of the impact estimates to other reasonable analytic methods. The size and statistical significance of such impacts were re-estimated using two different alterations in the original methodological approach: (1) trimming back the set of treatment group respondents to the response rate of the control group prior to subsampling nonrespondents and (2) clustering the standard errors of the observations on school attended instead of family (see appendix C).

1.3 Contents of This Report

This report is the sixth in a series of required annual reports to Congress and the final impact report from the evaluation. It presents the effects of the Program on students and families at least four years after they applied and had the chance of being awarded and using a scholarship to attend a participating private school. In presenting these impacts, we first provide information on the participation of students and schools in the OSP, including the patterns of and reasons for use and non-use of scholarships among students who were awarded them (chapter 2). The main impact results, both for the overall group and for important subgroups of applicants, are described in chapter 3; these findings address whether students who received a scholarship through the lotteries (and their parents) benefited at least four years later as a result of the offer or the actual use of an Opportunity Scholarship. Chapter 4 assesses the impacts of the Program on intermediate outcomes—such as parent aspirations and supports, student motivation and engagement, school instructional characteristics, and the school environment. This analysis is an attempt to develop hypotheses about the mechanisms through which private school vouchers may or may not lead to higher student achievement or better outcomes for students. The final chapter (chapter 5) provides information regarding how DC schools have been changing in response to

[19] This year's power analyses combined observation numbers (i.e., *N*s) from the actual 2009 respondent sample with assumptions regarding the strength of relationships in the data drawn from an earlier experimental analysis of a privately funded K-12 scholarship program in DC. Because that earlier analysis did not include an assessment of educational attainment, the power estimates regarding the attainment analysis presented here rely more heavily on assumptions than do the power estimates regarding achievement, and thus should be viewed as less precise than the estimates of analytic power regarding achievement.

the Program. A later, second volume to this report will explore additional hypotheses for the pattern of impact results.

In the end, the findings in this report are a reflection of the particular Program elements that evolved from the law passed by Congress and the characteristics of the students, families, and schools, both public and private, that exist in the Nation's capital. The same program implemented in another city might yield different results, and a different scholarship program administered in Washington, DC, might also produce different outcomes.

2. School and Student Participation in the OSP

In interpreting the impacts of the Opportunity Scholarship Program (OSP) presented in later chapters, it is useful to examine the characteristics of the private schools that participate in the Program and the extent to which students offered scholarships (the treatment group) move into and out of them. These characteristics can best be viewed in the context of how the participating private schools look in comparison to the public schools most of the control group and some of the treatment group attend. This chapter describes the differences between the treatment and control groups' experiences, while a later one (chapter 4) explores the hypothesis that the OSP had an impact on these factors. The final chapter (chapter 5) explores the response of the public and private schools in the District of Columbia to the Program, by examining their OSP-related enrollment losses and gains and the specific activities they reported undertaking to retain or attract students.

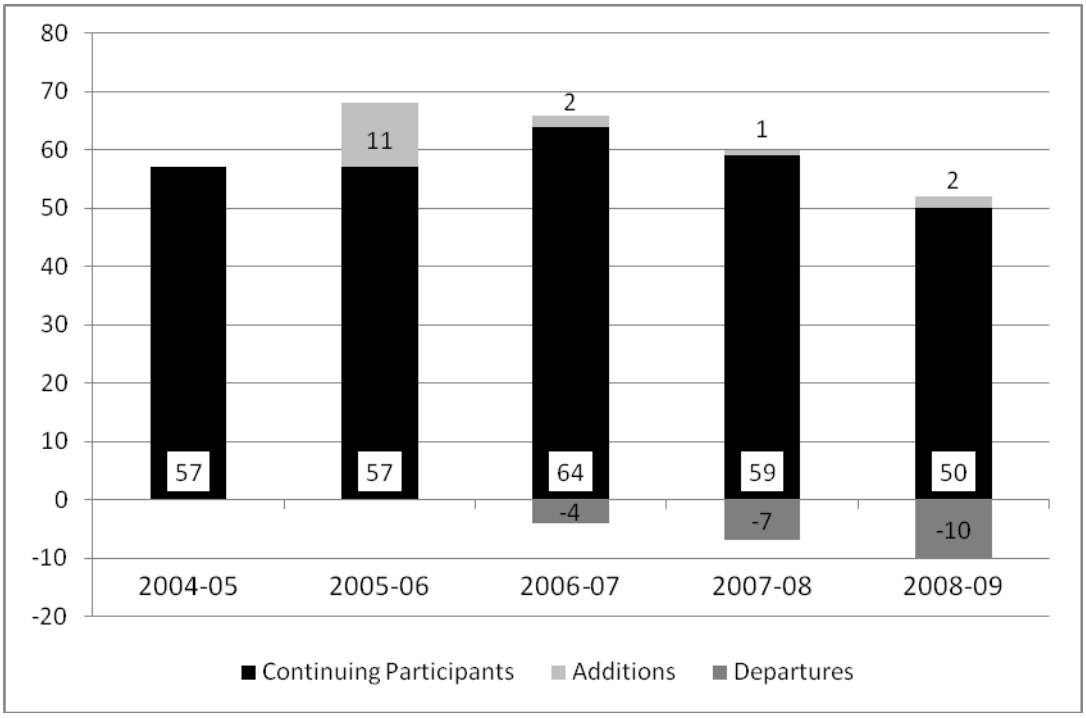
2.1 School Participation

The private schools participating in the OSP represent the choice set available to parents whose children received scholarships. A total of 52[20] of 90 private schools in the District of Columbia were participating in the Program at the start of the 2008-09 school year, the final year for which data were collected for this evaluation (figure 2-1).[21] That count of participating private schools was down from the peak of 68 achieved in 2005-06. Among the 22 schools that participated at any point but left the Program are 7 Catholic schools that, in their last year of the Program (2007-08), enrolled 112 treatment group students; these schools converted to become public charter schools in 2008-09 and therefore no longer could be OSP voucher recipients.

[20] While, technically, 56 individual campuses were participating in the OSP from the start of the 2008-09 school year, the research team treats four of the schools with dual campuses as single entities because they have one financial office that serves both campuses, following the classification practice used by the National Center for Education Statistics in its Private School Survey.

[21] This figure represents a net loss of nine schools since the prior year. Eleven schools stopped participating while two new schools participated for the first time in 2008-09. The total number of private schools operating in DC declined from 109 in 2004-05 to 90 in 2008-09.

Figure 2-1. Number of Participating OSP Private Schools, 2004-05 through 2008-09



SOURCE: OSP School Directory information, 2004-05, 2005-06, 2006-07, 2007-08, 2008-09, WSF.

A total of 73 different private schools participated in the OSP at some point during the first five years of the Program. Among them, 52 percent consistently participated in all five years, while 48 percent partially participated, including the seven former Catholic private schools that operated as public charter schools in 2008-09 (table 2-1). Among the participating schools:

- Consistent participants enrolled an average of 23 OSP students per school per year; partial participants enrolled an average of 16 students per school per year. Catholic schools that converted to public charter schools, a subset of partial participants, enrolled an average of 41 OSP students per school per year in the Program.

- Forty-six percent of schools that partially participated charged tuition in excess of the $7,500 maximum scholarship award, while 37 percent of consistent participants charged more than this amount. No Catholic schools that converted to public charter schools charged above $7,500.

Table 2-1. Features of Private Schools Participating in the OSP by Participation Status, 2004-05 through 2008-09

	Percent of All Participating Schools	Average School Size	Average Yearly Scholarship Users	Percent with Tuition Over $7,500
Consistent participants	52.1	251.0	23.4	36.8
All partial participants	47.9	238.2	16.3	45.7
Catholic to charter conversions	9.6	176.3	41.0	0.0
Other partial participants	38.4	262.2	10.1	57.1

NOTES: Average school size is the average of the school size data from the National Center for Education Statistics' Private School Survey, 2005-06 and 2007-08. Average tuition represents the average tuition charged across all years that a school participated in the Program.

SOURCES: OSP School Directory information, 2004-05, 2005-06, 2006-07, 2007-08, 2008-09, WSF. National Center for Education Statistics' Private School Survey, 2005-06 and 2007-08.

Among the participating schools in 2008-09:[22]

- Fifty-four percent (28) were faith-based, with a majority of them (15) the parochial schools of the Catholic Archdiocese of Washington;
- Fifty percent charged an average tuition above the OSP's scholarship cap of $7,500;[23]
- The average participating school had a total student population of 286 students in 2008-09;
- Twenty-three percent served high school students;[24]
- The average minority percentage among the student body was 66 percent; and
- The average student/teacher ratio was 9:4.

Schools Attended by Scholarship Users in 2008-09

Not all of the schools that agreed to participate in the Program served OSP students every year. During the 2008-09 school year, which represented five years after application for cohort 1 and four years after for cohort 2, OSP students were enrolled in 40, and the impact sample's treatment students in

[22] Information was obtained for all 52 participating schools from records of the WSF regarding whether the schools were faith-based, charged tuition above $7,500, and served high school. The data regarding school size (valid N = 43), percent minority students (valid N = 39) and student/teacher ratio (valid N = 39) were drawn from the National Center for Education Statistics' Private School Survey, last administered in 2007-08.

[23] For schools that charge a range of tuitions, the midpoint of the range was selected.

[24] Schools were classified as serving high school students if they enrolled students in any grade 9-12.

38, of the 52 schools available to them.[25] Since participating schools varied in how many slots they committed to the Program, OSP students were clustered in certain schools; this was also true of the students in the impact sample's treatment group.

The schools that offered the most slots to OSP students, and in which OSP students and the impact sample's treatment group were clustered, have characteristics that differed somewhat from the typical participating OSP school. Fourteen percent of treatment group students were attending a school that charged tuition above the statutory cap of $7,500 during the year represented by this evaluation report (table 2-2), even though 50 percent of participating schools charged tuitions above that cap in 2008-09. Although 54 percent of all participating schools were faith-based (29 percent part of the Catholic Archdiocese of Washington), 80 percent of the treatment group attended a faith-based school, with most of them (53 percent) attending the 15 participating Catholic parochial schools (figure 2-2).

Table 2-2. Features of Participating OSP Private Schools Attended by the Treatment Group in 2008-09

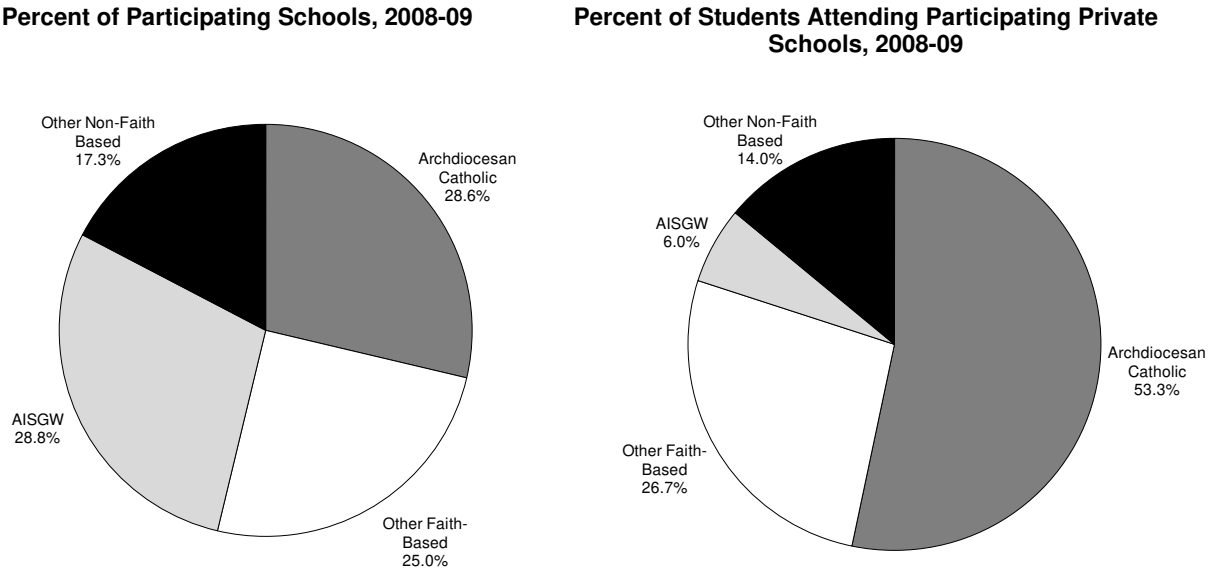
Characteristic	Weighted Mean	Highest	Lowest	Valid *N*
Charging over $7,500 tuition (percent of treatment students attending)	14.2%	NA	NA	38
Tuition	$7,252	$29,607	$4,500	38
Enrollment	292.1	1,097	16	31
Student *N*	465			

NOTES: "Valid *N*" refers to the number of schools for which information on a particular characteristic was available. When a tuition range was provided, the mid-point of the range was used. The weighted mean was generated by associating each student with the characteristics of the school he/she was attending and then computing the average of these student-level characteristics.

SOURCES: OSP School Directory information, 2008-09, WSF; National Center for Education Statistics' Private School Survey, 2007-08.

[25] The source for student enrollment in participating schools is the WSF OSP payment file for 2008-09.

Figure 2-2. Religious Affiliation of Participating Schools



NOTES: School *N* = 52 for percent of participating private schools in 2008-09. School *N* = 38 and Student *N* = 465 for percent of students in the treatment group attending participating private schools in 2008-09. AISGW is an acronym for the Association of Independent Schools of Greater Washington.

SOURCES: OSP School Directory information, 2004-05, 2005-06, 2006-07, 2007-08, Washington Scholarship Fund; OSP payment file for 2008-09, Washington Scholarship Fund.

Schools Attended by the Treatment Group in Relation to Those of the Control Group in 2008-09

While the characteristics of the participating private schools are important considerations for parents, how those characteristics differ from the public school options available to parents matters more. How different are the school conditions? Students in the treatment and control groups did not differ significantly regarding the proportion attending schools that offered a computer lab (95 vs. 91 percent), separate library (77 vs. 79 percent), gyms (68 and 71 percent), individual tutors (58 percent vs. 63 percent), music programs (93 vs. 91 percent), and after-school programs (91 vs. 88 percent) (table 2-3). However, there were statistically significant differences between students in the treatment and control groups in some aspects of the schools they attended:[26]

- Students in the treatment group were less likely than those in the control group to attend a school that offered special programs for students who may be academically challenged; these include programs or services for non-English speakers (32 vs. 57 percent) and for students with learning problems (75 vs. 90 percent);
- Students in the treatment group were less likely to be in schools with special programs for advanced learners (38 vs. 49 percent); and

[26] Characteristics are considered significantly different if the difference between them was statistically significant at the .05 level or higher.

- Students in the treatment group were less likely than those in the control group to attend a school with a cafeteria facility (76 vs. 91 percent), a nurse's office (50 vs. 82 percent), counselors (77 vs. 87 percent), and art programs (84 vs. 92 percent).

Table 2-3. Characteristics of School Attended by the Impact Sample, Year of Application and 2008-09

Percentage of Students Attending a School with:	Baseline Year			2008-09		
	Treatment	Control	Difference	Treatment	Control	Difference
Separate Facilities:						
Computer lab	72.02	71.87	.16	94.72	91.47	3.25
Library	80.00	77.15	2.85	77.22	78.98	-1.75
Gym	60.24	60.45	-.21	67.93	71.03	-3.09
Cafeteria	86.26	87.52	-1.27	75.87	90.96	-15.09**
Nurse's office	87.52	89.33	-1.81	49.85	82.45	-32.61**
Percent missing	7.05	7.12	-.07	1.09	4.67	-3.59
Programs:						
Special program for non-English speakers	45.95	40.21	5.75	32.01	57.41	-25.40**
Special program for students with learning problems	65.35	65.80	-.45	75.21	90.41	-17.90**
Special program for advanced learners	37.14	31.92	5.23	37.72	48.88	-11.17**
Counselors	79.68	77.50	2.19	77.42	87.14	-9.73**
Individual tutors	36.93	38.00	-1.07	58.49	62.83	-4.34
Music program	68.87	69.38	-.51	93.32	90.63	2.69
Art program	70.75	67.43	3.33	83.80	92.10	-8.30**
After-school program	83.07	83.22	-.15	91.45	88.33	3.13
Percent missing	7.29	7.47	-.18	1.09	4.67	-3.59
Sample size (unweighted)	1,060	586	474	1,060	586	474

**Statistically significant at the 99 percent confidence level.

NOTES: Data are weighted. For a description of the weights, see appendix A. Baseline year means presented here differ from those presented in previous reports due to the exclusion of grade-outs.

SOURCES: OSP applications, Impact Evaluation Parent Survey (for school attended), and Impact Evaluation Principal Survey.

2.2 Student Participation

The degree to which students initially and consistently used their scholarships provides some signal of the attractiveness of the OSP to parents and the ability of the Program and its participating schools to accommodate their needs. A total of 2,454 students who applied to the OSP in the first two years of Program operation were offered scholarships, with 1,387 of them in the impact sample's treatment group. By 2008-09, a total of 94 members of the treatment group were no longer eligible to receive scholarships because they had "graded out" of the Program, which means that they would have moved beyond 12th grade.

With the exclusion of the older students who had graded-out of Program eligibility, 1,293 treatment group students could have used a scholarship during the 2008-09 school year. However, as has been true in other programs, not all students offered a scholarship actually used it to enroll in a private school. Understanding the extent to which, and why, parents and students chose not to take advantage of the scholarship offer is important for program improvement and the assessment of program impacts.

Treatment group students were not the only participants in the study who attended private schools. Based on parent self-reports in year four or year five, 23 percent of the members of the control group attended a private school at some point during the evaluation. This "crossover" of control group members to the kind of situation that the treatment is intended to deliver (i.e. private schooling) is common in social experiments and is viewed simply as part of the counterfactual condition represented by the control group. That is, some students in the study would have attended private schools even absent the Program. The outcomes from control group students who attended private schools are always included on the control side of the comparison in the impact analyses presented in chapter 3. An estimated 2.9 percent of control group members were admitted to private schools along with siblings who had an Opportunity Scholarship. These students represent a special case of "program-enabled" crossover that is not part of the natural counterfactual and are adjusted for in the analysis of the impacts of using a scholarship in chapter 3. The Instrumental Variable analyses of the effects of private schooling presented in appendix D are the only analyses in this report that treat private school control group members differently from their control group peers who remained in public schools.

Patterns of Scholarship Use Among K-12 Treatment Students

According to rules determined by the Program operator (the WSF), once a student was offered an OSP scholarship, he or she could use it at any time until they graduated from high school.

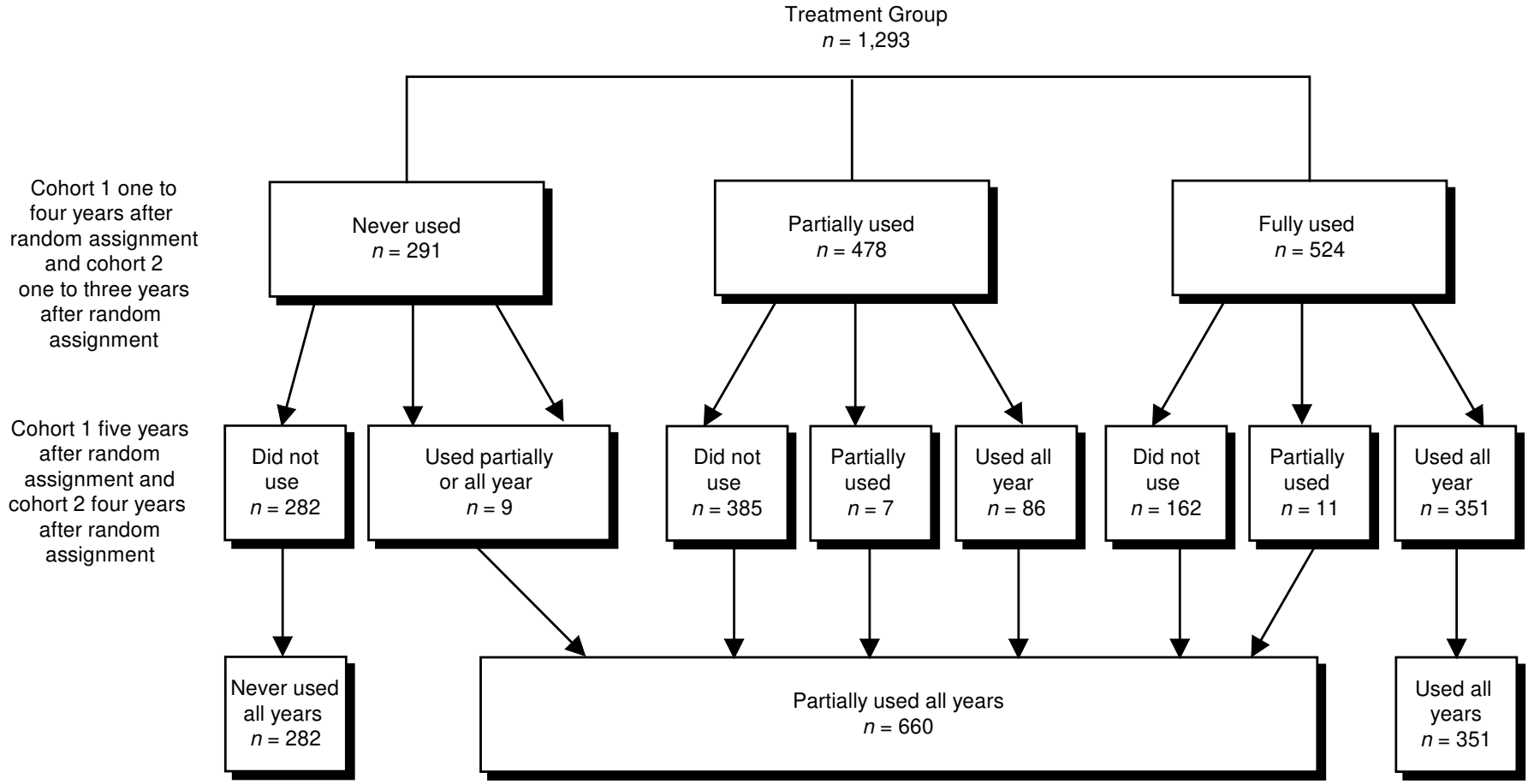
Looking across the members of the impact sample's treatment group who had four (cohort 1) or five (cohort 2) years of potential Program participation (figure 2-3; figure 2-4):

- 282 out of 1,293 (22 percent) never used the OSP scholarships offered to them;
- 660 treatment students (51 percent) used their scholarships, but not consistently, during the school years after the scholarship award. Among these students are an estimated 147 who may have been forced by circumstances to stop using their scholarship. Students could become "forced-decliners" because the school they continued to attend converted from a participating Catholic school to a public charter school (confirmed for 35 treatment students),[27] their family income grew to exceed the Program's income limit (confirmed for 21 treatment students), their family moved out of DC (confirmed for 29 students), or they may have faced a lack of space for them in a participating high school when they transitioned from 8th to 9th grade (estimated for 62 treatment students).[28] Among the students who partially used their scholarship over at least four years after random assignment, 17 percent (9 percent of eligible treatment group students overall) used their OSP scholarship in 2008-09.
- The remaining 351 treatment group students (27 percent) used their scholarship during all years available to them after the scholarship lottery.

[27] Based upon survey data, 35.9 percent of 97 treatment group students who used a scholarship to attend one of these Catholic schools in grades K-7 in 2007-08 continued to attend the same school when it converted to a public charter school in 2008-09.

[28] The estimate of the number of students forced to decline their scholarships due to the lack of high school slots was calculated by comparing the higher rate of scholarship continuation for 7th graders moving to 8th grade with the lower rate of scholarship continuation for 8th graders moving to 9th grade. The difference between those two continuation rates, applied to the number of OSP students moving from 8th to 9th grade, generates the estimate of forced decliners due to high school slot constraints of 62 (20 in year two plus 30 in year three plus 12 new cases in 2008-09). It is impossible to know for certain if all 62 of these students declined to use the scholarship solely or primarily because of high school slot constraints, and not for other reasons, or if some treatment students were forced to decline their scholarship at the very start due to high school slot constraints. It also is impossible to know if some students declined to even attempt to renew their scholarships because they knew their family exceeded the income limit, or how many treatment students moved out of DC and never informed the evaluators that they had "moved out" of Program eligibility. Therefore, the total estimate of 147 forced decliners for 2008-09 is simply an estimate based on the limited data available.

Figure 2-3. Scholarship Usage by K-12 Treatment Students Through 2008-09

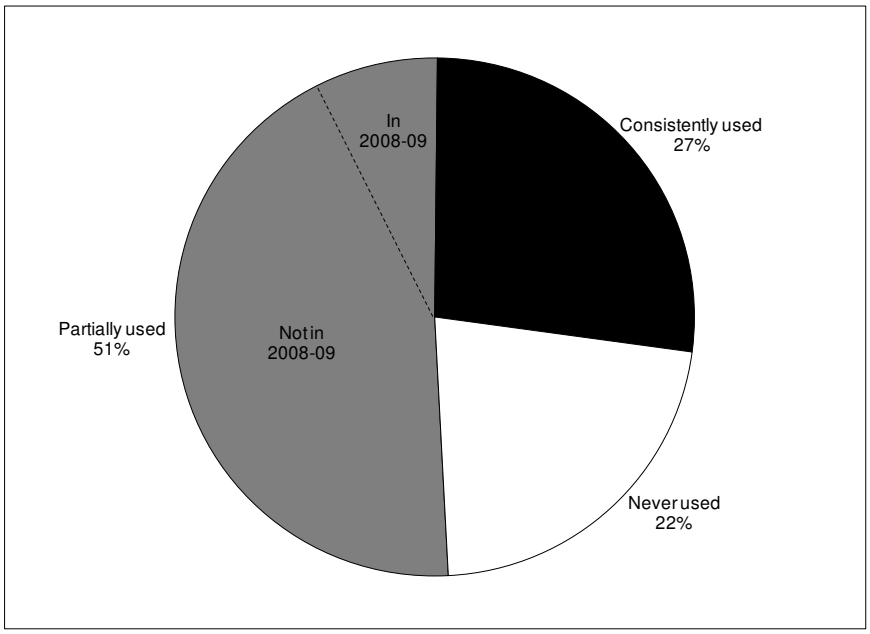


NOTES: Students were identified as scholarship users based upon information from WSF's payment files. Because some schools use a range of tuitions and some students had alternative sources of funding, students were classified as full users if WSF made payments on their behalf that equaled at least 80 percent of the school's annual tuition. Otherwise students were identified as partial users (1 percent to 79 percent of tuition paid) or nonusers (no payments).

By the 2008-09 school year, 94 treatment group students had "graded out" and are not included in this figure. Of these grade-outs, 53 never used their scholarships, 22 partially used their scholarships, and 19 fully used their scholarships over all of the years that were available to them.

SOURCE: WSF's payment files.

Figure 2-4. Proportion of K-12 Treatment Group Using Their OSP Scholarship Award Through 2008-09



NOTES: Students were identified as scholarship users based upon information from WSF's payment files. Because some schools use a range of tuitions and some students had alternative sources of funding, students were classified as full users if WSF made payments on their behalf that equaled at least 80 percent of the school's annual tuition. Otherwise students were identified as partial users (1 percent to 79 percent of tuition paid) or nonusers (no payments).

By the 2008-09 school year, 94 treatment group students had "graded out" and are not included in this figure. Of these grade-outs, 53 never used their scholarships, 22 partially used their scholarships, and 19 fully used their scholarships over all of the years that were available to them.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

SOURCE: WSF's payment files.

Reasons for Not Participating Among the Treatment Group

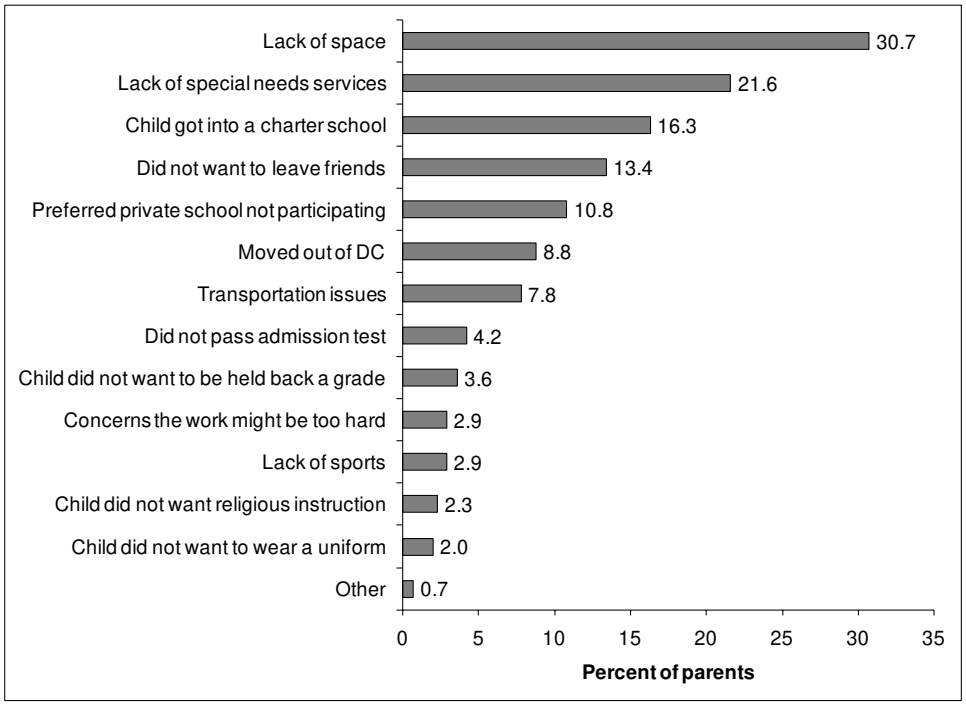
Students who were offered a scholarship could decline to participate in the OSP either initially or at any point during the evaluation follow-up period. Among those who completed surveys, the subgroup of parents of treatment group students who had never used their scholarships by 2008-09 cited a variety of reasons for not participating in the Program despite having the opportunity to do so (figure 2-5). The most common reasons given by parents of students who never used their scholarship were:

- Lack of available space in the private school they wanted their child to attend (31 percent);
- Unable to find a participating school that offered services for their child's special needs (22 percent);

- Child was accepted into a public charter school (16 percent);
- The child did not want to leave his/her friends in public school (13 percent); and
- The private school the child wanted to attend was not participating (11 percent).

Among students who initially used a scholarship but then left the Program, the most common reasons for leaving were that the child was admitted to a preferred public charter school (22 percent), a lack of space at their preferred private school (19 percent), and that the family moved out of DC (15 percent) (figure 2-6).

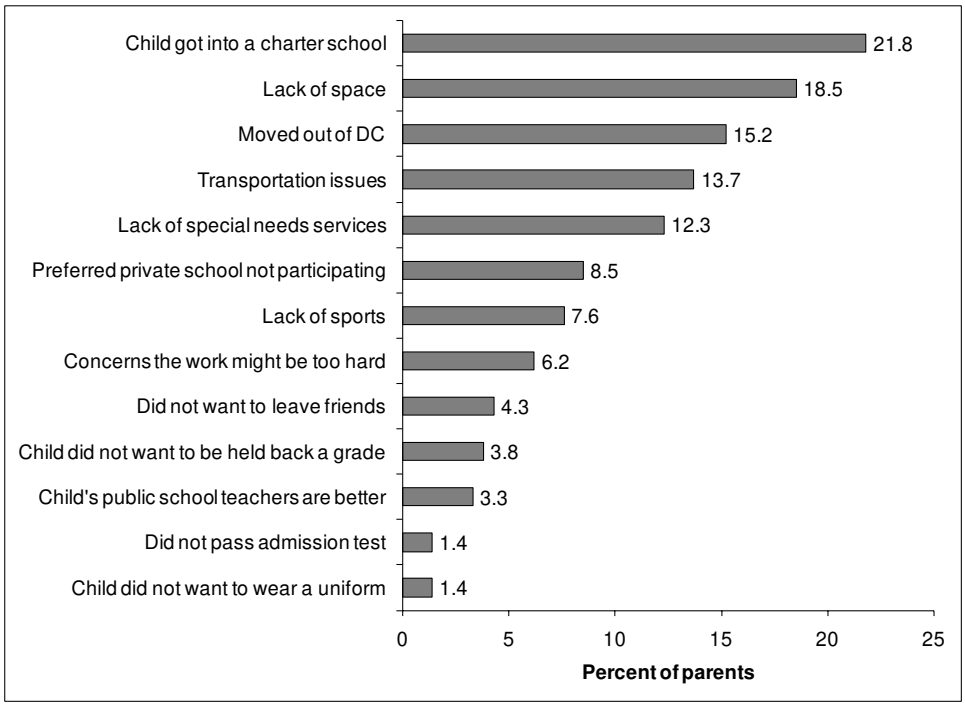
Figure 2-5. Reasons Given by Parents of Treatment Students for Never Using an OSP Scholarship



NOTES: Responses are unweighted. Respondents were able to select multiple responses each year, and some respondents participated in data collection for multiple years. Percentages given represent the sum of all responses obtained across years one through four of data collection (i.e., longitudinal responses) divided by the sum of all respondents (N = 306) across all of those same years (i.e., longitudinal respondents). As a result, this figure includes initial responses from parents of students who subsequently graded out of the Program. Categories with responses from fewer than three parents in any year are collapsed into the "Other reasons" category for confidentiality purposes.

SOURCE: Impact Evaluation Parent Surveys.

Figure 2-6. Reasons Given by Parents of Treatment Students for Not Continuing to Use an OSP Scholarship



NOTES: Responses are unweighted. Respondents were the parents of treatment students who used a scholarship in a previous year but not in a subsequent year (N = 211). The reasons for not using were drawn from the parent responses the first year after their child stopped using a scholarship. Respondents appear in the data only one time (i.e., unique respondents), though they may have provided multiple reasons for not continuing to use a scholarship. This figure includes initial responses from parents of students who subsequently graded out of the Program.

SOURCE: Impact Evaluation Parent Surveys.

Overall Movement Into and Out of Private and Public Schools

Where did students who declined to participate in the OSP attend school instead? Children in the treatment group who never used the OSP scholarship offered to them, or who did not use the scholarship consistently, could have remained in or transferred to a public charter school or a traditional DC public school, or enrolled in a non-OSP-participating private school. The same alternatives were available to students who applied to the OSP, were entered into the lottery, but were never offered a scholarship (the impact sample's control group); they could remain in their current DC public school (traditional or charter), enroll in a different public school, or try to find a way to attend a participating or nonparticipating private school. These choices could affect program impacts because traditional public, public charter, and private schools are presumed to offer different educational experiences and because previous studies suggest that switching schools has an initial short-term negative effect on student achievement (Hanushek, Kain, and Rivkin 2004).

The members of the impact sample were all attending DC public schools or were rising kindergarteners in the year they applied to the OSP. Of the students who were not entering kindergarten, approximately three-fourths were attending traditional DC public schools, while the remaining one-fourth were attending public charter schools. At least four years after random assignment, there was substantial variation across educational sectors (table 2-4).

Table 2-4. Percentage of the Impact Sample Still in K-12 by Type of School Attended: At Baseline and 2008-09

	Baseline		2008-09		
	Public		Public		
	Traditional	Charter	Traditional	Charter	Private
Treatment	75.1	24.9	26.6	18.4	55.0
Control	74.6	25.4	53.2	35.3	11.5
Difference	.5	-.5	-26.6	-16.9	43.5

NOTES: Baseline year means presented here differ from those presented in previous reports due to the exclusion of grade-outs. The longitudinal statistics presented in this table exclude data from students who were rising kindergarteners at baseline to reduce the risk of compositional bias across the years examined. As a result, the type of school attended reported here may vary slightly from other cross-sectional descriptions of school attended found in this report. Student N = 1,375. Percent missing baseline: Treatment = 4.7, Control = 7.9; percent missing year four or year five: Treatment = 19.4, Control = 25.5. Data are unweighted and represent actual responses. Given the rates of missing data, readers are cautioned against drawing firm conclusions.

Results for cohort 1 are based on five years after random assignment and for cohort 2 on four years after random assignment.

SOURCES: Program applications and Impact Evaluation Parent Surveys.

Based on data from survey respondents still in grades K-12 in 2008-09:[29]

- Twenty-seven percent of the treatment group and 53 percent of the control group attended a traditional public school;
- Eighteen percent of the treatment group and 35 percent of the control group were enrolled in public charter schools; and
- Fifty-five percent of the treatment group and 12 percent of the control group attended a private school.

These descriptive data regarding the types of school attended at baseline and four or five years after application to the OSP are limited to the sample of parents who identified their child's school in follow-up surveys or in response to telephone inquiries, which was approximately 62 percent of both the treatment and control groups. Readers are cautioned not to draw conclusions about the impact of the OSP in causing these patterns of school-sector enrollments.

[29] The subset of survey respondents in the treatment group is disproportionately comprised of treatment users; that is why the rates of treatment group members attending private schools presented here are significantly higher than the overall scholarship usage rates presented in other sections of the report. It is necessary to rely on survey respondents—in both the treatment and control groups—for the descriptive comparison provided here because WSF's payment files, which are used to calculate the Program-wide scholarship usage rates, do not contain any information on the types of schools attended by treatment nonusers or control group members.

These data show how assignment to treatment (being offered a scholarship) is not perfectly correlated with private school attendance and that assignment to the control group (no scholarship offer) does not necessarily entail attendance at a traditional public school. A number of school choices are available in DC to parents who seek alternatives to their neighborhood public school, and many members of the control group availed themselves of school choice options even if they were not awarded an Opportunity Scholarship. Similarly, among the treatment group still in grades K-12 in 2008-09, 45 percent of them chose not to use their scholarship. Twenty-seven percent of the K-12 treatment group attended a traditional public school in 2008-09, and 18 percent attended a public charter school.

The enrollment patterns of students who attended schools designated SINI from 2003-05 is a special focus of this evaluation, given that Congress assigned SINI students to be the highest service priority of the OSP (Section 306), and the evaluation covered student applicants in 2004 and 2005. Among the applicant parents of students in the impact sample that were still in grades K-12 and who provided the identity of their child's school (table 2-5):

- Fifty-three percent of the treatment and 44 percent of the control parents reported that, at the time they applied to the Program, their child was attending a school designated in need of improvement between 2003 and 2005 (SINI 2003-05).

- At least four years after random assignment, 20 percent of treatment group students were attending schools designated SINI between 2003 and 2005, while the percentage of control group students in such schools was 37 percent.

Table 2-5. Percentage of the Impact Sample Attending Schools Identified Between 2003 and 2005 as in Need of Improvement (SINI): Baseline and 2008-09

	Baseline		2008-09		
	SINI 2003-05 Schools	Not SINI 2003-05 Schools	SINI 2003-05 Schools	Not SINI 2003-05 Schools	Private
Treatment	53.0	47.0	20.2	24.8	55.0
Control	43.8	56.2	36.7	51.8	11.5
Difference	9.3	-9.3	-16.5	-27.0	43.5

NOTES: Schools were identified as SINI 2003-05 if they were officially designated as in need of improvement under the *Elementary and Secondary Education Act* between 2003 and 2005. Baseline year means presented here differ from those presented in previous reports due to the exclusion of grade-outs. The longitudinal statistics presented in this table exclude data from students who were rising kindergarteners at baseline to reduce the risk of compositional bias across the years examined. As a result, the type of school attended reported here may vary slightly from other cross-sectional descriptions of schools attended found in this report. Student *N* = 1,375. Percent missing baseline: Treatment = 4.7, Control = 7.9; percent missing year four or year five: Treatment = 19.4, Control = 25.5. Data are unweighted and represent actual responses. Given the rates of missing data, readers are cautioned against drawing firm conclusions.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

SOURCES: Program applications and Impact Evaluation Parent Surveys.

3. Impacts on Key Outcomes At Least Four Years After Application to the Program

The statute that authorized the District of Columbia Opportunity Scholarship Program (OSP) mandated that the Program be evaluated with regard to its impact on student test scores and safety, as well as the "success" of the Program, which, in the design of this study, includes satisfaction with school choices. The shorter term effects of the OSP on these measures, after one, two, or three years, have been examined in prior reports. However, the statute also required an assessment of the Program's impact on students' educational attainment. This educational outcome is evaluated for the first time in this report because it required four or more years from application to the OSP for a sufficiently large number of students to "age up" to status as potential high school graduates.

This chapter presents the longer run impacts of the Program on the full set of outcomes in 2009, which was four (for cohort 2) or five (for cohort 1) years after families and students applied to the OSP and could begin their first school year in the Program.[30] These impacts are cumulative, in that they represent students' entire experience up until that point in time. The first section summarizes the analytic methods used to determine the results and the techniques used to display them. The second section provides a brief review of the shorter run impacts, as reported previously (Wolf et al. 2007, Wolf et al. 2008, 2009). Section 3 presents the updated impacts of the program on students' educational outcomes in terms of achievement and attainment. The fourth and fifth sections describe the effects on student and parent views of school safety and satisfaction. The final section provides a brief summary of the chapter findings.

3.1 Analytic and Presentation Approaches

The sections in this chapter convey a substantial amount of information, through tables, figures, and their associated statistics.

[30] Cohort 2 represents 86 percent of the sample of students with test scores in 2009, while the earlier cohort 1 comprises 14 percent. The analysis presented in this chapter pools the cohorts in order to provide a combined impact. See appendix A.3 for more detail.

Impacts of the Program: the Scholarship Offer and Scholarship Use

For each key outcome that is a focus of the evaluation, this chapter presents the impacts of being awarded (i.e. "offered") a scholarship and of using a scholarship because both are included in the study's research questions (see table 3-1 and chapter 1). The first impacts are derived straight from the randomization of applicants into treatment and control groups (the "Intent to Treat" or ITT analysis) through the lotteries. The second set of results (the "Impact on the Treated" or IOT analysis) takes the impacts from the ITT analysis but adjusts them by the rate of scholarship nonuse, effectively re-scaling the ITT impacts across only the treatment students who actually used their scholarship at any point during the evaluation. Since the IOT analysis involves simply a straightforward mathematical adjustment of the ITT results, and not a new statistical comparison of the treatment and control groups, the significance test and resulting "*p* value" from the ITT analysis is applied to the IOT results as well. Appendix sections A.8 and A.9 provide a more detailed description of the analytic methods used for both types of analyses, based on the receipt of a scholarship offer and based on the actual use of a scholarship. For information about the relationship between attending private school (with or without an Opportunity Scholarship) and outcomes see appendix D.

Table 3-1. Overview of the Analytic Approaches

Research Question	Approach
• What is the impact of being awarded (offered) an OSP scholarship?	ITT Analysis; includes never users, partial users, and full users as members of the treatment group. We compare the outcomes of students randomly assigned to receive the offer of a scholarship (treatment group) with the outcomes of students randomly assigned to not receive the offer (control group). The difference in outcomes is the impact of being offered a scholarship.
• What is the impact of using an OSP scholarship to attend a participating private school?	IOT Analysis Drawing on the impacts of being offered a scholarship, we use a simple computational technique to net out two groups of students: (1) those who received a scholarship offer but declined to ever use it (the "never users") and (2) those who never received a scholarship offer but who, by virtue of having a sibling with an OSP scholarship, wound up in a participating private school (the "program-enabled crossover").[31]

[31] The "never users" comprise 22 percent of the non-graded out impact sample treatment group (as reported in chapter 2) but 15 percent of the treatment group who participated in data collection in 2009. This difference reflects the well-established survey research finding that sample members who decline to participate in a program are least likely to participate in data collection about it (e.g., Hanushek 1999). To calculate impacts on scholarship users, we use the percentages of data respondents, and in the analysis we control for the nonresponse phenomenon using weights based on students' characteristics at application. The percentage of "program-enabled crossover" is also based on treatment group members who participate in data collection because only respondents can provide information about the school they are currently attending. In 2009, these crossovers made up 2.9 percent of the control group.

For a longitudinal program evaluation such as this one, it is important that the study respondents whose data inform the analysis of impacts after at least four years accurately reflect the initial group of students who applied to the Program and were randomly assigned through lotteries ("the impact sample"). We found that study respondents in the final year of data collection were statistically similar to the overall impact sample on 13 of 14 baseline (pre-Program) characteristics before nonresponse weights were applied—a difference predicted by chance—and indistinguishable from each other on all these features after applying weights (see appendix A, pp. A-31 and A-33). Thus, the impact results presented here do not appear affected by participant nonresponse and do appear to be generalizable to the original sample of students in the study.

Subgroups for which Impacts are Estimated

The results of primary interest pertain to the impact of the OSP on all of the students and parents in the impact sample. A secondary set of results across various student subgroups of policy interest also is discussed. The participant subgroups that are analyzed in this study were designated prior to the collection and analysis of Program impacts, with the designation based on their use in previous evaluations of scholarship programs or importance to contemporary policy discussions about educational improvement. They are:

- **Whether students attended a SINI school prior to application to the Program.** The Program statute designates such students as the highest service priority for the OSP, making the question of whether Program impacts vary based on SINI status a central component of the evaluation. Previous studies of scholarship programs have considered whether achievement impacts differ for students who apply from higher quality or lower quality schools (Barnard et al. 2003; Mayer et al. 2002, appendix E).

- **Whether students were relatively lower performing or relatively higher performing at baseline.** Previous scholarship evaluations have examined whether achievement impacts vary based on initial student performance levels, suggesting that such programs could have a greater effect on lower performers because they have the most to gain from a change, or on higher performers since they might be better prepared to benefit from a private school environment (Howell et al. 2006, p. 155).

- **Student gender.** Researchers have argued that girls and boys learn differently (Gilligan 1993; Summers 2001), and therefore, educational interventions might have differential effects on students based on their gender.

Analyses by student subgroups have lower statistical power to detect impacts and are less precise because they parse the overall data into smaller groups. In previous evaluation reports, two additional subgroup pairings of students were analyzed: (1) whether students were in grades K-8 or 9-12

at the time of application and (2) whether students were in application cohort 1 (applied in 2004) or application cohort 2 (applied in 2005). However, all of the students in grades 9-12 at the time of application and a significant portion (32 percent) of students in cohort 1 had aged up or "graded-out" of the Program by the 2008-09 school year and therefore were not eligible for data collection. The remaining subsamples of students in the grade 9-12 and cohort 1 subgroups were too small to conduct a reliable analysis of these subgroup pairings (see appendix section A.2 for statistical power calculations for these and the other subgroup impacts).

In the analysis of impacts by subgroup, two types of comparisons are made. First, the difference between the outcomes of treatment and control group students are compared within each subgroup. In other words, the Program impact is estimated just for cases within a subgroup. Second, the Program impact for a subgroup (e.g., SINI 2003-05 students) is compared with the Program impact for its opposing subgroup (e.g., not SINI 2003-05 students) and significance tests are conducted on that difference. This tells us whether we can have confidence that the Program is having a differential effect on groups of students. We found no evidence of clear differential effects.

Understanding the Tables and Figures

In presenting the results that follow, we provide a variety of information about the average outcomes (means) for the treatment and control groups and any difference between them (i.e., the programmatic impact) that is drawn from the regression equations described in appendix section A.8:[32]

- The text and tables include effect sizes (ES) to translate each impact into a standard metric and allow the reader to assess whether the size of the impact might be considered meaningful, whether or not it is statistically significant.[33]

- The *p*-values in the tables give a sense of the extent to which we can be certain that an estimated impact of the Program is reliable and not a chance finding. The smaller the *p*-value, the more confidence we can have that an observed impact is due to the treatment and not merely due to chance. Any result with a *p*-value higher than .05 is characterized as "not statistically significant," consistent with the traditional standard of 95 percent confidence used in most evaluation research.

---

[32] Readers interested in the results based on unadjusted subgroup means can find them in appendix E.

[33] Specifically, the effect sizes are computed as a percentage of a standard deviation for the control group after four years. In the cases where outcomes are for a particular subgroup of students, effect sizes are computed as a percentage of a standard deviation for the control group students within the respective subgroup. Since the outcomes of the experimental control group signal what would have happened to the treatment group in the absence of the intervention, a standard deviation in the distribution of the control group outcomes represents an especially appropriate gauge of the magnitude of any treatment impacts observed. The power analysis (see appendix section A.2) forecasts that this final impact analysis will contain sufficient data to correctly identify an overall reading or math impact of the offer of a scholarship of .13 standard deviations or higher if such an impact actually exists. Subgroup ITT impacts are estimated to be detectable at various sizes, ranging from .15 to .23 standard deviations.

- Whenever the results from each subgroup pair (e.g., males and females) are presented in tables, the difference between the average impact on each subgroup is also presented, in a row labeled "Difference," and the difference is subjected to the same statistical test for significance (i.e., *T*-test) as the impacts on the subgroups themselves. If the difference (i.e., interaction effect) is not statistically significant, it means we cannot be certain that there are real differential impacts across subgroups, even though one subgroup (e.g., females) may demonstrate a significant Program impact and the other subgroup (males) may not (see appendix section A.8 for more details about the correct interpretation of the subgroup results).

- A statistical test was administered to the results drawn from multiple comparisons of treatment and control group members across the six different subgroups to identify any statistically significant findings that could be due to chance, or what statisticians refer to as "false discoveries" (Benjamini and Hochberg 1995; Schochet 2007, p. 5) (appendix B).[34] These adjustments are made independently for each year of this evaluation based upon the cross-sectional analysis of outcomes each year. Throughout this report, the phrase "may have had an impact" is used to caution readers regarding statistically significant impacts that may have been false discoveries.

- The impact results from the primary analysis were subject to sensitivity tests that included a sample trimmed to exactly equalize the treatment and control response rates at the level before response conversion efforts were applied to initial nonrespondents in both the treatment and control group (trimmed sample) and the clustering of student observations on the school attended instead of family (clustering on current school).[35] These analyses were conducted to assess how robust the main estimates are to specific modifications in the analytic approach (appendix C). Because they were conducted as a robustness check on the results of the primary analysis, and not as alternatives to that analysis, no adjustments were made for multiple comparisons in the estimations that make up the sensitivity analysis.

3.2 Impacts Reported Previously (Through Year 3)

The evaluation of the impact of the OSP is a longitudinal study, in that it tracks the outcomes of a sample of students over multiple years of their potential participation in the Program. Three earlier reports described impacts one, two, and three years after students applied to the OSP and were randomly assigned by lottery to either the treatment or control group.[36] Each year's impacts provide the cumulative

---

[34] The estimates of the treatment impacts on parent and student perceptions of safety were not adjusted for multiple comparisons, since each was estimated using a single safety index. Although the treatment impact on perceptions of parent and student satisfaction was estimated using three measures for each of the two samples, two of those measures ("percent assigning the school a grade of A or B" and "average grade assigned to school") are the exact same outcome data classified two alternative ways. Finally, only a single measure of parent and student satisfaction—the percentage assigning a grade of A or B—is used in the primary analysis of Program impacts presented here, reducing the danger of chance false discoveries in that specific outcome domain.

[35] Because subsampling for nonresponse conversion was not used in the case of the follow-up survey on educational attainment, the trimmed sample sensitivity test was not applied to the attainment analysis. Instead, a calculation was made regarding how different the results from nonrespondents would have had to have been for the true difference between the treatment and control on attainment to be 0.

[36] See Wolf et al. 2007, table ES-2, Wolf et al. 2008, xxii-xxiv, Wolf et al. 2009, xxvi-xxx.

effects of the Program on the students who provided data that year, not incremental year-to-year impacts. Thus, the earlier studies described the shorter term impacts of the Program while the later results represent estimates of longer term impacts.

Leading up to this report we found that the full sample of students awarded scholarships performed at similar levels in math compared to those not given scholarships, but after three years they scored significantly higher in reading achievement. This difference was equivalent to about three (impact of the scholarship offer) or four (impact of scholarship use) months of additional learning compared to what the control group experienced over the 30 months of schooling that was measured. The positive effect on reading test scores was observed for 5 of 10 subgroups of students (an extra 4 to 19 months of learning) but was not observed for the high-priority SINI 2003-05 subgroup (or the subgroups of males, students who entered the Program with lower achievement, students entering the Program in high school, or cohort 2). Across the first three years, parents consistently rated their child's school as safer and gave it a higher grade if their child was offered or used a scholarship. There were no impacts on students' own views of school safety and satisfaction during the first three years.

3.3 Impacts on Student Educational Outcomes After At Least Four Years

The statute identifies students' academic achievement as the primary focus of the evaluation and another student outcome, educational attainment, as an important secondary one (Sec. 309(a)(4)).

In summary, the analysis of the longer-term effects of the OSP on these outcomes revealed:

- No statistically significant impacts on overall student achievement in reading and math after at least four years (table 3-2), although sensitivity tests suggest the reading impacts could be positive under alternative estimation methods (appendix C, table C-1).

- No significant impacts on reading or math test scores for students who came from SINI 2003-05 schools, the subgroup of students for whom the statute gave top priority (table 3-3). Nor was there evidence of achievement impacts for male students and students who applied to the Program with relatively lower levels of academic performance (table 3-3). There was no evidence of math achievement impacts for any subgroups examined (table 3-3).

- Positive Programmatic impacts in reading achievement for the remaining three of six subgroups examined: (1) participants who applied from not SINI 2003-05 schools, (2) those who applied to the Program with relatively higher levels of academic performance, and (3) female students (table 3-3). However, these findings should be

interpreted with caution as multiple comparison tests suggest they could be false discoveries.

- The offer of an OSP scholarship raised students' probability of graduating by 12 percentage points and the use of a scholarship by 21 percentage points, based on parent self-reports (table 3-5). Statistically higher graduation rates were also found for SINI 2003-05 students, students who applied to the Program with relatively higher levels of academic performance, and female students if they were offered or used an OSP scholarship (table 3-5).

Impacts on Achievement for the Full Sample of Students

The mandate to assess the Program's effect on student test scores is consistent with the stated purpose of the Program and the priority Congress placed on having the OSP serve students from low-performing schools. Academic achievement as a measure of Program success is also well aligned with parents' stated priorities in choosing schools (Wolf et al. 2005, p. C-7).

The primary analysis indicates no statistically significant overall impact of the Program on reading or math achievement after at least four years. That is, the average reading and math test scores of the treatment group as a whole were not significantly different from those of the control group as a whole (table 3-2).[37]

Table 3-2. Impact Estimates of the Offer and Use of a Scholarship on the Full Sample: Academic Achievement, 2008-09

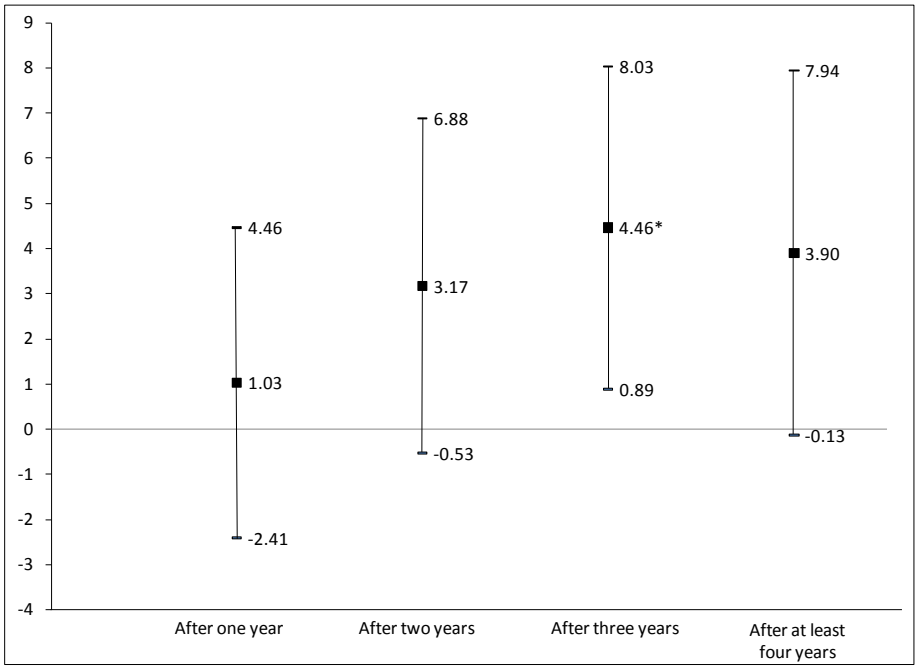
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Reading	649.15	645.24	3.90	.11	4.75	.13	.06
Math	644.06	643.36	.70	.02	.85	.03	.71

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are in terms of standard deviations. Valid *N* for reading = 1,328; math = 1,330. Separate reading and math sample weights used.

The achievement impacts at least four years after random assignment can be viewed clearly and placed in the context of impacts estimated in prior years in figures 3-1 and 3-2. The 95 percent confidence interval for the regression-adjusted difference of 3.90 scale score points between the treatment and control groups in reading at least four years after random assignment ranges from -.13 to

[37] Appendix E contains a parallel set of results tables that include the raw (unadjusted) group means as well as additional statistical detail regarding the impact estimates.

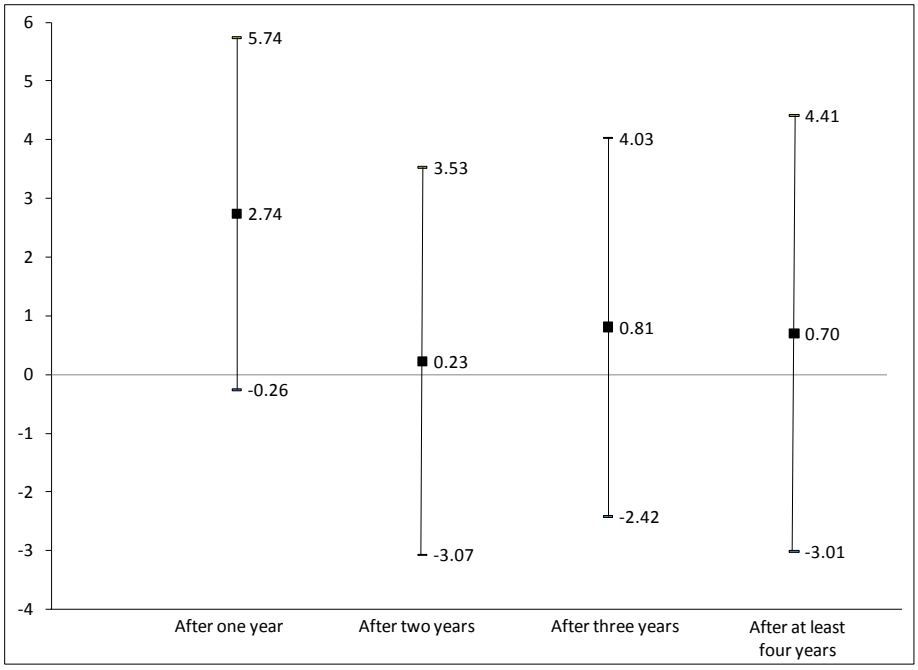
Figure 3-1. Impact of the OSP on Reading Achievement Overall, by Years After Application



* Statistically significant at the 95 percent confidence interval.

NOTES: The dark squares are the regression-adjusted impact estimates (difference in means), and the vertical lines are the 95 percent confidence interval around the impacts.

Figure 3-2. Impact of the OSP on Math Achievement Overall, by Years After Application



NOTES: The dark squares are the regression-adjusted impact estimates (difference in means), and the vertical lines are the 95 percent confidence interval around the impacts.

7.94.[38] Because the lower end of the confidence interval is not greater than the value for a zero impact (the horizontal axis), we cannot be confident that the overall reading impact is positive as opposed to zero or negative. Likewise, because the 95 percent confidence interval for the regression adjusted difference of .70 scale score points between the treatment and control groups in math at least four years after random assignment ranges from -3.01 to 4.41, we are uncertain if the general math impact is positive, zero, or negative.

The reading and math achievement impacts at least four years after random assignment appear to be similar to the impacts for the prior year, but differences between each year's cumulative impact estimates, which would indicate whether an additional year of participation in the Program improved student performance, have not been tested to determine their statistical significance. Out of the eight impact estimates presented in figures 3-1 and 3-2, one, reading achievement after three years, indicates a statistically certain Program impact.[39]

As in previous years, we re-calculated impacts using two alternate estimation approaches. The first approach used exactly the same analytic approach as the main analysis but applied that approach to a sample of respondents for which enough of the "latest-to-respond" treatment group students were excluded from the sample so that the treatment group response rate exactly matched the pre-subsample control group response rate of 63.5 percent. The second approach used the exact same respondent sample as the main analysis but generated an alternative set of robust standard errors by clustering on the school the student attended instead of on the student's family. The two sensitivity tests yielded results that were consistent with the math findings from the main analysis but different regarding overall reading impacts. The overall impact estimate for reading was larger (4.80 scale score points) and statistically significant when drawn from the trimmed sample; the overall reading estimate was also statistically significant when the statistical model was modified to control for similarities among students who attended the same school as opposed to controlling for similarities among students who are from the same family (appendix C, table C-1). These tests indicate that the effects of the OSP on reading achievement after at least four years depend on the particular way it is estimated.

---

[38] The scale score mean and standard deviation (SD) for the SAT-9 norming population varies by grade and is 463.8 (SD = 38.5) for kindergarteners tested in the spring, compared to 652.1 (SD = 39.1) for fifth graders and 703.6 (SD = 36.5) for students in 12th grade. Because scale scores on a given test are vertically equated, it is possible to combine observations across grades in the estimation of overall test score impacts. The estimation of those impacts, through regression equations, and the calculation of effect sizes produce average effects and effect sizes across the entire group that are weighted by the proportion of students in each grade.

[39] The samples of students who provided outcome data each year varied somewhat. Therefore, readers are cautioned not to draw conclusions about differences in impacts across the years. The difference in samples across the years was greatest between the sample after three years and the sample at least four years, due to the fact that 296 students had graded-out of the study at least four years after random assignment.

Impacts on Achievement at the Subgroup Level

The offer of a scholarship, and the use of a scholarship, had a statistically significant positive impact on reading achievement at least four years after random assignment for one-half of the student subgroups, including at least two subgroups who applied with a relative advantage in academic preparation. There were no impacts on math achievement for any of the six subgroups examined, as was true for the full impact sample (tables 3-3, 3-4). The subgroups with positive reading impacts include:

- Students in the treatment group who had not attended a SINI 2003-05 school prior to the Program. These students scored an average of 5.8 scale score points higher (3.5 months of additional learning) in reading than students in the control group from not SINI 2003-05 schools (the impact of the offer of a scholarship); the calculated impact of using a scholarship for this group was 7.0 scale score points (4.2 months of additional learning).

- Students in the treatment group who entered the Program in the higher two-thirds of the applicant test-score performance distribution scored an average of 5.2 scale score points higher in reading (3.9 months of additional learning) than similar students in the control group; the impact of using a scholarship for this group was 6.1 scale score points (4.6 months).

- Female students in the treatment group, who scored an average of 5.3 scale score points higher in reading (3.4 months) than female students in the control group; the impact of using a scholarship was 6.2 scale score points (4.0 months).

The sensitivity tests confirmed the main analysis subgroup results (see appendix C, table C-1). However, because multiple comparison tests indicate that it is possible that these positive subgroup effects are false discoveries and because the difference between each of these groups and their corresponding pair (e.g., SINI 2003-05 vs. not SINI 2003-05) was not statistically significant, the subgroup impacts should be interpreted with caution (see appendix B, table B-1).

There was no evidence of a subgroup impact in reading for students who applied from a school designated SINI between 2003 and 2005—the highest service priority for the Program according to the statute. The analysis also did not show evidence of subgroup impacts for students who entered the Program in the lower one-third of the applicant test-score performance distribution or for male students.

Table 3-3. Impact Estimates of the Offer and Use of a Scholarship on Subgroups At Least Four Years After Application: Academic Achievement

	Reading						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement: Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
SINI 2003-05	657.49	656.41	1.08	.03	1.33	.04	.76
Not SINI 2003-05	643.25	637.45	5.80*	.16	6.99*	.19	.02
Difference	14.24	18.96	-4.72	-.13			.27
Lower performance	629.45	628.27	1.18	.04	1.54	.05	.74
Higher performance	657.79	652.61	5.18*	.15	6.08*	.18	.04
Difference	-28.34	-24.34	-4.00	-.11			.35
Male	642.78	640.33	2.45	.07	3.07	.09	.44
Female	654.64	649.38	5.27*	.15	6.24*	.18	.05
Difference	-11.86	-9.05	-2.81	-.08			.50

	Math						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement: Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
SINI 2003-05	656.67	657.05	-.38	-.01	-.47	-.02	.90
Not SINI 2003-05	635.31	633.89	1.42	.04	1.71	.05	.56
Difference	21.36	23.16	-1.80	-.05			.65
Lower performance	632.39	631.16	1.24	.04	1.61	.05	.71
Higher performance	649.08	648.59	.49	.01	.58	.02	.83
Difference	-16.69	-17.44	.75	.02			.85
Male	639.59	640.70	-1.11	-.04	-1.38	-.04	.68
Female	648.04	645.64	2.40	.07	2.84	.08	.37
Difference	-8.44	-4.94	-3.50	-.10			.35

*Statistically significant at the 95 percent confidence level.

† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are in terms of standard deviations. Total valid *N* for Reading = 1,328, including: SINI 2003-05 *N* = 520, Not SINI *N* = 808, Lower performance *N* = 435, Higher performance *N* = 893, Male *N* = 649, Female *N* = 679. Total Valid *N* for Math = 1,330, including SINI 2003-05 *N* = 516, Not SINI 2003-05 *N* = 814, Lower performance *N* = 435, Higher performance *N* = 895, Male *N* = 649, Female *N* = 681.

Table 3-4. Estimated Impacts in Months of Schooling From the Offer and Use of a Scholarship for Statistically Significant Reading Impacts After At Least Four Years

Student Achievement: Reading	Impact of the Scholarship Offer (ITT)		Impact of Scholarship Use (IOT)	
	Effect Size	Months of Schooling	Effect Size	Months of Schooling
Not SINI 2003-05	.16	3.49	.19	4.21
Higher performance	.15	3.90	.18	4.58
Female	.15	3.37	.18	4.00

NOTES: Treatment impacts were converted to months based upon the average monthly increase in reading scale scores for the control group across all years.

Two exploratory analyses of the impact of the OSP on student test scores were conducted. An examination of the impacts on a newly constructed pair of student subgroups—those whose parents listed "academic quality" as their highest priority in choosing a school and those whose parents did not—is presented in appendix F. An analysis of the achievement impacts of the OSP across the entire outcome distribution of test scores was performed using quantile regression methods. Appendix G presents a discussion of those findings.

Impacts on Attainment for the Full Sample and Subgroups of Students

The evaluation considers the outcome of educational attainment because the OSP statute requires such an assessment,[40] and prior research has linked enrollment in private schools, particularly Catholic schools, to higher rates of high school graduation as well as higher college enrollment rates (Evans and Schwab 1995; Grogger and Neal 2000; Neal 1997; Warren 2010). This evaluation is the first to use random assignment to estimate the causal relationship between a school voucher program or private schooling and educational attainment, thus providing a more rigorous estimate than previous studies.

Educational attainment is most commonly measured as achievement of certain educational milestones, most notably high school graduation, college enrollment, and college graduation. Because 51 percent of the impact sample applied to enter the Program in grades K-5, and another 31 percent applied to enter grades 6-8, by four or five years later few of these students were potential candidates for high school graduation. However, combining those older middle school applicants with the 18 percent of the impact sample that entered the Program in grades 9-12 provided sufficient data to determine the impact of the OSP on high school graduation rates (see appendix section A.2 for statistical power calculations for

[40] Title III of Division C of the *Consolidated Appropriations Act*, 2004, P.L. 108-199, Section 309(4)D.

this outcome). There were too few students in the sample, however, for whom college enrollment could be confirmed at the time of data analysis.

This attainment analysis focused on students in the impact sample who were forecasted to have been seniors in high school during or before the 2008-09 school year and therefore had the opportunity to graduate prior to the summer of 2009.[41] A parent follow-up survey asked parents whether these students had received high school diplomas by the end the 2008-09 school year. Differences in parental reports of student educational attainment between the treatment and control groups were then measured via the evaluation's standard regression method of analysis (see appendix sections A.8 and A.9).

The attainment impact analysis revealed (table 3.5):

- Overall, a positive impact on high school graduation of 12 percentage points for students who received the offer of a scholarship. The high school graduation rate was 82 percent for the treatment group compared to 70 percent for the control group. Using a scholarship increased the graduation rate by 21 percentage points.

- A positive impact of 13 percentage points for students who came from SINI 2003-05 schools, the subgroup of students for whom the statute gave top priority. The high school graduation rate was 79 percent for the SINI 2003-05 members of the treatment group compared to 66 percent for the SINI 2003-05 members of the control group. Using a scholarship increased the graduation rate by 20 percentage points.

- The offer of an OSP scholarship led to a positive impact for students who applied to the Program with relatively higher levels of academic performance (14 percentage points) and female students (20 percentage points). Using a scholarship to attend a participating private school increased the graduation rate by 25 and 28 percentage points, respectively.

- There was no statistically significant evidence of impacts on graduation rates for students who applied to the Program from not SINI 2003-05 schools, with relatively lower levels of academic performance, and male students.[42]

[41] As such, the sample for the impact analysis of the OSP on educational attainment differs from the overall impact sample used for other elements of the evaluation in that it is limited to older applicants. The attainment experiences of impact sample students below grade 8 at baseline are as yet unknowable and could conceivably differ from those reported for the older applicants here.

[42] None of the subgroup interaction effects themselves were statistically significant, which means that we cannot say with confidence that the impact of the OSP on educational attainment was different across the various subgroup pairs.

Table 3-5. Impact Estimates of the Offer and Use of a Scholarship on Students Forecasted in Grade 12 or Above by 2008-09: Percent with High School Diploma, 2008-09

High School Diploma	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		***p*-value of estimates**
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**.82**	**.70**	**.12****	**.26**	**.21****	**.46**	**.01**
SINI 2003-05	.79	.66	.13*†	.28	.20*†	.43	.01
Not SINI 2003-05	.89	.82	.07	.19	.21	.54	.46
Difference	-.10	-.16	.06	.13			.59
Lower performance	.60	.49	.12	.23	.20	.40	.12
Higher performance	.93	.79	.14*†	.35	.25*†	.61	.02
Difference	-.33	-.30	-.03	-.06			.80
Male	.71	.66	.07	.14	.14	.30	.26
Female	.95	.75	.20**†	.46	.28**†	.65	.01
Difference	-.24	-.08	-.15	-.34			.18

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

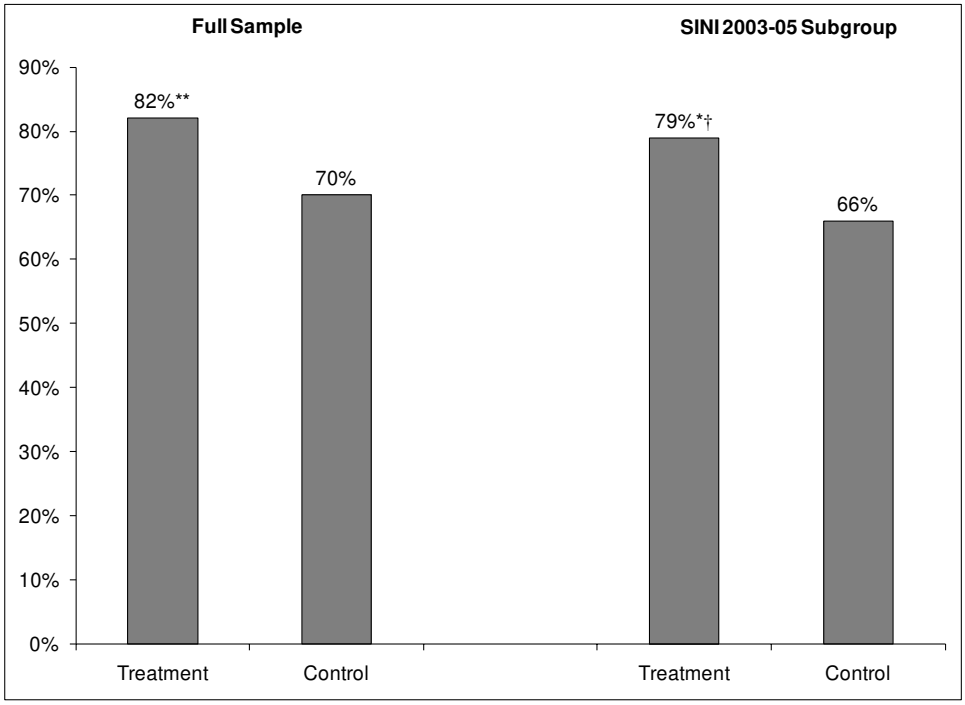
NOTES: Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid N = 316, including SINI 2003-05 N = 231, Not SINI 2003-05 N = 85, Lower performance N = 105, Higher performance N = 211, Male N = 167, Female N = 149. Sample weights used.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. High school graduation determined via parental self-reports.

All of these OSP attainment impact findings based on parent-provided data retained their significance after adjustments for multiple comparisons (see appendix B, table B-2) and through the sensitivity testing (appendix C, table C-2).[43] The attainment impacts of the OSP on the overall sample and the high-priority SINI 2003-05 subgroup of students can be viewed most clearly in figure 3-3.

[43] A total of 36.8 percent of the impact sample students who were targets of the attainment follow-up survey did not provide responses. This nonresponse rate was similar for both the treatment (37.1) and control (36.6) groups and so does not raise concern for bias in the analysis but does limit the generalizability of the findings (appendix A, table A-9). We calculated how different the impact of the Program on nonrespondents would have to have been in order for the high school graduation rates of all the treatments (respondents plus nonrespondents) and all the controls (respondents plus nonrespondents) to be identical. Given the impact of 12 percentage points on the 63.2 percent of the impact sample that responded, the impact on the 36.8 percent that did not respond would, instead, have had to have been a decrease of 21 percentage points in order for the total impact to be exactly zero (appendix C, table C-1).

Figure 3-3. High School Graduation Rates of the Treatment and Control Groups for the Overall Sample and the SINI 2003-05 Subgroup, 2008-09



*Statistically significantly at the 95 percent confidence level.

**Statistically significantly at the 99 percent confidence level.

† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* = 316, including SINI 2003-05 *N* = 231, Not SINI 2003-05 *N* = 85. High school graduation determined via parental self-reports.

3.4 Impacts on Reported Safety and an Orderly School Climate

School safety is a valued feature of schools for the families who applied to the OSP. A total of 17 percent of cohort 1 parents at baseline listed school safety as their most important reason for seeking to exercise school choice—second only to academic quality (48 percent) among the available reasons (Wolf et al. 2005, p. C-7). A separate study of why and how OSP parents choose schools, which relied on focus group discussions with participating parents, found that school safety was among their most important educational concerns (Stewart, Wolf, and Cornman 2005, p. v).

There are no specific tests to evaluate the safety of a school as there are for evaluating student achievement. There are various indicators of the relative orderliness of the school environment, such as the presence or absence of property destruction, cheating, bullying, and drug distribution, to name a few (see appendix A.3 for more information). Students and parents can be surveyed regarding the extent

to which such indicators of disorder are or are not a problem at their or their child's school. The responses then can be consolidated into an index of safety and an orderly school climate and analyzed, as we do here and has been done in other school choice studies.

In summary, the analysis suggests that:

- Overall, treatment group parents rated their child's school as significantly safer and more orderly than did control group parents (table 3-6).
- There is no evidence that parents of students who applied to the Program from SINI 2003-05 schools, or four of the other five subgroups, viewed their child's school as safer if they had been awarded or used an OSP scholarship (table 3-6). There was a positive impact on perceptions of school safety for parents of students who did not apply from SINI 2003-05 schools (table 3-6).
- Treatment and control group students, overall or within subgroups, had comparable views on their schools' safety and climate (table 3-7).

Parent Self-Reports

Overall, the parents of students offered an Opportunity Scholarship in the lottery subsequently reported their child's school to be safer and more orderly than did the parents of students in the control group. The impact of the offer of a scholarship on parental perceptions of safety and an orderly school climate was .48 on a 10-point index of indicators of school safety and orderliness, an effect size of 0.14 standard deviations (table 3-6; figure 3-4 for a visual display). The impact of using a scholarship was .58 on the index, with an effect size of .17 standard deviations. These findings persisted through the sensitivity tests: that is, the statistical significance of the findings did not change as a result of the different models (see appendix C, table C-3).

This impact of the offer of a scholarship on parental perceptions of safety and an orderly school climate was observed for one of the subgroups of students examined: parents of students from not SINI 2003-05 schools experienced a positive increase of .22 standard deviations (table 3-6). There was no evidence of an impact on parents of students from SINI 2003-05 schools, parents of students who entered the Program with relatively higher and lower levels of academic achievement, and parents of male and female students (table 3-6). The not SINI 2003-05 subgroup impact on parental views of school safety remained statistically significant after adjustments to account for multiple comparisons (see appendix B, table B-3) and in both sensitivity tests (appendix C, table C-3).

Table 3-6. Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Parent Perceptions of Safety and an Orderly School Climate, 2008-09

	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Safety and an Orderly School Climate: Parents	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Full sample	**7.94**	**7.47**	**.48***	**.14**	**.58***	**.17**	**.02**
SINI 2003-05	7.45	7.35	.10	.03	.13	.04	.77
Not SINI 2003-05	8.28	7.55	.73**†	.22	.88**†	.27	.01
Difference	-.83	-.20	-.63	-.19			.16
Lower performance	7.84	7.36	.48	.14	.61	.18	.20
Higher performance	7.99	7.51	.48	.15	.57	.17	.06
Difference	-.15	-.15	-.00	-.00			.99
Male	8.05	7.67	.38	.12	.48	.14	.20
Female	7.87	7.31	.56	.17	.67	.20	.06
Difference	.18	.36	-.18	-.05			.68

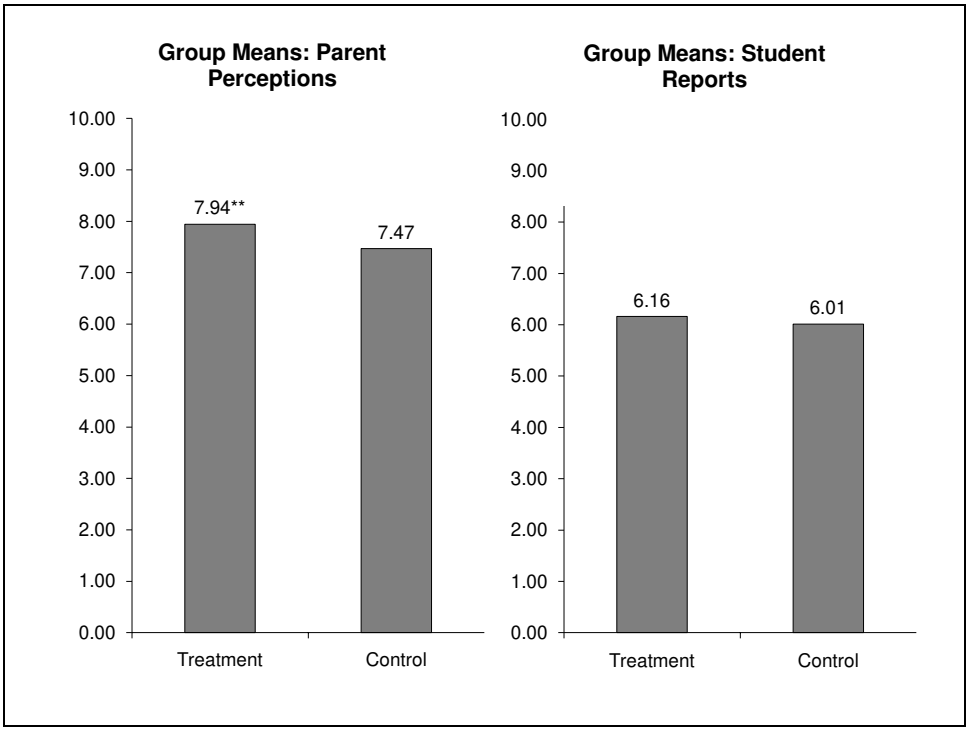
*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Effect sizes are in terms of standard deviations. Valid *N* = 1,224, including: SINI 2003-05 *N* = 472, Not SINI 2003-05 *N* = 752, Lower performance *N* = 400, Higher performance *N* = 824, Male *N* = 597, Female *N* = 627. Parent survey weights were used.

Figure 3-4. Parent Perceptions and Student Reports of Safety and an Orderly School Climate, 2008-09



**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Parent perceptions are based on a 10-point scale; student reports are based on an 8-point scale. For parent perceptions, valid $N = 1,224$; parent survey weights were used; the 10-point index of indicators of school safety and an orderly environment includes the absence of property destruction, tardiness, truancy, fighting, cheating, racial conflict, weapons, drug distribution, drug/alcohol use, and teacher absenteeism. For student reports, valid $N = 1,054$; student survey weights were used; the survey was given to students in grades 4-12; the means represent the absence of incidents on an 8-item index for student reports of students being a victim of theft, drug-dealing, assaults, threats, bullying or taunting, or had observed weapons at school. Means are regression adjusted using a consistent set of baseline covariates.

Student Self-Reports

The students in grades 4-12 who completed surveys paint a different picture about school safety at their school than do their parents. While parent safety was a measure of parental perceptions, the student index of school climate and safety asked students if they personally had been a victim of theft, drug-dealing, assaults, threats, bullying, or taunting or had observed weapons at school. On average, reports of school climate and safety by students offered scholarships through the lottery were not statistically different from those of the control group (table 3-7; figure 3-4 for a visual display). That is, there was no evidence of an impact from the offer of a scholarship or the use of a scholarship on students' reports. No statistically significant findings were evident across the subgroups analyzed. Nor did the sensitivity tests conducted lead to a different set of overall findings (see appendix C, table C-4).

Table 3-7. Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Student Reports of Safety and an Orderly School Climate, 2008-09

Safety and an Orderly School Climate: Students	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		*p*-value of estimates
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**6.16**	**6.01**	**.15**	**.07**	**.18**	**.09**	**.33**
SINI 2003-05	6.27	6.01	.26	.12	.32	.15	.21
Not SINI 2003-05	6.08	6.01	.06	.03	.08	.04	.76
Difference	.19	-.00	.20	.10			.51
Lower performance	6.08	5.80	.28	.12	.37	.17	.36
Higher performance	6.19	6.11	.09	.04	.10	.05	.62
Difference	-.11	-.31	.20	.10			.58
Male	6.04	5.93	.12	.05	.14	.06	.66
Female	6.25	6.07	.18	.09	.21	.11	.32
Difference	-.21	-.15	-.06	-.03			.85

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Effect sizes are in terms of standard deviations. Valid N = 1,054, including: SINI 2003-05 N = 502, Not SINI 2003-05 N = 552, Lower performance N = 333, Higher performance N = 721, Male N = 505, Female N = 549. Student survey weights were used. Survey given to students in grades 4-12.

3.5 Impacts on School Satisfaction

Economists have long used customer satisfaction as a proxy measure for product or service quality (see Johnson and Fornell 1991). While not specifically identified as an outcome to be studied, it is an indicator of the "success of the Program in expanding options for parents," which Congress asked the evaluation to consider.[44] Satisfaction is also an outcome studied in the previous evaluations of K-12 scholarship programs, all of which concluded that parents tend to be significantly more satisfied with their child's school if they have had the opportunity to select it (see Greene 2001, pp. 84-85).

Satisfaction of both parents and students was measured by the percentage that assigned a grade of A or B to their child's or their school.[45] In summary, the analysis suggests that at least four years after random assignment:

- Treatment group parents overall were more likely (8 percentage points) to give their child's school a high grade than were control group parents (table 3-8).

[44] Section 309 of the *District of Columbia School Choice Incentive Act of 2003*.

[45] Satisfaction impacts based on the full A-F grade scale as well as a 12-item satisfaction scale are provided in appendix E.

- Parents of students who applied to the Program from SINI 2003-05 schools were not significantly more likely to grade their child's school highly if their child had been awarded an Opportunity Scholarship (table 3-8). The same was true for parents of male students (table 3-8).

- Treatment group parents of the remaining four subgroups of students were significantly more likely (8-12 percentage points) to report a high grade for their child's school, although three of these estimates could be false discoveries (table 3-8).[46]

- There were no significant impacts of the OSP on student satisfaction with their school, overall or for any subgroup (table 3-9).

Parent Self-Reports

At least four years after random assignment, parents overall were more satisfied with their child's school if they had been offered a scholarship and if their child used a scholarship to attend a participating private school. A total of 76 percent of treatment parents assigned their child's school a grade of A or B in 2009 compared with 68 percent of control parents—a difference of 8 percentage points (impact of the offer of a scholarship) (table 3-8; figure 3-5 for a visual display); the impact of using a scholarship was a difference of 10 percentage points in parent's likelihood of giving their child's school a grade of A or B. The effect sizes of these impacts were .18 and .22, respectively (table 3-8).

For each of the six subgroups of parents, those in the treatment group were more satisfied than their counterparts in the control group. These differences, however, were not statistically significant at the subgroup level for parents of scholarship students from SINI 2003-05 schools or who were male—two groups that also did not demonstrate significant achievement gains from the Program, though the SINI 2003-05 students did show significant benefits in terms of educational attainment (section 3.3). Parents of students from not SINI 2003-05 schools, parents of students who had higher and lower test-score performance at baseline, and parents of female students were significantly more likely to give their child's school a grade of A or B if they were in the treatment group. The effect sizes ranged from .17 to .27 standard deviations for the offer of, and from .20 to .32 standard deviations for the use of, a scholarship in these groups.

[46] None of the subgroup interaction effects themselves were statistically significant, which means that we cannot say with confidence that the impact of the OSP on parent reports of school satisfaction was different across the various subgroup pairs.

Table 3-8. Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Parent Reports of Satisfaction with Their Child's School, 2008-09

Parents Who Gave Their School a Grade of A or B	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		***p*-value of estimates**
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**.76**	**.68**	**.08****	**.18**	**.10****	**.22**	**.00**
SINI 2003-05	.69	.66	.03	.07	.04	.08	.51
Not SINI 2003-05	.81	.69	.12**†	.27	.15**†	.32	.00
Difference	-.12	-.02	-.10	-.21			.12
Lower performance	.70	.61	.10*	.20	.12*	.25	.05
Higher performance	.79	.71	.08*	.17	.09*	.20	.03
Difference	-.08	-.10	.02	.04			.75
Male	.75	.67	.08	.17	.10	.21	.06
Female	.77	.68	.09*	.19	.10*	.22	.03
Difference	-.02	-.01	-.01	-.01			.90

*Statistically significant at the 95 percent confidence level.

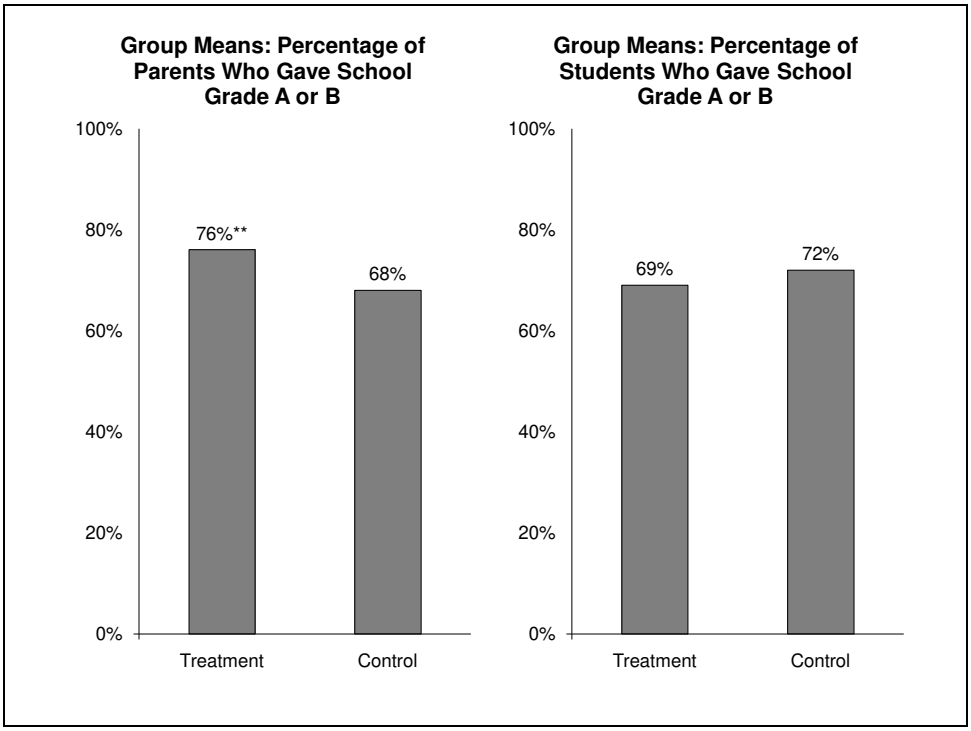
**Statistically significant at the 99 percent confidence level.

† = subgroup impact result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid N = 1,227, including: SINI 2003-05 N = 475, Not SINI 2003-05 N = 752, Lower performance N = 402, Higher performance N = 825, Male N = 594, Female N = 633. Parent survey weights were used.

With the exception of the impact for parents of students from not SINI 2003-05 schools, adjustments for multiple comparisons suggest that the subgroup impacts on parent satisfaction may be false discoveries (see appendix B, table B-4). The school satisfaction impacts estimated for both parents and students, in the overall sample and for subgroups, were confirmed by the sensitivity tests in all but three cases. The parent satisfaction impact on parents of students with lower test score performance at baseline was not significant in the trimmed sample analysis, and the impact on the satisfaction of parents of students with higher test score performance at baseline was not significant when clustering on school attended. The trimmed analysis did produce a positive and significant impact for male students (appendix C, table C-5).

Figure 3-5. Parent and Student Reports of School Satisfaction, 2008-09



**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. For parent reports, valid *N* = 1,227; parent survey weights were used. For student reports, valid *N* = 1,001; student survey weights were used; the survey was given to students in grades 4-12. Means are regression adjusted using a consistent set of baseline covariates.

Student Self-Reports

As was true with the school safety and climate measures, students had a different view of their schools than did their parents. At least four years after random assignment, there were no significant differences between the treatment group and the control group in their likelihood of assigning their schools a grade of A or B (table 3-9; figure 3-5 for a visual display).[47] Student reports of school satisfaction were statistically similar between the treatment and control groups for all six subgroups examined. These results were confirmed by both sensitivity tests.

[47] Only students in grades 4-12 were administered surveys, so the satisfaction of students in early elementary grades is unknown.

Table 3-9. Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Student Reports of Satisfaction with Their School, 2008-09

Students Who Gave Their School a Grade of A or B	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		*p*-value of estimates
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**.69**	**.72**	**-.03**	**-.06**	**-.03**	**-.07**	**.43**
SINI 2003-05	.66	.66	.00	.00	.00	.00	1.00
Not SINI 2003-05	.71	.76	-.05	-.12	-.06	-.15	.29
Difference	-.05	-.10	.05	.11			.46
Lower performance	.71	.67	.04	.08	.05	.11	.54
Higher performance	.68	.73	-.06	-.13	-.07	-.16	.14
Difference	.04	-.06	.09	.20			.18
Male	.70	.73	-.03	-.07	-.04	-.09	.55
Female	.68	.71	-.02	-.05	-.03	-.06	.62
Difference	.01	.02	-.01	-.02			.91

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid N =1,001, including: SINI 2003-05 N = 478, Not SINI 2003-05 N = 523, Lower performance N = 319, Higher performance N = 682, Male N = 480, Female N = 521. Student survey weights were used. Survey given to students in grades 4-12.

3.6 Chapter Summary

This chapter presents the estimated impacts of the OSP at least four years after the random assignment of students to treatment or control groups. These longer run estimates indicate no clear evidence of a longer term effect on achievement for students overall or for the high priority group of students who applied from SINI 2003-05 schools. Although reading impacts were observed for the not SINI 2003-05, higher baseline performance, and female subgroups of students, those findings should be treated with caution because they may be false discoveries. A variety of factors, such as the length of scholarship use, that could potentially or partially explain the pattern of achievement results that we report here will be explored in a later, second volume to this final evaluation report.

The OSP had a significant positive impact on parent-reported high school graduation rates. Overall, 82 percent of those offered scholarships graduated compared to 70 percent of those who were not offered scholarships, a difference of 12 percentage points. The impact on graduation of actually using a scholarship to attend private school was 21 percentage points. Similar benefits extended to the high-priority SINI 2003-05 students, those who were higher performing when they entered the Program, and female participants. These results provide support for prior research suggesting that private schools

provide students with an educational climate that encourages school completion either through the faculty and school environment or by having similarly motivated and achieving peers (Evans and Schwab 1995; Grogger and Neal 2000; Neal 1997; Warren 2010).

Parents whose children were offered or used a scholarship were more satisfied and viewed their child's school as safer than parents whose children were not offered a scholarship. There were some exceptions: parents of two groups of students for whom there was no evidence of academic benefits, those from SINI 2003-05 schools or who were male, were not significantly more satisfied with their child's schooling if they were offered or used a scholarship. Across the board, students did not rate their schools differently regarding safety or satisfaction whether they did or did not receive an OSP scholarship.

4. Exploratory Analysis of OSP Intermediate Outcomes At Least Four Years After Random Assignment

Whatever effect the Opportunity Scholarship Program (OSP) has on key outcomes, researchers and policymakers have long been interested in understanding the *mechanisms* by which voucher programs may or may not benefit students (e.g., Howell and Peterson 2006, pp. 158-166; Wolf and Hoople 2006). There are a variety of theoretical hypotheses in the literature about how programs like the OSP might positively affect achievement, such as: (1) participating students are exposed to a group of peers who better facilitate learning (Benveniste 2003; Hoxby 2000; Nielsen and Wolf 2001), (2) school organization or instruction is different (Chubb and Moe 1990), (3) parents and students develop different expectations for their success (Akerlof and Kranton 2002; Bryk, Lee, and Holland 1993), (4) the school community surrounding students is more comprehensive and nurturing (Brandl 1998; Coleman and Hoffer 1987), and (5) parents become more involved in that school community (Coulson 1999). The conceptual basis for these hypotheses depends on two important linkages: (1) access to a voucher alters the educational experiences or behaviors mentioned above, and (2) those differences lead to better student outcomes. However, so far there has been little research that empirically tests these relationships (Hess and Loveless 2005).

The first section of this chapter presents methods used to examine what, if any, elements of students' educational environment (that is, intermediate outcomes) changed as a result of the OSP. Section 2 reviews findings from the two previous analyses of intermediate outcomes. These analyses were made two and three years after the offer of a scholarship to students in the treatment group. The third and final section presents intermediate outcomes for the current evaluation year, examining the impact of the offer of a scholarship on students' educational conditions and experiences at least four years after random assignment.

4.1 Selection and Construction of Intermediate Outcomes

This chapter examines how the first set of linkages in the hypothesized causal chain may be playing out for the OSP. The investigation explores the question, "Did the OSP change the daily educational life or experiences of participating students?" While part of this question was examined descriptively in chapter 2, the analysis here estimates the actual impact of the Program on a set of

variables that we call "intermediate outcomes" because they may be influenced by the Program, but they themselves are not an "end outcome" as identified in the OSP statute.[48]

A variety of educational conditions, attitudes, and behaviors might be affected by the OSP and, in turn, affect student achievement. In crafting the parent, student, and principal surveys for the evaluation, we included questions that provide measures of 24 factors that could plausibly be intermediate outcomes of the OSP and mediators of its impacts on student test scores. These measures were identified from the body of theory and prior research on the predictors of educational achievement and on differences between public and private schools (see appendix H for greater detail concerning the identification and construction of these variables). These 24 educationally important factors fall into four conceptual groups: Home Educational Supports, Student Motivation and Engagement, Instructional Characteristics, and School Environment. The impact of the Program—the offer of a scholarship—was estimated on each of the 24 indicators for the overall sample of students using the same analytic model used to estimate the impacts reported in chapter 3.[49]

Because this analysis of the possible intermediate outcomes of the offer of a scholarship involves multiple comparisons, statistical adjustments are made to reduce the threat of false discoveries (Benjamini and Hochberg 1995). Data on these adjustments are provided in appendix B.

4.2 Intermediate Outcomes Across Evaluation Years

The evaluation of the impact of the OSP is a longitudinal study with respect to both student outcomes as measured by test scores and intermediate outcomes obtained from survey data. With respect to intermediate outcomes, it collects and analyzes survey data from both treatment and control groups over multiple years of their potential participation in the scholarship program.

Two previous reports have described impacts on the 24 intermediate outcomes two and three years after students applied to the OSP and were randomly assigned by lottery to either the treatment or control group (Wolf et al. 2008, pp. 55-64; Wolf et al. 2009, pp. 53-66). One way to assess the results is to consider that if the offer of a scholarship has had a persistent or somewhat persistent impact on any of

[48] See Hatry (2001) for a discussion of the difference between "intermediate" and "end" outcomes.

[49] Intermediate outcomes with an approximately continuous distribution were estimated using the Ordinary Least Squares version of our analytic model. Intermediate outcomes with a binary distribution were estimated using the Logit or "probability estimation" version of our analytic model. Intermediate outcomes with ordinal distributions were estimated using an Ordered Logit version of our analytic model. In all three cases, the set of explanatory variables in the model is identical to the set used to estimate the main study impacts presented in chapter 3 (see appendix section A.3 for the list of covariates and section A.8 for details regarding the analytic strategy).

the intermediate outcomes examined in this chapter, there may be an increased possibility that these outcomes could mediate the relationship between the offer of a scholarship and student achievement. Likewise, outcomes which show varying or insignificant differences across years between the treatment and control groups may be less likely to be mediating this relationship.

Out of the 24 variables we identified as intermediate outcomes with potential for influencing achievement, the OSP had a statistically significant impact on fewer than half of them each year (table 4-1). After two years, the Program had a significant effect on 11 and after three years on 9 of these outcomes. After at least four years, the offer of a scholarship affected 8 of the 24 potential mediators.

For the most part, the pattern of Program effects on student educational conditions over time was not consistent. Over those years of potential Program participation, impacts on three outcomes were both statistically significant and the same sign (that is, in the same direction) in each year. Students with a scholarship experienced a lower likelihood that their school offered special programs for students with learning problems. Students in the treatment group also were less likely, in every year, to attend a school with special programs for English language learners. Finally, the students offered scholarships consistently experienced a school smaller in size than that attended by the control group.

Three other intermediate outcomes were similarly affected by the scholarship offer in two years but not all three. Those somewhat consistent Program effects were a reduced likelihood of attending a school that makes tutors available; an increased likelihood of attending a school with enrichment programs such as art, music, and drama; and a reduction in the proportion of the student body that was non-white. Eleven of the 24 possible mediators were only significantly affected by the OSP in one of the three years or were affected differently by the Program in different years.

Table 4-1. ITT Impacts on Intermediate Outcomes: Significant Impacts in Effect Sizes After Two, Three, and at Least Four Years

Mediators:	After Two Years	After Three Years	After at Least Four Years
Home Educational Supports			
Parental Involvement			
Parent Aspirations	.12*		
Out-of School Tutor Usage		-.14**†	
School Transit Time	.25**†		
Student Motivation and Engagement			
Student Aspirations			
Attendance			1.32*#
Tardiness			
Reading for Fun		-.16*	
Engagement in Extracurricular Activities			
Frequency of Homework (days)			
Instructional Characteristics			
Student-Teacher Ratio	-.29**†		
Teacher Attitude			
Ability Grouping			.28**†
Availability of Tutors	-.32**†	-.38**†	
In-School Tutor Usage	.13*†		
Programs for Learning Problems	-.66**†	-.36**†	-.52**†
Programs for English Language Learners	-.66**†	-.61**†	-.57**†
Programs for Advanced Learners		.27*†	-.53**†
Before/After School Programs			
Enrichment Programs	.19*†	.23**†	
School Environment			
Parent-School Communication			.24**†
School Size	-.43**†	-.29**†	-.36**†
Percent Non-White	-.39**†		-.14*†
Peer Classroom Behavior	.16*		
Total Number of Significant Effects	11	8	8

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

† = intermediate outcome result remained statistically significant after adjustments for multiple comparisons.

Odds ratio.

NOTES: Some of the intermediate outcomes in this table are binary variables, representing the presence (1) or absence (0) of the experience (out-of-school tutor usage; reads for fun; ability grouping in classes; availability of tutors in school; programs for students with learning problems, English language learners, or for advanced learners; the availability of before- or after-school programs). Other variables are based on Item Response Theory (IRT) scales taken from responses to survey questions. These include parental involvement (range = .75 to 7.52), teacher attitude (range = .47 to 10.32), and peer classroom behavior (range = 3.23 to 12.70). Parent aspirations (range = 11 to 19), student aspirations (range = 11 to 19), student-teacher ratio (range = .90 to 35.80), school size (range = 16 to 1,604), and percent non-white (range = .10 to 1) are continuous variables. School transit time (range = 1 to 6), attendance and tardiness (range = 0 to 3) are ordinal categorical variables. Engagement in extracurricular activities (range = 0 to 5), frequency of homework (range = 0 to 5), enrichment programs (range = 0 to 3), and parent-school communication (range = 1 to 4) are count variables. For greater detail about the construction and interpretation of these variables, see appendix H.

4.3 Impact of the OSP on Intermediate Outcomes At Least Four Years After Random Assignment

Over the longer run, as measured at least four years after students potentially entered the Program, the OSP may have had a significant impact on eight intermediate outcomes (table 4-1, far right column; tables 4-2 – 4-5):

- ***Less frequent attendance at school.*** Based on parent surveys, students in the treatment group were absent from school for more days than were students in the control group. For example, parents reported that 39 percent of the treatment group compared to 46 percent of the control group had perfect attendance during the school year (no absences). This result could be a false discovery, so it should be interpreted with caution.

- ***Less access to special programs for students who are at both ends of the academic distribution.*** Students offered a scholarship experienced a lower likelihood that their school provided special programs for students with learning problems (ES = -.52), special programs for children who were English language learners (ES = -.57), or programs for advanced learners (ES = -.53) compared to control group students.[50]

- ***More access to ability grouping.*** Students offered a scholarship experienced a higher likelihood that their school grouped students by academic ability (ES = .28), compared to control group students.

- ***More parent-school communication.*** Students offered a scholarship attended schools with more frequent parent-school communication as reported by the schools (ES = .24) than did students in the control group. Parent-school communication was measured by surveying school principals about the use of letters and reports to notify parents of student grades and behavior, as well as school newsletters.

- ***Smaller schools.*** Treatment group students also attended schools that were smaller than those attended by control group students by an average of 136 students (ES = -.36), as measured by total student enrollment.

- ***Fewer non-white students***. Although the overall impact sample students attended schools that were overwhelmingly non-white, averaging 94 percent non-white classmates, the schools attended by the treatment group were less non-white than were the schools attended by the control group by 2 percentage points (ES = -.14).

[50] As reported in chapter 2, 58 percent of schools attended by the treatment group and 63 percent of schools attended by the control group made tutors available at school; 32 percent of schools attended by the treatment group and 57 percent of schools attended by the control group offered special programs for non-English speakers; and 75 percent of schools attended by the treatment group and 90 percent of schools attended by the control group offered special programs for students with learning problems.

Table 4-2. ITT Impacts on Intermediate Outcomes as Potential Mediators: Home Educational Supports, 2008-09

Mediators:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parental Involvement	2.80	2.87	-.07	-.04	.54
Parent Aspirations	17.29	17.30	-.01	-.00	.94
Out-of-school Tutor Usage	.11	.08	.03	.11	.09
School Transit Time	2.71	2.68	.02	1.02#	.84
Under 10 minutes	*.22*	*.22*	*-.00*		
11-20 minutes	*.31*	*.31*	*-.00*		
21-30 minutes	*.20*	*.19*	*.00*		
31-45 minutes	*.14*	*.14*	*.00*		
46 minutes to an hour	*.10*	*.10*	*.00*		
More than one hour	*.04*	*.04*	*.00*		

Odds ratio.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Parental involvement = 1,228; Parent aspirations = 1,184; Out-of-school tutor usage = 1,180; School transit time = 1,235. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimate for the dichotomous variable "Out-of-school tutor usage" is reported as the marginal effect. Impact estimate for the ordered categorical variable "School transit time" was obtained by ordered logit.

Out-of-school tutor usage is a binary variable, representing the presence (1) or absence (0) of the experience. Parental involvement (range = .75 to 7.52) is based on an IRT scale taken from responses to the parent survey. Parent aspirations (range = 11 to 19) is a continuous variable. School transit time is an ordinal categorical variable (range = 1 to 6). For greater detail about the construction and interpretation of these variables, see appendix H.

Table 4-3. ITT Impacts on Intermediate Outcomes as Potential Mediators: Student Motivation and Engagement, 2008-09

Mediators:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Student Aspirations	16.72	16.82	-.10	-.05	.57
Absence from School	1.06	.78	.28*	1.32#	.03
No days absent	*.39*	*.46*	*-.07*		
1-2 days absent	*.39*	*.37*	*.03*		
3-4 days absent	*.16*	*.13*	*.03*		
5 or more days	*.06*	*.05*	*.01*		
Tardiness	.46	.54	-.08	.93#	.59
No days tardy	*.65*	*.63*	*.02*		
1-2 days tardy	*.23*	*.24*	*-.01*		
3-4 days tardy	*.07*	*.08*	*-.00*		
5 or more days	*.05*	*.05*	*-.00*		
Reads for Fun	.38	.43	-.05	-.11	.17
Engagement in Extracurriculars	2.33	2.28	.05	.04	.61
Frequency of Homework (days)	3.80	3.75	.05	.03	.67

* Statistically significant at the 95 percent confidence level.

Odds ratio.

† = intermediate outcome result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Student aspirations = 1,012; Attendance = 1,192; Tardiness = 1,182; Reads for fun = 1,070; Engagement in extracurricular activities = 1,016; Frequency of homework = 1,046. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for the ordered categorical variables "Attendance" and "Tardiness" were obtained by ordered logit.

Reading for fun is a binary variable, representing the presence (1) or absence (0) of the experience. Student aspirations (range = 11 to 19) is a continuous variable. Attendance (range = 0 to 3) and tardiness (range = 0 to 3) are ordinal categorical variables. Engagement in extracurricular activities (range = 0 to 5) and frequency of homework (range = 0 to 5) are count variables. For greater detail about the construction and interpretation of these variables, see appendix H.

Table 4-4. ITT Impacts on Intermediate Outcomes as Potential Mediators: Instructional Characteristics, 2008-09

Mediators:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Student-Teacher Ratio	11.87	11.81	.06	.01	.85
Teacher Attitude	2.81	2.83	-.02	-.01	.91
Ability Grouping	.80	.67	.13**†	.28	.00
Availability of Tutors	.58	.63	-.06	-.11	.20
In-school Tutor Usage	.28	.25	.03	.08	.23
Programs for Learning Problems	.76	.91	-.15**†	-.52	.00
Programs for English Language Learners	.29	.58	-.28**†	-.57	.00
Programs for Advanced Learners	.42	.67	-.25**†	-.53	.00
Before-/After-School Programs	.91	.88	.03	.08	.11
Enrichment Programs	2.67	2.62	.05	.08	.30

**Statistically significant at the 99 percent confidence level.

† = intermediate outcome result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Student/teacher ratio = 904; Teacher attitude = 1,060; Ability grouping = 893; Availability of tutors = 946; In-school tutor usage = 1,197; Programs for learning problems = 951; Programs for English language learners = 951; Programs for advanced learners = 953; Before-/after-school programs = 953; Enrichment programs = 953. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for the dichotomous variables "Ability grouping," "Availability of tutors," "In-school tutor usage," and "Before-/after-school programs," are reported as marginal effects. Regression models for Before-/after-school programs omit the grade level 4 dummy variable because it predicted success perfectly.

Some of the intermediate outcomes in this table are binary variables, representing the presence (1) or absence (0) of the experience (ability grouping in classes; availability and use of tutors in school; programs for students with learning problems, English language learners, or for advanced learners; the availability of before- or after-school programs). Teacher attitude (range = .47 to 10.32) is a variable based on an IRT scale taken from responses to survey questions. Student-teacher ratio (range = .90 to 35.80) is a continuous variable. Enrichment programs (range = 0 to 3) is a count variable. For greater detail about the construction and interpretation of these variables, see appendix H.

Table 4-5. ITT Impacts on Intermediate Outcomes as Potential Mediators: School Environment, 2008-09

Mediators:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parent-School Communication	3.21	3.00	.21**†	.24	.00
School Size	408.23	544.62	-136.39**†	-.36	.00
Percent Non-White	.93	.95	-.02*†	-.14	.04
Peer Classroom Behavior	8.16	8.05	.11	.05	.51

**Statistically significant at the 99 percent confidence level.

† = intermediate outcome result remained statistically significant after adjustments for multiple comparisons.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for School communication policies = 936; School size = 1011; Percent non-white = 994; Peer classroom behavior = 1,061. Separate weights were used for items from parent surveys, student surveys, and principal surveys.

Peer classroom behavior (range = 3.23 to 12.70) is based on an IRT scale taken from responses to survey questions. School size (range = 16 to 1,604) and percent non-white (range = .10 to 1.00) are continuous variables. Parent-school communication (range = 1 to 4) is a count variable. For greater detail about the construction and interpretation of these variables, see appendix H.

4.4 Chapter Summary

This chapter presents the results of an experimental analysis of the longer run impacts of the OSP on specific features of students' educational experience and environment, viewed as "intermediate outcomes" of the OSP. The analysis determined that, as a result of being offered an Opportunity Scholarship, students attended smaller schools with fewer non-white students and more frequent parent-school communications that were more likely to use ability grouping but less likely to have programs for students with a variety of special conditions including giftedness, learning problems, and difficulties learning the English language. The parents of students in the treatment group tended to report more student absences than did the parents of students in the control group, though that difference may be a false discovery, and therefore should be interpreted with caution.

Of these eight factors that were affected by the Program after at least four years, three of the results were also found in all prior years of conducting this type of analysis. The lower likelihood that treatment group students attend a school with special programs for students with learning difficulties or for students who are English language learners has been consistent across the last three years of impact estimates, as has the smaller average size of the schools attended by the treatment group.

Perhaps it is not surprising that we find only a few candidates for possible factors that a program like the OSP can affect consistently. Most of our measures are based on principal, parent, and student surveys, and the group of individuals who complete the surveys each year varies somewhat. Differences in the schools that impact sample students attended may have been largest between year three and this final year of the analysis because 296 students had graded-out of the study; many of the remainder had moved from elementary to secondary schools; and 35 of them had continued to attend formerly Catholic parochial schools that had converted to public charter schools. These substantial shifts in the demographic characteristics of the students in the study and the nature of the schools that no longer participate in the Program could explain some of the inconsistencies in the intermediate outcomes observed, especially in the last two years of the evaluation. Alternatively, given that no overall achievement impacts of the OSP were observed in this final year of the impact evaluation, we also might not expect the Program to demonstrate clear impacts on many of these intermediate measures if they do, in fact, influence achievement.

5. Exposure, Awareness, and Response of DCPS and Private Schools to the OSP

The OSP statute mandates that the evaluation examine the effects of the Program not only on students but also on schools in the District of Columbia.[51] School choice theory hypothesizes that a thriving private school scholarship program provides competition to the public schools and could generate improvements to the public school system, the private school system, or both (e.g., see Chubb and Moe 1990; Henig 1994). Such systemic changes could take place if significant percentages of students in the public school system, or in specific schools, apply for, receive, and use scholarships to transfer to private schools. Systemic changes also could occur in the private sector, if private schools adjust program operations to better attract or retain scholarship students.

In practice, two features of choice programs are necessary if any system-wide change in public schools is likely. First, the threat of losing students to the DC OSP must be significant (e.g., Armour and Peiser 1998). The threat need not be that students actually take the scholarship (at least initially), but that the program is available to a large share of the student population (e.g., Greene and Foster 2002; Hoxby 2003). Based on demographic statistics provided to the research team by DCPS, we estimate that at least 67 percent of the students in District neighborhood public and public charter schools are eligible for the OSP.[52] The second feature is that the loss of students to the scholarship program also entails a loss of funding (Hoxby 2003). However, the law that established the OSP ensured that DCPS would gain, rather than lose, funds, and district officials were not given information to determine how many students left individual public schools as a result of the Program.[53]

Irrespective of Program features, establishing a strong causal link between student achievement outcomes and the implementation of the OSP is not possible. Many changes were underway in the DC public and private school systems between 2004 and 2009, the period of evaluation data collection and the main operations of the Program. New academic standards and assessments were

[51] P.L. 108-199, Section 309 (a) (4).

[52] This estimate was generated by dividing the total number of students confirmed as participating in the federal lunch program (44,740) by the total number of students in the DCPS database (66,868). The federal lunch program has an income ceiling that matches that of the OSP (185 percent of the poverty line), although participation in the lunch program tends to decline as students age, probably because of the desire to avoid a poverty stigma. Thus, the actual percentage of DC public school students eligible for the OSP could be higher than 67 percent.

[53] The appropriations law that established the OSP and that initially funded the Program provided approximately $13 million for the OSP, approximately $13 million for DC charter schools, and approximately $13 million for the traditional public schools in DCPS, with an additional $1 million available for OSP program administration and evaluation. The current FY 2010 version of this statute provides $13 million for the OSP, $20 million for DC public charter schools, and $42 million to DCPS. Because of the confidentiality provisions in the law, neither WSF nor IES could reveal information about which or how many students left individual DC schools.

introduced in 2005-06 (the DC Comprehensive Assessment System, known as the DC-CAS); the public schools received a new Schools Chancellor in 2007; there was significant turnover in principals (Turque 2009); the charter school sector expanded (Haynes and Labbe 2007); and the financial circumstances of some DC private schools (particularly those run by the Archdiocese) led to modifications, including closures and conversions to charter schools (Labbe 2007). Any of these changes to the education landscape occurring simultaneously with the OSP could potentially alter student outcomes.

Because of these factors, our approach to assessing the effects on DC schools is to describe how "exposed" public and private schools were to the OSP and the ways in which they responded. In particular, the analysis examines (1) the extent to which DC public schools faced potential or actual student departures due to the OSP and the reliance of private schools on OSP enrollments, (2) the level of awareness of the OSP among DCPS and private school principals, and (3) whether principals reported making any changes to their school's operations in order to retain (public school) or attract and retain (private school) students. The analysis draws on surveys of all public and private school principals in DC and on data from WSF about the enrollments of all students with OSP scholarships between 2004 and 2009, not just the evaluation's impact sample. These data allow for a more complete picture of the OSP and its potential for affecting the public and private schools. Public schools are examined in this section and private schools in the following section.

DC Public Schools

Exposure to the OSP

As a first step, it is important to understand how exposed public schools were to the Program, on average and across individual schools. This can be measured by student scholarship application and use rates at public schools. The "rate" is the cumulative number of OSP students who applied as well as the number who used a scholarship to transfer out of a public school as a percentage of school enrollment in a given year.[54] After the first year, students who transferred out joined continuing participants in the Program.

[54] The analysis was done using data from 2004-05 to 2008-09 and includes all five cohorts of scholarship users, not only those in cohorts 1 and 2 who are members of the impact sample. The figures do not include rising kindergarten students or cohort 1 students who were in a private school when they were awarded a scholarship. The exposure measures are calculated as the total number of participants in a school across all five cohorts combined, divided by the school enrollment in the 2007-08 school year. This calculation was done for scholarship applicants and scholarship users. The 2007-08 enrollment figures (for all schools with a school ID) were used as the denominator because they are the most recent available from the CCD. The total number of participants from all cohorts was used as the numerator because exposure is cumulative and builds over time.

In terms of public school exposure to the OSP (table 5-1), we found:

- Nearly 6 percent of students in DC public schools applied to the OSP between 2004 and 2009, as a percentage of enrollment in 2007-08; half of them (3 percent) obtained a scholarship and transferred out to attend a participating private school;
- Schools that were designated SINI between 2003 and 2005 were more exposed to the OSP than were schools not designated SINI in those years; both the application rates (roughly 5.9 percent versus 5.4 percent) and scholarship use rates (3.4 percent versus 2.9 percent) were higher among SINI 2003-05 than not SINI 2003-05 schools;[55] and
- Charter schools were less affected by Program application (2 percent) and scholarship use (1 percent) than were traditional public schools (16 percent and 9 percent, respectively).[56]

Table 5-1. Average Percent of DC Students Who Applied or Used a Scholarship Cumulatively by 2008-09, by SINI 2003-05 Status of the Schools They Were Attending at Application and by Type of Public School

OSP Students as Percentage of Student Body, by Type of School	Applied for a Scholarship	Used a Scholarship to Transfer Out
All DC Public Schools (*N*=75,409)	5.7%	3.2%
SINI 2003-05 (*N*=37,540)	5.9%	3.4%
Not SINI 2003-05 (*N*=37,869)	5.4%	2.9%
Traditional Public (*N*=55,382)	16.4%	9.2%
Charter (*N*=20,027)	1.8%	1.0%

NOTES: Numerator is all eligible applicants in cohorts 1 through 5. Denominator is all students enrolled in the district (traditional public and charter schools), 2007-08. Data are student-level averages. Includes all public schools, not just schools attended by students from the impact sample. *N* represents the total number of students in the category.

SOURCES: Enrollment from the CCD, 2007-08. Applied and used from Program applications and WSF payment files, cohorts 1 through 5.

The level of exposure to the OSP varied across DC public schools.[57] At one end of the spectrum were schools that saw no students apply to the Program (3 percent of schools) or use a scholarship to transfer out (15 percent of schools). At the other end of the distribution, the maximum rate of application was 31.6 percent, and for students ever using a scholarship it was 21.4 percent. In terms of exposure to the OSP, the top 25 percent of schools had an average application rate of 15.5 percent and usage/transfer rate of 9.4 percent.

[55] The difference between SINI 2003-05 and not SINI 2003-05 is statistically significant for applied, $\chi^2(1, N = 75{,}409) = 10.44$, $p < .01$ and for used a scholarship, $\chi^2(1, N = 75{,}409) = 17.91$, $p < .01$.

[56] The difference between traditional public and charter schools is statistically significant for applied, $\chi^2(1, N = 75{,}409) = 5806.87$, $p < .01$,and for used a scholarship, $\chi^2(1, N = 75{,}409) = 3218.24$, $p < .01$.

[57] Application exposure had mean = 7.7, standard deviation = 5.8, skewness = 1.2, and kurtosis = 1.8. Scholarship usage exposure had mean = 4.3, standard deviation = 3.9, skewness = 1.3, and kurtosis = 2.4. Shapiro-Wilk tests for normality indicate that neither exposure measure is normally distributed, $p < .01$.

To put these figures in context, the DC Office of the State Superintendent of Education (OSSE) reported the average monthly student mobility rate in the district was about 2 percent from the fall of 2007 to the spring of 2008.[58] Taken across a 10-month school year, that monthly rate translates into an annual average mobility rate of 20 percent. Thus, the cumulative exposure of DCPS across five cohorts of students using OSP scholarships represents less than one-fifth of the average annual mobility of students in the district. Given these figures, OSP-related transfers to private schools may not have been distinguishable from the larger share of other student departures.

Awareness of and Response to the OSP

As noted above, theory suggests that principals might make changes to their schools if they were threatened by a loss of OSP students. To examine this hypothesis, through surveys we collected information on how aware public school principals were of the Program and what, if any, modifications they reported making in response. In addition, we asked public school principals to provide their own estimates of student departures for the OSP, and compared them to our administrative data on actual Program-related losses at schools.

Among the public schools, we found:

- Between 2006 and 2009, approximately 70 percent of principals in DCPS had heard about the OSP through a newspaper, talking to other principals, or through other means (table 5-2). In the earlier years of the Program, higher proportions of charter school principals (85 percent in 2006 and 92 percent in 2007) than traditional public school principals (69 and 63 percent in each year, respectively) were aware of the Program; by 2008 the levels of awareness between the two types of public schools was comparable.[59]

- Between 2006 and 2009, between two-thirds and three-quarters of principals reported that they did not think they had lost any students to the OSP (table 5-3). Fewer than 4 percent of principals each year estimated OSP-related transfers of more than 10 students.

- Some principals who reported having no student losses to the OSP did, in fact, have students leave with scholarships. In 2006, nearly two-thirds of the public school principals (63 percent) who reported having no OSP departures did actually lose some (on average, 4) students to the Program (tables 5-3 and 5-4). The rate of the public school principals who reported having no OSP departures but did actually lose some

[58] "A student is defined as 'mobile' if the student attended a different school or was not enrolled in the snapshot from the prior month." http://www.osse.dc.gov/seo/lib/seo/dc_student_mobility_report2008_06_10.pdf

[59] The difference between traditional public and charter schools is not statistically significant for 2006, $\chi^2(1, N = 134) = 3.47$, $p = .06$; for 2007 the difference is statistically significant, $\chi^2(1, N = 123) = 8.31$, $p < .01$; for 2008 the difference is not statistically significant, $\chi^2(1, N = 132) = 0.04$, $p > .05$; and for 2009 the difference is not statistically significant, $\chi^2(1, N = 168) = 0.06$, $p > .05$.

students declined by nearly half and remained around one-third of those surveyed for the next three years (table 5-3).[60] The average number of students leaving for the Program in these cases remained fairly steady (table 5-4).

- Twenty-eight percent of public school principals reported making any changes to their operations in order to keep students who might be interested in the OSP or private schools in general (figure 5-1). The modification principals most commonly reported was encouraging greater parental involvement in school activities (86 percent of those making changes; 24 percent of all principals), while the change least commonly reported was altering class size (29 percent of those making changes, 8 percent overall).

Table 5-2. Public School Principal-Reported Awareness of the OSP, Over Time

Aware of OSP	2006	2007	2008	2009
Overall	73%	69%	71%	70%
Traditional Public Schools	69%	63%	71%	69%
Charter Schools	85%	92%	73%	71%
Total Respondents	(134)	(123)	(132)	(168)

NOTES: Responses are unweighted. Respondents were able to select multiple responses. The survey question is "How did you hear about the DC Opportunity Scholarship Program, the federally funded scholarship program for students from low-income DC families administered by Washington Scholarship Fund (WSF)?" Principals who responded "I have not heard of the program" were coded as being unaware of the Program. All other responses indicated awareness of the Program.

SOURCE: Impact Evaluation Public School Surveys, 2005-06 to 2008-09. Includes all public schools, not just schools attended by students from the impact sample.

Table 5-3. Overall: Principal-Reported Estimates of Student Departure for the OSP, Over Time

Student Loss (number)	2006	2007	2008	2009
None	67%	76%	74%	76%
1 to 10	31%	21%	22%	21%
11 or More	2%	3%	4%	3%
% of principals who reported "None," but actually did have students leave	63%	32%	36%	29%
Total Respondents	(154)	(123)	(133)	(168)

NOTES: Responses are unweighted. Respondents were asked to select a single response. The survey question is "How many students do you think left your school as a result of the DC Opportunity Scholarship Program for the given school year?" Principals were asked to select one of the following responses: None, 1-6, 6-10, 11-20, 21-30, More than 30, or Don't know. For principals with losses, the numerator is the number of principals reporting no departures, but actually having departures, and the denominator is the number of principals reporting no departures in that year.

SOURCE: Impact Evaluation Public School Surveys, 2005-06 to 2008-09. Includes all public schools, not just schools attended by students from the impact sample.

[60] The difference between 2006 and 2009 is statistically significant, $\chi^2(1, N = 231) = 27.07$, $p < .01$.

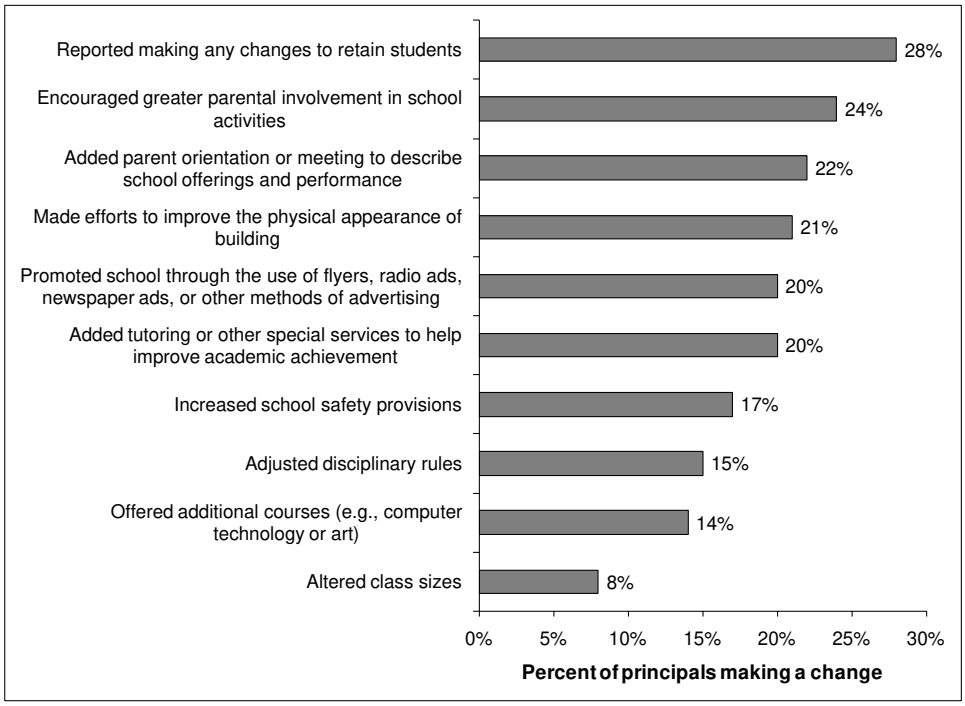
Table 5-4. Departures of OSP Scholarship Users from DC Public Schools, Over Time

Measure	Initially Used Scholarship			
	2006	2007	2008	2009
Average number of students leaving overall	3	1	1	1
Average number of students leaving among principals who reported *no* student departures	4	3	4	3

NOTES: The average number of students departing is based on the cohort corresponding to the survey year (2006 = cohort 2; 2007 = cohort 3; 2008 = cohort 4; 2009 = cohort 5). Initially used is the exposure measure because of the time dependence of each survey administration.

SOURCES: Estimates of student losses from the Impact Evaluation Public School Surveys, 2008-09. Number of students who ever used a scholarship from Program applications and WSF payment files, cohorts 2 through 5. Includes all public schools, not just schools attended by students from the impact sample.

Figure 5-1. Overall: Principal Reports of Specific Changes to Retain Students, 2008-09



NOTES: Responses are unweighted. Respondents were able to select multiple responses. The survey question is "In the past five years or since you became principal, have you made any changes specifically to encourage students interested in private schools (or the Opportunity Scholarship Program) to remain enrolled in your school?" If the principal answered yes, then the principal was asked to indicate which (of the following) changes were made. For all percentages, the numerator is the number of principals who answered "yes" to making a change and the denominator is the total number of survey respondents (*N*=168).

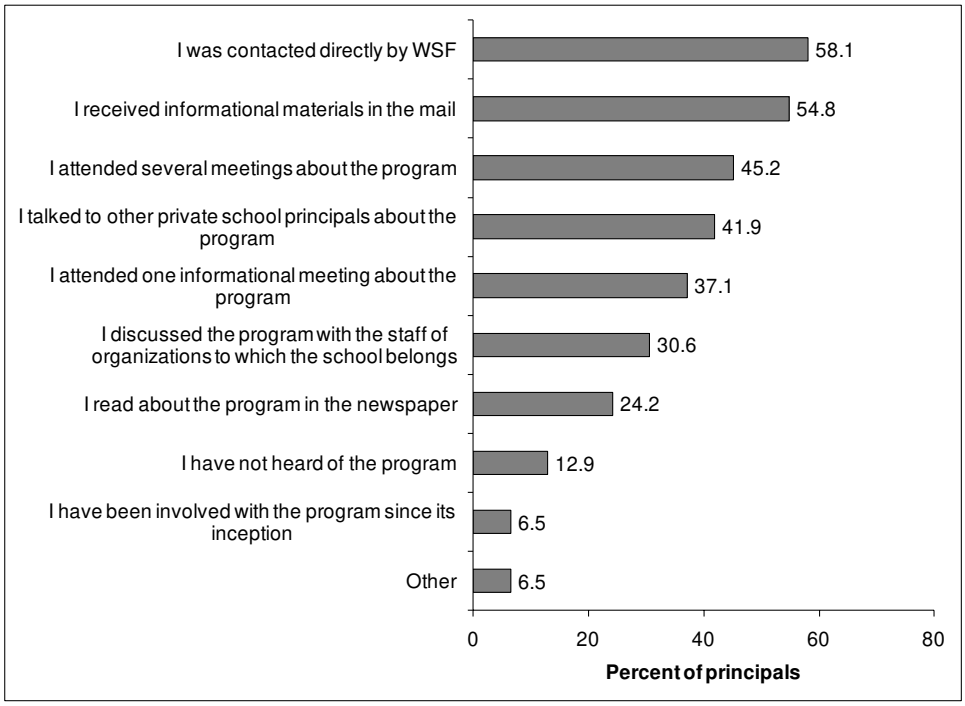
SOURCE: Impact Evaluation Public School Surveys, 2008-09.

Private Schools

Awareness of the OSP

The principal survey administered by the evaluation team in spring 2009 asked all private school principals in DC if they were aware of the OSP and how they learned of its existence. Nearly 60 percent reported that they had been contacted directly by WSF. Principals also received informational materials in the mail, had attended meetings, and had discussed the Program with other private school principals. Thirteen percent of the private school principals that responded to the survey indicated that they had not heard of the Program. (figure 5-2).

Figure 5-2. Private School Principal Awareness of the OSP, 2008-09



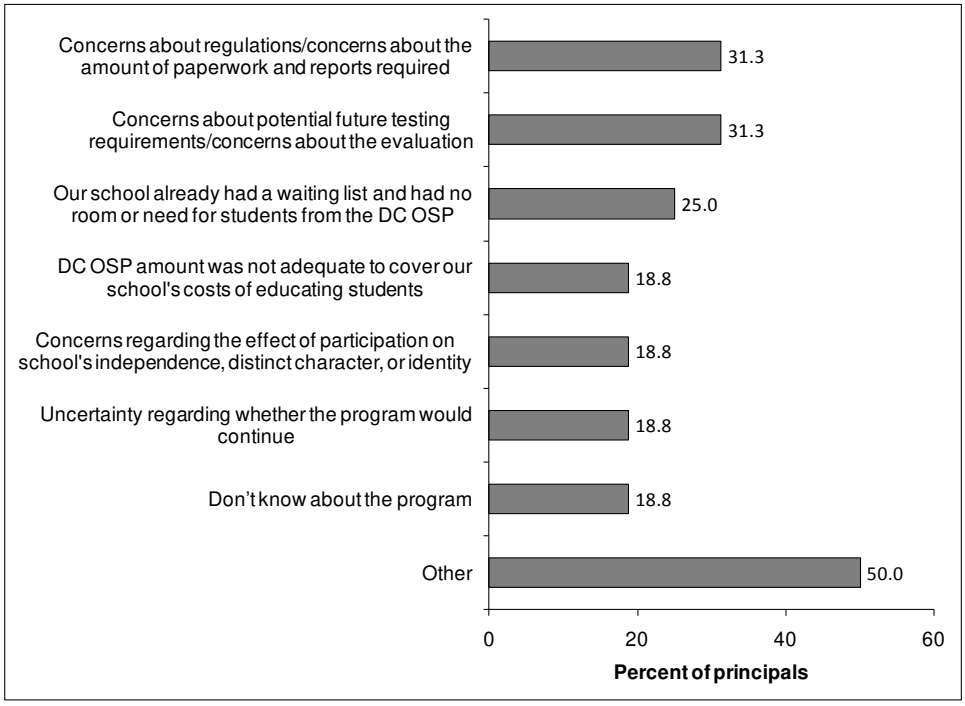
NOTES: Responses are unweighted. Respondents (valid $N = 62$) were able to select multiple responses. Categories with responses from fewer than three principals are collapsed into the "Other reasons" category for confidentiality purposes.

SOURCE: Impact Evaluation Principal Surveys, 2008-09.

Principals of schools that never participated in the OSP cited a variety of reasons for not participating (figure 5-3). Among them were concerns about the amount of regulations and paperwork required (31 percent) and worries about testing requirements and the evaluation (31 percent). Twenty-five percent reported that they did not have room to accommodate OSP students. Additionally, principals at non-participating schools reported concerns about the size of the OSP scholarships (19 percent), how the

Program might alter the distinct identity of their school (19 percent), and uncertainty about the longevity of the Program (19 percent).

Figure 5-3. Reasons Private School Principals Gave for Not Participating, 2008-09



NOTES: Responses are unweighted. Respondents (valid *N* = 16) were able to select multiple responses. Categories with responses from fewer than three principals are collapsed into the "Other reasons" category for confidentiality purposes.

SOURCE: Impact Evaluation Principal Surveys, 2008-09.

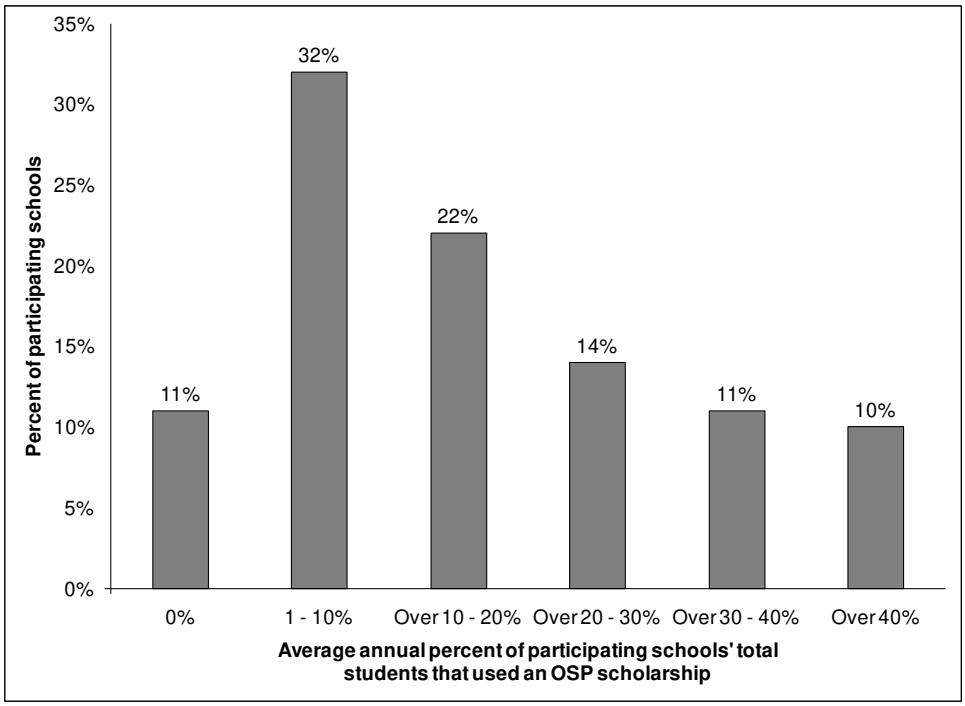
Participating Private Schools' Exposure to the OSP

The degree to which participating private schools were exposed to the OSP can be measured by scholarship enrollment rates at participating private schools. In this case, the "rate" is the average number of OSP students who used a scholarship to enroll in a private school each year a school participated expressed as a percentage of each school's total annual enrollment. Of the schools that have fully or partially participated in the OSP since 2004:

- On average, OSP students made up 16 percent of participating private schools' student populations;
- Thirty-five percent of participating private schools had OSP populations that were greater than 20 percent of their enrollment (figure 5-4);

- Twenty-two percent of schools had OSP populations that made up over 10 to 20 percent of their enrollment (figure 5-4);
- Thirty-two percent of schools had OSP populations that were 10 percent or less of their total enrollment (figure 5-4); and
- Eleven percent of participating schools did not enroll any OSP students (figure 5-4).[61]

Figure 5-4. Percent of Participating Private Schools by OSP Enrollment Rate



NOTES: Valid school $N = 63$.

SOURCES: Average school size is the average of the school size data from the National Center for Education Statistics' Private School Survey, 2005-06 and 2007-08. Students were identified as scholarship users based upon information from WSF's payment files.

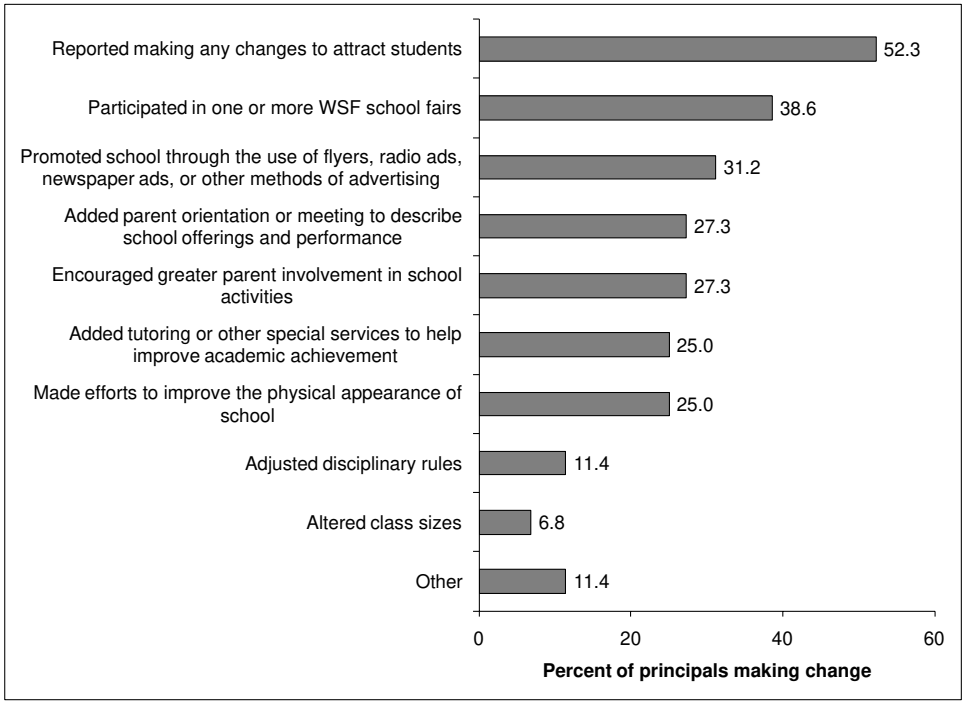
Participating Private Schools' Response to the OSP

The survey administered by the evaluation team also asked private school principals if they had made any changes specifically to encourage OSP students to enroll. Fifty-two percent of participating private school principals reported that they had made specific changes to encourage OSP enrollment (figure 5-5). Principals indicated that they had attended WSF school fairs (39 percent); promoted their schools through flyers, radio ads, newspaper ads, or other methods of advertising (31 percent); added

[61] Private schools were deemed by the WSF, the Program operator, as participating if they agreed to take OSP voucher students even if no OSP students were admitted.

parent orientations or meetings (27 percent); or encouraged greater parent involvement in school activities (27 percent).

Figure 5-5. Participating Private School Responses to the OSP, 2008-09



NOTES: Responses are unweighted. Respondents (valid $N = 44$) were able to select multiple responses. Categories with responses from fewer than three principals are collapsed into the "Other reasons" category for confidentiality purposes.

SOURCE: Impact Evaluation Principal Surveys, 2008-09.

References

Abrevaya, Jason. "The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes," *Empirical Economics* 2001, 26(1): 247-257.

Akerlof, George. A., and Robert E. Kranton. "Identity and Schooling: Some Lessons for the Economics of Education." *Journal of Economic Literature* 2002, 40: 1167-1201.

Angrist, Joshua, Guido Imbens, and Donald B. Rubin. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 1996, 91: 444-455.

Armour, David J., and Brett M. Peiser. "Interdistrict Choice in Massachusetts." In Paul E. Peterson and Bryan C. Hassle, eds., *Learning From School Choice* (pp. 157-186). Washington, DC: The Brookings Institution, 1998.

Arum, Richard. "Do Private Schools Force Public Schools to Compete?" *American Sociological Review* 1996, 661(1): 29-46.

Ballou, Dale, and Michael Podgursky. "Teacher Recruitment and Retention in Public and Private Schools." *Journal of Policy Analysis and Management* 1998, 17(3): 393-417.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 2003, 98: 299-323.

Basset, Gilbert W., Jr., and Hsiu-Lang Chen. "Portfolio Style: Return-based Attribution Using Quantile Regression," *Empirical Economics* 2001, 26(1): 293-305.

Bauch, Patricia A., and Ellen B. Goldring. "Parent Involvement and School Responsiveness: Facilitating the Home-School Connection in Schools of Choice." *Educational Evaluation and Policy Analysis* 1995, 17: 1-21.

Benjamini, Yoav, and Yosef Hochberg. "Controlling for the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B (Methodological)* 1995, 57(1): 289-300.

Benveniste, Luis. *All Else Equal: Are Public and Private Schools Different?* New York: Routledge Falmer, 2003.

Bitler, Marianne, Jonah Gelbach and Hilary Hoynes. What Mean Imputs Miss: Distributional Effects of Welfare Reform Experiments, *American Economic Review* 2006, 96(4): 988-1012.

Bloom, Howard S. "Accounting for No-Shows in Experimental Evaluation Designs." *Evaluation Review* 1984, 8(2): 225-246.

Booker, Kevin, Tim R. Sass, Brian Gill, and Ron Zimmer. *Going Beyond Test Scores: Evaluating Charter School Impact on Educational Attainment in Chicago and Florida*. Working paper WR-610-BMG. Santa Monica, CA: RAND Education, August 2008.

Boruch, Robert, Dorothy de Moya, and Brooke Snyder. "The Importance of Randomized Field Trials in Education and Related Areas." *Evidence Matters: Randomized Trials in Education Research*, Frederick Mosteller and Robert Boruch, editors. Washington, DC: The Brookings Institution Press, 2002.

Brandl, John E. *Money and Good Intentions Are Not Enough*. Washington, DC: The Brookings Institution Press, 1998.

Bryk, Anthony S., Valerie E. Lee, and Peter B. Holland. *Catholic Schools and the Common Good*. Cambridge, MA: Harvard University Press, 1993.

Card, David, and Alan Krueger. "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *Journal of Political Economy* 1992, 100(1): 1-40.

Chubb, John E., and Terry M. Moe. *Politics, Markets, and America's Schools*. Washington, DC: The Brookings Institution Press, 1990.

Cohen, Peter A., James A. Kulik, and Chen-Lin C. Kulik. "Educational Outcomes of Tutoring: A Meta-Analysis of Findings." *American Educational Research Journal* 1982, 19(2): 237-248.

Coleman, James S., and Thomas Hoffer. *Public and Private High Schools: The Impact of Communities*. New York: Basic, 1987.

Coleman, James S., and others. *Equality of Educational Opportunity*. U.S. Department of Health, Education, and Welfare, Office of Education. Washington, DC: U.S. Government Printing Office, 1966.

Coleman, James S. *Equality and Achievement in Education*. Boulder, CO: Westview Press, 1990.

Cook, Thomas D., and Donald T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Co., 1979.

Coulson, Andrew. "Comparing Public, Private, and Market Schools: The International Evidence." *Journal of School Choice* 2009, 3(1): 31-54.

Coulson, Andrew. *Market Education: The Unknown History*. New Brunswick, NJ: Transaction Publishers, 1999.

District of Columbia Public Schools. "District of Columbia Public Schools Progress Report: 2007-2008 School Year." Washington, DC: Author.

Dolton, Peter, Oscar D. Marcenaro, and Lucia Navarro. "The Effective Use of Student Time: A Stochastic Frontier Production Function Case Study." *Economics of Education Review* 2003, 22(6): 547-560.

Eide, Eric, and Mark H. Showalter. "The Effect of School Quality on Student Performance: A Quantile Regression Approach," *Economic Letters* 1998, 58(3): 345-350.

Evans, William N., and Robert M. Schwab. "Finishing High School and Starting College: Do Cathloic Schools Make a Difference?" *The Quarterly Journal of Economics* 1995, 110: 941-974.

Fan, Xitao and Michael Chen. "Parental Involvement and Students' Academic Achievement: A Meta-Analysis." *Educational Psychology Review 2001*, 13(1): 1-22.

Fisher, Ronald A. *The Design of Experiments*. Edinburgh: Oliver and Boyd, 1935.

Gill, Brian, P. Mike Timpane, Karen E. Ross, Dominic J. Brewer, and Kevin Booker. *Rhetoric Versus Reality: What We Know and What We Need to Know About Vouchers and Charter Schools*. Santa Monica, CA: RAND Education, 2007.

Gilligan, Carol. *In a Different Voice: Psychological Theory and Women's Development*. Cambridge, MA: Harvard University Press, 1993.

Greene, Jay P., and Greg Forster. "Rising to the Challenge: The Effect of School Choice on Public Schools in Milwaukee and San Antonio." *Civil Bulletin No. 27*. New York City: Manhattan Institute for Policy Research. October 2002, 5-6.

Greene, Jay P. "Vouchers in Charlotte." *Education Matters* 2001, 1(2): 55-60.

Grogger, Jeffrey, and Derek Neal. "Further Evidence on the Effects of Catholic Secondary Schooling," *Brookings-Wharton Papers on Urban Affairs* 2000: 151-193.

Gruber, Kerry J., Susan D. Wiley, Stephen P. Broughman, Gregory A. Strizek, and Marisa Burian-Fitzgerald. *Schools and Staffing Survey, 1999-2000: Overview of the Data for Public, Private, Public Charter, and Bureau of Indian Affairs Elementary and Secondary Schools.* Washington, DC: U.S. Department of Education, 2002.

Hambleton, Ronald K., Hariharan Swaminathan, and Jane H. Rogers. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications, 1991.

Hanushek, Eric. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review* 1971, 61(2): 280-288.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. "New Evidence About Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement." NBER Working Paper No. 8471; January 2002. Available online at [http://www.nber.org/papers/w8741].

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. "Disruption Versus Tiebout Improvement: The Costs and Benefits of Switching Schools." *Journal of Public Economics* 2004, 88:1721-1746.

Harris, Judith Rich. *The Nurture Assumption: Why Children Turn Out The Way They Do*. New York: Free Press, 1998.

Hatry, Harry P. "What Types of Performance Information Should Be Tracked?" *Quicker, Better, Cheaper? Managing Performance in American Government*, Dall W. Forsythe, editor. Albany, NY: The Rockefeller Institute Press, 2001.

Haynes, V. Dion, and Theola Labbe. "A Boom for D.C. Charter Schools." *The Washington Post*, April 25, 2007.

Heckman, James J. "Identification of Causal Effects Using Instrumental Variables: Comment." *Journal of the American Statistical Association* 1996, 91: 459-462.

Henderson, Anne T., and Nancy Berla. *A New Generation of Evidence: The Family is Critical to Student Achievement*. Washington, DC: Center for Law and Education, 1994.

Henig, Jeffrey. *Rethinking School Choice: Limits of the Market Metaphor*. Princeton, MA: Princeton University Press, 1994.

Hess, Frederick M., and Tom Loveless. "How School Choice Affects Student Achievement." *Getting Choice Right*, Julian R. Betts and Tom Loveless, editors. Washington, DC: The Brookings Institution Press, 2005.

Hoffer, Thomas, Andrew M. Greeley, and James S. Coleman. "Achievement Growth in Public and Catholic Schools." *Sociology of Education* 1985, 58(2): 74-97.

Howell, William G., and Paul E. Peterson, with Patrick J. Wolf and David E. Campbell. *The Education Gap: Vouchers and Urban Schools.* Revised Edition, Washington, DC: The Brookings Institution Press, 2006.

Howell, William G., and Paul E. Peterson. "Uses of Theory in Randomized Field Trials: Lessons from School Voucher Research on Disaggregation, Missing Data, and the Generalization of Findings." *American Behavioral Scientist* 2004, 47(5): 634-657.

Howell, William G., Patrick J. Wolf, David E. Campbell, and Paul E. Peterson. "School Vouchers and Academic Performance: Results from Three Randomized Field Trials." *Journal of Policy Analysis and Management* 2002, 21(2): 191-217.

Hoxby, Caroline M. *Peer Effects in the Classroom: Learning from Gender and Race Variation*. National Bureau of Economic Research Working Paper 7867, Cambridge, MA, August 2000.

Hoxby, Caroline M. "School Choice and School Competition." *Swedish Economic Policy Review* 2003, 10: 11-67.

Johnson, Michael D., and Claes Fornell. "A Framework for Comparing Customer Satisfaction Across Individuals and Product Categories." *Journal of Economic Psychology* 1991, 12(2): 267-286.

Kling, Jeffrey R., Jens Ludwig, and Lawrence F. Katz, "Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment." *Quarterly Journal of Economics* 2005, 120(1): 87-130.

Kmenta, Jan. *Elements of Econometrics*. Second Edition, New York: Macmillan, 1986.

Krueger, Alan B., and Pei Zhu. "Another Look at the New York City School Voucher Experiment." *American Behavioral Scientist* 2004a, 47(5): 658-698.

Krueger, Alan B., and Pei Zhu. "Inefficiency, Subsample Selection Bias, and Nonrobustness: A Response to Paul E. Peterson and William G. Howell." *American Behavioral Scientist* 2004b, 47(5): 718-728.

Labbe, Theola. "Seven Catholic Schools to Be Converted to Charters." *The Washington Post*, November 6, 2007.

Lamdin, Douglas J. "Evidence of Student Attendance as an Independent Variable in Education Production Functions." *Journal of Educational Research* 1996, 89(3): 155-162.

Lee, David S. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." Working Paper No. 11721, Cambridge, MA: National Bureau Of Economic Research, 2005.

Lee, Valerie E., and Anthony S. Bryk. "Curriculum Tracking as Mediating the Social Distribution of High School Achievement." *Sociology of Education* 1988, 61(2): 78-94.

Lee, Valerie E., Robert F. Dedrick, and Julia B. Smith. "The Effect of the Social Organization of Schools on Teachers' Efficacy and Satisfaction." *Sociology of Education* 1991, 64(3): 190-208.

Lee, Valerie E. and Susanna Loeb. "School Size in Chicago Elementary Schools: Effects on Teachers' Attitudes and Students' Achievement." *American Educational Research Journal* 2000, 37(1): 3-31.

Liang, Kung-Yee, and Scott L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika* 1986, 73(1): 13-22.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell. *School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program*. MPR Reference No. 8404-045. Cambridge, MA: Mathematica Policy Research, 2002.

McNeal, Ralph B. Jr. "Extracurricular Activities and High School Dropouts." *Sociology of Education*, Jan. 1995, 68(1): 62-80.

Mulkey, Lynn M., Robert L. Crain, and Alexander J.C. Harrington. "One-Parent Households and Achievement: Economic and Behavioral Explanations of a Small Effect." *Sociology of Education* 1992, 65(1): 48-65.

Mullis, I.V.S., M.O. Martin, E.J. Gonzalez, and A.M. Kennedy. *Progress in International Reading Literacy Study 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary Schools*, Chestnut Hill, MA: Boston College, 2003.

Natriello, G., and Edward L. McDill. "Performance Standards, Student Effort on Homework, and Academic Achievement." *Sociology of Education* 1986, 59(1): 18-31.

Neal, Derek. "The Effects of Catholic Secondary Schooling on Educational Achievement." *Journal of Labor Economics* 1997, 15(1): 98-123.

Nielsen, Laura B., and Patrick J. Wolf. "Representative Bureaucracy and Harder Questions: A Response to Meier, Wrinkle, and Polinard." *The Journal of Politics* 2001, 63(2): 598-615.

Nye, Barbara, Larry V. Hedges, and Spyros Konstantopoulos. "The Effects of Small Class Sizes on Academic Achievement: The Results of the Tennessee Class Size Experiment." *American Educational Research Journal* 2000, 37(1): 123-151.

Peterson, Paul E., and William G. Howell. "Efficiency, Bias, and Classification Schemes: A Response to Alan B. Krueger and Pei Zhu. *American Behavioral Scientist* 2004a, 47(5): 699-717.

Peterson, Paul E., and William G. Howell. "Voucher Research Controversy: New Looks at the New York City Evaluation." *Education Next* 2004b, 4(2): 73-78.

Plank, Stephen, Kathryn S. Schiller, Barbara Schneider, and James S. Coleman. "Effects of Choice in Education," in Edith Rasell and Richard Rothstein (eds.), *School Choice: Examining the Evidence* (pp. 111-134). Washington, DC: Economic Policy Institute, 1993.

Reardon, Sean F., and John T. Yun. *Private School Racial Enrollments and Segregation*. Cambridge, MA: Harvard Civil Rights Project, 2002. Available online at [http://www.civilrightsproject.ucla.edu/research/deseg/Private_Schools.pdf].

Ritter, Gary W. *The Academic Impact of Volunteer Tutoring in Urban Public Elementary Schools: Results of an Experimental Design Evaluation.* Ann Arbor, MI: Bell & Howell Information and Learning Company, 2000.

Rouse, Cecilia Elena. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program." *Quarterly Journal of Economics* 1998, 113(2): 553-602.

Rouse, Cecilia Elena, and Lisa Barrow. "School Vouchers and Student Achievement: Recent Evidence, Remaining Questions." *Annual Review of Economics*, Volume 1, 2009.

Rumberger, Russell W., and Gregory J. Palardy. "Does Segregation Still Matter? The Impact of Student Composition on Academic Achievement in High School." *Teachers College Record* 2005, 107(9): 1999-2045.

Rutter, Michael, Barbara Maughan, Peter Mortimore, and Janet Ouston. *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children*. Cambridge, MA: Harvard University Press, 1979.

Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn. "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment." *Journal of Human Resources* 2006, 41(4): 649-691.

Sander, William. "Private Schools and Public School Achievement." *The Journal of Human Resources* 1999, 34(4): 697-709.

Schneider, Mark, Paul Teske, and Melissa Marschall. 2000. *Choosing Schools: Consumer Choice and the Quality of American Schools*. Princeton, NJ: Princeton University Press, 2000.

Schneider, Mark, and Jack Buckley. "What Do Parents Want From Schools? Evidence From the Internet." *Educational Evaluation and Policy Analysis* 2002, 24(2): 133-144.

Schochet, Peter Z. *Guidelines for Multiple Testing in Experimental Evaluations of Educational Interventions*, Revised Draft Report. MPR Reference No: 6300-080. Cambridge, MA: Mathematica Policy Research, 2007.

Singh, Kusum, Patricia G. Bickley, Paul Trivette, Timothy Z Keith, Patricia B. Keith, and Eileen Anderson. "The Effects of Four Components of Parental Involvement on Eighth-Grade Student Achievement: Structural Analysis of NELS-88 Data." *School Psychology Review* 1995, 24(2): 299-317.

Spector, Paul E. *Summated Rating Scale Construction: An Introduction.* Newbury Park, CA: Sage Publications, 1991.

Sui-Chu, Esther H., and J. Douglas Willms. "Effects of Parental Involvement on Eighth-Grade Achievement." *Sociology of Education* 1996, 69(2): 126-141.

Turque, Bill. "Rhee's Two-Page Framework Spells Out Teaching Guidelines. *The Washington Post*, August 23, 2009.

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. *Head Start Impact Study: First Year Findings.* Washington, DC: Author, 2005. Available online at [http://www.acf.hhs.gov/programs/opre/hs/impact_study/reports/first_yr_finds/first_yr_finds.pdf].

Warren, John Robert. *Graduation Rates for Choice and Public School Students in Milwaukee, 2003-2008*. Milwaukee, WI: School Choice Wisconsin, February, 2010.

Wasley, Patricia A. "Small Classes, Small Schools: The Time Is Now." *Educational Leadership* 2002, 59(5): 6-11.

Wayne, Andrew J., and Peter Youngs. "Teacher Characteristics and Student Achievement Gains: A Review." *Review of Educational Research* 2003, 73(1): 89-122.

What Works Clearinghouse. *What Works Clearinghouse Evidence Standards for Reviewing Studies*. U.S. Department of Education, Institute for Education Sciences. September 2006. Available online at: [http://ies.ed.gov/ncee/wwc/pdf/study_standards_final.pdf].

White, Halbert. "Maximum Likelihood Estimation of Misspecified Models. *Econometrica* 1982, 50(1): 1-25.

Witte, John F. *The Market Approach to Education: An Analysis of America's First Voucher Program.* Princeton, NJ: Princeton University Press, 2000.

Wolf, Patrick. "School Voucher Programs: What the Research Says About Parental School Choice." *Brigham Young University Law Review*, 2008:2.

Wolf, Patrick J., and Daniel S. Hoople. "Looking Inside the Black Box: What Schooling Factors Explain Voucher Gains in Washington, DC." *Peabody Journal of Education* 2006, 81: 7-26.

Wolf, Patrick J., Paul E. Peterson, and Martin R. West. *Results of a School Voucher Experiment: The Case of Washington, D.C. After Two Years.* Paper delivered at the National Center for Education Statistics 2001 Data Conference, Mayflower Hotel, Washington, DC: July 25-27, 2001. Available online at [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=313822].

Wolf, Patrick, Babette Gutmann, Nada Eissa, Michael Puma, and Marsha Silverberg. *Evaluation of the DC Opportunity Scholarship Program: First Year Report on Participation.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance. Washington, DC: U.S. Government Printing Office, 2005. Available online at [http://ies.ed.gov/ncee/].

Wolf, Patrick, Babette Gutmann, Michael Puma, and Marsha Silverberg. *Evaluation of the DC Opportunity Scholarship Program: Second Year Report on Participation.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2006-4003. Washington, DC: U.S. Government Printing Office, 2006. Available online at [http://ies.ed.gov/ncee/].

Wolf, Patrick, Babette Gutmann, Michael Puma, Lou Rizzo, Nada Eissa, and Marsha Silverberg. *Evaluation of the DC Opportunity Scholarship Program: Impacts After One Year.* U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2007-4009. Washington, DC: U.S. Government Printing Office, 2007. Available online at [http://ies.ed.gov/ncee/].

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. *Evaluation of the DC Opportunity Scholarship Program: Impacts After Two Years*. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2008-4023. Washington, DC: U.S. Government Printing Office, 2008. Available online at [http://ies.ed.gov/ncee/].

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. *Evaluation of the DC Opportunity Scholarship Program: Impacts After Three Years*. U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2009-4050. Washington, DC: U.S. Government Printing Office, 2009. Available online at [http://ies.ed.gov/ncee/].

Wu, Fang and Sen Qi. "Longitudinal Effects of Parenting on Children's Academic Achievement in African-American Families." *The Journal of Negro Education* 2006, 75(3): 415-430.

Appendix A
Research Methodology

This appendix describes the central features of the evaluation's research design, the sources and treatment of data (including why and how the data were adjusted to maintain sample balance), and how the data were analyzed in order to identify Program impacts.

A.1 Defining the "Treatment" and the "Counterfactual"

The primary purpose of this evaluation is to assess the impact of the DC Opportunity Scholarship Program (OSP), where impact is defined as the difference between outcomes observed for scholarship awardees and what *would have been observed for these same students had they **not** been awarded a scholarship.* Although it is impossible to observe the same individuals in these two different situations, if random assignment is well implemented, the students who were offered scholarships will not differ in any systematic or unmeasured way from the group of nonawardees, except for the fact that they were offered scholarships. More precisely, there may be some nonprogrammatic differences between the two groups, but the expected or average value of these differences is zero because they are the result of mere chance. Under this design, a simple comparison of outcomes for the two groups yields an unbiased estimate of the effect of the treatment condition, in this case an unbiased estimate of the impact of the award of an OSP scholarship on various outcomes of interest.

It is important, however, to keep in mind the precise definition of the treatment and what it is being compared to because it is the difference in outcomes under these two conditions that leads to the estimated impact of the Program.

- The ***treatment*** is the award or offer of an OSP scholarship, which is all the Program can do. The Program does not compel students to actually use the scholarship or make them move from a public to a private school. Therefore, the Program's estimated average impact includes the reality that some students who are offered a scholarship will, in fact, be disinclined to use it (what we refer to as "decliners").

- This offer of a scholarship is compared to the ***counterfactual*** or control group condition, which is defined as applying for but not being awarded an OSP scholarship. Students randomized into this group are **not** prevented from moving to a private school on their own, if the family opts to use its own resources or if the student is able to obtain another type of scholarship from an entity other than Washington Scholarship Fund (WSF). Such independent access to a private school education, or to a non-OSP

scholarship, is **not** a violation of random assignment but a correct reflection of what probably would have happened in the absence of the new Program, i.e., that some students in the applicant pool would have found a way to attend a private school on their own.

While these two study conditions and their comparison represent the main impact analysis approach, often called the Intent to Treat (ITT) analysis, the evaluation also provides separate estimates of the impact of the OSP on that subset of children who actually used the scholarship, referred to as estimated Impact on the Treated (IOT). These different analyses are described below in separate sections of this appendix.[1]

A.2 Study Power

The goals of statistical power analysis and sample size estimation are to determine how large a sample is needed to make accurate and reliable statistical judgments and how likely it is that a statistical test will detect effects of a given magnitude. Formally, power is the probability of rejecting the null hypothesis (the initial assumption that the treatment has no effect) if the treatment does, in fact, have a non-zero effect on the outcomes of interest. Power is typically estimated at the early stages of a study, based on assumptions regarding the amount of data (i.e., the planned sample sizes) and the strength of relationships within those data. Power estimates establish reasonable expectations, prior to actual data collection, regarding how large true programmatic effects would need to be in order for the data and analysis to reveal them.

Before presenting the results of our power analysis for this study, several key points are worth noting:

- The results of the power analysis are presented in terms of minimum detectable effects (MDEs), which are a simple way to express the statistical precision or "power" of an impact study design. Intuitively, an MDE is the smallest program impact or "effect size" that could be measured with confidence given random sampling and statistical estimation error. Study power itself is much like the power of a microscope—the greater the power, the smaller the objects that can be detected. Thus, MDEs of a small fraction of a standard deviation (SD), such as 0.10 SD, signal greater study power (i.e., an ability to "see" relatively small program effects) than do larger MDEs, such as 0.30 SD.

- Although this evaluation examines a variety of outcomes, including student test scores in every year post-baseline, in this report we present the test score power analysis numbers only for the data collected at least four years after random assignment. Power

[1] In addition, the evaluation estimates the relationship between attending a private school, regardless of whether an OSP scholarship is used, and key outcomes. The methodological approach and results of that analysis are provided in appendix E.

estimates for the test score analyses from earlier study years are available elsewhere (e.g., Wolf et al. 2009, appendix A2). An estimate of the power of the study's attainment analysis also is provided here, since this is the only impact report in the series that has examined that specific outcome.

- Central to analytic power is the sample size of study participants *who actually provide outcome information in a given year*. In order to produce highly precise power estimates at least four years into this longitudinal study, here we use the actual counts of student observations obtained from the data collection at least four years after random assignment. Sample size is one of three parameters of these power estimates that are fixed based upon actual numbers from this evaluation.

- The second parameter of the power analysis that we set based on actual data from the evaluation is the sibling rate. A majority of the students in the impact sample (56 percent) have siblings who also are participating in the evaluation. The test scores of children from the same family tend to be correlated with each other because siblings share some of the same genes and experience similar home environments that affect learning. Thus, the power analysis that we conducted adjusts for the fact that test-score clustering within families reduces the amount of independent information that siblings contribute to the evaluation.[2]

- If all else is equal, power is greatest when the treatment and control groups are the same size. The third parameter of the power analysis that we set based on actual conditions is the treatment/control sample ratio which, in the case of the test score analysis, is 1.65 overall but varies by subgroup. Because neither the overall nor subgroup samples have actual treatment/control ratios close to 1.00, our analysis will have slightly less power than a study with a comparable number of participants equally distributed across the treatment and control conditions.

- For the estimation of power in the test score analysis, the power analysis takes account of the estimated correlation between baseline test scores and outcome test scores, derived from a previous experimental analysis.[3] By including baseline test scores in the statistical estimation of outcome test scores, analysts make the estimation of the impact of the treatment on the outcome more precise, thus increasing power. Since we cannot determine, from data outside of the study, the correlation between baseline test scores and the likelihood of high school graduation, the test score correlation was not factored into the power forecasts for the attainment analysis.

- These power estimates do **not** account for the reality that some students in the treatment group who are offered the scholarship decline to use it (referred to as "no shows" or "decliners" in the experimental literature). Assuming that the Program has no impact on the students who decline to use a scholarship, each study participant who is a treatment decliner generates outcome data that have the practical effect of reducing

---

[2] Specifically, the power estimates generated here assume an intraclass correlation of 0.1 for siblings in the study. The number of family members in the study averaged 1.6 student participants per family cluster, and that information was incorporated into the power analysis as well, since the primary analysis produces robust standard errors by clustering on family.

[3] A "proxy" correlation between baseline and outcome test scores is drawn from a previous similar study to enable us to forecast study power independent of the actual relationships between variables in the outcome data. The use of actual data, as opposed to close proxies, limits one's ability to classify a study as "under" or "adequately" powered as the ability of an actual analysis to detect a significant effect is indistinguishable from its actual identification or not of that effect.

the ITT impact estimate toward zero. Thus, experimental evaluations of programs that experience high levels of "no shows" may fail to report statistically significant programmatic impacts simply because fewer than expected members of the treatment group actually use the programmatic treatment.[4]

- Finally, the following are the key assumptions used in the power calculations:

  α the statistical significance level, set equal to 0.05 (i.e., 95 percent confidence);

  $(1-\beta)$ the power of the test, set at 0.80;

  σ the standard deviation for an outcome of interest, in this case, set at 20 for the student test scores;

  ρ the correlation between a given student's test scores at baseline and outcome at least four years after random assignment, set at 0.57; and

  ζ the correlation between sibling outcomes, set at 0.5 in the case of test scores and 0.1 in the case of high school graduation.

The assumptions above regarding test score standard deviations and correlations are drawn from the actual data obtained from the previous experimental evaluation of the privately funded WSF program, 1998-2001 (see Wolf, Peterson, and West 2001). Though characterized as assumptions, they are likely to be more accurate than mere educated guesses because they are based on actual data from a similar analysis. A review of the literature suggests that 0.5 is representative of the degree to which sibling test scores are correlated. The MDEs are estimated for reading impacts, but would be approximately similar for math impacts as well.

Data from actual respondents are the main driver of study power. By the time of the spring 2009 data collection events, 296 members of the impact sample were forecasted to have exceeded 12th grade based on their grade upon application to the Program. Since students were always tested in their forecasted grade (calculated by adding the years since application to their grade upon application), and the SAT-9 does not have a test for grades 13 or above, these students were not testable and thus were classified as grade-outs for the purposes of the 2009 test score analysis. Grade-outs indirectly reduce study power by shrinking the size of the pool of respondent targets that might ultimately produce actual outcome data to inform the analysis. In this case, the grade-outs shrunk the potential respondent pool by 12.8 percent compared to outcome year 1, when there were no grade-outs. Under such circumstances, either response rates need to increase proportionally to the extent to which the target respondent pool has

[4] Low treatment usage rates do not reduce the analytic power of ITT estimates. They make findings of program impact less likely because they reduce the size of the average impact of the program across the entire treatment group of users and non-users. Thus, a high-powered analysis is likely to detect programmatic impacts even under conditions of moderate levels of program attrition because such an analysis will be able to detect relatively small average treatment effects.

shrunk, or study power will decline. The test score response rate of effectively 69.5 percent in 2009 was similar to the response rates in previous years, when the target pool was not reduced by grade-outs, so power for the 2009 analysis of OSP impacts necessarily will be reduced.

Examining student achievement in 2009, which was four years after random assignment for cohort 2 and five years after random assignment for cohort 1, the study has sufficient power to detect an overall test score impact of .13 of a standard deviation or higher (table A-1). The MDEs for analyzing test score impacts at the subgroup level range between .15 and .23 of a standard deviation. This evaluation should be able reliably to detect test score impacts from the OSP that are greater than or equal to these thresholds.

To place these estimated effect sizes in context, an effect of 0.13 to 0.23 of a standard deviation equates to a Normal Curve Equivalent (NCE) difference of 2.73 to 4.84 NCE points.[5] Converting NCEs to a change in percentile ranks depends on where on the overall distribution the observed change occurs. For example, if the control group was, on average, at the 20th percentile, a gain of 2.73 NCEs would bring it up to about the 24th percentile.

Table A-1. Minimum Detectable Effects: Student Achievement in Reading, Overall and by Subgroup, 2008-09

	Sample Size				MDE
Impact Sample	Treatment	Control	Treatment/ Control Ratio	Total	
All K-12 (2008-09)	863	474	1.82	1337	.13
Subgroup					
School					
SINI 2003-05	363	158	2.30	521	.21
Not SINI 2003-05	500	316	1.58	816	.16
Performance					
Lower	295	143	2.06	438	.23
Higher	331	568	0.58	899	.15
Gender					
Male	260	424	0.61	684	.17
Female	214	439	0.49	653	.19

NOTES: Estimates at 80 percent power using a two-tailed hypothesis test at the 0.05 level of statistical significance.
Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

[5] The standard deviation of the SAT-9 is 21.06 NCEs.

This evaluation examining impacts at least four years after random assignment also examines Program impacts on educational attainment, operationally defined as the rate of graduation from high school. In the overall impact sample, 500 students were forecasted to have completed 12th grade by the summer of 2009 based on their grade level and age upon application to the Program. A total of 296 of the students had graded-out of the test-score element of the study prior to the 2008-09 school year, while the remaining 204 targets of the attainment study were forecasted to be completing 12th grade that year. Responses to the follow-up parent survey on educational attainment were received regarding 316 of these students. The power analysis indicates that the attainment study is powered to detect an overall impact of .26 standard deviations or higher (table A-2). The attainment study power across the subgroups of the impact sample with more than 105 actual observations ranges from .31 to .39 standard deviations. Previous non-experimental research on the effect of Catholic schools on rates of high school graduation has suggested positive effects of 10 to 18 percentage points (Booker, Sass, Gill, and Zimmer 2008, p. 4). In the case of this impact analysis, attainment effects of 10-18 percentage points would equate to effect sizes of .22-.39 standard deviations. Effect sizes within much of that range would be detectable for the overall sample and the larger four of the six subgroup samples.

Table A-2. Minimum Detectable Effects: Attainment, Overall and by Subgroup, 2008-09

	Sample Size				MDE
Impact Sample	Treatment	Control	Treatment/ Control Ratio	Total	
All K-12 (2008-09)	127	189	0.67	316	.26
Subgroup					
School					
SINI 2003-05	97	134	0.72	231	.31
Not SINI 2003-05	30	55	0.55	85	N/A
Performance					
Lower	48	57	0.84	105	N/A
Higher	79	132	0.60	211	.33
Gender					
Male	70	97	0.72	167	.36
Female	57	92	0.62	149	.39

NOTES: Estimates at 80 percent power using a two-tailed hypothesis test at the 0.05 level of statistical significance. Subgroups labeled "N/A" for MDE contained too few observations to support a reliable power estimate.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

In summary, the power analysis shows that we are able to estimate treatment effects of reasonable magnitudes at least four years after random assignment. The analysis suggests that this experimental study will be powered, at the 80 percent level, to achieve the impact analysis goals of

determining whether the Program significantly influences test score and educational attainment outcomes for all randomly assigned participants as well as many of the policy-relevant subgroups of participants.

A.3 Sources of Data, Outcome Measures, and Baseline Covariates

Sources of Data

Comparable data were collected for each student in the impact sample regardless of whether the student was in cohort 1 or 2 or was randomly assigned to the treatment or control group. However, the temporal separation of the two study cohorts leads to the relationship between the actual timing of data collection and the impact analysis samples analyzed through the course of the evaluation. For the year one through year three impact analyses, the cohort 1 and 2 data were aligned so that they represented a similar elapsed time after random assignment (one, two, and three years), which for the two cohorts actually occurred in different years (table A-3). Originally the final year of data collection for the evaluation was to be the spring of 2008. As a result, no cohort 1 data were collected in the spring of 2009, as it was expected that no comparable "year four" data would be collected on cohort 2 in the spring of 2009. Congress voted to extend both the Program and the evaluation in the late spring of 2008, after data collection for that year had ended. The evaluation team subsequently collected data on both cohort 1 and cohort 2 participants in the spring of 2009. Because no year four data were collected on cohort 1, the data on both cohorts—cohort 1 after five years and cohort 2 after four years—were aligned to match the year of data collection (2009) for this final impact analysis (shaded in table A-3). It is not expected that the mixing of year five outcomes from cohort 1 with year four outcomes from cohort 2 will bias the analysis because:

1. It was done for both the treatment and control groups;
2. The year five outcomes from cohort 1 represent only 14 percent of the test scores in the impact sample for this final year evaluation; and
3. Four years of treatment exposure is only 20 percent less than five years of treatment exposure.

Still, it should be noted that the data analyzed in this final year of the OSP evaluation include a mixture of mostly year four outcome data from cohort 2 with some year five outcome data from cohort 1.

Table A-3. Alignment of Cohort Data with Impact Years

Annual Impact	Cohort 1 (Spring 2004 Applicants)	Cohort 2 (Spring 2005 Applicants)
	Spring 2004 (baseline)	Spring 2005 (baseline)
One year after random assignment	Spring 2005 (one year post-lottery)	Spring 2006 (one year post-lottery)
Two years after random assignment	Spring 2006 (two years post-lottery)	Spring 2007 (two years post-lottery)
Three years after random assignment	Spring 2007 (three years post-lottery)	Spring 2008 (three years post-lottery)
At least four years after random assignment	Spring 2009 (five years post-lottery)	Spring 2009 (four years post-lottery)

NOTE: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Only the final year outcome data from cohort 1 and cohort 2 were used in the analysis that informs this report. Although the fact that outcome data were collected annually would enable evaluators to examine multiple years of outcome data longitudinally, such an approach was not taken for this particular report for several reasons. First, the statute guiding the evaluation states that Congress should be informed regarding study outcomes on an annual basis. Second, as an experimental evaluation with baseline controls, each annual estimation of impacts in this study already is cumulative in the sense that the impacts in year t+four for any respondent in that year subsume any impacts of the Program on that participant from year t+three, t+two, and t+one (regardless of whether that specific participant even responded to data collection in earlier years). Third, the process of consolidating all of the outcome data over five years and preparing it for longitudinal data analysis would have substantially delayed the production and release of this report of the Program's impacts after four or more years. Because a longitudinal analysis of all the annual data from the evaluation could be performed as an additional method for informing policymakers about the impacts of the OSP, the evaluation team is preparing such an analysis for release in the future as a separate policy report.

The full data collection activity each year included the following separate sources of information:

- **Student assessments.** Baseline measures of student achievement in reading and math for public school applicants came from the Stanford Achievement Test 9th Edition (SAT-9) standardized assessment administered by the District of Columbia Public Schools (DCPS) as part of its spring testing program for cohort 1 and from the SAT-9 standardized assessment administered by the evaluation team in the spring for cohort 2.[6] Each spring after the baseline year, the evaluation team administered the SAT-9 to all cohort 1 and 2 students who were offered a scholarship, as well as to all members of the control group who did not receive a scholarship.[7] The testing took place primarily on Saturdays, during the spring, in locations throughout DC arranged by the evaluators. The testing conditions were similar for members of the treatment and control groups, and the test administrators hired and trained by the evaluation team did not know whether specific students were members of the treatment or control groups. The standardized testing in reading and math provided the outcome measures for student achievement. The sample-wide response rates for these data collection instruments were 83 percent for the baseline year and 69.5 percent for the follow-up assessments in 2008-09.[8]

- **Parent surveys.** The OSP application included baseline surveys for parents applying to the Program. These surveys were appended to the OSP application form and therefore were completed at the time of application to the Program.[9] Each spring after the baseline year, surveys of parents of all applicants were conducted at the Saturday testing events, while parents were waiting for their children to complete their outcome testing. The parent surveys provided the self-reported outcome measures for parental satisfaction and safety. Other topics included reasons for applying, school involvement, educational climate, and curricular offerings at the school. The response

---

[6] For cohort 1 at baseline, students in grades not tested by DCPS were contacted by the evaluation team and asked to attend Saturday testing events where the SAT-9 was administered to them. Fill-in baseline test scores were obtained for 70 percent of the targeted students. Combined with the scores received from DCPS, baseline test scores were obtained from 76 percent of the cohort 1 impact sample in reading and 77 percent in math. In the school year for which cohort 2 families applied for the OSP, the DCPS assessment program was in transition, and fewer grades were tested. As a result, the evaluation team attempted to administer the SAT-9 to all eligible applicants entering grades kindergarten through 12 at Saturday testing sessions in order to obtain a comprehensive and comparable set of baseline test scores for this group. Baseline test scores were obtained from 68 percent of the cohort 2 impact sample in reading and 79 percent in math. Baseline test score response rates in reading were 79 percent for the cohort 1 treatment group and 73 percent for the cohort 1 control group, a difference of 6 percentage points. In math, the cohort 1 treatment response rate at baseline was 80 percent—7 percentage points above the control rate of 73 percent. For cohort 2, baseline test score response rates were higher for the treatment group than for the control group in reading—71 percent compared to 63 percent—and in math—84 percent for the treatment group versus 72 percent for the control group. For the combined cohort impact sample, the baseline response rates in reading were 73 percent for the treatment group and 67 percent for the control group. In math, the combined cohort response rate was 83 percent for the treatment group and 75 percent for the control group.

[7] Although the SAT-9 is not available for students below first grade, Stanford Achievement does offer similar tests that are vertically equated to the SAT-9 for younger students. We administered these tests—the SESAT 1 for rising kindergarteners and the SESAT 2 for current kindergarteners (i.e., rising first graders).

[8] See section A.5 for a discussion of the treatment of incomplete test score data.

[9] The levels of response to the baseline parent surveys varied somewhat by item. All study participants provided complete baseline data regarding characteristics that were central to the determination of eligibility and priority in the lottery, such as family income and grade level. Response rates were very high (98-99 percent) for baseline survey items associated with the basic demographic characteristics of participating students, such as age, race, ethnicity, and number of siblings. Baseline survey response rates were lower (85-86 percent) for items concerned with the education and employment status of the child's mother. The baseline survey response rates for the treatment and control groups did not differ systematically.

rate for this data collection instrument was 100 percent for the baseline year and 66 percent for the follow-up surveys in 2008-09.

- **Student surveys.** Each spring after the baseline year, surveys of students in grades 4 and above were conducted at the outcome testing events. The student surveys provided the self-reported outcome measures for student satisfaction and safety. Additional topics included attitude toward school, school environment, friends and classmates and individual activities. For the follow-up data collection in 2008-09, the survey response rate among students in grade 4 or higher was 66.5 percent.

- **Principal surveys.** Each spring, surveys of principals of all public and private schools operating in the District of Columbia were conducted. Topics included self-reports of school organization, safety, climate, principals' awareness of and response to the OSP, and, for private school principals, why they were or were not participating in the OSP. In 2008-09, the response rate for the public school principal survey was 75 percent and the response rate for the private school principal survey was 72 percent.

In addition, for this impact report, in the summer of 2009 a follow-up survey of educational attainment was administered by telephone to the parents of students in the impact sample forecasted to have graduated from high school by June of that year. The survey asked a series of questions about the student's educational and employment status, most importantly whether the student had received a high school diploma.

Outcome Measures

Congress specified in the Program statute that the rigorous evaluation study possible impacts regarding academic achievement, educational attainment, school safety, and satisfaction. For this report, impact estimates were produced for all four of these outcome domains: (1) academic achievement in reading and math (two measures), (2) educational attainment (one measure), (3) parent self-reports of school safety (one measure) and student self-reports of school safety (one measure), and (4) parental self-reports of satisfaction (one measure) and student self-reports of satisfaction (one measure). All outcome data were obtained from impact sample respondents in the spring or summer of 2009 and include the following:

- **Academic outcomes.** The academic outcomes used in these analyses are assessments of student academic achievement in reading/language arts and mathematics derived from the administration of the SAT-9 by Westat-trained staff.[10] Like most norm-referenced tests, the SAT-9 includes subtests within the reading and math domains in most grades; e.g., in grades 3-8, the reading test comprises reading vocabulary and reading comprehension, while the math test consists of math problemsolving and math procedures. This norm-referenced test is designed to measure how a student's

[10] The law requires the evaluation to use as its academic achievement measure the same assessment DCPS was using the first year the OSP was implemented, which was the SAT-9.

performance compares with the scores of other students who took the test for norming purposes.[11] Each student's performance is measured using scale-scores that are derived from item response theory (IRT) item-pattern scoring methods, which use all of the information contained in a student's pattern of item responses to compute an individual's score. These scores have an additional property called "vertically equating," which allows scores to be compared across a grade span (e.g., K-12) to measure changes over time.

- **Educational attainment.** The educational attainment measure used in this analysis is a dichotomous measure of whether a student graduated from high school with a regular diploma. Parents of students forecasted to have completed 12th grade by June of 2009, based on their grade-level and age at baseline, were contacted and asked if their child had in fact obtained a high school diploma. Students whose parents answered "yes" were coded "1" for the variable "attained a high school diploma" and students whose parents answered "no" were coded "0" for that outcome.
- **Parent self-reports of safety and an orderly school climate.** Parents were asked about the perceived seriousness of a number of problems at their child's school commonly associated with danger and rule-breaking. The specific items, all drawn from the surveys used in previous experimental evaluations of scholarship programs, were:
  - Property destruction;
  - Tardiness;
  - Truancy;
  - Fighting;
  - Cheating;
  - Racial conflict;
  - Weapons;
  - Drug distribution;
  - Drug and alcohol use; and
  - Teacher absenteeism.

  Parents were asked to label these conditions as "very serious," "somewhat serious," or "not serious" at their child's school. Responses to these items subsequently were categorized as "yes" (very or somewhat serious) or "no" (not serious). The number of "yes" responses for each parent were then summed to create a parental danger index or count that ranged from 0 to 10. Finally, the index was reverse coded to transform it from a "danger" measure to a "safety" (i.e., lack of danger) measure.[12]

[11] The norming sample for the SAT-9 included students from the Northeastern, Midwestern, Southern, and Western regions of the United States and is also representative of the Nation in terms of ethnicity, urbanicity, socio-economic status, and students enrolled in private and Catholic schools. The norming sample is representative of the Nation, but not necessarily of DC or of low-income students. Scale scores are vertically integrated across grades, so that scores tend to be higher in the upper grades and lower in the lower grades. For example, the mean and standard deviation (SD) for the norming population is 463.8 (SD=38.5) for kindergarteners tested in the spring, compared to 652.1 (SD=39.1) for 5th graders and 703.6 (SD=36.5) for students in 12th grade. (*Stanford-9 Technical Data Report*. San Antonio TX: Harcourt Educational Measurement. Harcourt Assessment, Inc. 1997.)

[12] Previous experimental evaluations of scholarship programs used summary scales to measure parental satisfaction, as we do below, but generally presented parental and student danger outcomes and student satisfaction outcomes for the individual items that we list here. We have created scales of satisfaction and indexes of danger concerns because the outcome patterns for the individual items tend to be generally consistent and, under such conditions, scaling them or combining them in indices tends to generate more reliable results.

- **Student self-reports of safety and an orderly school climate.** Students were asked how often (never, once or twice, three times or more) various adverse events had occurred to them this school year. The student danger indicators, drawn from previous scholarship program evaluations, included instances of:

  – Theft;
  – Robbery;
  – Being offered drugs;
  – Physical assault;
  – Threats of physical harm;
  – Observations of weapons being carried by other students;
  – Bullying; and
  – Taunting.

  Responses to these items were categorized as "yes" (at least once) or "no" (never) to create a count of the number of reported events that ranged from 0 to 8. The index was reverse coded to transform it from a "danger" measure to a "safety" (i.e., lack of danger) measure. [13]

- **Parental self-reports of satisfaction.** Parent satisfaction with their child's school was measured three ways because previous evaluations of scholarship programs have used multiple indicators of participant satisfaction (see Mayer et al. 2002; Witte 2000). The three measures are (1) the percentage of parents who assigned their child's school a grade of A or B, (2) average rating of school on a 5-point scale, and (3) average score on a 12-item school satisfaction index. To avoid multiple comparisons in the analysis of satisfaction impacts, a single measure—the percentage of parents who graded their child's school A or B—was used in the impact analysis presented in chapter 3. Impacts on the other measures of satisfaction as well as the responses to individual items of the satisfaction scale are presented in appendix E.

  To generate the primary measure of school satisfaction, parents were asked "What overall grade would you give this child's current school?" A response of "F" was assigned the value "1," a "D" was assigned a "2," and so on up to a value of "5" for an "A." Observations with the value "5" or "4" were then recoded "1" and all other values were recoded "0" for the binary variable "graded school A or B" used in the main analysis. The original, full-grade scale was preserved, and the impact of the Program on that measure of parent satisfaction is presented in appendix E, table E-7.

  In addition, parents were asked "How satisfied are you with the following aspects of your child's school?" and to rate each of the following dimensions on a 4-point scale ranging from "very dissatisfied" to "very satisfied:"

  – Location of school;
  – School safety;
  – Class sizes;
  – School facilities;
  – Respect between teachers and students;
  – How much teachers inform parents of students' progress;

[13] As a count of discrete items, the student school danger index and the similar index from parent reports were not subject to internal consistency checks using Cronbach's Alpha. The sum of item counts lacks multi-dimensional features of scale items, such as both direction and degree, which generate the data patterns necessary to produce consistency ratings.

- How much students can observe religious traditions;
- Parental support for the school;
- Discipline;
- Academic quality;
- Racial mix of students; and
- Services for students with special needs.

The responses to this set of items were combined into a single parent satisfaction scale using maximum likelihood IRT. IRT is a procedure that draws upon the complete pattern of responses to a set of questions in order to develop a reliable gauge of the respondent's level of a "latent" or underlying trait, in this case satisfaction (Hambleton, Swaminathan, and Rogers 1991). (See section A.4 below for a more detailed description of IRT.) The consistency and reliability of scaled measures of traits such as satisfaction can be determined by a rating statistic called Cronbach's Alpha (Spector 1992). The completed parent satisfaction scale exhibited very high reliability with a Cronbach's Alpha of .93.[14] The impact of the Program on the parent satisfaction with school scale is presented in appendix E, table E-8. Program impacts on individual scale items appear in appendix E, table E-14.

- **Student self-reports of satisfaction.** Students were also asked to grade their school using the same question asked of parents, and two outcomes were created—a grade range and a dichotomous variable—as discussed above for parents. The results of the analysis of the impact of the Program on a student's likelihood of assigning his or her school a grade of A or B appear in chapter 3 as the primary measure of students satisfaction with their schools. The impact of the OSP on the average grade given across the full grade range appears in appendix E, table E-10.

  Students were also asked to rate 17 specific aspects of their current school on a 4-point scale. The individual items covered the following general topics:

  - Behavior and discipline;
  - Academic quality;
  - Social supports and interactions; and
  - Teacher quality.

  A single composite satisfaction scale was created for students using the same IRT procedures used to create the parent satisfaction scale. (See section A.4 below for a more detailed description of IRT.) The student scale also exhibited a high level of reliability; it had a Cronbach's Alpha of .85. The impact of the Program on the student satisfaction with school scale is presented in appendix E, table E-11. Program impacts on individual scale items appear in appendix E, table E-15.

Baseline or "Preprogram" Covariates

In addition to the collection of outcome data for each study participant, various personal, family, and educational characteristics of the students in the impact sample were obtained prior to random

[14] J.C. Nunnally is credited with developing the widely accepted standard that a Cronbach's Alpha above .70 demonstrates an acceptable degree of internal consistency for a multi-item scale (Spector 1992, p. 32).

assignment via the application form (including a parent survey) and administration of the SAT-9 in reading and math. Such "baseline" covariates are important in the context of an experimental evaluation because they permit researchers to (1) verify the integrity of the random assignment, (2) inform the generation of appropriate nonresponse weights, and (3) include the covariates in regressions to improve the precision of the estimations of treatment impacts and adjust for any baseline differences across the treatment and control groups.[15] The covariates that are most useful in performing each of these three functions are those that previous research has linked to the study outcomes of interest (Howell et al. 2006, p. 212).[16] These variables regularly are included in regression models designed to estimate educational outcomes such as test scores, or, in the case of the SINI indicator, are especially important to this particular evaluation:[17]

- Student's baseline reading scale score;
- Student's baseline math scale score;
- Student attended a school designated SINI 2003-05 indicator;
- Student's age (in months) at the time of application for an Opportunity Scholarship;
- Student's forecasted entering grade for the next school year;
- Student's gender—male indicator;
- Student's race—African American indicator;
- Special needs indicator—whether the parent reported that the student has a disability;
- Mother has a high school diploma indicator (GED not included);
- Mother has a four-year college degree indicator;
- Mother employed either full or part time indicator;
- Household income—reported total annual income;
- Total number of children in student's household; and
- Stability—the number of months the family has lived at its current address.

---

[15] Analysts tend to agree that baseline covariates are useful in these ways within the context of an RCT, although some of them disagree regarding which of the three functions of preprogram covariates is most important. For a spirited exchange on this question, see Howell and Peterson 2004; Krueger and Zhu 2004a, 2004b; Peterson and Howell 2004a, 2004b; Howell et al. 2006, pp. 237-254).

[16] Previous analysts of voucher experiments have used a similar set of baseline covariates to estimate attendance at outcome data collection events and therefore inform student-level nonresponse weights.

[17] This list of baseline covariates is almost identical to the one that Krueger and Zhu (2004a, p. 692) used in one of their re-analyses of the data from the New York City voucher experiment. The only differences include alternate measures of the same characteristic (e.g., our measure of student disability includes English language learners, whereas Krueger and Zhu included a separate indicator for English spoken at home) or variables that we were not able to measure at baseline (e.g., mother's religion and mother's place of birth).

A.4 IRT Analysis Used to Create Scales

Questionnaire Items

Two separate satisfaction scales were created, one for parents and one for students, using responses to the parent and student surveys, respectively. The parent scale was created from the following question consisting of 12 individual items:

Q9. How satisfied are you with the following aspects of this child's current school?
(✓ Check one box per row)

	Very dissatisfied	Dissatisfied	Satisfied	Very satisfied
a. Location of school	❑1	❑2	❑3	❑4
b. School safety	❑1	❑2	❑3	❑4
c. Class sizes	❑1	❑2	❑3	❑4
d. School facilities	❑1	❑2	❑3	❑4
e. Respect between teachers and students	❑1	❑2	❑3	❑4
f. How much teachers inform parents of students' progress	❑1	❑2	❑3	❑4
g. How much students can observe religious traditions'	❑1	❑2	❑3	❑4
h. Parental support for the school	❑1	❑2	❑3	❑4
i. Discipline	❑1	❑2	❑3	❑4
j. Academic quality	❑1	❑2	❑3	❑4
k. Racial mix of students	❑1	❑2	❑3	❑4
l. Services for students with special needs	❑1	❑2	❑3	❑4

The student scale was created from two different questions consisting of 17 items:

Q11. Do you agree or disagree with these statements about your school?
(✓ Check one box on each row)

	Agree strongly	Agree	Disagree	Disagree strongly
Students are proud to go to this school	❑1	❑2	❑3	❑4
There is a lot of learning at the school	❑1	❑2	❑3	❑4
Rules of behavior are strict	❑1	❑2	❑3	❑4
When students misbehave, they receive the same treatment	❑1	❑2	❑3	❑4
I don't feel safe	❑1	❑2	❑3	❑4
People at my school are supportive	❑1	❑2	❑3	❑4
I feel isolated at my school	❑1	❑2	❑3	❑4
I enjoy going to school	❑1	❑2	❑3	❑4

Q13. Do you agree or disagree with these statements about the students and teachers in your school? (Check one box on each row)

	Agree strongly	Agree	Disagree	Disagree strongly
Students				
a. Students behave well with the teachers	❑1	❑2	❑3	❑4
b. Students neglect their homework	❑1	❑2	❑3	❑4
c. In class, I often feel made fun of by other students	❑1	❑2	❑3	❑4
d. Other students often disrupt class	❑1	❑2	❑3	❑4
e. Students who misbehave often get away with it	❑1	❑2	❑3	❑4
Teachers				
f. Most of my teachers really listen to what I have to say	❑1	❑2	❑3	❑4
g. My teachers are fair	❑1	❑2	❑3	❑4
h. My teachers expect me to succeed	❑1	❑2	❑3	❑4
i. Some teachers ignore cheating when they see it	❑1	❑2	❑3	❑4

Prior to scale construction, all items were coded to create a consistent direction of satisfaction, i.e., that a value of 4 indicated that the respondent was most satisfied with the particular dimension of his/her school.

Scale Development and Scoring

The two scales were developed, and scores assigned to individual parents and students, using a statistical procedure called maximum likelihood IRT (see Hambleton, Swaminathan, and Rogers 1991). IRT has gained increasing attention in the development of standardized academic tests and, most recently, in the development of scales measuring a wide variety of "subjective traits" such as satisfaction with treatment and individual perceptions of health status and overall quality of life.

The basic idea of IRT is to model a relationship between a hypothesized underlying trait or construct, which is unobserved, and an individual's responses to a set of survey questions or items on a test. Common educational examples are a student's reading and math ability as measured by an achievement test. In the current situation, the underlying trait of interest is the student's or parent's "satisfaction" with the child's school. The results of the IRT analysis can be used to determine the extent

to which the items included in the scale (or test) are good measures of the underlying construct, and how well the items "hang together" (show common relationships) to characterize the underlying, and unobserved, construct.

IRT models were applied to a fixed and consistent set of survey items to generate each scale each year of the evaluation. This approach, as opposed to selecting the scale items empirically through factor analysis, was used for several reasons. First, the survey items for each scale were pre-selected from survey scales used in previous experimental evaluations of scholarship programs (Howell et al. 2006; Mayer et al. 2002). As a result, the items already had been field tested, and any program impacts involving the scales themselves could be compared directly to the results of previous studies. Second, using a fixed set of scale items rendered the scales compositionally consistent across evaluation years, so that evaluation results from a scaled outcome in one year of the study could be compared to evaluation results from the same scaled outcome in other years of the study. Moreover, reliability tests conducted on the scales at the conclusion of the IRT process confirmed that they met or exceeded the traditional threshold for scale reliability.

In IRT models, the underlying trait or construct of interest (e.g., an individual's reading ability) is designated by theta (θ). Individuals with higher levels of θ have a higher probability of getting a particular test item correct or, in our case, a higher probability of agreeing with a particular item in the satisfaction scale, than do individuals with lower levels of θ. The modeled relationship between θ and the individual test or questionnaire items is typically based on a two-parameter logistic function: (1) the first parameter is the item difficulty, which captures individual differences in their ability to get an item correct (or in their satisfaction), and (2) the second parameter is the slope, or discrimination, parameter, which captures how well a particular item differentiates between individuals on the underlying construct or trait. In other words, the IRT model estimates the probability of getting a particular item correct on a test (or agreeing with a statement on an attitude scale) conditional on an individual's underlying trait level, i.e., the higher a person's trait level, the greater the probability that the person will agree with the item or provide a correct answer. For example, if the following statement is presented, "Students behave well with the teachers," then students with higher levels of satisfaction (our θ in this example) will have higher probabilities for agreeing with this statement.

More traditional methods of creating scales often involve just counts of individual item-level responses. This approach assumes that each item is equally related to the underlying trait. IRT, on the other hand, uses all of the available information contained in an individual's responses to all of the test or survey questions and uses the difficulty and discrimination parameters to estimate an individual's test or scale score. As a result, two individuals can have the same summed score (e.g., the same number of

correct test items), but they may have very different IRT scores if they had a different pattern of responses. For example, if this were a test of academic ability, one student might answer more of the highly discriminating and difficult items than another student and would receive a higher IRT-derived score than another student who answered the same number of items but scored correctly on items with lower difficulty.

Another important advantage of IRT models is that they can produce reliable scale estimates even when an individual fails to respond to particular items; that is, the model yields the same estimate of the individual's score regardless of missing data.

A.5 Treatment of Incomplete Test Score Data

Like most norm-referenced standardized tests, the SAT-9 includes subtests within the reading and math domains in most grades; for example, the Reading Comprehension subtest is one component of the reading test battery. Ideally, students complete each subtest within a given domain, and their total or composite score for that domain is the average of their performance on the various subtests. The composite score is superior to any specific subtest score as a measure of achievement in reading or math because it represents a more comprehensive gauge of mastery of domain skills and content and also draws upon more test items in calculating the achievement score. When available, composite scores for a domain are preferred to subtest scores alone.

Some students provided some, but not all, outcome subtest scores within the reading and math domains at least four years after random assignment because they either missed or skipped entire subtests. This included 52 students in reading and 20 students in math.[18] The total number of individual students who provided incomplete test score data was 66, since six students provided only subtest scores in both reading and math.

When the problem of incomplete test scores first emerged during the initial stages of the evaluation, the research team conducted an analysis to determine how closely subtest reading and math scores correlated with composite scores for the over 1,600 respondents for whom both subtest and composite scores were available. The correlations between subtest and composite scores within particular domains and grades were very strong, ranging from a low of r =.79 to a high of r = .92.[19] Given such high levels of correlations, and consistent with the principle of bringing as many observations as possible to

[18] In grades 9-12, the SAT-9 includes only a single mathematics test with no subsections.

[19] Figures are for bivariate correlations using Pearson's *R*.

the test score impact analysis, a decision was made to substitute subtest scores for the composite scores in all cases where only the subtest scores were available. At least four years after random assignment, these 66 cases were considered respondents for the purposes of calculating the test score nonresponse weights and were therefore included in the test score impact analysis.

A.6 Imputation for Missing Baseline Data

One difficulty that arose regarding the baseline data was the extent to which data were missing. Although some important baseline covariates (e.g., family income, grade, race, and gender) were available for all students, other baseline covariates contained some missing values. Importantly, nearly 20 percent of math scores and 29 percent of reading scores were not obtained at baseline.[20] To deal with this occurrence, missing baseline data were imputed by fitting stepwise models to each covariate using all of the available baseline covariates as potential predictors. Predicted values were then generated, and imputation was done using a "nearest neighbor" procedure in which a "donor" was found for each "recipient" in a way that minimized the difference between the predicted value for the recipient and the actual value for the donor across all potential donors.[21] For example, if a particular student was missing a value for the total number of children in the student's household, a regression estimation predicted the likely number of children in the student's household (e.g., 2.8) based on all known characteristics of the student, and another student in the study was located with a known value (e.g., 3) for number of children in the household that closely matched the value the data predicted the student might have. That donor student's value was then imputed as the recipient's value for that characteristic.[22]

A.7 Sampling and Nonresponse Weights

Sampling weights were used in the impact analyses to account for the fact that the study sample was selected differently in the two years of initial OSP implementation, as well as across different priority groups and grade bands. Conducting the analyses without weights would run the risk of confusing the effect of the treatment with compositional differences between the treatment and control groups due to

---

[20] In some of these cases, students did not come for the required baseline testing. In other cases, they attended the testing but did not attempt to answer enough questions on one or more of the subsections of the test to be assigned a valid test score.

[21] The stepwise regressions and imputations that made up the imputation procedure were done in an iterative cycle, in that "current" imputations were used in fitting the stepwise model, and then that stepwise model was used to generate a new set of imputations. This imputation-regression-imputation cycle went through the set of baseline covariates in a cyclical sequence, and this was continued until convergence resulted (i.e., no change in imputations or model fits between cycles). To initiate the procedure (i.e., to get the first set of imputations), an initial set of imputations was computed via a simple hot deck procedure. The final result of this algorithm was an efficient set of imputations that respected the underlying patterns in the data as were picked up by the stepwise regression procedures, while providing a set of imputations with distributional patterns similar to those of the real values.

[22] For continuous variables (e.g., baseline score), a residual was taken from a hot deck procedure (a random draw from all residuals from the model) and added to the predicted value from the recipient.

the fact that certain kinds of eligible applicants had higher or lower probabilities of being awarded a scholarship. The sampling weights consist of two primary parts: (1) a "base weight," which is simply the inverse of the probability of being selected to treatment (or control) and (2) an adjustment for differential nonresponse to data collection for impact sample members still eligible for the Program.

Base Weights

The base weight is the inverse of the probability of being assigned to either the treatment or control group. For each randomization stratum *s* defined by cohort, SINI status, and grade band, *p* is designated as the probability of assignment to the treatment group and *1-p* the probability of being assigned to the control group.

First, designate the treatment and control groups as *t* and *c*, respectively, and let *i* represent an individual student. Then Y_{sit} represents a particular outcome (e.g., a reading test score) for a particular student in the population pool if the student was assigned to the treatment group, and Y_{sic} the outcome for a particular student in the population pool if the student was assigned to the control group.

The population totals can then be written as:

$$Y_c = \sum_{s=1}^{8} \sum_{i=1}^{N_s} Y_{sic} \qquad Y_t = \sum_{s=1}^{8} \sum_{i=1}^{N_s} Y_{sit}$$

where Y_c, for example, corresponds to the population total achieved if every member of the population pool does not receive the treatment, and Y_t corresponds to the population pool if every member of the population receives the treatment. Under the null hypothesis of no treatment effect, $Y_c = Y_t$ and $Y_t - Y_c$ is defined to be the effect of treatment, but this difference cannot be directly observed for any particular student as no student can be in both treatment and control groups. However, utilizing the randomization from the treatment assignment process, we can generate unbiased estimators of Y_t and Y_c as follows (with n_s equal to the number of treatment group members in stratum *s*):

$$\hat{Y}_c = \sum_{s=1}^{8} \sum_{i=1}^{N_s - n_s} \frac{y_{sic}}{1 - p_s} \qquad \hat{Y}_t = \sum_{s=1}^{8} \sum_{i=1}^{n_s} \frac{y_{sit}}{p_s}$$

Writing w_{sc} and w_{st} as the base weights for stratum s and control and treatment group respectively, $w_{sc} = (1 - p_s)^{-1}$ and $w_{st} = p_s^{-1}$, we can write

$$\hat{Y}_c = \sum_{s=1}^{8} \sum_{i=1}^{N_s - n_s} w_{sc} y_{sic} \qquad \hat{Y}_t = \sum_{s=1}^{8} \sum_{i=1}^{n_s} w_{st} y_{sit}$$

The values of these base weights are then assigned to the participants in each stratum (table A-4).

Table A-4. Base Weights by Randomization Strata

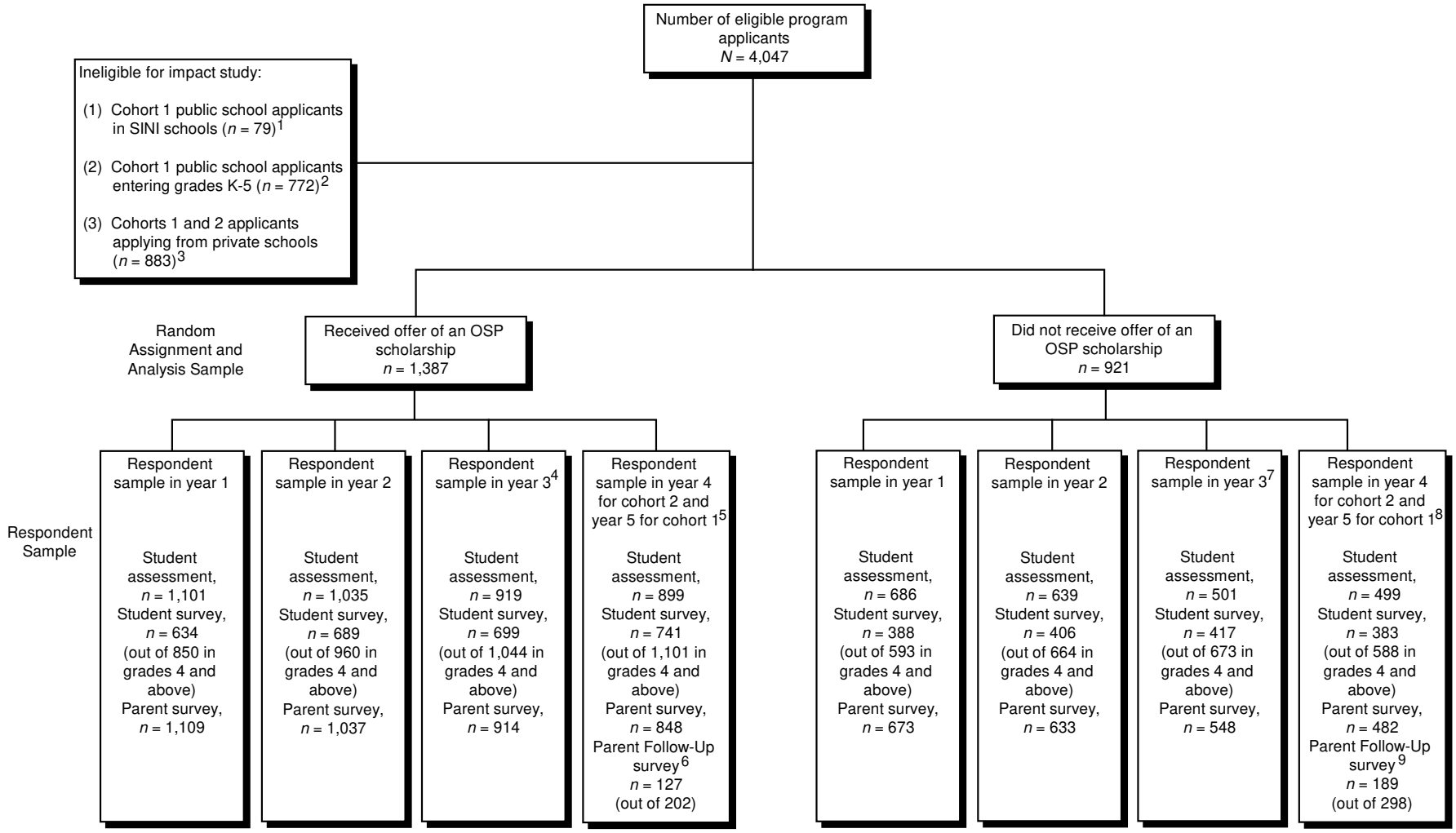
Stratum	Cohort	SINI Status	Grade Band	Treatment Sampling Rate (%)	Base Weight for Control Group	Base Weight for Treatment Group
1	Cohort 1	Not SINI 2003-05	6th to 8th	75.89	4.15	1.32
2	Cohort 1	Not SINI 2003-05	9th to 12th	28.21	1.39	3.54
3	Cohort 2	SINI 2003-05	K to 5th	78.34	4.62	1.28
4	Cohort 2	SINI 2003-05	6th to 8th	75.00	4.00	1.33
5	Cohort 2	SINI 2003-05	9th to 12th	38.14	1.62	2.62
6	Cohort 2	Not SINI 2003-05	K to 5th	59.05	2.44	1.69
7	Cohort 2	Not SINI 2003-05	6th to 8th	55.33	2.24	1.81
8	Cohort 2	Not SINI 2003-05	9th to 12th	28.57	1.40	3.50

Adjustments for Nonresponse

The members of the treatment and control groups were offered similar inducements to cooperate in outcome data collection. Treatment students were invited to data collection events to renew their scholarships, and their parents were given a small cash payment for their time and transportation costs in responding. Control students were made eligible for follow-up scholarship lotteries, and their parents were provided with a compensation payment for attending follow-up data collection sessions. The initial base weights were adjusted for nonresponse, where a "respondent" was considered a student with reading or mathematics test data at least four years after random assignment (figure A-1).[23] Similar adjustments were made for response to the student survey and to the parent survey, which had very different response patterns to those of the test assessments, resulting in four distinct sets of weights. The use of these adjustments helps control nonresponse response bias by compensating for different data collection response rates across various demographic groups of students organized within classification "cells." In effect, the nonresponse adjustment factor "spreads the weight" of the nonresponding students over the

[23] Students were required to have produced at least one complete subtest score in the relevant domain (i.e., reading or math) to be counted as a respondent for that domain.

Figure A-1. Flow of Cohort 1 and Cohort 2 Applicants From Eligibility Through Analysis: At Least Four Years After Application and Random Assignment



[1]The program operator offered a scholarship to all eligible public school applicants in cohort 1 applying from SINI schools.
[2]The program operator awarded scholarships to all eligible public school applicants in cohort 1 entering grades K-5 because there were sufficient slots in private schools to accommodate all the applicants in these grades.
[3]The evaluation design is intended to estimate the impact of giving students the opportunity to attend private school, so applicants to the Program who were already in private schools were excluded from the study.
[4]Fifty-four members of the treatment group were no longer grade eligible three years after random assignment. These "grade-outs" were not invited to data collection events.
[5]Ninety-four members of the treatment group were no longer grade eligible by 2008-09. These "grade-outs" were not invited to data collection events.
[6]The Parent Follow-Up survey was administered to parents of students who had turned 16 as of June 30, 2009.
[7]Thirty-one members of the control group were no longer grade eligible three years after random assignment. These "grade-outs" were not invited to data collection events.
[8]Two-hundred two members of the control group were no longer grade eligible by 2008-09. These "grade-outs" were not invited to data collection events.
[9]The Parent Follow-Up survey was administered to parents of students who had turned 16 as of June 30, 2009.

responding students in that cell, so that they represent not only students who responded (i.e., themselves), but also students who were like them in relevant ways but did not respond to outcome data collection.[24] This maintains the same mix of the impact sample across classification cells as would have been present had there been no nonresponse (see Howell et al. 2006, pp. 209-216; U.S. Department of Health and Human Services 2005). As a last step, the nonresponse-adjusted base weights were trimmed. Trimming prevents extremely large weights from unduly inflating the estimated variances and thus reducing the precision of the impact estimates.[25]

Grade-outs

The OSP is limited to students attending kindergarten through grade 12. Participants are no longer eligible for a scholarship after completing 12th grade. Moreover, the SAT does not have a version for students beyond 12th grade, so the evaluation team did not have a test instrument to administer to any study participants who had exceeded 12th grade. By the time data collection began in the spring of 2009, a total of 296 students comprising 12.8 percent of the impact sample were forecasted to have completed 12th grade based on their grade level and age upon application to the Program and thus had "graded-out" of the evaluation.[26] Since graded-out students were not invited to testing events, and could not be tested even if they attended such sessions, they are all study nonrespondents for purposes of spring 2009 data collection. Including grade-outs in the original impact sample, 57.9 percent of the sample provided valid test scores at least four years after random assignment, including 62.2 percent of the treatment group and 51.5 percent of the control group (table A-5). Response rates based on the original impact sample were especially low for cohort 1, as 32 percent of that group had graded out of the study by 2008-09. Because grade-outs were no longer targets of data collection for testing and parent and student surveys, the

[24] To determine the factors used to create the nonresponse adjustment cells, both logistic regression (with response or not as the dependent variable) and a software package called CHAID (Chi-squared Automatic Interaction Detector) were used to determine which of the available baseline variables were correlated with the propensity to respond. The available baseline variables from which predictors of response propensity were drawn included family income, mother's job status, mother's education, disability status of the child, race, grade, gender, and baseline test score data (both reading and math). Stepwise logistic regression was first used to select a set of characteristics generally predictive of response (using the SAS procedure PROC LOGISTIC with a 20 percent level of significance entry cutoff). These stepwise procedures were done separately within each of the eight sampling strata. The CHAID program (now a part of the SPSS statistical software package) was then used to define a set of cells with differing response rates within each sampling stratum, using the set of characteristics for the sampling stratum coming from the PROC LOGISTIC models. Cells with fewer than six observations were not allowed. The nonresponse cells nested within the sampling strata and within treatment status. The nonresponse adjustment for each respondent in the cell was equal to the reciprocal of the base-weighted response rate within the cell.

[25] The trimming rule was that any weights that were larger than 4.5 times the median weight (with medians computed separately within the treatment and control groups) were trimmed back to be equal to 4.5 times the median weight. This procedure affected only a very small number of cases. Such trimming is standard procedure and is done as a matter of course in the National Assessment of Educational Progress (NAEP) assessment sample weighting.

[26] "Forecasted grade" is the grade a student was confirmed to be in at the time of Program application plus the number of years since application. Forecasted grade always was used in place of the actual grade the student was enrolled in during a given year so that differential rates of grade retention across the treatment and control groups would not result in lower grade-level tests being administered to one group compared with the other group and thus bias the analysis of test score impacts.

response rate calculations that exclude grade-outs more accurately represent the actual level of response at least four years after random assignment. The remaining discussion of response rates will rely exclusively on calculations that exclude grade-outs.

Table A-5. Test Score Response Rates as a Percentage of Original Impact Sample, 2008-09

	Original Impact Sample Members	Grade-Outs	2008-09 Impact Sample	2008-09 Actual Respondents	Response Rate by Original Sample
Cohort 1 C	193	112	81	43	22.3
Cohort 1 T	299	44	255	152	50.8
Cohort 2 C	728	90	638	431	59.2
Cohort 2 T	1,088	50	1,038	711	65.3
Cohort 1 total	492	156	336	195	39.6
Cohort 2 total	1,816	140	1,676	1142	62.9
C total	921	202	719	474	51.5
T total	1,387	94	1,293	863	62.2
Combined total	2,308	296	2,012	1337	57.9

NOTE: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Response to Data Collection

For the 2008-09 data collection, after four invitations to attend testing events, the evaluation team had obtained responses from 64 percent of the treatment group and 63 percent of the control group (table A-6).

Recently, a technique was developed to help reduce nonresponse bias in longitudinal impact analyses. Nonresponse subsampling is a strategy to reduce the differences between the characteristics of baseline and outcome samples by way of random sampling and nonresponse conversion. After the regular period of outcome data collection is over, a subsample of nonrespondents is drawn and subjected to intensive efforts at nonresponse conversion. If initial nonresponse was significantly higher in one experimental group compared with the other, as was the case in years 1 through 3 for this evaluation, then the subsample can be drawn exclusively from the underresponded group (e.g., controls). Since the initial nonresponse rates were similar between the treatment and control groups in the final year of data collection, both groups were subsampled.

Table A-6. Test Score Response Rates Before Drawing Subsample, 2008-09

	Impact Sample Members	Pre-sample Respondents	Response Rate (%)
Cohort 1 C	81	41	50.6
Cohort 1 T	255	135	52.9
Cohort 2 C	638	410	64.3
Cohort 2 T	1,038	691	66.6
Cohort 1 total	336	176	52.4
Cohort 2 total	1,676	1,101	65.7
C total	719	451	62.7
T total	1,293	826	63.9
Combined total	2,012	1277	63.5

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Each initial subsampled nonrespondent who converts to a respondent by providing outcome data counts as one more respondent for purposes of the "actual" response rate but counts as 1/sampling rate (r) respondents for purposes of the "effective" response rate. Through a simple weighting algorithm, the random sampling permits the respondent to also "stand in" for members of the initial nonrespondent group who were not selected for the subsample but who presumably would have converted to respondent status if they had been selected to receive the intensive recruiting efforts and incentives that were the conversion "treatment." In other words, the proportion of subsampled nonrespondents that converts represents themselves as well as the same proportion of nonsampled nonrespondents.

This technique was applied for the spring 2009 data collection, as it had been in the previous years of the evaluation, to increase the outcome response rates and obtain similar effective response rates for the treatment and control groups. The initial data gathering effort was followed by a targeted intensive recruitment of initial nonrespondents. A random sample of 367 of the 735 initial nonrespondents was drawn (50 percent),[27] and the selected participants were offered a larger turnout incentive and greater flexibility and convenience in an attempt to "convert" as many as possible from nonrespondent to respondent status. A total of 61 initial nonrespondents (16 percent) were converted to respondents as a result of this effort, 24 members of the control group and 37 members of the treatment group (table A-7).

[27] Nonrespondents included 159 from cohort 1 (40 from the control group and 119 from the treatment group) and 576 from cohort 2 (229 from the control group and 347 from the treatment group). The random sample of 367 consisted of 88 from cohort 1 (25 from the control group and 63 from the treatment group) and 279 from cohort 2 (109 from the control group and 170 from the treatment group).

Table A-7. Subsample Conversion Response Rates for Test Score Outcomes, 2008-09

	Subsample Members	Actual Response Conversions	Actual Conversion Rate (%)
Control total	134	24	.18
Treatment total	233	37	.16
Total	367	61	.17

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

These "converted" cases were weighted more heavily than the other observations in the outcome sample, by a factor of two, to account for the complementary set of initial nonrespondents who were not randomly selected for targeted conversion efforts but who would have responded if they had been targeted (see Kling, Ludwig, and Katz 2005; Sanbonmatsu et al. 2006).[28] The weights ensure that each converted member of the subsample represents him or herself as well as another study participant: a nonrespondent like him or her who would have converted had he/she been included in the subsample. As a result of implementing this approach, the response rate for outcome testing in math and reading for the control group increased to an effective rate of 69.4 percent and for the treatment group increased to 69.5 percent. The difference in the response rates across the control and treatment groups was 0.1 percentage points (table A-8). Response rates to the follow-up survey on educational attainment were 63.4 percent for the control group and 62.9 percent for the treatment group (table A-9).[29] The parent survey had an effective response rate of 67 percent for the control group and 65.6 percent for the treatment group (table A-10). Response rates for the student survey were effectively 65.1 for the control group and 67.3 for the treatment group (table A-11). The response rate for the public school principal survey was 74.7 percent and 72.2 percent for the private school principal survey (table A-12).

The What Works Clearinghouse (WWC) considers a Randomized Control Trial (RCT) such as this evaluation to meet evidence standards for claims of causality without reservations if study sample attrition is neither severe overall nor significantly different across the treatment and control groups. Even if an RCT suffers from one or both of these sample attrition problems, it is still classified as meeting

[28] For example, the Moving to Opportunity Section 8 housing voucher experimental evaluation obtained an initial year one response rate of 78 percent. Evaluators then drew a random sample of 30 percent of the initial nonresponders and subjected them to intense recruitment efforts that resulted in nearly half of them responding, thereby increasing their response rate to 81 percent. The evaluators then assumed that the second-wave respondents were similar to the half of the larger nonrespondent group that they did not pursue aggressively and thus estimated and reported an "effective response rate" of 90 percent, even though actual data were obtained for only 81 percent of the respondents.

[29] Nonrespondent subsampling was not conducted on the target sample for the attainment outcome because no conversion incentives were available to offer them. For the attainment analysis, the effective response rate was the actual response rate.

Table A-8. Response Rates for Reading and Math After Drawing Subsample, Actual and Effective, 2008-09

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	81	43	53.1	44	54.6
Cohort 1 T	255	152	59.6	167	65.5
Cohort 2 C	638	431	67.6	454	71.2
Cohort 2 T	1,038	711	68.5	732	70.5
Cohort 1 total	336	195	58.0	211	62.9
Cohort 2 total	1,676	1,142	68.1	1,186	70.8
C total	719	474	65.9	499	69.4
T total	1,293	863	66.7	899	69.5
Combined total	2,012	1,337	66.5	1,398	69.5

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Table A-9. Response Rates for Parent Follow-Up Surveys: 2008-09

	Impact Sample Members	Respondents	Response Rate (%)
Cohort 1 C	135	79	58.5
Cohort 1 T	111	66	59.5
Cohort 2 C	163	110	67.5
Cohort 2 T	91	61	67.0
Cohort 1 total	246	145	58.9
Cohort 2 total	254	171	67.3
C total	298	189	63.4
T total	202	127	62.9
Combined total	500	316	63.2

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. Parents of these "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Table A-10. Response Rates for Parent Survey, Actual and Effective, 2008-09

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	81	40	49.4	41	50.9
Cohort 1 T	255	137	53.7	151	59.3
Cohort 2 C	638	415	65.0	441	69.1
Cohort 2 T	1,038	675	65.0	697	67.1
Cohort 1 total	336	177	52.7	192	57.3
Cohort 2 total	1,676	1,090	65.0	1,138	67.9
C total	719	455	63.3	482	67.0
T total	1,293	812	62.8	848	65.6
Combined total	2,012	1,267	63.0	1,330	66.1

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Table A-11. Response Rates for Student Survey, Actual and Effective, 2008-09

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	81	41	50.6	42	52.2
Cohort 1 T	255	151	59.2	166	65.1
Cohort 2 C	507	327	64.5	341	67.2
Cohort 2 T	846	561	66.3	575	68.0
Cohort 1 total	336	192	57.1	208	62.0
Cohort 2 total	1,353	888	65.6	916	67.7
C total	588	368	62.6	383	65.1
T total	1,101	712	64.7	741	67.3
Combined total	1,689	1,080	63.9	1124	66.5

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

Table A-12. Response Rates for Principal Surveys, 2008-09

	Total	Respondents	Response Rate (%)
Public schools	225	168	74.7
Private schools	90	65	72.2
Combined total	315	233	74.0

evidence standards without reservation if the study demonstrates that the treatment and control group have remained approximately equivalent despite the study attrition or that acceptable methods have been used to re-equate the study samples (What Works Clearinghouse 2006, pp. 6-7). In practice, the WWC considers overall sample responses that are below 70 percent, or rates that differ between the treatment and control group by more than 5 percentage points, as constituting a possible attrition problem. The response rates of the treatment and control groups were well within the 5-point differential WWC standard for all the outcome measures examined in this report. The overall response rates for the 2008-09 data collection of 69.5 percent (student tests), 63.2 percent (attainment survey), 66.1 percent (parent survey), and 66.5 percent (student survey) were just short of the WWC standard of 70 percent. The response rate of 74 percent for the principal surveys was above the WWC standard.

When some participants in a longitudinal evaluation such as this one fail to provide outcome data in a particular year, researchers often generate and apply nonresponse weights to re-balance the treatment and control groups on important observable baseline characteristics. In this study, the nonresponse weights were designed to account for the characteristics of non-subsampled initial nonrespondents and also for the characteristics of subsampled nonrespondents that never converted to respondent status. A secondary analysis of the characteristics of the 2008-09 respondent sample, with and without these weight adjustments, confirms that the analysis weights succeeded in adjusting the outcome sample that was used for analysis to reflect the conditions of the impact sample at baseline (table A-13). The respondent sample differed from the baseline sample regarding 1 of 14 demographic measures tested (percent of mothers with a 4-year college degree) prior to administration of the nonresponse weights but regarding none of the measures after the application of the nonresponse weights. An additional analysis confirmed that the treatment respondent sample did not differ significantly from the control respondent sample on any of the 14 baseline covariates either before or after the application of the nonresponse weights. Thus, the analytic weights succeeded in re-balancing the outcome sample as designed, and the evaluation continues to meet the WWC evidence standards. Still, an additional sensitivity test was added to the analysis of the outcome with the lowest response rate (educational attainment) to calculate how different the unobserved treatment-control difference among nonrespondents would have to be compared

with the observed treatment-control difference among respondents in order for the true treatment-control difference across the entire impact sample to be 0 (see appendix C).

Table A-13. Comparison of the 2008-09 Outcome Sample Characteristics to the Baseline Sample With and Without Analysis Weights

Baseline Characteristics	Full Sample with Base Weights	Pre-subsample Base Weights	Difference	*p*-value	Full Sample with Base Weights	Tested with Full Weights	Difference	*p*-value
Reading scale score	567.49	562.64	4.84	0.16	567.49	568.73	-1.24	0.70
Math scale score	560.78	555.72	5.07	0.15	560.78	561.30	-0.52	0.88
SINI_YR0	27.18	28.04	-0.01	0.45	27.18	27.71	-0.01	0.59
Age (months)	114.51	112.05	2.46	0.06	114.51	114.57	-0.06	0.96
Baseline entering grade	4.20	4.01	0.19	0.07	4.20	4.21	-0.01	0.92
Percent male	48.54	48.34	0.20	0.87	48.54	48.38	0.00	0.89
Percent African American	87.10	87.19	-0.09	0.91	87.10	87.89	-0.01	0.28
Percent special needs	12.14	11.70	0.44	0.60	12.14	12.39	0.00	0.74
Percent mother four-year degree	5.66	4.54	1.12	0.05*	5.66	5.29	0.00	0.46
Percent mother high school diploma	80.84	81.15	-0.31	0.76	80.84	81.37	-0.01	0.54
Percent mother full-time job	60.34	59.55	0.79	0.53	60.34	59.31	0.01	0.35
Income	17,112.22	17,702.55	-590.33	0.11	17,112.22	17,365.71	-254.00	0.48
Number of children	2.85	2.93	-0.08	0.11	2.85	2.92	-0.07	0.19
Months of residential stability	72.01	75.71	-3.70	0.22	72.01	73.82	-1.81	0.53
Sample size	2,012	1,267			2,012	1,328		

The final student-level weights for the analysis were equal to:

$$W_i = (1/p_i) * (X_i) * (NR_j) * (TR_i),$$

where p_i is the probability of selection to treatment or control for student *i,* X_i is the special factor for initial nonrespondents (with X_i equal to 2.0 for this set and equal to 1 otherwise), NR_j is the nonresponse adjustment (the reciprocal of the response rate) for the classification cell to which student *i* belongs, and TR_i is the trimming adjustment (usually equal to 1, but in some cases equal to 4.5 times median cutoff divided by the untrimmed weight).

Subgroup Sample Sizes and Response Rates

Because this evaluation examines Programmatic impacts across a predefined set of participant subgroups, study response rates and subsequent analytic sample sizes are presented for each of those subgroups and for all four primary data collection instruments (student tests, parent follow-up survey of attainment, parent surveys, and student surveys).

At least four years after random assignment, the subgroup-level effective response rates for student test scores ranged from a low of 67.9 percent for students who attended SINI schools 2003-05 to a high of 70.5 percent for students who attended not SINI schools 2003-05 (table A-14).

Table A-14. Response Rates for Test Scores, by Subgroup, 2008-09

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI 2003-05	808	549	67.9
Not SINI 2003-05	1,204	849	70.5
Lower performance	675	463	68.6
Higher performance	1,337	935	69.9
Male	1,002	686	68.5
Female	1,010	711	70.4

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade-outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

The subgroup response rates for student attainment in 2008-09 ranged from a low of 55.9 percent for students who attended not SINI schools 2003-05 to a high of 66.4 percent for students who attended SINI schools 2003-05 (table A-15).

Table A-15. Response Rates for the Parent Follow-Up Survey, by Subgroup, 2008-09

	Impact Sample Members	Respondents	Response Rate (%)
SINI 2003-05	348	231	66.4
Not SINI 2003-05	152	85	55.9
Lower performance	167	105	62.9
Higher performance	333	211	63.4
Male	234	149	63.7
Female	266	167	62.8

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

At least four years after random assignment, the subgroup-level effective response rates for parent surveys ranged from a low of 63.8 percent for participants in the SINI 2003-05 subgroup to a high of 67.6 percent for their counterparts from not SINI 2003-05 schools (table A-16).

Table A-16. Response Rates for the Parent Survey, by Subgroup, 2008-09

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI 2003-05	808	516	63.8
Not SINI 2003-05	1,204	814	67.6
Lower performance	675	442	65.4
Higher performance	1,337	888	66.4
Male	1,002	650	64.9
Female	1,010	680	67.3

NOTES: A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. Parents of these "grade-outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

At least four years after random assignment, the subgroup-level effective response rates for student surveys ranged from a low of 64.2 percent for participants in the lower performing subgroup to a high of 68.1 percent for females (table A-17).

Table A-17. Response Rates for the Student Survey, by Subgroup, 2008-09

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI 2003-05	805	541	67.2
Not SINI 2003-05	884	583	66.0
Lower performance	563	362	64.2
Higher performance	1,126	763	67.7
Male	833	541	64.9
Female	856	583	68.1

NOTES: Student surveys administered to students in grades 4-12. A total of 296 students initially in the impact sample were no longer grade eligible for the Program at least four years after random assignment. These "grade-outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment.

A.8 Analytical Model for Estimating the Impact of the Program, or the Offer of a Scholarship (Experimental Estimates)

To estimate the extent to which the Program has an effect on participants, this study first compares the outcomes of the two experimental groups created through random assignment. These outcomes are referred to as Intent to Treat or ITT impact estimates. The only completely randomized, and therefore strictly comparable, groups in the study are those students who were offered scholarships (the treatment group) and those who were not offered scholarships (the control group) based on the lottery. The random assignment of students into treatment and control groups should produce groups that are similar in key characteristics, both those we can observe and measure (e.g., family income, prior academic achievement) and those we cannot (e.g., motivation to succeed or benefit from the Program). A comparison of these two groups is the most robust and reliable measure of Program impacts because it requires the fewest assumptions and least effort to make the groups similar except for their participation in the OSP.

Overall Program Impacts

Because the RCT approach has the important feature of generating comparable treatment and control groups, we used a common set of analytic techniques, designed for use in social experiments, to estimate the Program's impact on test scores and the other outcomes listed above. These analyses began with the estimate of simple mean differences using the following equation, illustrated using the test score of student *i* in year *t* (Y_{it}):

(1) $Y_{it} = \alpha + \tau T_{it} + \varepsilon_{it}$ if t>k (period after Program takes effect),

where T_{it} is equal to 1 if the student *has the opportunity to participate* in the OSP (i.e., the award rather than the actual use of the scholarship) and is equal to 0 otherwise. Equation (1) therefore estimates the effect of the **offer** of a scholarship on student outcomes. Under this ITT model, all students who were randomly assigned by virtue of the lottery are included in the analysis, regardless of whether a member of the treatment group used the scholarship to attend a private school or for how long.

Proper randomization renders experimental groups approximately comparable, but not necessarily identical. In the current study, some modest differences, almost all of which are not significant, exist between the treatment group and the control group counterfactual at baseline.[30] The basic regression model can, therefore, be improved by adding controls for observable baseline characteristics to increase the reliability of the estimated impact by accounting for minor differences between the treatment and control groups at baseline and improving the precision of the overall model. This yields the following equation to be estimated:

(2) $Y_{it} = \alpha + \tau T_{it} + X_i \gamma + \delta_R R_{it} + \delta_M M_{it} + \varepsilon_{it}$.

where X_i is a vector of student and/or family characteristics measured at baseline and known to influence future academic achievement, and R_{it} and M_{it} refer to **baseline** reading and mathematics scores, respectively (each of the included covariates are described below). In this model, τ—the parameter of sole interest—represents the effect of scholarships on test scores for students in the Program, conditional on X_i and the baseline test scores. The δ's reflect the degree to which test scores are, on average, correlated over time. With a properly designed RCT, baseline test scores and controls for observable characteristics that predict future achievement should improve the precision of the estimated impact.

[30] For example, although the average test scores of the cohort 1 and cohort 2 treatment and control groups in reading and math are all statistically comparable, in all four possible comparisons (cohort 1 reading, cohort 1 math, cohort 2 reading, cohort 2 math) the control group average baseline score is higher. That is, on average the members of the control group began the experiment with slightly higher reading and math test scores than the members of the treatment group. The control group baseline test score advantage for cohort 1 reading, cohort 2 reading, cohort 1 mathematics, and cohort 2 mathematics was 4.7, 8.4, 4.1, and 8.7 respectively, using only the actual test scores obtained at baseline. The corresponding four differences were 4.1, 7.0, 3.7, and 1.6 when the imputations of the missing baseline test scores (see section A.6) are added to the sample. Thus, after imputation, the differences between treatment and control group baseline scores were attenuated. A joint *f*-test for the significance of the pattern of test score differences at baseline was not significant for the pre-imputation data (i.e., actual scores with missing data for some observations) but was significant after the baseline data were completed by replacing missing scores with imputed scores. This apparent anomaly is a result of the larger sample sizes after imputation, which reduces the standard errors across the board, thereby increasing the precision of the statistical test and the resulting likelihood of a statistically significant result. To deal with this difference in test scores across the treatment condition at baseline, we simply include the post-imputation baseline test scores in a statistical model that produces regression-adjusted treatment impact estimates. Controlling for baseline test scores in this way effectively transforms the focus of the analysis from one on achievement levels after one year, which could be biased by the higher average baseline test scores for the control group, to one on comparative achievement gains after one year from whatever baseline the individual student performed at to start the experiment. Because including baseline test scores in regression models both levels the playing field in this way and increases the precision of the estimate of treatment impact, it is a common practice in education evaluations generally and school scholarship experiments particularly.

Adjustment for Differences in Days of Exposure to School

A final important covariate to include in this model is the number of days from September 1 to the date of outcome testing for each student.[31] This "days until test" variable, signified by DT in the equation below, controls for the fact that test scores were obtained over a four-month period each spring and that a student's ability to perform on the standardized tests can be affected by the length of time he/she has been exposed to schooling. The DT variable was further interacted with elementary school status (i.e., K-5) because younger students tend to gain relatively more than older students from additional days of schooling.[32] Thus, the models that produced the regression-adjusted impact estimates for this analysis took the general form:[33]

$$(3)\ Y_{it} = \alpha + \tau T_{it} + X_i \gamma + \delta_R R_{it} + \delta_M M_{it} + \delta_{DT} DT_{it} + \varepsilon_{it}.$$

The same set of baseline covariates were used in all impact regression models, regardless of whether the outcomes being estimated were student achievement, student attainment, school satisfaction, school safety, or any of the intermediate outcomes.[34] The days-to-test variable was included in all models except for the estimation of educational attainment, since the attainment data were collected during a different time period (summer 2009) from the collection of the data on all the other outcomes (spring 2009).

In summary, our analytic models for estimating the impacts of the OSP on the overall study sample comprised three main elements in addition to the standard items of an outcome (Y) variable, intercept, and error term. Those three elements were (1) a dummy variable for treatment defined as random assignment to the offer of a scholarship or not, (2) a vector of baseline covariates including test

---

[31] September 1st was chosen as a common reference date because most private schools approximately follow the DCPS academic calendar, and September 1st fell within the first week of schooling in fall of both 2004 and 2005.

[32] The actual statistical results confirmed the validity of this assumption, as the effect of the DT variable on outcome test scores was positive and statistically significant for K-5 students but indistinguishable from zero for grades 6-12 students.

[33] The possibility of a nonlinear relationship of DT with the outcome variables was examined through the use of a categorized version of the DT variable, with one category level including students with DT below the median value, one level with DT in the third quartile (median to 75th percentile), and one level with DT in the fourth quartile (75th percentile to maximum). This allows for a quadratic relationship (down-up-down for example) in the regression estimation if such a relationship exists. The regression with the nonlinear DT component did not provide a better fit to the data than the regression modeling a simple linear slope. As a result, the simpler model was used.

[34] After the initial impacts were obtained in the year one impact analysis, a second set of estimates were run to test the sensitivity of the results to the set of covariates included in the model. This sensitivity model used only cohort, grade, special needs, number of children in the household, African American race, baseline reading, baseline math, and days until test as control variables, as these variables tended to be significant predictors of test score outcomes in the first set of models. No important differences regarding test score impacts were found (Wolf et al. 2007, pp. 43, 49-50). As a result and upon the recommendation of our Expert Advisory Panel, the limited covariate model was subsequently dropped from the sensitivity testing.

scores and student demographics, and (3) a control variable for the number of days the student was exposed to schooling prior to providing outcome data in a given year.

Subgroup ITT Impacts

In addition to estimating overall Program impacts, this study was interested in the possibility of heterogeneous impacts (i.e., separate impacts on particular subgroups of students). Subgroup impacts were estimated by augmenting the basic analytic equation (3) to allow different treatment effects for different types of students, as follows:

$$(4)\ Y_{ikt} = \mu + \tau_T T_{ikt} + \tau_P P_i + \tau_B P_i * T_{ikt} + \sum_{j=2}^{b} \varphi^j_{is} + X_{ik} \gamma + \delta_R R_{it} + \delta_M M_{it} + \delta_{DT} DT_{it} + \varepsilon_{ik,t}$$

where *P* is an index for whether a student is a member of a particular subgroup (the *P* must be part of the X's). These models were used to estimate impacts on the separate components of the subgroup (e.g., impacts on males and females separately), and the difference in impacts between the two groups. To better understand the treatment parameters (τ's). Table A-18 presents each parameter and its derivation, along with its interpretation. τ_T is derived by setting the treatment dummy equal to 1, and the subgroup dummy =0. As such, it is the treatment effect across all students with a scholarship. τ_P is derived by setting the treatment dummy off (no scholarship in the lottery), and the subgroup dummy on (say, males) and, therefore, represents the mean outcome for males in the control group. Finally, τ_B is derived by setting the treatment dummy and the subgroup dummy (say males) equal to 1 and, therefore, represents the mean impact of the scholarship on the subgroup, *relative to* the effect of treatment on the omitted subgroup.

Table A-18. Interpretation of Equation Parameters

Parameter	Interpretation
τ_T [P=0;T=1]	Basic (overall) treatment effect
τ_P [P=1;T=0]	Mean outcome (test score) for subgroup control
τ_B [P=1;T=1]	Marginal treatment effect on subgroup

These analyses of possible heterogeneous impacts across subgroups are conducted within the context of the experimental ITT design. Thus, as with the estimation of general Program-wide impacts, any subgroup-specific impacts identified through this approach are understood to have been caused by the treatment. The ability to reliably identify separate impacts, however, depends on the sample sizes within each subgroup.

Four criteria were applied in generating the list of student subgroups that would be the focus of the analysis of possible heterogeneous effects of the Program. Those criteria were:

1. Is the subgroup based upon a student characteristic emphasized in the Program statute?
2. Does the theoretical or research literature suggest that such a characteristic should matter regarding the impact of the Program on student outcomes?
3. Would the bifurcation of the overall sample based on the presence or absence of the characteristic leave subgroups of viable size for analytic purposes?
4. Is the total set of subgroups small enough that adjustments for multiple comparisons are not likely to be severe?

Consideration of the first three criteria influenced which subgroup pairs were chosen. Consideration of the fourth criterion influenced how many subgroup pairs were chosen.

Initially, five important student characteristics were identified, the presence or absence of which generated a total of five subgrouping pairs or 10 subgroups in total. Those five student characteristics involved the SINI status of the previous public school attended, baseline test score performance, gender, grade level, and cohort. Four or more years after random assignment, the subgroupings based on grade level and cohort were no longer viable, since most or all of the baseline high school and cohort 1 participants had aged out of the study. Consequently, for this final year of the evaluation, subgroup impacts were estimated for the following groups:[35]

- Applied from a school designated SINI in 2003-05—yes and no;
- Academically lower performing student at the time of baseline testing (i.e., bottom one-third of the test score distribution) and higher performing (top two-thirds);[36]
- Gender—male and female.

Significance tests were conducted both on the treatment-control differences within each subgroup (the subgroup impact) as well as on the difference in impacts across subgroups (the interaction effect). None of the interaction effects presented in this report were statistically significant. That means, from a scientific standpoint, there are no significant differences in the impact of the OSP on students from SINI 2003-05 schools compared to students from not SINI 2003-05 schools, lower performing compared

[35] In the year one through year three impact reports, outcomes also were estimated at the subgroup level for students entering grades K-8 versus grades 9-12 at baseline and cohort 1 versus cohort 2. By 2008-09, the entire grade 9-12 subgroup and most of the cohort 1 subgroup had graded out of the study, so it was no longer feasible to estimate separate subgroup impacts for those two subgroup pairs.

[36] The lower third of the baseline performance distribution was chosen because preliminary power analyses suggested it would be the most disadvantaged performance subgroup that would include a sufficient number of members to reveal a distinctive subgroup impact if one existed.

to higher performing students, or females compared to males. Some subgroups demonstrate Programmatic impacts at the subgroup level; that is, the coefficient τ_P was statistically significant when estimated for the specific subgroup in question. None of the subgroup showed impacts that, on average, were significantly different from the impacts experienced by its paired subgroup; that is, the interaction term τ_B was always not significant. In research terms, the statistical tests on the interaction variables indicate that the impacts of the OSP are homogeneous and not heterogeneous.

Computation of Standard Errors

In computing standard errors, it is necessary to factor in the stratified sample design, clustering of student outcomes within individual families, and nonresponse adjustments. As a consequence, all of the impact analyses were completed using sampling weights in STATA.[37] The effects of family clustering, which is not part of the sample design, but which may have a measurable effect on variance, were taken into account using robust regression calculations (i.e., "sandwich" variance estimates) (see Liang and Zeger 1986; White 1982).[38] By clustering the individual error terms on family unit, our models control for possible special autocorrelation among siblings. This statistical modeling feature does not render the regressions fixed effects models. They remain random effects models but with robust standard errors obtained through clustering.

Tests were run to determine if the impact findings were sensitive to the decision to adjust for clustering within families rather than within schools. These results are reported in appendix C.

A.9 Analytical Model for Estimating the Impact of Using a Scholarship

Although the ITT analysis described above is the most reliable estimate of Program impacts, it cannot answer the full set of questions that policymakers have about the effects of the Program. For example, policymakers may be interested in estimates of the impact of the OSP on students and families that actually use an Opportunity Scholarship. The Bloom adjustment, which simply re-scales the experimental impacts over the smaller population of treatment users, is used to generate such an Impact

[37] There is also a positive effect on variance (a reduction in standard errors) from the stratification. This effect will not be captured in the primary analyses, making the resultant variance estimators conservative.

[38] We also examined the effect on the standard errors of the estimates of clustering on the school students were currently attending. Baseline school clustering reduced the standard errors of the various impact estimates by an average of 2 percent, compared to an average reduction of less than 1 percent due to clustering by family. These results indicate that the student outcome data are almost totally independent of the most likely sources of outcome clustering. This may appear to be counterintuitive, since formally accounting for clustering among observations usually increases variance in effects; however, since the randomization cut across families and baseline schools, it is possible that family and school clusters served as the equivalent of random-assignment blocks, as most multi-student families and schools contained some treatments and some controls. Such circumstances normally operate to reduce variance in subsequent impact estimates, as the within-cluster positive correlation comes into the calculation of the variance of the treatment-control difference with a minus sign.

on the Treated (IOT) estimate, with a slight modification necessitated by special circumstances of the OSP.

Impact of Using a Scholarship

For the scholarship awardees in the OSP impact sample that provided final year outcome test scores, 86 percent had used a scholarship for all or part of the four or five years after they were offered a scholarship by lottery. We view these "ever users" as the proper focus of the IOT analysis because doing so requires that we make no assumptions regarding how much exposure to the OSP treatment actually counts. Treatment group members are classified as users if they ever used a scholarship for any amount of time. The 14 percent of the treatment students in the final year respondent sample who did not use their scholarships are treated the same as scholarship users for purposes of determining the effect of the offer of a scholarship, so as to preserve the integrity of the random assignment, even though scholarship decliners likely experienced no impact from the Program. Fortunately, there is a way to estimate the impact of the OSP on the average participant who actually used a scholarship, or what we refer to as the IOT estimate. This approach does not require information about why 14 percent of the individuals declined to use the scholarship when awarded, or how they differ from other families and children in the sample. But if one can assume that decliners experience zero impact from the scholarship Program, which seems reasonable given that they did not use the scholarship, it is possible to avoid these kinds of assumptions about (or analyses of) selection into and out of the Program.

This is possible by using the original comparison of **all** treatment group members to **all** control group members (i.e., the ITT estimates described above) but re-scaling it to account for the fact that a known fraction of the treatment group members did not actually avail themselves of the treatment and therefore experienced zero impact from the treatment. The average treatment impact that was generated from a mix of treatment users and nonusers is attributed only to the treatment users, by dividing the average treatment impact by the proportion of the treatment group who used their scholarships. For this report, depending on the specific outcome being rescaled, this "Bloom adjustment" (Bloom 1984) will increase the size of the ITT impacts that were statistically significant by 14-72 percent, since the percentage of treatment users among the population of students that provided valid scores on the various test and survey outcomes ranged from 58-88 percent.[39]

[39] The Bloom adjustment is generated by dividing the ITT estimate by the usage rate for that outcome. Any number that is divided by .70 will generate a dividend that is 43 percent larger. Any number that is divided by .90 will generate a dividend that is 11 percent larger.

Adjustment for Program-Induced Crossover

In the current evaluation, conventional Bloom adjustment may not be sufficient to accurately estimate the impact of using the OSP scholarship. It is conceivable that the design of the OSP and lotteries made it possible for some control group members to attend participating private schools, above and beyond the rate at which low-income students would have done so in the absence of the Program. Statistical techniques that take this "program-enabled crossover" into account are necessary for testing the sensitivity of the evaluation's impact estimates.

In a social experiment, even as some students randomized into the treatment group will decline to use the treatment, some students randomized into the control group will obtain the treatment outside of the experiment. For example, in medical trials, this control group "crossover" to the treatment can occur when the participants in the control group purchase the equivalent of the experimental "treatment" drug over the counter and use it as members of the treatment group would. The fact that crossovers have obtained the treatment does not change their status as members of the control group—just as treatment decliners forever remain treatments—for two reasons: (1) changing control crossovers to treatments would undermine the initial random assignment, and (2) control crossover typically represents what would have happened absent the experimental program and therefore is an authentic part of the counterfactual that the control group produces for comparison. If not for the medical trial, the control crossovers would have obtained the similar drug over the counter anyway. Therefore, under normal conditions, any effect that the crossover to treatment has on members of the control group is factored into the ITT and Bloom-adjusted IOT estimates of impact as legitimate elements of the counterfactual.

In the case of the OSP experiment, control crossover takes place in the form of students in the control group attending private school. Among the members of the control group for whom we knew their school attended in the years covered by this report, 23.1 percent reported attending a private school at some time since the Program began. This crossover rate is in the higher end of the range reported for previous experimental evaluations of privately funded scholarship programs (Howell et al. 2006, p. 44).[40] The crossover rate also is higher for control group students with siblings in the treatment group (26.4 percent) compared to those without treatment siblings (20.2 percent).[41] At outcome data collection events, some parents of control group students commented to evaluation staff that their control-group child was

[40] First-year control group crossover rates in the previous three-city experiment were 18 percent in Dayton, OH; 11 percent in Washington, DC; and just 4 percent in New York City. Among those three cities, the average tuition charged by private schools is lowest in Dayton and highest in New York, a fact that presumably explains much of the variation in crossover rates.

[41] Because program oversubscription rates varied significantly by grade, random assignment took place at the student and not the family level. As a result, nearly half the members of the control group have siblings who were awarded scholarships.

accepted into a participating private school free-of-charge because he or she had a treatment group sibling who was using a scholarship to attend that school, and private schools were inclined to serve a whole family. Thus, apparently some of the control crossover that is occurring in the OSP could be properly characterized as "Program-enabled" and not a legitimate aspect of the counterfactual.

The data suggest that 2.9 percent of the control group were likely able to enroll in a private school because of the existence of the OSP. This hypothesis is derived from the fact that 20.2 percent of the control group students without treatment siblings are attending private schools, whereas 23.1 percent of the control group overall is in private schools. Since the 20.2 percent rate for controls without treatment siblings could not have been influenced by "Program-enabled crossover," we subtract that "natural crossover rate" from the overall rate of 23.1 percent to arrive at the hypothesized Program-enabled crossover rate of 2.9 percent. To adjust for the fact that this small component of the control group may have actually received the private-schooling treatment by way of the Program, the estimates of the impact of scholarship use in chapter 3 include a "double-Bloom" adjustment.[42] We rescale the pure ITT impacts by an amount equal to the treatment decliner rate (~15 percent), as described above plus the estimated Program-enabled crossover rate (~2.9 percent) to generate the IOT estimates.

The double-Bloom adjustment that we use to generate the IOT estimate in this analysis differs both conceptually and operationally from the Instrumental Variable (IV) analysis of the effect of private schooling presented in appendix E. The double-Bloom IOT calculation measures the effect of ever using *a scholarship* to attend a private school, whereas the IV estimates the effect of private schooling by way of any form of financing. The IOT treats private school attendance by members of the control group as a natural aspect of the counterfactual condition, whereas the IV factors the rate of such control group "crossover" into the estimation of the private schooling variable in the first stage of the analysis. The IOT merely mathematically adjusts the results of the treatment and control group comparison that generated the ITT, whereas the IV analysis produces a new and distinct comparison of private school attenders in the final year of the analysis to public school attenders in the final year of the analysis. In sum, the double-Bloom IOT calculation is carefully calibrated to capture only effects that could only have been caused by the scholarship program, whereas the IV analysis is mere broadly designed to estimate any effects that could be attributable to private schooling in general.

[42] IOT estimates for student attainment do not incorporate an adjustment for program-induced crossover because the data for the attainment sample do not show evidence of a program-induced crossover effect.

Appendix B
Benjamini-Hochberg Adjustments for Multiple Comparisons

The following series of tables (tables B-1 through B-7) present the original *p*-values from the significance tests conducted in the analysis for all outcome domains in which multiple comparisons were made that produced statistically significant results. The sources of the multiple comparisons were either various subgroups of the impact sample (chapters 3), or the multiple comparisons made within the conceptual groupings of mediating effects (chapter 4). In both cases, Benjamini-Hochberg adjustments were made to reduce the probability of a false discovery given the number of multiple comparisons in a given set and the pattern of outcomes observed. The adjusted false discovery rate appears in the far-right column of each table. False discovery rate *p*-values at or below .05 indicate results that remained statistically significant after adjusting for multiple comparisons.

The *p*-values were not adjusted for the estimations of the treatment impact on the full study sample within the six domains that make up the primary analysis: student achievement, student attainment, parent perceptions of safety, student perceptions of safety, parent satisfaction with school, and student satisfaction with school. These six outcome domains were specified in advance as the foci of the evaluation, and indexes and scales were used to consolidate information from multiple items into discreet measures—two approaches that have been acknowledged as appropriate for reducing the danger of false discoveries in evaluations (Schochet 2007). Moreover, no statistically significant treatment impacts were observed in reading or in math, student reports of school climate and safety, or student satisfaction at least four years after random assignment, so there could not have been false discoveries in those domains. Significant impacts for the entire sample were observed regarding student attainment, parental perceptions of school climate and safety, and parental satisfaction with their child's school, but they were not the result of multiple comparisons. In chapter 4, no statistically significant impacts were found across the indicators of home educational supports and student motivation. Thus, there could not have been false discoveries across those domains, and no adjustments for multiple comparisons were applied to those particular results.

Table B-1. Multiple Comparisons Adjustments, Reading, 2008-09

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI 2003-05	.76	.76
Not SINI 2003-05	.02*	.10
Lower performance	.74	.76
Higher performance	.04*	.10
Male	.44	.66
Female	.05*	.10

*Statistically significant at the 95 percent confidence level.

Table B-2. Multiple Comparisons Adjustments, Student Attainment, 2008-09

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI 2003-05	.01*	.03*
Not SINI 2003-05	.46	.46
Lower performance	.12	.18
Higher performance	.02*	.04*
Male	.26	.31
Female	.01**	.03*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-3. Multiple Comparisons Adjustments, Parental Perceptions of Safety and an Orderly School Climate, 2008-09

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI 2003-05	.77	.77
Not SINI 2003-05	.01**	.04*
Lower performance	.20	.24
Higher performance	.06	.12
Male	.20	.24
Female	.06	.12

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-4. Multiple Comparisons Adjustments, Parent Satisfaction: Parents Gave Their Child's School a Grade of A or B, 2008-09

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI 2003-05	.51	.51
Not SINI 2003-05	.00**	.01**
Lower performance	.05*	.07
Higher performance	.03*	.06
Male	.06	.07
Female	.03*	.06

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-5. Multiple Comparisons Adjustments, Student Motivation and Engagement, 2008-09

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student aspirations	.57	.67
Attendance	.03*	.16
Tardiness	.59	.67
Reads for fun	.17	.52
Engagement in extracurricular activities	.61	.67
Frequency of homework (days)	.67	.67

*Statistically significant at the 95 percent confidence level.

Table B-6. Multiple Comparisons Adjustments, Instructional Characteristics, 2008-09

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/teacher ratio	.85	.91
Teacher attitude	.91	.91
Ability grouping	.00**	.00**
Availability of tutors	.20	..33
In-school tutor usage	.23	.33
Programs for learning problems	.00**	.00**
Programs for English language learners	.00**	.00**
Programs for advanced learners	.00**	.00**
Before-/after-school programs	.11	.22
Enrichment programs	.30	.38

**Statistically significant at the 99 percent confidence level.

Table B-7. Multiple Comparisons Adjustments, School Environment, 2008-09

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/school communication	.00**	.01**
School size	.00**	.00**
Percent non-White	.04*	.05
Peer classroom behavior	.51	.51

**Statistically significant at the 99 percent confidence level.

Appendix C
Sensitivity Testing

In any evaluation, decisions are made about how to handle certain data or analysis issues (e.g., nonresponse differentials, sampling weights, etc.). While there are some commonly accepted approaches in research and evaluation methodology, sometimes there are multiple approaches, and any could be acceptable. The evaluation team chose its approach in consultation with a panel of methodology experts before analyzing the data and seeing the results. However, in an effort to be both transparent and complete, each presentation of analyses is followed by a discussion of the sensitivity testing conducted to determine how robust the estimates are to specific changes in the analytic approach. These different specifications include:

- **Trimmed sample.** As discussed in appendix section A.7, at the conclusion of initial data collection each year, in most cases the response rates for the treatment and control groups were below 70 percent and differed from each other by 5 percentage points or more, with the control group demonstrating the lower rate of response. To reach or at least better approximate the WWC standards for sample response, we subsampled initial nonrespondents, subjecting them to extensive nonresponse conversion efforts and more heavily weighting the data of converted participants. The subsampling approach is designed to reduce the risk of nonresponse bias at the cost of reduced precision in the estimation of impacts due to the large weighting factors that result. The trimmed sample sensitivity testing is designed to determine what impacts would be estimated if this tradeoff of less efficiency in exchange for expected bias reduction were not made. The initial response rate of the control group, prior to any subsampling and nonresponse conversion efforts, is established as a benchmark. All observations that were obtained through nonresponse conversion are excluded. Finally, a sufficient number of observations from the "latest-to-respond" members of the treatment group are excluded so that the "trimmed" treatment group sample has exactly the same response rate as the "trimmed" control group sample. The response rates for the treatment and control groups are thereby artificially and exactly equalized at the initial control group response rate, and the analysis is re-run on that sample.[1] In the case of the test scores, the initial response rate of the treatment group (64 percent) was higher than the initial response rate of the control group (63 percent) by 1 percentage point. To generate the trimmed sample in this case, in effect the "latest 1 percent of the treatment group members to respond" were dropped from the sample until the treatment response rate matched the control group's pre-subsample response rate of 63 percent.

[1] Trimming response samples in order to test the sensitivity of the primary results is not new to the field of evaluation (e.g., Lee 2005); however, unlike Lee we have evidence, in the form of response date, regarding which members of the treatment group likely were the "marginal respondents" who only turned out for data collection because they were in the treatment group. The subgroup of respondents we trim away from the sample are those that most likely would not have responded to initial data collection had they been randomly assigned to the control group.

- **Clustering on school currently attending.** Robust standard errors are generated for the primary analysis by clustering on family units, which ensures that the analysis is sensitive to the potential correlation of error terms from students within the same family. The possibility that error terms are correlated at the school level is taken into account with an analysis that generates a different set of robust standard errors by clustering on the school each student is attending. This approach produces a more generalizable set of results, since different school choice programs are likely to generate different amounts and patterns of student clustering at the school level than the specific pattern observed in the DC OSP; however, that greater level of generalizability can come at the cost of study power and analytic efficiency in measuring the impacts from this particular program, especially if large numbers of study participants are clustered in a small number of schools.

- **Attainment sample sensitivity tolerance.** The follow-up survey regarding student educational attainment generated similar response rates of 63 percent for both the treatment and control groups. No subsampling of initial nonrespondents was conducted, meaning no trimmed sample sensitivity analysis was possible. The attainment study response rate of 63 percent was below the rate of 70 percent for experiments that the What Works Clearinghouse uses to declare that an evaluation has fully met standards for causal inference. Even when treatment and control group response rates are identical, as they were in the case of the OSP attainment analysis, it is possible that nonrespondents could bias the impact estimates if the treatment nonrespondents differed significantly from the control nonrespondents on the outcome measure, in this case, educational attainment. As an additional test of the sensitivity of the attainment impact estimate, we calculate how much lower the high school graduation rate of nonrespondents in the treatment group would have to have been compared with the graduation rate of nonrespondents in the control group, in order for the estimated OSP attainment impact to be, in actuality, 0 or no impact.

Sensitivity Testing of Main Impact Analysis Models

Here we subject the findings from the overall analysis of the impact of the offer of a scholarship on achievement, attainment, safety, and satisfaction outcomes to the sensitivity analysis of using only the trimmed sample and clustering on school attended instead of family. We also assess the impacts from the exploratory subgroup analyses using these same sensitivity tests.

Sensitivity Checks for ITT Impacts on Reading and Math Achievement

The sensitivity test produced changes in the overall findings for reading (table C-1). Both the trimmed sample and clustering on current school approaches produced positive and statistically significant reading impacts. The three subgroup reading impacts that were statistically significant in the main analysis were also statistically significant under the alternative specifications. The sensitivity test produced no changes in the overall findings for math.

Table C-1. Test Score ITT Impact Estimates and *P*-Values with Different Specifications, 2008-09

	Original Estimates		Trimmed Sample		Clustering on Current School	
Student Achievement Groups	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample: reading	3.90	.07	4.80*	.02	3.90*	.03
Full sample: math	.70	.71	.27	.89	.70	.71
Not SINI 2003-05: reading	5.80*	.03	5.60*	.03	5.80*	.02
Higher performance: reading	5.18*	.04	6.24*	.01	5.18*	.02
Female: reading	5.27*	.05	5.87*	.03	5.27*	.05

*Statistically significant at the 95 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Impacts are displayed in terms of scale scores. Original estimates valid *N* for reading = 1,328; math = 1,330. Trimmed sample valid *N* for reading = 1,267; math = 1,269. Separate reading and math sample weights were used.

Sensitivity Checks for ITT Impacts on Student Attainment

The sensitivity test that involves the use of robust regression analysis that clusters on students' current school in place of the clustering by family did not produce changes in the overall or subgroup findings for student attainment (table C-2). Additionally, the final column in the table presents what the treatment impact for nonrespondents would need to be for the impacts to equal zero.

Table C-2. ITT Impact Estimates on Student Attainment: Percent with High School Diploma, 2008-09

	Original Estimates		Clustering on Current School		Treatment Impact for Nonrespondents Necessary for a Zero Effect
Student Achievement Groups	Impact	*p*-value	Impact	*p*-value	
Full sample	.12**	.01	.12*	.02	-.21
SINI 2003-05	.13*	.01	.13*	.04	-.22
Higher performance	.14*	.02	.14*	.03	-.24
Female	.20**	.01	.20**	.01	-.34

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Impact estimates are reported as marginal effects in terms of percentiles. Valid *N* = 316. Sample weights used.

Sensitivity Checks for ITT Impacts on Parent Perceptions of Safety and an Orderly School Climate

The overall Programmatic impacts on parental reports of safety and an orderly school climate discussed in chapter 3 were consistent across analytic approaches (table C-3). The positive impact of the Program on the not SINI 2003-05 subgroup was also consistent with the primary analysis.

Additionally, the trimmed sample analysis produced positive and statistically significant impacts for the higher baseline performance and female subgroups. The impact for the female subgroup was also significant when clustering on current school.

Table C-3. Parent Perceptions of Safety and an Orderly School Climate: ITT Impact Estimates and *P*-Values with Different Specifications, 2008-09

School Safety and Climate: Parents	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample	.48*	.02	.67**	.00	.48*	.05
Not SINI 2003-05	.73**	.01	.82**	.00	.73**	.00
Higher performance	.48	.06	.74**	.00	.48	.07
Female	.56	.06	.81**	.01	.56*	.04

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Original estimates valid *N* = 1,224. Trimmed sample valid *N* = 1,165. Parent survey weights were used.

Sensitivity Checks for ITT Impacts on Student Reports of Safety and an Orderly School Climate

The primary analysis discussed in chapter 3 found no treatment impact on students' perceptions of a safe school climate. This result is consistent across different analytic approaches (table C-4). Regardless of how the data were analyzed, responses of those offered a scholarship did not differ significantly from control group students' perception of school safety.

Table C-4. Student Reports of Safety and an Orderly School Climate: ITT Impact Estimates and *P*-Values with Different Specifications, 2008-09

School Safety and Climate: Students	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample	.15	.33	.15	.33	.15	.33

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Original estimates valid *N* = 1,054. Trimmed sample valid N = 1,005. Student survey weights were used.

Sensitivity Checks for ITT Impacts on Parent Reports of School Satisfaction

The finding of a positive impact of the Program on parent satisfaction for the full sample and for four of six subgroups was not sensitive to different analytic approaches with two exceptions (table C-5). The positive impact of the Program on parents' likelihood of grading their child's school A or

B, which was statistically significant for the lower performance at baseline subgroup in the primary analysis, loses significance when estimated using the smaller trimmed sample. The positive impact of the Program on parents' likelihood of grading their child's school A or B for the higher performance at baseline subgroup loses significance when clustering on current school. Additionally, the impact for the male subgroup of students is positive and statistically significant when using the trimmed sample.

Table C-5. Parent Satisfaction ITT Impact Estimates and *P*-Values with Different Specifications, 2008-09

Parent Gave School Grade of A or B	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample	.08**	.00	.10**	.00	.08*	.02
Not SINI 2003-05	.12**	.00	.11**	.00	.12**	.00
Lower performance	.10*	.05	.09	.06	.10*	.04
Higher performance	.08*	.03	.10**	.01	.08	.07
Male	.08	.06	.09*	.03	.08	.08
Female	.09*	.03	.10*	.01	.09*	.04

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Original estimates valid *N* for school grade = 1,227. Trimmed sample valid *N* for school grade = 1,167. Parent survey weights were used. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Sensitivity Checks for ITT Impacts on Student Reports of School Satisfaction

The results of the primary analysis found no Programmatic impact on overall student self-reports of satisfaction. That finding is consistent across the different methodological approaches (table C-6). In every specification, there are no differences in the likelihood of a student grading his/her school A or B.

Table C-6. Student Satisfaction ITT Impact Estimates and *P*-Values with Different Specifications, 2008-09

Outcome	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Student Gave School Grade of A or B	-.03	.43	-.02	.52	-.03	.45

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Original estimates valid *N* for school grade = 1,001. Trimmed sample valid *N* for school grade = 951. Student survey weights were used. Impact estimates are reported as marginal effects. Survey given to students in grades 4-12.

Appendix D
Relationship Between Attending a Private School and Key Outcomes

Scholarship programs such as the Opportunity Scholarship Program (OSP) are designed to expand the opportunities for students to attend private schools of their parents' choosing. As such, policymakers have been interested in the outcomes that are associated with private schooling, whether via the use of an Opportunity Scholarship or by other means. However, efforts to estimate the effects of private schooling involve statistical techniques (called Instrumental Variable or "IV" analysis) that deviate somewhat from the randomized trial, and researchers are divided on how closely these techniques approximate an estimate of experimental "impact" (Angrist, Imbens, and Rubin 1996, pp. 444-455 and 468-472; Heckman 1996, pp. 459-462). Because of this debate, it is important to distinguish these analytic results from the estimated impacts of the award or use of an OSP scholarship and to treat these findings with some caution.

D.1 Instrumental Variables Method and Results

This appendix uses IV analysis to examine the relationship between private schooling and outcomes among members of the treatment and control groups. Such an analysis is conceptually distinct from estimating the Impact on the Treated (IOT) by way of the Bloom or "double-Bloom" adjustments since it examines outcome patterns in both treatment and control groups that could be the results of exposure to private schooling. We limit the IV estimations of the effects of private schooling to only the impacts found to be statistically significant in the intent to treat (ITT) analysis presented in chapter 3.[1] Because this element of the evaluation is merely supplemental to the analysis of ITT and IOT impacts of the Program, no adjustments are made to the significance levels of the IV estimates of the effects of private schooling to account for multiple comparisons.

In practice, instrumental variable analysis involves running two stages of statistical regressions to arrive at unbiased estimates of the effects of private schooling on a particular outcome (Howell et al. 2006, pp. 49-51). In the first stage, the results of the treatment lottery and student characteristics at baseline are used to estimate the likelihood that individual students attended a private

[1] Due to the smaller sample size and missing data rates on type of school attended, we do not use IV analysis on the results for student attainment.

school at least four years after random assignment. In the second stage, that estimate of the likelihood of private schooling operates in place of an actual private schooling indicator to estimate the effect of private schooling on outcomes.[2] In cases like this experiment, the IV procedure will generate estimates of the effect of private schooling that will be slightly larger than the double-Bloom IOT impact estimates. Since the IV process tends to be inefficient and place greater demands upon the data (Kmenta 1986, p. 684), special attention must be paid to the significance levels of IV estimates, as some experimental impacts that are statistically significant at the ITT stage lose their significance when subjected to IV analysis.

The two-stage IV process that was implemented for this analysis, similar to the one used by Mayer et al. (2002) in their evaluation of a scholarship program, took the following general form:

$$P_i = \alpha_0 + \alpha_1 T_1 + \alpha_2 X_i + \varepsilon_{pi}$$
$$y_i = \beta_0 + \beta_1 P_i + \beta_2 X_i + \varepsilon_{yi} \qquad \text{(D.1)}$$

where T_1 is the lottery instrument and equals 1 if the student was offered a scholarship and 0 otherwise; X_i is the vector of baseline descriptive variables that we include in all regression estimations as control variables; P_i equals 1 if the student attended a private school in the analysis year,[3] in this case year four (cohort 2) or year five (cohort 1) and 0 otherwise; y_i is the outcome of interest; ε_{pi} and ε_{yi} are random error terms that capture the effects of unobserved factors that influence both private-school attendance and the outcome; and α 's and β 's are parameters or vectors of parameters to be estimated. The parameter of most interest is β_1 because it shows the impact of attending private school in year four or year five on the outcome.

As such, the estimation of the effect of private schooling on outcomes presented here is different from the estimation of the impact of scholarship treatment on the treated (IOT) generated via a

---

[2] A careful consideration of how the lottery instrument actually operates reveals why IV estimates with lottery instruments generate unbiased estimates of program effects. In the first stage of the analysis, the lottery variable assigns the same probability of private school attendance to each member of the treatment group and to each member of the control group, regardless of whether they actually attended a private school. A self-selected and elite subgroup of treatments and controls may have enrolled in private schools, but the lottery instrument essentially is ignorant to that fact. Since the lottery instrument distinguishes only between treatments and controls (who were randomly assigned) and cannot distinguish between private school enrollees and nonprivate school enrollees (who were self-selected), the use of the lottery as the instrumental variable in this analysis generates unbiased estimates of the effects of private schooling.

[3] The, P_i variable was coded based on school attended in the year of analysis and not based on whether the student ever attended a private school during the course of the evaluation. Coding the variable based on "ever private" would have created a serious missing data problem since any students who ever failed to respond to data collection would have to be treated as "unknown" or "missing" for purposes of classifying them as "ever private." This problem of missing longitudinal data on private school attendance does not affect the IOT estimates presented in the main text that were obtained via simple Bloom adjustment because (1) since treatment and control status is established at baseline and constant throughout the evaluation and (2) whether a treatment group student ever used an Opportunity Scholarship throughout the study is known precisely.

customized Bloom adjustment and presented in the main body of the report. For example, in cases where a treatment group student attended a private school but without the assistance of a scholarship, P_i would equal 1 for purposes of the IV analysis but 0 for purposes of the IOT analysis since the student is not being treated by the scholarship treatment. Similarly and more commonly, in cases where a control group student attended a private school without the assistance of a scholarship (or a sibling with a scholarship), P_i again would equal 1 for purposes of the IV analysis but 0 for purposes of the IOT analysis. This distinction is necessary because the treatment being evaluated is the scholarship program, not private schooling. Private schooling experiences obtained outside of the OSP are handled as a naturally occurring aspect of the counterfactual condition for purposes of evaluating the impact of the Program but are treated similarly to private schooling experiences obtained through the Program for purposes of this supplemental analysis of the effect of private schooling on student outcomes.

Applying IV analytic methods to the experimental data from the evaluation, we find a statistically significant relationship between enrollment in a private school at least four years after random assignment and the following outcomes for groups of students and parents (table D-1):

- Though the ITT results in chapter 3 found statistically significant treatment impacts in reading achievement for the subgroups of students who applied from not SINI 2003-05 schools and for students who applied with relatively higher academic performance, these same effects were not significant through the IV estimation of the outcomes of private schooling.

- Reading achievement for female students who were enrolled in private school at least four years after random assignment was 12.7 scale score points higher than that of like students who were not in private school at least four years after random assignment.

- Parental perceptions of school climate and safety were .41 standard deviations higher for those enrolled in private schools at least four years after random assignment than for those with children in public schools. For the not SINI 2003-05 subgroup, parents of students enrolled in a private school at least four years after random assignment had perceptions of school safety that were .54 standard deviations higher.

- Parents of students who attended private schools were more likely (20 percentage points) to give their child's school a grade of A or B than if the child was in a public school.

- The subgroup impacts reported in chapter 3 on parental satisfaction were larger and remained statistically significant through the IV estimation in three out of four cases. Parents of students were more likely to give their child's school a grade of A or B if they were in the not SINI 2003-05 subgroup (29 percentage points higher), the subgroup with relatively lower performance at baseline (37 percentage points higher), and the female subgroup (19 percentage points higher).The impact on the satisfaction of parents of students who applied with relatively higher academic performance, however, was not statistically significant in the IV analysis.

Table D-1. Private Schooling Effect Estimates for Statistically Significant ITT Results

Outcomes	IV Regression Estimate	*p*-value	Effect Size
Student Achievement			
Not SINI 2003-05: reading	9.24	.10	.26
Higher performing: reading	8.93	.11	.27
Female: reading	12.73*	.04	.37
School Safety and Climate: Parents			
Full sample	1.40**	.01	.41
Not SINI 2003-05	1.75**	.00	.54
School Satisfaction: Parents Gave School a Grade of A or B			
Full sample	.20**	.01	.42
Not SINI 2003-05	.26**	.00	.55
Lower performance	.37*	.03	.74
Higher performance	.14	.06	.30
Female	.19*	.04	.40

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for reading = 1,252. Reading sample weights were used. Difference displayed in terms of scale scores. Valid *N* for school safety = 1,175. Parent survey weights were used. Valid *N* for school grade = 1,178. Parent survey weights were used. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects. Effect sizes are in terms of standard deviations.

D.2 Sensitivity Testing of Instrumental Variable Analysis Models

As with the results of the offer of a scholarship reported in chapter 3, we subject the results of the original IV estimation of private schooling effects to two sensitivity tests involving different methodological approaches (table D-2).[4]

- The finding that reading achievement for female students who were enrolled in private school at least four years after random assignment was higher than that of like students who were not in private school is not sensitive to different analytic methods.
- The same two subgroups of students with significant ITT reading impacts that did not show statistically significant reading impacts in the main IV analysis (the not SINI 2003-05 and higher baseline performance subgroups) also did not demonstrate significant reading effects when using the trimmed sample or when clustering on current school attended.

[4] For a description of the sensitivity tests, see appendix C.

- The finding that parental perceptions of school climate and safety were higher for those who enrolled their child in private school—for the overall sample and for the not SINI 2003-05 subgroup—is not sensitive to different analytic methods.

- The overall finding that parental satisfaction is higher for those who enrolled their child in a private school is not sensitive to different analytic methods.

- The subgroups with significant parental satisfaction impacts in the main IV analysis were not sensitive to different analytic approaches. Additionally, the impact of attending a private school for parents of students who entered the Program with relatively higher baseline performance was significant when using the trimmed sample.

Table D-2. Private Schooling Achievement Effects and *P*-Values with Different Specifications, 2008-09

	Original IV Estimate		Trimmed Sample		Clustering on Current School	
Outcomes	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Student Achievement						
Not SINI 2003-05: reading	9.24	.10	9.34	.07	9.24	.11
Higher performing: reading	8.93	.11	9.69	.06	8.93	.08
Female: reading	12.73*	.04	14.22*	.01	12.73*	.05
School Safety and Climate: Parents						
Full sample	1.40**	.01	1.75**	.00	1.40*	.01
Not SINI 2003-05	1.75**	.00	1.92**	.00	1.75**	.00
School Satisfaction: Parents Gave School a Grade of A or B						
Full sample	.20**	.01	.22**	.00	.20*	.01
Not SINI 2003-05	.26**	.00	.22**	.00	.26**	.00
Lower performance	.37*	.03	.35*	.01	.37*	.02
Higher performance	.14	.06	.17*	.02	.14	.09
Female	.19*	.04	.19*	.02	.19*	.03

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for reading = 1,250; trimmed sample valid *N* = 1,198. Reading sample weights were used. Difference displayed in terms of scale scores. Valid *N* for school safety = 1,175; trimmed sample valid *N* = 1,126. Parent survey weights were used. Valid *N* for school grade = 1,178; trimmed sample valid *N* = 1,128. Parent survey weights were used. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

In conclusion, the IV results show evidence of a positive impact of 12.7 scale score points for female students who attended private schools at least four years after random assignment. Parental perceptions of school climate and safety were .41 standard deviations higher for those enrolled in private schools at least four years after random assignment. Finally, parents of students who attended private

schools were more likely (20 percentage points) to give their child's school a grade of A or B than if the child was in a public school.

Appendix E
Detailed ITT Tables

Table E-1. Academic Achievement: ITT Impacts in Reading, 2008-09

	Mean Differences				Regression-Based Impact Estimates		
Student Achievement	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	645.92	645.24	0.67	.83	3.90	.06	.11
	(36.29)	(36.30)	(3.13)		(2.06)		(36.30)
Subgroups							
SINI 2003-05	658.10	656.41	1.68	.74	1.08	.76	.03
	(34.24)	(34.75)	(5.07)		(3.46)		(34.75)
Not SINI 2003-05	638.20	637.45	.74	.85	5.80*	.02	.16
	(36.32)	(36.62)	(3.92)		(2.50)		(36.62)
Difference	19.90	18.96	.94	.88	-4.72	.27	-.13
	(3.71)	(5.39)	(6.44)		(4.24)		(36.30)
Lower performance	629.43	628.27	1.15	.81	1.18	.74	.04
	(30.68)	(32.27)	(4.84)		(3.50)		(32.27)
Higher performance	654.70	652.61	2.09	.59	5.18*	.04	.15
	(34.72)	(33.49)	(3.83)		(2.49)		(33.49)
Difference	-25.28	-24.34	-.94	.88	-4.00	.35	-.11
	(3.60)	(5.19)	(6.17)		(4.23)		(36.30)
Male	641.63	640.33	1.30	.79	2.45	.44	.07
	(34.08)	(36.01)	(4.88)		(3.16)		(36.01)
Female	650.40	649.38	1.02	.81	5.27*	.05	.15
	(37.75)	(34.19)	(4.13)		(2.66)		(34.19)
Difference	-8.77	-9.05	.28	.97	-2.81	.50	-.08
	(3.57)	(5.24)	(6.47)		(4.13)		(36.30)

NOTE: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for reading = 1,328. Reading sample weights used.

Table E-2. Academic Achievement: ITT Impacts in Math, 2008-09

	Mean Differences				Regression-Based Impact Estimates		
Student Achievement	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	641.00	643.36	-2.36	.46	.70	.71	.02
	(31.62)	(33.80)	(3.19)		(1.89)		(33.80)
Subgroups							
SINI 2003-05	656.90	657.05	-.15	.97	-0.38	.90	-.01
	(26.33)	(28.45)	(4.60)		(3.01)		(28.45)
Not SINI 2003-05	630.96	633.89	-2.93	.49	1.42	.56	.04
	(34.36)	(36.08)	(4.20)		(2.45)		(36.08)
Difference	25.94	23.16	2.78	.66	-1.80	.65	-.05
	(3.47)	(5.24)	(6.25)		(3.90)		(33.80)
Lower performance	630.69	631.16	-.47	.93	1.24	.71	.04
	(24.25)	(30.11)	(5.46)		(3.31)		(30.11)
Higher performance	646.43	648.59	-2.16	.57	.49	.83	.01
	(32.33)	(32.75)	(3.83)		(2.29)		(32.75)
Difference	-15.75	-17.44	1.69	.80	.75	.85	.02
	(3.82)	(5.58)	(6.64)		(4.01)		(33.80)
Male	638.30	640.70	-2.40	.62	-1.11	.68	-.04
	(31.66)	(31.27)	(4.80)		(2.67)		(31.27)
Female	643.81	645.64	-1.83	.67	2.40	.37	.07
	(30.86)	(34.14)	(4.25)		(2.65)		(34.14)
Difference	-5.51	-4.94	-.57	.93	-3.50	.35	-.10
	(3.70)	(5.19)	(6.41)		(3.74)		(33.80)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for math = 1,330. Math sample weights used.

Table E-3. High School Graduation: ITT Impacts, 2008-09

High School Graduation	Mean Differences: Treatment (S.D./S.E.)	Mean Differences: Control (S.D./S.E.)	Mean Differences: T-C Difference (S.E.)	Mean Differences: *p*-value	Regression-Based Impact Estimates: Estimated Impact (S.E.)	Regression-Based Impact Estimates: *p*-value	Regression-Based Impact Estimates: Effect Size (S.D.)
Full sample	.82	.70	.12*	.03	.12**	.01	.26
	(.39)	(.46)	(.05)		(.05)		(.46)
Subgroups							
SINI 2003-05	.80	.66	.13*	.03	.13*	.01	.28
	(.40)	(.47)	(.06)		(.05)		(.47)
Not SINI 2003-05	.87	.82	.06	.59	.07	.46	.19
	(.34)	(.38)	(.11)		(.10)		(.38)
Difference	-.08	-.14	.07	.58	.06	.59	.13
	(.09)	(.06)	(.12)		(.10)		(.46)
Lower performance	.68	.49	.14	.07	.12	.12	.23
	(.47)	(.50)	(.08)		(.08)		(.50)
Higher performance	.89	.79	.13	.07	.14*	.02	.35
	(.31)	(.41)	(.07)		(.06)		(.41)
Difference	-.26	-.26	.00	.99	-.03	.80	-.06
	(.09)	(.08)	(.11)		(.10)		(.46)
Male	.75	.66	.08	.25	.07	.26	.14
	(.43)	(.47)	(.07)		(.06)		(.47)
Female	.89	.75	.18*	.03	.20**	.01	.46
	(.31)	(.44)	(.09)		(.07)		(.44)
Difference	-.17	-.07	-.11	.35	-.15	.18	-.34
	(.08)	(.07)	(.13)		(.13)		(.46)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid $N = 316$. Sample weights used.

Table E-4. Parental Perceptions of School Climate and Safety: ITT Impacts, 2008-09

Parental Perceptions of School Danger	Mean Differences: Treatment (S.D./S.E.)	Mean Differences: Control (S.D./S.E.)	Mean Differences: T-C Difference (S.E.)	Mean Differences: *p*-value	Regression-Based Impact Estimates: Estimated Impact (S.E.)	Regression-Based Impact Estimates: *p*-value	Regression-Based Impact Estimates: Effect Size (S.D.)
Full sample	7.93	7.46	.47*	.03	.48**	.02	.14
	(3.25)	(3.36)	(.21)		(.21)		(3.36)
Subgroups							
SINI 2003-05	7.41	7.32	.09	.81	.10	.77	.03
	(3.52)	(3.45)	(.38)		(.35)		(3.45)
Not SINI 2003-05	8.27	7.56	.71**	.01	.73**	.01	.22
	(3.01)	(3.30)	(.26)		(.26)		(3.30)
Difference	-.85	-.24	-.62	.18	-.63	.16	-.19
	(.27)	(.39)	(.46)		(.45)		(3.36)
Lower performance	7.75	7.35	.41	.30	.48	.20	.14
	(3.39)	(3.48)	(.39)		(.37)		(3.48)
Higher performance	8.02	7.51	.51	.05	.48	.06	.15
	(3.17)	(3.31)	(.26)		(.26)		(3.31)
Difference	-.27	-.17	-.10	.83	-.00	.99	-.00
	(.27)	(.40)	(.47)		(.45)		(3.36)
Male	8.05	7.66	.39	.20	.38	.20	.12
	(3.09)	(3.29)	(.30)		(.30)		(3.29)
Female	7.81	7.32	.49	.11	.56	.06	.17
	(3.41)	(3.41)	(.31)		(.30)		(3.41)
Difference	.24	.34	-.10	.82	-.18	.68	-.05
	(.24)	(.35)	(.43)		(.43)		(3.36)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Effect sizes are in terms of standard deviations. Valid $N = 1,224$. Parent survey weights used.

Table E-5. Student Reports of School Climate and Safety: ITT Impacts, 2008-09

Student Perceptions of School Danger	Mean Differences: Treatment (S.D./S.E.)	Mean Differences: Control (S.D./S.E.)	Mean Differences: T-C Difference (S.E.)	Mean Differences: *p*-value	Regression-Based Impact Estimates: Estimated Impact (S.E.)	Regression-Based Impact Estimates: *p*-value	Regression-Based Impact Estimates: Effect Size (S.D.)
Full sample	6.18	6.02	.16	.32	.15	.33	.07
	(1.75)	(2.03)	(.16)		(.15)		(2.03)
Subgroups							
SINI 2003-05	6.30	6.02	.28	.21	.26	.21	.12
	(2.13)	(2.13)	(.22)		(.21)		(2.13)
Not SINI 2003-05	6.09	6.02	.07	.74	.06	.76	.03
	(1.96)	(1.96)	(.21)		(.21)		(1.96)
Difference	.21	-.00	.21	.49	.20	.51	.10
	(.16)	(.26)	(.31)		(.30)		(2.03)
Lower performance	6.17	5.82	.35	.24	.28	.36	.12
	(1.82)	(2.24)	(.30)		(.30)		(2.24)
Higher performance	6.18	6.11	.07	.71	.09	.62	.04
	(1.71)	(1.92)	(.18)		(.17)		(1.92)
Difference	-.01	-.29	.28	.42	.20	.58	.10
	(.19)	(.30)	(.35)		(.35)		(2.03)
Male	6.06	5.95	.11	.67	.12	.66	.05
	(1.86)	(2.25)	(.26)		(.26)		(2.25)
Female	6.29	6.07	.22	.24	.18	.32	.09
	(1.61)	(1.86)	(.19)		(.18)		(1.86)
Difference	-.22	-.12	-.10	.75	-.06	.85	-.03
	(.17)	(.27)	(.32)		(.32)		(2.03)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Effect sizes are in terms of standard deviations. Valid N = 1,054. Student survey weights used. Survey given to students in grades 4-12.

Table E-6. Parental Satisfaction: ITT Impacts on Parents Who Gave School a Grade of A or B, 2008-09

Parents Who Gave School a Grade of A or B	Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	.76	.68	.08**	.01	.08**	.00	.18
	(.43)	(.47)	(.03)		(.03)		(.47)
Subgroups							
SINI 2003-05	.70	.67	.03	.46	.03	.51	.07
	(.46)	(.47)	(.05)		(.05)		(.47)
Not SINI 2003-05	.80	.69	.12**	.00	.12**	.00	.27
	(.40)	(.46)	(.04)		(.04)		(.46)
Difference	-.10	-.02	-.09	.18	-.10	.12	-.21
	(.04)	(.05)	(.07)		(.07)		(.47)
Lower performance	.73	.61	.11*	.02	.10*	.05	.20
	(.44)	(.49)	(.05)		(.05)		(.49)
Higher performance	.78	.71	.07	.06	.08*	.03	.17
	(.42)	(.45)	(.04)		(.04)		(.45)
Difference	-.05	-.10	.04	.47	.02	.75	.04
	(.04)	(.05)	(.06)		(.06)		(.47)
Male	.75	.67	.08	.06	.08	.06	.17
	(.43)	(.47)	(.04)		(.04)		(.47)
Female	.77	.69	.09*	.03	.09*	.03	.19
	(.42)	(.46)	(.04)		(.04)		(.46)
Difference	-.02	-.01	-.01	.91	-.01	.90	-.01
	(.04)	(.04)	(.06)		(.06)		(.47)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for school grade = 1,227. Parent survey weights used. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Table E-7. Parental Satisfaction: ITT Impacts on Average Grade Parent Gave School, 2008-09

Average Grade Parent Gave School (5.0 Scale)	Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	4.04	3.83	.21**	.00	.21**	.00	.21
	(0.94)	(1.02)	(.07)		(.06)		(1.02)
Subgroups							
SINI 2003-05	3.92	3.86	.05	.60	.04	.69	.04
	(1.00)	(.97)	(.10)		(.10)		(.97)
Not SINI 2003-05	4.13	3.81	.32**	.00	.33**	.00	.32
	(.88)	(1.05)	(.08)		(.08)		(1.05)
Difference	-.21	.05	-.26*	.05	-.29*	.02	-.29
	(.08)	(.12)	(.13)		(.13)		(1.02)
Lower performance	3.96	3.66	.30*	.01	.27*	.02	.25
	(.98)	(1.10)	(.12)		(.12)		(1.10)
Higher performance	4.09	3.91	.18*	.02	.19*	.01	.19
	(.91)	(.97)	(.08)		(.07)		(.97)
Difference	-.14	-.25	.12	.42	.08	.54	.08
	(.08)	(.13)	(.14)		(.14)		(1.02)
Male	3.99	3.81	.18	.05	.18	.06	.17
	(.95)	(1.03)	(.09)		(.09)		(1.03)
Female	4.10	3.85	.25**	.01	.25**	.01	.25
	(.91)	(1.00)	(.09)		(.09)		(1.00)
Difference	-.11	-.04	-.07	.60	-.07	.58	-.07
	(.07)	(.11)	(.13)		(.13)		(1.02)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTE: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for school grade = 1,227. Parent survey weights used.

Table E-8. Parental Satisfaction: ITT Impacts on School Satisfaction Scale, 2008-09

School Satisfaction Scale	Mean Differences: Treatment (S.D./S.E.)	Mean Differences: Control (S.D./S.E.)	Mean Differences: T-C Difference (S.E.)	Mean Differences: *p*-value	Regression-Based Impact Estimates: Estimated Impact (S.E.)	Regression-Based Impact Estimates: *p*-value	Regression-Based Impact Estimates: Effect Size (S.D.)
Full sample	26.13	23.77	2.36**	.00	2.34**	.00	.28
	(7.58)	(8.46)	(.55)		(.53)		(8.46)
Subgroups							
SINI 2003-05	25.11	23.66	1.45	.09	1.50	.06	.19
	(7.51)	(7.98)	(.85)		(.81)		(7.98)
Not SINI 2003-05	26.79	23.85	2.94**	.00	2.91**	.00	.33
	(7.54)	(8.79)	(.71)		(.68)		(8.79)
Difference	-1.68	-.19	-1.49	.17	-1.41	.18	-.17
	(.63)	(.97)	(1.10)		(1.04)		(8.46)
Lower performance	25.20	22.95	2.25*	.02	2.30*	.01	.26
	(8.19)	(8.91)	(.98)		(.93)		(8.91)
Higher performance	26.62	24.11	2.51**	.00	2.36**	.00	.29
	(7.81)	(8.25)	(.67)		(.63)		(8.25)
Difference	-1.42	-1.16	-.26	.83	-.06	.96	-.01
	(.64)	(1.00)	(1.19)		(1.11)		(8.46)
Male	26.22	24.21	2.01**	.01	1.95**	.01	.24
	(7.29)	(8.21)	(.77)		(.74)		(8.21)
Female	26.04	23.42	2.62**	.00	2.69**	.00	.31
	(7.86)	(8.64)	(.81)		(.77)		(8.64)
Difference	.18	.79	-.61	.59	-.74	.49	-.09
	(.57)	(.97)	(1.13)		(1.09)		(8.46)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for parent satisfaction = 1,233. Parent survey weights used.

Table E-9. Student Satisfaction: ITT Impacts on Students Who Gave School a Grade of A or B, 2008-09

Students Who Gave School a Grade of A or B	Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	.69 (.46)	.72 (.45)	-.02 (.03)	.48	-.03 (.03)	.43	-.06 (.45)
Subgroups							
SINI 2003-05	.66 (.47)	.65 (.48)	-.01 (.05)	.86	.00 (.05)	.99	.00 (.47)
Not SINI 2003-05	.72 (.45)	.76 (.43)	-.04 (.05)	.41	-.05 (.05)	.29	-.12 (.43)
Difference	-.07 (.04)	-.10 (.06)	.03 (.07)	.65	.05 (.06)	.46	.11 (.45)
Lower performance	.72 (.45)	.67 (.47)	.04 (.06)	.47	.04 (.06)	.54	.08 (.47)
Higher performance	.68 (.47)	.74 (.44)	-.06 (.04)	.17	-.06 (.04)	.14	-.13 (.44)
Difference	.04 (.04)	-.07 (.06)	.09 (.06)	.16	.09 (.06)	.18	.20 (.45)
Male	.71 (.46)	.73 (.44)	-.02 (.05)	.65	-.03 (.05)	.55	-.07 (.44)
Female	.68 (.47)	.71 (.46)	-.03 (.05)	.55	-.02 (.05)	.62	-.05 (.46)
Difference	.03 (.04)	.02 (.06)	.01 (.07)	.95	-.01 (.07)	.91	-.02 (.45)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for school grade =1,001. Student survey weights used. Impact estimates reported for the dichotomous variable "students who gave school a grade of A or B" are reported as marginal effects. Survey given to students in grades 4-12.

Table E-10. Student Satisfaction: ITT Impacts on Average Grade Student Gave School, 2008-09

Average Grade Student Gave School (5.0 Scale)	Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	3.94	4.01	-.06	.42	-.07	.30	-.08
	(.97)	(.94)	(.08)		(.07)		(.94)
Subgroups							
SINI 2003-05	3.81	3.90	-.10	.33	-.09	.38	-.10
	(.99)	(.90)	(.10)		(.10)		(.90)
Not SINI 2003-05	4.04	4.08	-.04	.71	-.06	.51	-.07
	(.94)	(.96)	(.11)		(.10)		(.96)
Difference	-.23	-.18	-.06	.70	-.02	.87	-.02
	(.09)	(.11)	(.15)		(.14)		(.94)
Lower performance	4.07	3.99	.08	.51	.09	.46	.10
	(.92)	(.91)	(.12)		(.12)		(.91)
Higher performance	3.88	4.01	-.13	.16	-.15	.08	-.16
	(.99)	(.95)	(.09)		(.08)		(.95)
Difference	.19	-.02	.21	.16	.24	.09	.25
	(.09)	(.13)	(.15)		(.14)		(.94)
Male	4.00	4.03	-.03	.79	-.06	.61	-.06
	(.96)	(.95)	(.11)		(.11)		(.95)
Female	3.89	3.99	-.10	.35	-.09	.36	-.09
	(.97)	(.93)	(.11)		(.10)		(.93)
Difference	.11	.04	.07	.65	.03	.82	.04
	(.09)	(.12)	(.16)		(.15)		(.94)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for school grade = 1,001. Student survey weights used. Survey given to students in grades 4-12.

Table E-11. Student Satisfaction: ITT Impacts on School Satisfaction Scale, 2008-09

School Satisfaction Scale	Mean Differences: Treatment (S.D./S.E.)	Mean Differences: Control (S.D./S.E.)	Mean Differences: T-C Difference (S.E.)	Mean Differences: *p*-value	Regression-Based Impact Estimates: Estimated Impact (S.E.)	Regression-Based Impact Estimates: *p*-value	Regression-Based Impact Estimates: Effect Size (S.D.)
Full sample	34.12	34.11	.01	.98	-.09	.84	-.02
	(6.45)	(6.09)	(.46)		(.44)		(6.09)
Subgroups							
SINI 2003-05	33.94	34.68	-.73	.27	-.88	.62	-.15
	(6.56)	(5.82)	(.66)		(.64)		(5.82)
Not SINI 2003-05	34.24	33.68	.56	.37	.48	.42	.08
	(6.38)	(6.25)	(.63)		(.60)		(6.25)
Difference	-.29	1.00	-1.29	.16	-1.36	.12	-.22
	(.60)	(.73)	(.91)		(.88)		(6.09)
Lower performance	33.78	33.73	.05	.95	-.09	.91	-.01
	(5.86)	(6.05)	(.78)		(.74)		(6.05)
Higher performance	34.30	34.28	.02	.98	-.10	.86	-.02
	(6.74)	(6.10)	(.58)		(.56)		(6.10)
Difference	-.52	-.55	.03	.97	.01	.99	.00
	(.60)	(.76)	(.99)		(.93)		(6.09)
Male	35.02	34.44	.58	.36	.61	.33	.10
	(6.02)	(6.07)	(.63)		(.63)		(6.07)
Female	33.18	33.86	-.68	.29	-.72	.24	-.12
	(6.75)	(6.09)	(.64)		(.62)		(6.09)
Difference	1.84	.58	1.26	.16	1.33	.13	.22
	(.55)	(.71)	(.89)		(.88)		(6.09)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for student satisfaction =1,078. Student survey weights used. Survey given to students in grades 4-12.

Table E-12. Parental Perceptions of School Climate and Safety: ITT Impacts on Individual Items, 2008-09

Parental Safety: NOT Current School Problems	Mean Differences				Regression-Based Impact Estimates		
	Treatment	Control	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Kids destroying property	.79	.73	.06*	.03	.07*	.02	.15
			(.03)		(.03)		(.44)
Kids being late for school	.64	.57	.07*	.04	.07*	.03	.14
			(.03)		(.03)		(.49)
Kids missing classes	.73	.67	.06	.06	.06*	.04	.13
			(.03)		(.03)		(.47)
Fighting	.70	.63	.07*	.03	.07*	.03	.14
			(.03)		(.03)		(.48)
Cheating	.81	.74	.07**	.01	.08**	.00	.18
			(.03)		(.03)		(.44)
Racial conflict	.85	.85	-.00	.91	-.00	1.00	-.00
			(.02)		(.02)		(.36)
Guns or other weapons	.85	.81	.03	.15	.04	.11	.09
			(.02)		(.02)		(.39)
Drug distribution	.85	.81	.04	.12	.04	.06	.11
			(.03)		(.03)		(.39)
Drug and alcohol use	.84	.81	.03	.26	.03	.16	.08
			(.02)		(.02)		(.39)
Teacher absenteeism	.83	.78	.05*	.05	.05*	.04	.11
			(.03)		(.03)		(.47)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N*s for the individual items range from 1,190 to 1,203.

Table E-13. Student Reports of School Climate and Safety: ITT Impacts on Individual Items, 2008-09

Student Safety – Did NOT Happen This Year	Mean Differences				Regression-Based Impact Estimates		
	Treatment	Control	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Something stolen from desk, locker, or other place	.53	.59	-.06 (.04)	.12	-.07 (.04)	.07	-.14 (.49)
Taken money or things from me by force or threats	.89	.86	.03 (.03)	.22	.03 (.02)	.20	.08 (.35)
Offered drugs	.87	.85	.02 (.03)	.45	.02 (.02)	.26	.07 (.35)
Physically hurt by another student	.81	.80	.02 (.03)	.65	.02 (.03)	.57	.04 (.40)
Threatened with physical harm	.87	.82	.05 (.03)	.09	.04 (.03)	.09	.11 (.38)
Seen anyone with a real/toy gun or knife at school	.81	.76	.05 (.03)	.14	.06* (.03)	.04	.15 (.43)
Been bullied at school	.86	.81	.04 (.03)	.16	.04 (.03)	.18	.09 (.39)
Been called a bad name	.47	.49	-.02 (.04)	.69	-.03 (.04)	.48	-.06 (.50)

*Statistically significant at the 95 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N*s for the individual items range from 1,022 to 1,047.

Table E-14. Parental Satisfaction: ITT Impacts on Individual Scale Items, 2008-09

School Satisfaction Scale: Items (1-4 Scale)	Mean Differences				Regression-Based Impact Estimates		
	Treatment	Control	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Location	3.32	3.10	.22**	.00	.21**	.00	.25
			(.06)		(.05)		(.86)
Safety	3.27	3.04	.23**	.00	.23**	.00	.26
			(.06)		(.05)		(.91)
Class sizes	3.18	2.95	.23**	.00	.22**	.00	.25
			(.06)		(.06)		(.90)
School facilities	3.18	2.99	.19**	.00	.19**	.00	.23
			(.05)		(.05)		(.84)
Respect between teachers and students	3.19	3.01	.18**	.00	.19**	.00	.21
			(.06)		(.06)		(.88)
Teachers inform parents of students' progress	3.21	3.02	.19**	.00	.19**	.00	.21
			(.06)		(.06)		(.90)
Amount students can observe religious traditions	3.26	2.95	.31**	.00	.31**	.00	.28
			(.07)		(.07)		(1.09)
Parental support for the school	3.22	2.95	.27**	.00	.27**	.00	.31
			(.06)		(.05)		(.87)
Discipline	3.21	2.95	.26**	.00	.26**	.00	.28
			(.06)		(.06)		(.92)
Academic quality	3.21	3.00	.21**	.00	.20**	.00	.23
			(.06)		(.06)		(.87)
Racial mix of students	3.11	2.98	.13*	.02	.12*	.03	.14
			(.06)		(.06)		(.88)
Services for students with special needs	3.82	3.67	.15	.06	.15	.06	.12
			(.08)		(.08)		(1.19)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N*s for the individual items range from 1,166 to 1,210.

Table E-15. Student Satisfaction: ITT Impacts on Individual Scale Items, 2008-09

School Satisfaction Scale: Items (1-4 Scale)	Mean Differences				Regression-Based Impact Estimates		
	Treatment	Control	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Students are proud to go to this school	2.91	3.00	-.09 (.07)	.19	-.11 (.06)	.07	-.14 (.83)
There is a lot of learning at this school	3.34	3.35	-.01 (.05)	.88	.00 (.05)	.99	.00 (.68)
Rules of behavior are strict	3.22	3.23	-.01 (.07)	.89	-.01 (.07)	.94	-.01 (.86)
When students misbehave, they receive the same treatment	2.85	2.72	.13 (.08)	.11	.09 (.08)	.28	.09 (1.05)
I feel safe	3.05	3.01	.04 (.09)	.68	.04 (.08)	.59	.04 (1.05)
People at my school are supportive	3.11	3.09	.01 (.07)	.84	.00 (.06)	.97	.00 (.82)
I do not feel isolated at my school	3.14	3.11	.03 (.09)	.70	.04 (.08)	.59	.04 (1.00)
I enjoy going to school	3.17	3.23	-.06 (.07)	.40	-.07 (.07)	.32	-.08 (.84)
Students behave well with the teachers	2.78	2.75	.03 (.07)	.65	.00 (.07)	.98	.00 (.88)
Students do their homework	2.42	2.47	-.05 (.07)	.47	-.02 (.07)	.74	-.03 (.91)
I rarely feel made fun of by other students	3.02	2.97	.05 (.09)	.61	.07 (.08)	.36	.07 (1.09)
Other students seldom disrupt class	2.30	2.18	.12 (.08)	.13	.11 (.08)	.13	.11 (1.00)
Students who misbehave rarely get away with it	2.62	2.70	-.08 (.08)	.36	-.05 (.08)	.52	-.05 (1.04)
Most of my teachers really listen to what I have to say	3.10	3.08	.02 (.08)	.79	-.01 (.08)	.90	-.01 (.94)
My teachers are fair	3.00	3.00	.00 (.07)	.98	-.01 (.07)	.92	-.01 (.87)
My teachers expect me to succeed	3.53	3.52	.01 (.06)	.85	.02 (.06)	.78	.02 (.67)
Some teachers punish cheating when they see it	3.17	3.27	-.11 (.08)	.16	-.08 (.07)	.29	-.08 (.93)

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N*s for the individual items range from 961 to 1054.

Appendix F
Exploration of Whether Parents Get What They Seek From School Choice

This analysis examines why parents pursued school choice through the Opportunity Scholarship Program (OSP), based on their survey responses, and whether they appear to have found what they sought. Specifically, statistical models featuring interaction variables are used to determine if the students of parents who applied to the OSP primarily to enhance the academic experience of their child were more likely to experience Programmatic test score impacts than the students of parents who applied to the OSP for other reasons. Although the methods used to conduct this analysis were experimental, it is presented as an exploratory analysis because the question it addresses extends beyond the central evaluation question of what were the overall impacts of the OSP.

F.1 Research Question

The core question that motivates this analysis can be expressed in a number of ways. Technically, the question is whether the OSP has had heterogeneous achievement impacts on subgroups of students depending on whether a student's parents primarily sought academic quality in their child's school of choice. Put more colloquially, do parents tend to get what they choose for and not get what they don't choose for when choosing schools?

The question is central to the debate over how school choice programs are expected to work. The core theory surrounding market-based reforms in any policy area, including education, is that the industry will provide what customers demand. If parents do not seek school characteristics that are associated with socially desirable outcomes such as academic achievement or school safety, then programs designed to increase parental school choice are unlikely to be efficacious. Moreover, if parental customers do not receive what they seek from school choice, then the central motivation for choice is undermined.

Many evaluations of school choice programs have documented the reasons that parents give for seeking school choice and the most important characteristics of the schools that they seek. As John Witte wrote about the pioneering Milwaukee voucher program:

> The leading reasons given for participating in the program were the perceived educational quality and teaching approach and style in private schools. The disciplinary environment and the general atmosphere that parents associated with those schools were the next most important factors (Witte 2000, p. 62).

Few studies, however, have taken the additional step of examining whether parental schooling preferences are subsequently manifested in greater student exposure to the educational conditions that are sought or actual outcomes (e.g., achievement gains) that drove parental preferences.

In one study of public school choice in New York City, Mark Schneider and his colleagues found that parents with distinct schooling preferences, when given new school choice opportunities, tended to place their children in schools rated relatively high on the schooling characteristic they most preferred, be it test scores or safety (Schneider, Teske, and Marschall 2000, pp. 168-169).

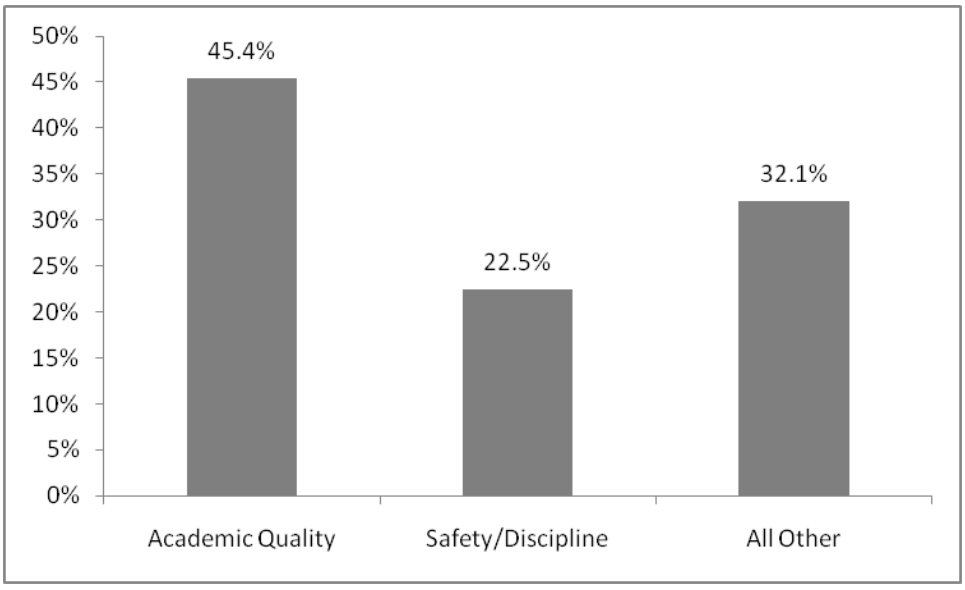
In another study of public school choice in Charlotte-Mecklenburg, Justine Hastings and her colleagues found that only the students of parents that expressed strong preferences for high-achievement schools clearly benefited academically one year after exercising school choice (Hastings, Kane, and Staiger 2009). No study of private school choice through vouchers or scholarships has yet explored the connection between what parents prefer and what outcomes they obtain from their school choice experience.

F.2 Data

The data for this analysis are taken from the OSP baseline application survey and the evaluation's analysis files for impact years 1, 2, 3, and at least four years. The key explanatory variable is the baseline application survey question that asked parents what is the single most important consideration in their choice of a school for their child. A total of 45.4 percent of parent respondents said that "academic quality" was their single most important consideration in choosing a school. "School safety" and "discipline," when combined into one thematic response category, were the reasons selected by 22.5 percent of respondents. The responses of the remaining 32.1 percent of parents were dispersed across the nine other answer categories (figure F-1).[1]

[1] The "other" responses and their frequencies were Location (13.8 percent), Class Size (6.7 percent), Services for Students with Special Needs (3.6 percent), Respect Between Teachers and Students (2.5 percent), Information Between School and Home (2.0 percent), Observance of Religious Tradition (1.2 percent), Parental Involvement (1.0 percent), School Facilities (0.9 percent), and Racial Mix of Students (0.5 percent).

Figure F-1. Parent Baseline Responses about the "Most Important Consideration" in Choosing a School



NOTES: Valid *N* = 1,774.

Data were missing regarding parental responses to the survey question for 534 students representing 23.1 percent of the impact sample. Since the students with missing data on this baseline question were approximately as likely to show up for outcome data collection as students without missing data on the question, the analysis sample for this exploratory analysis includes between 22 and 28 percent fewer observations across the impact years' samples as the analysis of impacts by subgroup presented in each year's impact reports. A smaller sample size means that this exploratory subgroup analysis will have less analytic power, and a greater chance of finding no significant impact when an impact actually exists, than the subgroup analysis in the main body of the report.

In order to maximize the available analytic power for this subgroup analysis, student observations were classified as "academic quality choosers" if the academic quality response category had been selected by their parents at baseline (45.4 percent of the sample) and "not academic quality choosers" if any alternative responses were selected (56.6 percent). The impact of the OSP on the academic outcomes for these two subgroups of student participants is the focus of this exploratory analysis.

F.3 Analysis

Once the data had been classified into subgroups based on the most important characteristic sought in schools, this exploratory analysis used the same format as the evaluation's standard analysis of heterogeneous impacts by subgroup (appendix section A.8). A single regression estimation is used to determine the separate impacts of the OSP on achievement outcomes for each student subgroup (e.g., academic quality choosers and not academic quality choosers) and for the difference in impact between the subgroups. The estimation of subgroup impacts took the following analytic form:

$$(1)\ Y_{ikt} = \mu + \tau T_{ikt} + \tau_B P_i * T_{ikt} + \sum_{j=2}^{b} \varphi^{j}_{is} + X_{ik}\gamma + \delta_R R_{it} + \delta_M M_{it} + \varepsilon_{ik,t}$$

where Y is test score outcomes in year t and P is an index for students whose parents most sought academic quality. The treatment effect τ in this specification is the average impact for students whose parents did not choose primarily based on academic quality whereas τP represents the difference between the treatment impacts on students whose parents did choose on academic quality compared to the impacts on those whose parents did not. X is the evaluation's standard set of demographic covariates and R and M are baseline reading and math scores. The variables of interest in the estimation are the impact of treatment for each of the individual subgroups separately and the interaction between being an "academic priority" student and treatment. A statistically significant coefficient on the interaction variable would indicate that the treatment impact was different for "academic priority" and "non-academic priority" choosers. A statistically significant coefficient for the treatment impact on academic quality choosers ($\tau + \tau P$) would signal achievement impacts from the OSP specifically on the subgroup of students whose parent's viewed academic quality as the most important feature of their child's school.

These analyses were performed separately on outcome data from years 1, 2, 3, and at least four years after random assignment. The outcomes estimated were reading and math achievement scores on the Stanford Achievement Test. The complete set of results provide some exploratory information regarding the question: "Did parents who value academic quality most of all get what they chose for in the OSP quickly, slowly, or not at all?"

F.4 Results

The academic impacts of the OSP on the test score outcomes of students did not differ significantly depending on whether or not a student's parent selected "academic quality" as his/her most desired school feature. Although the achievement impacts of the Program were larger for academic

choosers compared to non-academic choosers in seven of eight instances, none of the impact differences across the two subgroups were statistically significant at the 95 percent confidence level (table F-1).

Two of the 16 subgroup estimates of OSP achievement impacts were statistically significant and both of them involved subgroups of students whose parents said that they most valued academic achievement in schools. In year one, the math outcomes for academic choosers were 6.0 scale score points higher if they were in the treatment compared to the control group (ES = .18). The math gain due to scholarship use was 7.9 scale score points (ES = .24) for academic choosers in year one. In year three, academic choosers experienced a reading impact from the OSP of 5.4 scale score points (ES = .16), which amounted to a gain of 6.3 scale score points (ES = .19) from scholarship use. No subgroup achievement impacts were observed for non-academic choosers in any subject in any year, and no statistically significant OSP impacts on achievement for either subgroup were observed in the final year of the analysis, at least four years after random assignment.

Table F-1. Impact Estimates of the Offer and Use of a Scholarship on Academic Chooser Subgroups Across All Evaluation Years: Academic Achievement

	Reading						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Year 1							
Academic choosers	610.09	606.63	3.46	.10	4.56	.14	.23
Not academic choosers	608.10	607.09	1.01	.03	1.32	.04	.73
Difference	2.00	-.46	2.46	.07			.59
Year 2							
Academic choosers	624.43	623.01	1.42	.04	1.75	.05	.64
Not academic choosers	614.64	612.96	1.68	.05	2.11	.06	.61
Difference	9.79	10.05	-.26	-.01			.95
Year 3							
Academic choosers	642.30	636.87	5.43*	.16	6.32*	.19	.05
Not academic choosers	629.77	626.84	2.93	.08	3.45	.10	.36
Difference	12.53	10.03	2.50	.07			.52
At least 4 years							
Academic choosers	653.35	648.21	5.14	.17	6.04	.20	.10
Not academic choosers	644.55	642.32	2.24	.06	2.79	.08	.46
Difference	8.80	5.89	2.90	.09			.52

	Math						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Year 1							
Academic choosers	597.27	591.24	6.02*	.18	7.93*	.24	.02
Not academic choosers	593.87	592.15	1.72	.05	2.26	.07	.51
Difference	3.40	-.90	4.30	.13			.27
Year 2							
Academic choosers	618.63	617.40	1.23	.04	1.52	.05	.68
Not academic choosers	606.97	608.61	-1.65	-.05	-2.06	-.06	.56
Difference	11.66	8.78	2.88	.09			.49
Year 3							
Academic choosers	637.57	634.18	3.38	.12	3.93	.14	.16
Not academic choosers	622.98	624.70	-1.72	-.05	-2.02	-.06	.53
Difference	14.58	9.48	5.10	.16			.15
At least 4 years							
Academic choosers	646.48	642.43	4.05	.14	4.76	.16	.16
Not academic choosers	639.05	642.48	-3.43	-.10	-4.26	-.13	.26
Difference	7.73	-.05	7.48	.23			.08

*Statistically significant at the 95 percent confidence level.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Means are regression adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are in terms of standard deviations. Total valid *N* for Year 1 reading = 1,190, Year 1 math = 1,272; valid *N* for Year 2 reading = 1,226, Year 2 math = 1228; valid *N* for Year 3 reading = 1,126, Year 3 math = 1,135; valid *N* for Year 4 Reading = 1,015, Year 4 math = 1,014.

Appendix G
To What Extent Are Treatment Effects of the OSP Observed Across the Outcome Test-Score Distribution? Quantile Regression Analysis of the OSP

This exploratory analysis builds on the identification of the impact of the offer of the OSP on the average test score of students and examines the impacts across the entire outcome distribution of test scores. The primary focus of the main OSP evaluation on the mean impacts of the Program ignores the intense interest in impacts across the distribution of students on relevant characteristics, most notably academic performance. To examine that question, researchers typically assign students to different subgroups (such as weak and strong performers), and re-estimate the basic outcome regression within subgroups. Estimating mean impacts within subgroups is certainly a legitimate and useful way of testing for heterogeneous program impacts, and it has been part of the OSP impact evaluation over the past four years. An alternative, and arguably more comprehensive, means of examining heterogeneity in treatment effects is through quantile methods.

G.1 Research Question

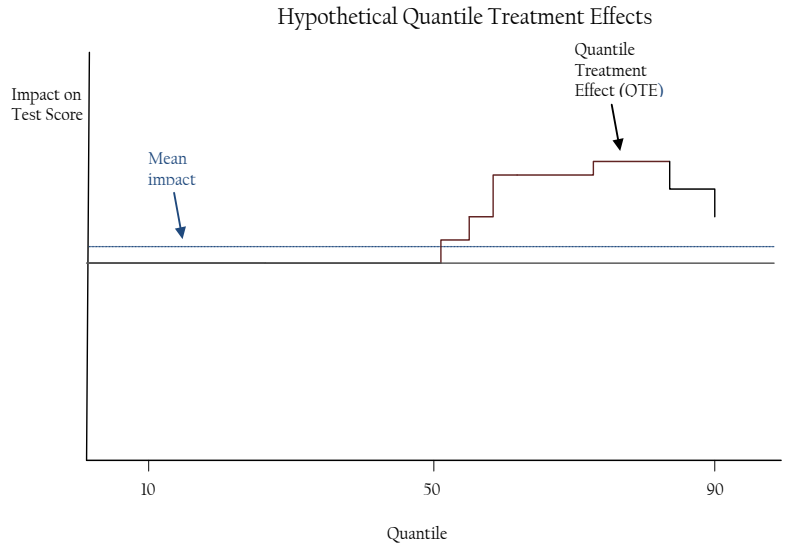
A full understanding of the impact of education interventions is informed by examining not only the average impact but also impacts across the entire distribution of key student characteristics. If test score gains are observed, are they concentrated among the best performers, the worst performers, or spread out so that all students perform better? The research question analyzed here, therefore, is to what extent are the treatment effects of the OSP observed on the overall distribution of test scores?

The methodological approach used is based on quantile regressions, which model the conditional distribution of the outcome variable (test score, in this case) as a function of observed characteristics. The approach is well established and has been applied to school quality (class size) and academic performance as well as the labor supply effects of welfare-to-work programs, CEO compensation and firm equity, and infant birth-weight and maternal behavior (Abrevaya 2001; Bassett and Chen 2001; Bitler, Gelbach and Hoynes 2006; Eide and Showalter 1998).

To appreciate the value of quantile regressions, consider the case in which the impact of a (private school) scholarship offer on math test scores is close to zero. Typically, this result is interpreted

as no impact, on average, for all students. This is because the treatment effect is assumed to be constant for all students. But, the average treatment effect could mask a great deal of heterogeneity across the (test score) distribution. In fact, a near zero average treatment effect is entirely consistent with very different treatment impacts across the distribution. Figure G-1 shows a case in which only students above the median show any kind of response to the offer of a scholarship. Another distribution might show negative treatment effects at the bottom and positive at the top. In both cases, the mean impact could be (near) zero.

Figure G-1. Hypothetical Quantile Treatment Effects



Estimating quantile treatment effects (QTEs) across the distribution of (reading and math) test scores allows us to capture this type of heterogeneity. It is then possible to assess whether the impact of the scholarship offer is constant across the distribution, or whether the offer leads to larger improvements in, say, reading in some parts of the distribution than in others.

G.2 Framework

This section introduces the concept of quantiles and develops the basic framework for interpreting quantile treatment effects. Consider the results of a standardized examination and a student who performs better than a share τ and worse than a share $(1-\tau)$ of his/her peers. This student scores at the τ^{th} quantile of the test distribution. Thus, the student whose test score is at the median performs better than half of students and worse than the other half of students. A student scoring better than 25 percent of the students and worse than 75 percent of his/her peers is said to be at the 0.25 quantile.

Average and Quantile Treatment Effects

The basic treatment effect estimated thus far is the average difference in test scores (ψ) across treatment status between students offered a scholarship and those not offered a scholarship. Let that difference be given by ($\psi_{\mathbf{T}}$-$\psi_{\mathbf{X}}$). The quantile treatment effect (QTE) at a given quantile $\boldsymbol{\tau}$ is given by the analogous difference in test score: ($\psi_{\mathbf{T},\boldsymbol{\tau}}$ $\psi_{\mathbf{X},\boldsymbol{\tau}}$). This is simply the difference in test scores across treatment status at the $\boldsymbol{\tau}$ quantile. This exercise could be repeated for each and any quantile across the distribution: the QTE at the 0.75 quantile would be generated by the difference in the 75th percentile of the treatment group test-score distribution and the 75th percentile of the control group test-score distribution. In principle, this exercise could be repeated for each of the 99 centiles of the distribution, and for different subgroups.

Quantile Regression

Quantile regressions generalize this concept and estimate treatment effects at different points along the distribution, what are referred to as conditional quantile functions. These models express the quantiles of the conditional distribution of test scores, in this case, as functions of observed covariates. In contrast, ordinary least squares (OLS) estimates conditional mean functions. It is important to note that the difference between OLS and quantile regression is not one of conditionality. Both can be and usually are conditioned on observable characteristics. Rather, the difference is that quantile regression estimates are distinct treatment impacts at various points along the distribution; whereas OLS only permits the estimation of a single average impact across the distribution.

The specification of the empirical model is therefore identical to the base model used in the OSP evaluation (appendix section A.8):

$$(1)\ Y_{it} = \alpha + \tau\, T_{it} + X_i \gamma + \delta_R R_{it} + \delta_M M_{it} + \delta_{DT} DT_{it} + \varepsilon_{it}.$$

where X_i is a vector of student and/or family characteristics measured at baseline and known to influence future academic achievement, and R_{it} and M_{it} refer to **baseline** reading and mathematics scores, respectively (each of the included covariates are described below). In this model, τ—the parameter of sole interest—represents the effect of scholarships on test scores for students in the Program, conditional on X_i and the baseline test scores.

Using this framework, we can ask, at any chosen quantile, how different are the corresponding scores of students offered scholarships and those not offered scholarships, given a

specification of the other conditioning variables. The idea therefore is to compare the coefficient $\hat{\tau}$ (on the treatment dummy) using OLS with the $\hat{\tau}'s$ estimated at different quantiles: $\hat{\tau}_{0.1}$ (10th percentile), $\hat{\tau}_{0.5}$ (median), etc. If the data show little difference in the coefficients across the distribution, showing that $\hat{\tau}$ is essentially unchanged, we would conclude that the impacts are homogenous and evenly spread out. If, however, the coefficients on the offer variable show significant impacts at some points and not others, then we could conclude the impact of the OSP is heterogeneous and localized over some range of outcomes.

Interpretation of Quantile Estimates

QTEs allow us to tell a story about impacts on the distribution, which may or may not match up with individual students. To appreciate this point and to interpret QTEs appropriately, recall that QTEs are estimated by differencing outcomes at particular points along the test score distribution. In other words, QTEs tell us how the "test-score distribution" changes when we assign OSP scholarships randomly. These estimates do not necessarily identify the impact of treatment for a given student. If we find, for example, that the OSP offer raises the lower decile of the test-score distribution, this does not necessarily mean that a low-performing student (at the lower decile without the OSP) is now a better student. Rather, all we could infer is that low-performing students in the OSP are better off than they would have been without the OSP. This distinction is subtle and has to do with whether the OSP preserves a student's rank in the test-score distribution. If the intervention is rank preserving, an increase in the lower decile makes those who would have been low-performers, in fact, better students in absolute terms. This is because students' relative status is unchanged. Otherwise, all we could say is that low-performing students, whoever they may be, are better off in relative terms due to the intervention.[91]

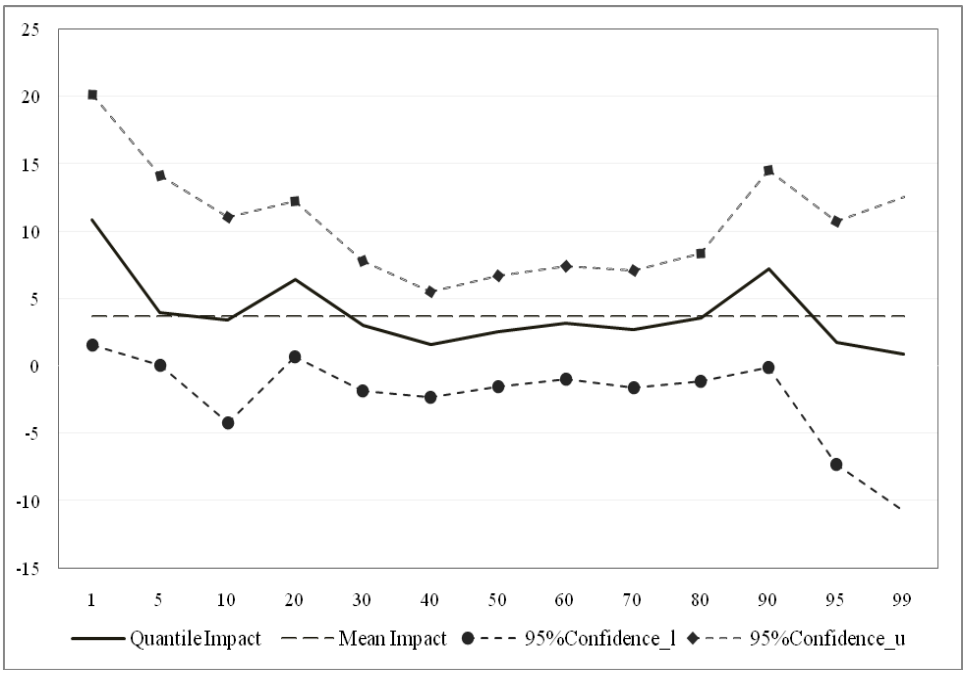
G.3 Results

Figures G-2 and G-3 graph the mean impact of the OSP, along with the quantile treatment effects (QTEs) and their associated 95 percent confidence interval. Broadly, the QTEs for reading and math suggest somewhat similar stories about who might have been affected by the OSP, although the QTEs were generally found to be statistically insignificant at conventional levels. In other words, one

[91] The impact on the distribution will map onto the impact on students only if an intervention is rank-preserving. Simply put, that means the intervention does not alter the ranking or ordering of individual students along the test score distribution. If it is not the case, all that can be said is that at a specific quantile, students offered an OSP scholarship perform better (or worse) than those not offered a scholarship from the same quantile, whoever they may be. More formally, rank preservation is an assumption that a student's spot in the test-score distribution is invariant to program assignment. Random assignment does not guarantee that because any student can move up and down the distribution for various reasons (say, if ability does not map 1:1 onto rank in the distribution). So long as this movement can occur for non-program reasons, the assumption fails to hold, and we are limited in how much we can say from the evidence. For example, we could not say low-performing students (measured by their baseline test scores) are performing better because of the OSP.

cannot reject that the impact of the OSP is consistent across the distribution. There are caveats to this finding, however. Recall that the mean impact of the OSP in year four was 3.90 (ES = 0.11) for reading and not statistically significant; this is shown in the flat dashed line of figure G-2 for reference purposes. The QTEs for reading have observably larger impacts around the 1st, 20th, and 90th quantiles and observably smaller impacts at most other points, but these impacts are statistically not different from each other. In fact, like the mean impact, most of these impacts estimated along the distribution are statistically insignificant.

Figure G-2. Quantile Impacts of the OSP on Reading After At Least Four Years

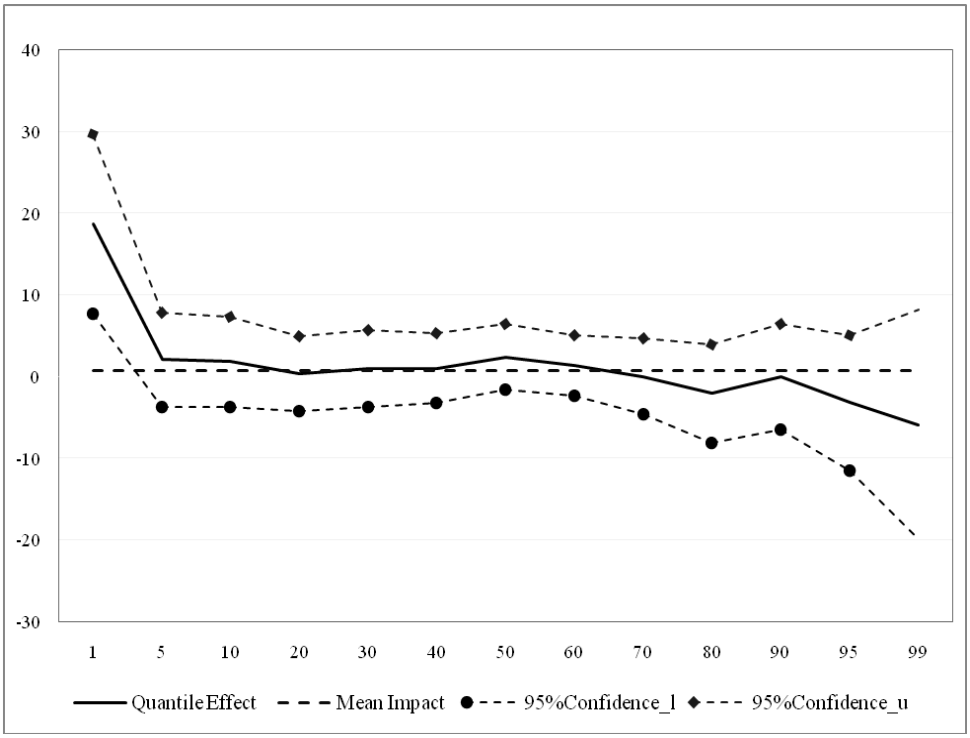


NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Impacts are regression adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Valid *N* = 1,328.

The mean impact of the OSP on math in year four was 0.70 (ES = 0.02), also not statistically significant. Again the mean impact is shown in the flat dashed line of figure G-3. The data show the lack of any meaningful impact on math test scores, which is entirely consistent across the test-score distribution, with the exception of the 0.01 quantile (the very lowest performers). The estimated impact at that point of the distribution is large, by any standard (18.65), and statistically significant. This exploratory examination does not identify the extent to which these results can be assigned to individual students. In addition, it is notable that the impact disappears entirely by the 5th percentile, suggesting that

the high impact at the very bottom may be anomalous. Finally, overall, we again cannot reject that the mean impact of the OSP is representative of the impact across the distribution.

Figure G-3. Quantile Impacts of the OSP on Mathematics After At Least Four Years



NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Impacts are regression adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Valid $N = 1{,}330$.

Appendix H
Intermediate Outcome Measures

An analysis of the impacts of the Opportunity Scholarship Program (OSP) on intermediate outcomes was conducted to determine if certain factors might be candidates for mediators of the impact of the treatment on student achievement. Previous research regarding possible influences on student achievement tends to focus on four general types of factors: educational supports provided in the home, the extent to which students are enthusiastic about learning and engaged in school activities, the nature of the instructional program delivered to students, and the general school environment. Twenty-four specific intermediate outcomes were identified and measured among each of these four categories, as described below.

H.1 Home Educational Supports

The first grouping of mediating factors is Home Educational Supports. As a general category this set of factors seeks to assess the impact that the OSP may have had on the educational supports provided by a student's family. The category contains four potential mediators: Parental Involvement, Parent Aspirations, Out-of-school Tutor Usage, and School Transit Time.

1. Parental Involvement

Parental involvement seeks to measure how active a parent is in his/her child's education. The variable is an Item Response Theory (IRT) scale composed of responses from the parent survey to three questions about how often during the school year the parent volunteered in school, attended a school organization meeting, or accompanied students on class trips. Parental involvement was chosen because it has been shown to vary between public and private schools (Bauch and Goldring 1995; Bryk et al. 1993; Witte 1993) and to have a relationship to student achievement (Fan and Chen 2001; Henderson and Berla 1994; Sui-Chu and Willms 1996; Wu and Qi 2006).

The parental involvement variable ranges from .75 to 7.52 with a mean of 2.83 and a standard deviation of 1.91. The Cronbach's Alpha for the parental involvement scale is .75.[1]

[1] Cronbach's Alpha is a measure of the consistency and reliability of a scale (Spector 1991). The critical value of Cronbach's Alpha is .70, above which a scale is considered to have a satisfactory level of reliability.

2. Parent Aspirations

Parent aspirations is a measure of how many years of education a parent expects his/her child to receive. Taken from the parent survey, the variable is treated as a continuous variable with the following values:

a. Some high school, but will not graduate=11
b. Complete high school=13
c. Attend a two-year college=14
d. Attend a four-year college=15
e. Obtain a certificate=15
f. Obtain a bachelor's degree=17
g. Obtain a master's degree or other higher degree=19.

Parent aspirations is one of two measures of educational aspirations used in the intermediate outcomes analysis, along with student aspirations. These factors were chosen for analysis because educational aspirations are associated with student achievement (Fan and Chen 2001; Natriello and McDill 1986; Singh et al. 1995; Wu and Qi 2006). The measure of parent aspirations ranges from 11 to 19. The mean of parent aspirations is 17.31, and the standard deviation is 2.25.

3. Out-of-school Tutor Usage

Out-of-school tutor usage, taken from the parent survey, is a measure of whether the student receives help on schoolwork from tutoring held outside of the child's school. Out-of-school tutor usage is one of two measures of tutor usage, along with in-school tutor usage. These measures were chosen because tutor usage has been shown to vary across public and private schools (Howell et al. 2006) and to be associated with student achievement (Cohen, Kulik, and Kulik 1982; Ritter 2000). As a dichotomous variable, out-of-school tutor usage can take the value of 0 or 1. The mean value of out-of-school tutor usage is .10, and the standard deviation is .30.

4. School Transit Time

School transit time seeks to measure the length of the school commute that a parent provides for his/her child. The variable is taken from the parent survey and is an ordinal variable with values assigned as:

a. Under 10 minutes= 1
b. 11-20 minutes=2
c. 21-30 minutes=3
d. 31-45 minutes=4
e. 46 minutes to an hour=5
f. More than one hour=6

This variable was chosen because it has been shown to be associated with student achievement (Dolton, Marcenaro, and Navarro 2003). Commuting time has a negative effect on student achievement because it is unproductive time that is not being spent on student learning. The school transit time variable ranges from 1 to 6 with a mean of 2.67 and a standard deviation of 1.39.

H.2 Student Motivation and Engagement

Student motivation and engagement is a grouping of potential mediators that seeks to measure the impact of the OSP on the personal investment of students in their own education. The category contains six components: Student Aspirations, Attendance, Tardiness, Reads for Fun, Engagement in Extracurricular Activities, and Frequency of Homework (measured in days per week).

1. Student Aspirations

Student aspirations is a measure of how many years of education the student expects to receive. Taken from the student survey, the variable is treated as a continuous variable with the following values:

a. Some high school, but will not graduate=11
b. Complete high school=13
c. Attend a 2-year college=14
d. Attend a 4-year college=15
e. Obtain a certificate=15
f. Obtain a bachelor's degree=17
g. Obtain a master's degree or other higher degree=19.

Student aspirations is one of two measures of educational aspirations, along with parent aspirations. These factors were chosen as potential mediators because educational aspirations have been shown to vary across public and private schools (Plank, Schiller, Schneider, and Coleman 1993) and to be associated

with student achievement (Natriello and McDill 1986; Singh, et al. 1995). The student aspirations variable ranges from 11 to 19 years of education. The mean of student aspirations is 16.82, and the standard deviation is 1.91.

2. Attendance

Attendance is a measure of how often the student has missed school. Attendance is an ordinal variable taken from the parent survey that measures how many school days the student missed in the preceding month:

a. None=0
b. 1-2 Day =1
c. 3-4 Days=2
d. 5 or more days=3

Attendance was chosen as a possible mediator because it has been shown to be associated with student achievement (Lamdin 1996). The attendance variable ranges from 0 to 3. Attendance has a mean of .81 and a standard deviation of .85.

3. Tardiness

Tardiness is a measure of how often the student has missed school. Taken from the parent survey and measuring how many days the student arrived late in the preceding month, tardiness is an ordinal variable with the following values:

a. None=0
b. 1-2 Days=1
c. 3-4 Days=2
d. 5 or more days=3

Tardiness was chosen as a possible mediator because it has been associated with student achievement (Mulkey, Crain, and Harrington 1992). The tardiness variable ranges from 0 to 3. Tardiness has a mean of .50 and a standard deviation of .81.

4. Reads for Fun

Reads for fun seeks to measure whether the student reads for personal enjoyment. The variable is taken from the student survey and is a dichotomous variable that equals 1 if the student claims to read for fun and 0 if not. The variable was chosen as a possible mediator because it has been shown to be associated with student achievement (Mulkey et al. 1992; Mullis, Martin, Gonzalez, and Kennedy 2003). Reading for fun has a mean of .38 and a standard deviation of .49.

5. Engagement in Extracurricular Activities

Engagement in extracurricular activities seeks to measure the student's involvement in programs that are not a required part of the school's educational program. Taken from the student survey, the variable is a count of the number of activities in which a student reports participating from a list of five items that include community service and volunteer work, boy or girl scouts, and other such activities. The variable was chosen as a possible mediator because it has been shown to be associated with student achievement (McNeal 1995). Engagement in extracurricular activities ranges from 0 to 5 with a mean of 2.33 and a standard deviation of 1.28.

6. Frequency of Homework

Frequency of homework measures how many nights during a typical week the student reported doing homework. Taken from the student survey, the variable is a count, from zero to five, of the number of school days per week that the student said that he or she typically works on homework. Frequency of homework was chosen because it has been shown to vary across public and private schools (Hoffer, Greeley, and Coleman 1985) and to be associated with student achievement (Natriello and McDill 1986; Rumberger and Palardy 2005; Rutter, Maughan, Mortimore, and Ouston 1979; Wolf and Hoople 2006). The mean of frequency of homework is 3.72, and the standard deviation is 1.55.

H.3 Instructional Characteristics

Instructional characteristics is a grouping of factors that seeks to capture features of the educational program experienced by students in the treatment group compared to those in the control group. There are 10 possible mediating factors in the category: Student/Teacher Ratio, Teacher Attitude, Ability Grouping, Availability of Tutors, In-School Tutor Usage, Programs for Students with Learning Problems, Programs for English Language Learners, Programs for Advanced Learners, Before- or After-School Programs, and Enrichment Programs.

1. Student/Teacher Ratio

Student/teacher ratio is the number of students at the child's school divided by the full-time equivalency of classroom teachers at the school. The variable is a continuous measure taken from the National Center for Educational Statistics' Common Core of Data (NCES CCD) and Private School Universe Survey (NCES PSS). Student/teacher ratio was chosen as a possible mediator because it has been shown to vary across public and private schools and to be associated with student achievement (Arum 1996; Nye, Hedges, and Konstantopoulos 2000). Student/teacher ratio ranges from .90 to 35.80. The mean of student/teacher ratio is 11.80, and the standard deviation is 3.83.

2. Teacher Attitude

Teacher attitude measures the extent to which students report being treated with consideration by their classroom teachers. Taken from the student survey, the variable is an IRT scale that combines student evaluations of four items involving how well teachers listen to them, are fair, expect students to succeed, and encourage students to do their best. Teacher attitude was chosen because it has been shown to differ across public and private schools (Ballou and Podgursky 1998; Gruber et al. 2002) and to be associated with student achievement (Card and Krueger 1992; Hanushek 1971; Wayne and Youngs 2003; Wolf and Hoople 2006). Teacher attitude ranges from .47 to 10.32 with a mean of 2.77 and a standard deviation of 1.97. The Cronbach's Alpha for teacher attitude is .78.

3. Ability Grouping

Ability grouping is a measure of the ways in which a school differentiates instruction based on student ability level. Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school differentiates instruction by either organizing classes with similar content but different difficulty levels or organizing classes with different content. The variable equals 0 if neither of these methods of differentiating instruction is used. Ability grouping was chosen as a possible mediator because it has been shown both to vary across public and private schools and to be associated with student achievement (Lee and Bryk 1988). Ability grouping has a mean of .76 and a standard deviation of .43.

4. Availability of Tutors

Availability of tutors measures whether the school a student attends has tutors available for its students. Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school makes tutors available to its students and 0 if not. Though not entirely comparable to the two

measures of tutor *usage* analyzed as possible mediators, this variable was chosen for similar reasons: availability of tutors has been shown to vary across public and private schools (Howell et al. 2006) and to be associated with student achievement (Cohen et al. 1982; Ritter 2000). Availability of tutors has a mean of .61 and a standard deviation of .49.

5. In-school Tutor Usage

In-school tutor usage is a measure of whether a child actually uses a tutor provided by the school. Taken from the parent survey, the measure is a dichotomous variable that equals 1 if the student uses a school-provided tutor and 0 if not. In-school tutor usage is one of two measures of tutor usage, along with out-of-school tutor usage, analyzed as possible mediators. These measures were chosen because tutor usage has been shown to vary across public and private schools (Howell et al. 2006) and to be associated with student achievement (Cohen et al. 1982; Ritter 2000). In-school tutor usage has a mean of .27 and a standard deviation of .45.

6. Programs for Students with Learning Problems

Programs for students with learning problems is an indicator of an affirmative response to a question in the principal survey about providing distinctive instructional activities for students with learning problems. This measure of special school programs was chosen for analysis because the availability of such programs has been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the principal gave an affirmative response when asked if his or her school offers such programs and 0 if not. The mean of this variable is .81, and the standard deviation is .39.

7. Programs for English Language Learners

Programs for English language learners is an indicator of an affirmative response to a question in the principal survey about providing special instruction for non-English speakers. This measure of special programs was chosen for analysis because the availability of such programs has been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the principal gave an affirmative response when asked if his or her school offers such programs and 0 if not. The mean of this variable is .45, and the standard deviation is .50.

8. Programs for Advanced Learners

Programs for advanced learners is a measure of whether a principal reports offering any of three items in the principal survey: Advanced Placement (AP) courses, International Baccalaureate (IB) programs, and special instructional programs for advanced learners or a gifted and talented program. The variable is one of four potential mediators that measure special school programs. These factors were chosen for analysis because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school reported offering any of the three types of programs and 0 if it reported offering none. The mean of this variable is .57, and the standard deviation is .50.

9. Before-/After-School Programs

Before- or after-school care programs was taken from the principal survey and is a dichotomous variable that equals 1 if the school offers a program for students either before or after school and equals 0 if not. The variable is one of four that measure the availability of special school programs. These programmatic variables were chosen for the mediator analysis because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). The mean of before/after school programs is .90, indicating that almost every student in the impact sample attended a school with a before- or after-school program, and the standard deviation is .30.

10. Enrichment Programs

Enrichment programs is a count of how many programs a school reports offering out of three items: foreign language programs, music programs, and arts programs. The variable is one of four that measures the availability of special school programs. These factors were chosen for analysis as possible mediators because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). The enrichment programs variable ranges from 0 to 3 with a mean of 2.65 and a standard deviation of .63.

H.4 School Environment

School environment is the final conceptual grouping of potential mediators of the OSP treatment. The category includes certain characteristics of schools that might influence achievement but

are not explicitly established by school policy. The category has four components: Parent/School Communication, School Size, Percent Non-White, and Peer Classroom Behavior.

1. School Communication Policies

School communication policies measures the number of distinct policies a school has regarding required school-parent communications. Taken from the principal survey, the variable is a count of the number of communication policies a school reports having out of four items: informing parents of their childs' grades halfway through the grading period, notifying parents when students are sent to the office the first time for disruptive behavior, sending parents weekly or daily notes about their child's progress, and sending parents a newsletter about what is occurring in their child's school or school system. School communication policies was chosen for analysis as a possible mediator because it has been shown to vary across public and private schools (Bauch and Goldring 1995; Howell et al. 2006) and to be associated with student achievement (Henderson and Berla 1994; Sui-Chu and Willms 1996). The variable for parent/school communication ranges from 1 to 4 with a mean of 3.13 and a standard deviation of .82.

2. School Size

School size is the total reported student enrollment in the attended school and is taken from the NCES CCD (http://nces.ed.gov/ ccd) and the NCES PSS (http://nces.ed.gov/surveys/pss). The variable was included in the analysis as a possible mediator because it has been shown to vary across public and private schools (Wasley 2002) and to be associated with student achievement (Lee and Loeb 2000; Sander 1999). School size ranges from 16 to 1,604. The mean of school size is 454.78, and the standard deviation is 359.42.

3. Percent Non-White

Percent non-white is the percentage of enrolled students at the attended school who were identified as American Indian/Alaska Native, Asian Pacific Islander, Black non-Hispanic, and Hispanic. The data for the variable were taken from the NCES CCD and PSS. The variable was included in the analysis as a possible mediator because it has been shown to vary across public and private schools (Plank et al. 1993; Reardon and Yun 2002; Schneider and Buckley 2002) and to be associated with student achievement (Coleman 1966; Coleman 1990; Hanushek, Kain, and Rivkin 2002; Nielsen and Wolf 2002). Percent non-white ranges from .10 to 1.00 with a mean of .93 and a standard deviation of .19.

4. Peer Classroom Behavior

Peer classroom behavior seeks to measure the degree to which the other students in the child's class are well behaved. Taken from the student survey, the variable is an IRT scale composed of student evaluations of five statements about their peers: whether students behave well with teachers, students neglect their homework, students tease them, other students often disrupt class, and students get away with bad behavior. Peer classroom behavior was chosen for the analysis as a possible mediator because it has been shown to vary across public and private schools (Harris 1998; Lee, Dedrick, and Smith 1991) and to be associated with student achievement (Card and Krueger 1992). Peer classroom behavior ranges from 3.23 to 12.70 with a mean of 8.26 and a standard deviation of 2.13. The Cronbach's Alpha for peer classroom behavior is .70.

H.5 Impacts on Intermediate Outcomes for Student Subgroups

Table H-1. Effect Sizes for Subgroups: Home Educational Supports (ITT), 2008-09

Subgroup:	Parental Involvement	Parent Aspirations	Out-of-school Tutor Usage	School Transit Time #
Overall Impact	**-.04**	**-.00**	**.11**	**1.03**
SINI 2003-05	.04	-.05	.16	1.17
Not SINI 2003-05	-.08	.03	.09	.94
Difference	.12	-.07	.06	1.26
Lower performance	-.08	.04	.07	.76
Higher performance	-.02	-.03	.13	1.18
Difference	-.06	.06	-.05	.64
Male	-.02	-.04	.08	1.01
Female	-.06	.02	.14	1.04
Difference	.04	-.06	-.06	.97

Effect sizes for categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher category response. For a complete description of the treatment and control group response patterns for these variables see appendix F.5.

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Parental Involvement = 1,228, including: SINI 2003-05 *N* = 472, Not SINI 2003-05 *N* = 756, Lower performance *N* = 404, Higher performance *N* = 824, Male *N* = 598, Female *N* = 630. Valid *N* for Parent Aspirations = 1,184, including: SINI 2003-05 *N* = 459, Not SINI 2003-05 *N* = 725, Lower performance *N* = 391, Higher performance *N* = 793, Male *N* = 558, Female *N* = 626. Out-of-school Tutor Usage *N* = 1,180, including: SINI 2003-05 *N* = 457, Not SINI 2003-05 *N* = 723, Lower performance *N* = 392, Higher performance *N* = 788, Male *N* = 565, Female *N* = 615. Valid *N* for School Transit Time = 1,235, including: SINI 2003-05 *N* = 476, Not SINI 2003-05 *N* = 759, Lower performance *N* = 410, Higher performance *N* = 825, Male *N* = 293, Female *N* = 642. Impact estimates are regression adjusted using a consistent set of baseline covariates. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for the dichotomous variable "Out-of-school Tutor Usage" are reported as marginal effects.

Table H-2. Effect Sizes for Subgroups: Student Motivation and Engagement (ITT), 2008-09

Subgroup:	Student Aspirations	Attendance#	Tardiness#	Reads for Fun	Engagement in Extra-curricular Activities	Frequency of Homework
Overall Impact	**-.05**	**1.32***	**.93**	**-.11**	**.04**	**.03**
SINI 2003-05	-.15	1.23	.98	-.14	-.03	.10
Not SINI 2003-05	.01	1.39*	.89	-.08	.09	-.03
Difference	-.14	.88	1.11	-.07	-.12	.14
Lower performance	-.04	1.66*	1.06	-.10	-.12	.13
Higher performance	-.05	1.19	.86	-.11	.12	-.02
Difference	.00	1.40	1.23	.01	-.24	.16
Male	-.11	1.27	1.00	-.12	-.03	.08
Female	.02	1.38	.86	-.10	.09	-.02
Difference	-.14	.92	1.15	-.01	-.12	.11

* Statistically significant at the 95 percent confidence level.

Effect sizes for categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher category response.

NOTES: Because these subgroup analyses are merely exploratory, tests for multiple comparisons were not conducted. Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Student Aspirations = 1,012, including: SINI 2003-05 *N* = 485, Not SINI 2003-05 *N* = 527, Lower performance *N* = 328, Higher performance *N* = 684, Male *N* = 485, Female *N* = 527. Valid *N* for Attendance = 1,192, including: SINI 2003-05 *N* = 455, Not SINI 2003-05 *N* = 737, Lower performance *N* = 392, Higher performance *N* = 800, Male *N* = 574, Female *N* = 618. Valid *N* for Tardiness = 1,182, including: SINI 2003-05 *N* = 456, Not SINI 2003-05 *N* = 726, Lower performance *N* = 384, Higher performance *N* = 798, Male *N* = 567, Female *N* = 615. Valid *N* for Reads for Fun = 1,070, including: SINI 2003-05 *N* = 511, Not SINI 2003-05 *N* = 559, Lower performance *N* = 342, Higher performance *N* = 728, Male *N* = 515, Female *N* = 555. Valid *N* for Engagement in Extracurricular Activities = 1,016, including: SINI 2003-05 *N* = 486, Not SINI 2003-05 *N* = 530, Lower performance *N* = 329, Higher performance *N* = 687, Male *N* = 486, Female *N* = 530. Valid *N* for Frequency of Homework = 1,046, including: SINI 2003-05 *N* = 500, Not SINI 2003-05 *N* = 546, Lower performance *N* = 330, Higher performance *N* = 716, Male *N* = 500, Female *N* = 546. Impact estimates are regression adjusted using a consistent set of baseline covariates. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for "Attendance" and "Tardiness" are derived from ordered logistic regression. Impact estimates for the dichotomous variable "Reads for Fun" are reported as marginal effects. Data regarding student aspirations, reads for fun, engagement in extracurricular activities, and frequency of homework were drawn from student surveys and therefore limited to students in grades 4-12.

Table H-3. Effect Sizes for Subgroups: Instructional Characteristics (ITT), 2008-09

Subgroup:	Student/ Teacher Ratio	Teacher Attitude	Ability Grouping	School Provides Tutors	In- School Tutor Usage	Programs for Learning Problems	Programs for English Language Learners	Programs for Advanced Learners	Before- or After-School Programs	Enrichment Programs
Overall Impact	**.01**	**-.01**	**.28****	**-.11**	**.08**	**-.52****	**-.57****	**-.53****	**.08**	**.08**
SINI 2003-05	-.03	.18	.21	-.03	.13	-.61**	-.64**	-.81**	.08	.08
Not SINI 2003-05	.05	-.12	.33**	-.17	.05	-.47**	-.52**	-.39**	.07	.08
Difference	-.07	.28	-.13	.14	.07	-.04	-.14	-.35	.02	.01
Lower performance	-.02	-.11	.26*	-.10	.02	-.82**	-.52**	-.65**	.13*	-.26
Higher performance	.03	.05	.29**	-.12	.11	-.45**	-.59**	-.48**	.04	.19*
Difference	-.04	-.17	-.01	.02	-.09	-.15	.07	-.19	.09	-.42*
Male	-.03	-.18	.27**	-.36**	.09	-.64**	-.57**	-.55**	-.02	-.06
Female	.06	.16	.29**	.09	.07	-.41**	-.58**	-.51**	.18**	.19
Difference	-.09	-.34	-.01	-.45**	.02	-.19	.00	-.06	-.29*	-.26

*Statistically significant at the 95 percent confidence level

**Statistically significant at the 99 percent confidence level.

NOTES: Because these subgroup analyses are merely exploratory, tests for multiple comparisons were not conducted. Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for Student/Teacher Ratio = 904, including: SINI 2003-05 *N* = 353, Not SINI 2003-05 *N* = 551, Lower performance *N* = 274, Higher performance *N* = 630, Male *N* = 452, Female *N* = 452. Valid *N* for Teacher Attitude = 1,060, including: SINI 2003-05 *N* = 504, Not SINI 2003-05 *N* = 557, Lower performance *N* = 335, Higher performance *N* = 726, Male *N* = 507, Female *N* = 554. Valid *N* for Ability Grouping = 893, including: SINI 2003-05 *N* = 341, Not SINI 2003-05 *N* = 552, Lower performance *N* = 262, Higher performance *N* = 631, Male *N* = 432, Female *N* = 461. Valid *N* for School Provides Tutors = 946, including: SINI 2003-05 *N* = 361, Not SINI 2003-05 *N* = 585 Lower performance *N* = 284, Higher performance *N* = 662, Male *N* = 462, Female *N* = 484. Valid *N* for In-School Tutor Usage = 1,197, including: SINI 2003-05 *N* = 461, Not SINI 2003-05 *N* = 736, Lower performance *N* = 398, Higher performance *N* = 799, Male *N* = 574, Female *N* = 623. Valid *N* for Learning Problems = 951, including: SINI 2003-05 *N* = 362, Not SINI 2003-05 *N* = 589, Lower performance *N* = 287, Higher performance *N* = 664, Male *N* = 464, Female *N* = 487. Valid *N* for Programs for English Language Learners = 951, including: SINI 2003-05 *N* = 362, Not SINI 2003-05 *N* = 589, Lower performance *N* = 287, Higher performance *N* = 666, Male *N* = 465, Female *N* = 488. Valid *N* for Programs for Advanced Learners = 953, including: SINI 2003-05 *N* = 330, Not SINI 2003-05 *N* = 536, Lower performance *N* = 244, Higher performance *N* = 622 Male *N* = 431, Female *N* = 435. Valid *N* for Before- or After-School Programs = 953, including: SINI 2003-05 *N* = 363, Not SINI 2003-05 *N* = 590, Lower performance *N* = 287, Higher performance *N* = 666, Male *N* = 465, Female *N* = 488. Valid *N* for Enrichment Programs = 953, including: SINI 2003-05 *N* = 363, Not SINI 2003-05 *N* = 590, Lower performance *N* = 287, Higher performance *N* = 666, Male *N* = 465, Female *N* = 488. Impact estimates are regression adjusted using a consistent set of baseline covariates. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for the dichotomous variables "School Provides Tutors" and "Ability Grouping" are reported as marginal effects. Data regarding Teacher Attitude and the Challenge of Classes were drawn from student surveys and therefore limited to students in grades 4-12.

Table H-4. Effect Sizes for Subgroups: School Environment (ITT), 2008-09

Subgroup:	School Communication Policies	School Size	Percent Non-White	Peer Classroom Behavior
Overall Impact	**.24****	**-.36****	**-.14***	**.05**
SINI 2003-05	.30*	-.30**	-.23	-.03
Not SINI 2003-05	.20*	-.40**	-.12	.10
Difference	.10	.08	.00	-.13
Lower performance	.36*	-.41**	-.09	.04
Higher performance	.20*	-.33**	-.16	.05
Difference	.17	-.14	.13	-.01
Male	.10	-.27**	-.28*	.17
Female	.38**	-.44**	-.04	-.06
Difference	-.28	.21	-.20	.23

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

NOTES: Because these subgroup analyses are merely exploratory, tests for multiple comparisons were not conducted. Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Valid *N* for School Communication Policies = 936, including: SINI 2003-05 *N* = 359, Not SINI 2003-05 *N* = 577, Lower performance *N* = 280, Higher performance *N* = 656, Male *N* = 459, Female *N* = 477. Valid *N* for School Size = 1,011, including: SINI 2003-05 *N* = 403, Not SINI 2003-05 *N* = 608, Lower performance *N* = 314, Higher performance *N* = 697, Male *N* = 494, Female *N* = 517. Valid *N* for Percent Non-White = 994, including: SINI 2003-05 *N* = 402, Not SINI 2003-05 *N* = 592, Lower performance *N* = 313, Higher performance *N* = 681, Male *N* = 487, Female *N* = 507. Valid *N* for Peer Classroom Behavior = 1,061, including: SINI 2003-05 *N* = 505, Not SINI 2003-05 *N* = 557, Lower performance *N* = 336, Higher performance *N* = 726, Male *N* = 508, Female *N* = 554. Impact estimates are regression adjusted using a consistent set of baseline covariates. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Data regarding Peer Classroom Behavior were drawn from student surveys and therefore limited to students in grades 4-12.

H.6 Impacts on Intermediate Outcomes for Ordinal Variables by Variable Category

Table H-5. Marginal Effects of Treatment: School Transit Time for Full Sample, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.22	-.00
11-20 minutes	.31	.31	-.00
21-30 minutes	.20	.19	.00
31-45 minutes	.14	.14	.00
46 minutes to an hour	.10	.10	.00
More than one hour	.04	.04	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .02. Effect is not statistically significant (*p*-value = .84).

Table H-6. Marginal Effects of Treatment: School Transit Time for SINI 2003-05 Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.18	.20	-.03
11-20 minutes	.29	.30	-.01
21-30 minutes	.21	.20	.01
31-45 minutes	.17	.15	.01
46 minutes to an hour	.12	.10	.01
More than one hour	.04	.04	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .16. Effect is not statistically significant (*p*-value = .44).

Table H-7. Marginal Effects of Treatment: School Transit Time for Not SINI 2003-05 Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.24	.23	.01
11-20 minutes	.32	.32	.01
21-30 minutes	.19	.19	-.00
31-45 minutes	.13	.14	-.01
46 minutes to an hour	.09	.09	-.00
More than one hour	.03	.03	-.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.07. Effect is not statistically significant (*p*-value = .66).

Table H-8. Marginal Effects of Treatment: School Transit Time for Lower Performing Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.17	.04
11-20 minutes	.31	.28	.02
21-30 minutes	.19	.20	-.02
31-45 minutes	.14	.17	-.02
46 minutes to an hour	.11	.13	-.02
More than one hour	.04	.05	-.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.27. Effect is not statistically significant (*p*-value = .22).

Table H-9. Marginal Effects of Treatment: School Transit Time for Higher Performing Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.24	-.03
11-20 minutes	.31	.32	-.02
21-30 minutes	.20	.19	.01
31-45 minutes	.15	.13	.01
46 minutes to an hour	.10	.08	.01
More than one hour	.03	.03	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .16. Effect is not statistically significant (*p*-value = .25).

Table H-10. Marginal Effects of Treatment: School Transit Time for Male Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.21	.22	-.00
11-20 minutes	.31	.31	-.00
21-30 minutes	.20	.19	.00
31-45 minutes	.15	.14	.00
46 minutes to an hour	.10	.10	.00
More than one hour	.04	.04	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .01. Effect is not statistically significant (*p*-value = .96).

Table H-11. Marginal Effects of Treatment: School Transit Time for Female Subgroup, 2008-09

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.22	-.01
11-20 minutes	.31	.31	-.00
21-30 minutes	.20	.20	.00
31-45 minutes	.14	.14	.00
46 minutes to an hour	.10	.09	.00
More than one hour	.03	.03	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .04. Effect is not statistically significant (*p*-value = .82).

Table H-12. Marginal Effects of Treatment: Parent-Reported Attendance for Full Sample, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.39	.46	-.07
1-2 days	.39	.37	.03
3-4 days	.16	.13	.03
5 or more days	.06	.05	.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .28. Effect is statistically significant (*p*-value = .03).

Table H-13. Marginal Effects of Treatment: Parent-Reported Attendance for SINI 2003-05 Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.39	.44	-.05
1-2 days	.39	.37	.02
3-4 days	.15	.13	.02
5 or more days	.06	.05	.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .20. Effect is not statistically significant (*p*-value = .33).

Table H-14. Marginal Effects of Treatment: Parent-Reported Attendance for Not SINI 2003-05 Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.38	.46	-.08
1-2 days	.39	.36	.03
3-4 days	.16	.12	.03
5 or more days	.07	.05	.02

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .33. Effect is statistically significant (*p*-value = .04).

Table H-15. Marginal Effects of Treatment: Parent-Reported Attendance for Lower Performing Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.35	.47	-.12
1-2 days	.41	.36	.05
3-4 days	.17	.12	.05
5 or more days	.07	.05	.02

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .51. Effect is statistically significant (*p*-value = .03).

Table H-16. Marginal Effects of Treatment: Parent-Reported Attendance for Higher Performing Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.41	.45	-.04
1-2 days	.38	.37	.02
3-4 days	.15	.13	.02
5 or more days	.06	.05	.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .17. Effect is not statistically significant (*p*-value = .25).

Table H-17. Marginal Effects of Treatment: Parent-Reported Attendance for Male Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.41	.47	-.06
1-2 days	.38	.36	.02
3-4 days	.15	.12	.02
5 or more days	.06	.05	.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .24. Effect is not statistically significant (*p*-value = .16).

Table H-18. Marginal Effects of Treatment: Parent-Reported Attendance for Female Subgroup, 2008-09

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.37	.45	-.08
1-2 days	.40	.37	.03
3-4 days	.16	.13	.03
5 or more days	.07	.05	.02

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .32. Effect is not statistically significant (*p*-value = .10).

Table H-19. Marginal Effects of Treatment: Parent-Reported Tardiness for Full Sample, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.65	.63	.02
1-2 days	.23	.24	-.01
3-4 days	.07	.08	-.00
5 or more days	.05	.05	-.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.08. Effect is not statistically significant (*p*-value = .59).

Table H-20. Marginal Effects of Treatment: Parent-Reported Tardiness for SINI 2003-05 Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.63	.63	.00
1-2 days	.24	.24	-.00
3-4 days	.08	.08	-.00
5 or more days	.05	.05	-.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.02. Effect is not statistically significant (*p*-value = .95).

Table H-21. Marginal Effects of Treatment: Parent-Reported Tardiness for Not SINI 2003-05 Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.66	.63	.03
1-2 days	.23	.24	-.02
3-4 days	.07	.08	-.01
5 or more days	.05	.05	-.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.12. Effect is not statistically significant (*p*-value = .50).

Table H-22. Marginal Effects of Treatment: Parent-Reported Tardiness for Lower Performing Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.56	.57	-.01
1-2 days	.28	.27	.01
3-4 days	.10	.09	.00
5 or more days	.07	.06	.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = .06. Effect is not statistically significant (*p*-value = .81).

Table H-23. Marginal Effects of Treatment: Parent-Reported Tardiness for Higher Performing Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.69	.66	.03
1-2 days	.21	.23	-.02
3-4 days	.06	.07	-.01
5 or more days	.04	.04	-.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.15. Effect is not statistically significant (*p*-value = .38).

Table H-24. Marginal Effects of Treatment: Parent-Reported Tardiness for Male Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.63	.63	.00
1-2 days	.24	.24	-.00
3-4 days	.07	.07	-.00
5 or more days	.05	.05	-.00

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.00. Effect is not statistically significant (*p*-value = .99).

Table H-25. Marginal Effects of Treatment: Parent-Reported Tardiness for Female Subgroup, 2008-09

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.66	.63	.03
1-2 days	.23	.25	-.02
3-4 days	.07	.08	-.01
5 or more days	.05	.05	-.01

NOTES: Results are for cohort 1 five years after random assignment and cohort 2 four years after random assignment. Ordered logit beta = -.15. Effect is not statistically significant (*p*-value = .45).