NCEE 2008-4023

U.S. DEPARTMENT OF EDUCATION

Evaluation of the DC Opportunity Scholarship Program

Impacts After Two Years

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE

Institute of Education Sciences

Evaluation of the DC Opportunity Scholarship Program

Impacts After Two Years

June 2008

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

Marsha Silverberg, Project Officer, Institute of Education Sciences

NCEE 2008-4023
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR
EDUCATION EVALUATIC
AND REGIONAL ASSISTA
Institute of Education Sciences

U.S. Department of Education
Margaret Spellings
Secretary

Institute of Education Sciences
Grover J. Whitehurst
Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

June 2008

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.

While permission to reprint this publication is not necessary, the citation should be: Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. *Evaluation of the DC Opportunity Scholarship Program: Impacts After Two Years* (NCEE 2000-4023). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

To order copies of this report,

- Write to ED Pubs, Education Publications Center, U.S. Department of Education, P.O. Box 1398, Jessup, MD 20794-1398.
- 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 800-437-0833.
- Fax your request to 301-470-1244.
- Order online at www.edpubs.org.

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

Page

List of Tables

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Figures

Acknowledgments

This report is the fourth 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, Abt Associates.

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, Sabria Hardy, and Kevin Jay of Westat. Expert editorial and production assistance was provided by Evarilla Cover and Saunders Freeland of Westat. Ryan Marsh of the University of Arkansas ably assisted with the intermediate outcomes analysis and the drafting of chapter 4 and 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 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 Technical Working Group 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 the Congress in January 2004, established the first federally funded, private school voucher program in the United States. As part of this legislation, the Congress mandated a rigorous evaluation of the impacts of the Program, now called the DC Opportunity Scholarship Program (OSP). This report presents findings from the evaluation on the impacts 2 years after families who applied were given the option to move from a public school to a participating private school of their choice.

The evaluation is based on a randomized controlled trial design that compares the outcomes of eligible applicants randomly assigned to receive (treatment group) or not receive (control group) a scholarship through a series of lotteries. The main findings of the evaluation so far include:

- **After 2 years, there was no statistically significant difference in test scores in general between students who were offered an OSP scholarship and students who were not offered a scholarship.** Overall, those in the treatment and control groups were performing at comparable levels in mathematics and reading (table 3).

- **The Program had a positive impact on overall parent satisfaction and parent perceptions of school safety, but not on students' reports of satisfaction and safety** (tables 4 and 5). Parents were more satisfied with their child's school and viewed the school as less dangerous if the child was offered a scholarship. Students had a different view of their schools than did their parents. Reports of dangerous incidents in school were comparable for students in the treatment and control groups. Overall, student satisfaction was unaffected by the Program.

- **This same pattern of findings holds when the analysis is conducted to determine the impact of *using* a scholarship rather than being *offered* a scholarship.** Twenty-six percent of students who were randomly assigned by lottery to receive a scholarship chose not to use it in either the first or second year. We use a common statistical technique to take those "never users" into account; it assumes that the students had zero impact from the OSP, but it does not change the statistical significance of the original impact estimates. Therefore, the positive impacts on parent views of school safety and satisfaction all increase in size, and there remains no impact on academic achievement and no overall impact on students' perceptions of school safety or satisfaction from using an OSP scholarship.

- **There were some impacts on subgroups of students, but adjustments for multiple comparisons indicate that these findings may be due to chance.** There were no statistically significant impacts on the test scores of the high-priority subgroup of students who had previously attended schools designated as in need of improvement (SINI). However, being offered or using a scholarship may have improved reading test scores among three subgroups of students: those who had not attended a SINI school when they applied to the OSP, those who had relatively higher pre-Program academic performance, and those who applied in the first year of Program implementation. The Program may also have had a positive impact on school satisfaction for students who had previously attended SINI schools. However, these findings were no longer statistically significant when subjected to a reliability test to adjust for the multiple comparisons of treatment and control group students across 10 subgroups; the results may be "false discoveries" and should therefore be interpreted and used with caution.

- **The second year impacts are generally consistent with those from the first year.**[1] The main difference is that after 1 year, the non-SINI and higher performing groups of students appeared to experience statistically significant positive impacts on math achievement, while in the second year the impacts were on reading achievement. Adjustments for multiple comparisons suggest that both sets of results may be false discoveries.

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 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 prescribed how scholarships would be awarded: (1) in a given year, if there are more eligible applicants than available scholarships or open slots in private schools, scholarships are to be awarded by random selection (e.g., by lottery), and (2) priority for scholarships is given first to students attending SINI public schools and then to families that lack the resources to take advantage of school choice options.

The Program is operated by the Washington Scholarship Fund (WSF). To date, there have been four rounds of applications to the OSP (table 1). Applicants in spring 2004 (cohort 1) and spring 2005 (cohort 2) represent the majority of Program applicants; the evaluation sample was drawn from these two groups.[2] There were a smaller number of applicants in spring 2006 (cohort 3) and spring 2007 (cohort 4) who were recruited and enrolled by WSF in order to keep the Program operating at capacity—approximately 2,000 students—each year.

[1] See Wolf, Gutmann, Puma, Rizzo, Eissa, and Silverberg 2007.

[2] Descriptive reports on each of the first 2 years of implementation and cohorts of students have been previously prepared and released (Wolf, Gutmann, Eissa, Puma, and Silverberg 2005; Wolf, Gutmann, Puma, and Silverberg 2006) and are available on the Institute of Education Sciences' website at http://ies.ed.gov/ncee.

Table 1. OSP Applicants by Program Status, Cohorts 1 Through 4, Years 2004-2007

	Cohort 1 (Spring 2004)	Cohort 2 (Spring 2005)	Total Cohort 1 and Cohort 2	Cohort 3 (Spring 2006) and Cohort 4 (Spring 2007)	Total, All Cohorts
Applicants	2,692	3,126	5,818	1,308	7,126
Eligible applicants	1,848	2,199	4,047	846	4,893
Scholarship awardees	1,366	1,088	2,454	846	3,300
Scholarship users in initial year of receipt	1,027	797	1,824	712	2,536
Scholarship users fall 2005	919	797	1,716	NA	1,716
Scholarship users fall 2006	788	684	1,472	333	1,805
Scholarship users fall 2007	678	581	1,259	671	1,930

NOTES: 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 receipt was fall 2004 for cohort 1, fall 2005 for cohort 2, fall 2006 for cohort 3, and fall 2007 for cohort 4.

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

Mandated Evaluation of the OSP

In addition to establishing the OSP, Congress mandated an independent evaluation of it be conducted, with annual reports on the progress of the study. The legislation indicated the evaluation should analyze the effects of the Program on various academic and non-academic outcomes of concern to policymakers and use ". . . the strongest possible research design for determining the effectiveness" of the Program. The current 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. 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). 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 et al. 2002; Mayer et al. 2002).

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 in this report. 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 3 years during the evaluation period. During the first 2 years of recruitment, WSF received applications from 5,818 students. Of these, approximately 70 percent (4,047 of 5,818) were eligible to enter the Program (table 1). Of the total pool of eligible applicants, 2,308 students who were rising kindergarteners or from public schools entered lotteries (492 in cohort 1; 1,816 in cohort 2), resulting in 1,387 students assigned to the treatment condition and 921 assigned to the control condition. These students constitute the evaluation's impact analysis sample and represent three-quarters of all students in cohorts 1 and 2 who were not already attending a private school when they applied to the OSP.

Data are collected from the impact sample each year, starting with the spring in which students applied to the OSP (baseline) and each spring thereafter. These data include assessments of student achievement in reading and mathematics using the Stanford Achievement Test version 9 (SAT-9),[3] surveys of parents, and surveys of students in grade 4 and above—all administered by the evaluation team in central DC locations on Saturdays or weekday evenings because neither the public nor private schools would allow data collection on their campuses during the school day. In addition, the evaluation surveys all DC public and private schools each spring in order to address the statute's interest in understanding how the schools are responding to the OSP.

Participation in the OSP

In interpreting the impacts of the OSP, 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) moved into and out of them during the first 2 years.

School Participation

The private schools participating in the OSP represent the choice set available to parents whose children received scholarships. That group of schools had mostly stabilized by the 2005-06 school year. The schools that offered the most slots to OSP students, and in which OSP students and the impact

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

sample's treatment group were clustered, have characteristics that differed somewhat from the average participating OSP school. Only 11.2 percent of treatment group students were attending a school that charged tuition above the statutory cap of $7,500 during their second year in the Program (table 2) even though 39 percent and 38 percent of participating schools charged tuitions above that cap in 2005-06 and 2006-07, respectively.[4] Although 55 percent of all participating schools were faith-based (35 percent were part of the Catholic Archdiocese of Washington), nearly 80 percent of the treatment group attended a faith-based school, with more than half of them (53 percent) attending the 23 participating Catholic parochial schools. The average OSP student in the treatment group attended a school with 196 students—somewhat smaller than the average of 236 (2005-06) and 242 (2006-07) students across the set of all participating OSP schools.

Table 2. Features of Participating Private Schools Attended by the Treatment Group in Year 2

Characteristic	Weighted Mean	Highest	Lowest	Valid *N*
Schools charging over $7,500 tuition (percent of OSP students attending)	11.2%	NA	NA	51
Archdiocesan Catholic schools	52.7%	NA	NA	51
Other faith-based schools	23.9%	NA	NA	51
Tuition	$5,928	$29,902	$3,500	51
Enrollment	196.4	1,056	20	50
Student *N*	841			

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.

SOURCE: OSP School Directory information, 2004-05, 2005-06, and 2006-07, WSF.

While the characteristics of the participating private schools are important considerations for parents, in many respects it is how the schools differ from the public school options available to them that matters most. In the second year after applying to the OSP, students in the treatment and control groups did not differ significantly regarding the proportion attending schools that offered computer labs (93 and 92 percent), libraries (83 and 87 percent), gyms (70 and 66 percent), and art programs (90 and 86 percent). Differences in school characteristics between the treatment and control groups 2 years after they applied to the OSP that were statistically significant at the .01 level included:

[4] The average tuition charged to these treatment group students who used their scholarships was $5,928 but varied between $3,500 and $29,902. The WSF reported that families in their second year of the Program were required to pay at least some money out-of-pocket for tuition in 164 cases where the tuition charged by the school exceeded the $7,500 cap.

- Students in the treatment group were more likely to attend schools that offered a music program (92 percent), an after-school program (97 percent), and special programs for advanced learners (45 percent) compared to students in the control group (84 percent, 94 percent, and 33 percent for each type of program, respectively).

- Students in the treatment group were less likely to attend a school that offered counselors (74 percent), tutors (63 percent), programs for non-English speakers (19 percent), and programs for students with learning problems (55 percent) than were students in the control group (89 percent, 73 percent, 50 percent, and 79 percent, respectively, for each offering).

Student Participation

As has been true in similar programs, not all students offered an OSP scholarship actually used it to enroll in a private school. For students assigned to the treatment group, during the first 2 years of the Program (figure 1):

- 26 percent (366 out of 1,387) of those offered an OSP scholarship never used it;

- 20 percent (271) used their scholarship during some but not all of the first 2 years after the award; and

- The remaining 54 percent (750 students) used their scholarship consistently for the entire 2 years after the lottery.

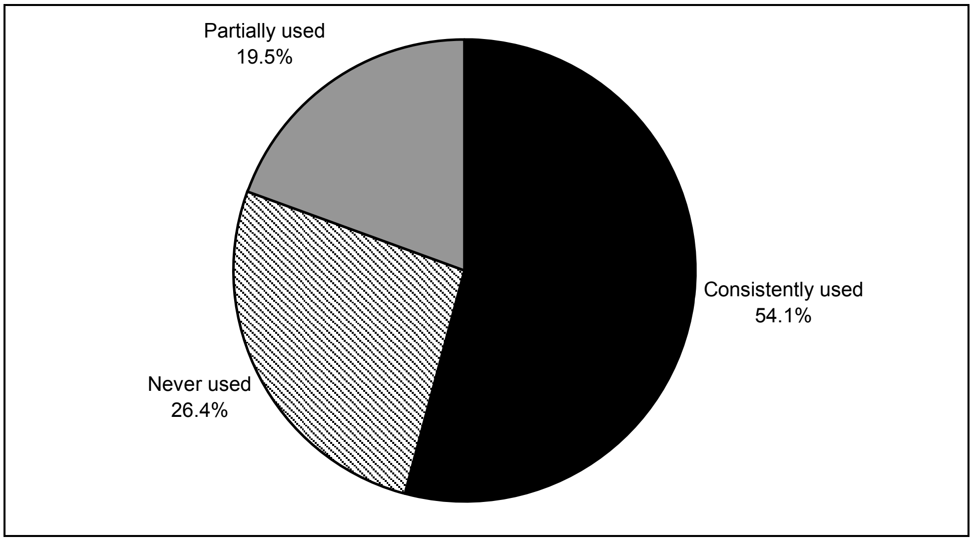
The reasons for not using the scholarship varied. The most common reasons cited by parents whose students declined the scholarship and completed surveys were (figure 2):

- Lack of available space in the private school they wanted their child to attend (29 percent of these parents);

- Participating schools did not offer services for their child’s learning or physical disability or other special needs (17 percent of these parents); and

- Child was accepted into a public charter school (16 percent of these parents).

Figure 1. Proportions of Treatment Group Students Who Experienced Various Categories of Usage in First 2 Years



NOTES: Data are not weighted. Valid *N* = 1,387. 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 non-users (no payments).

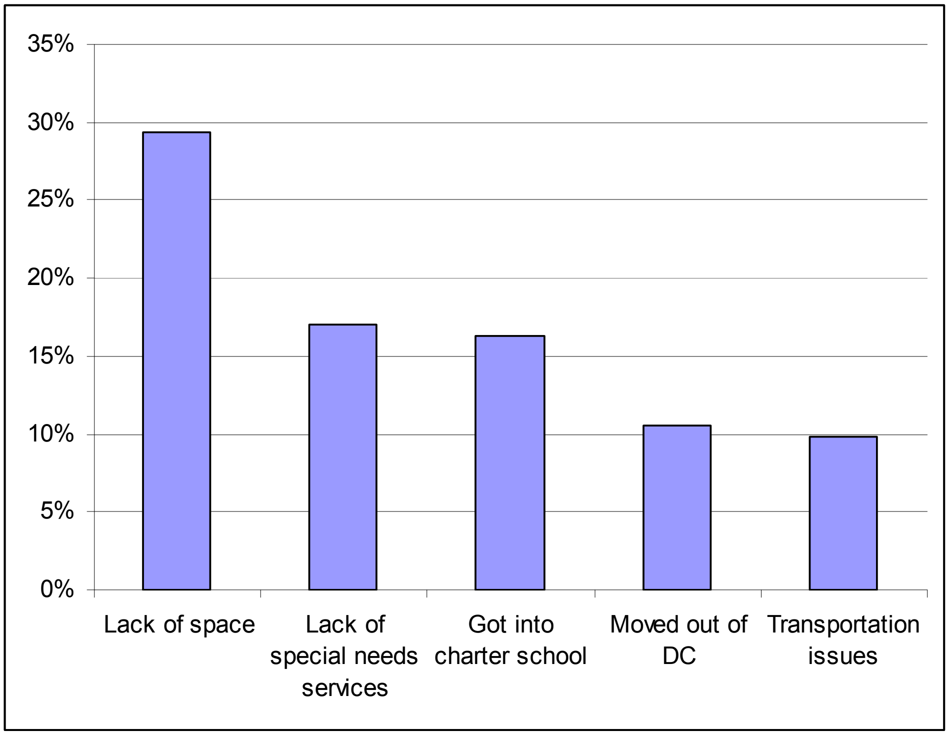
SOURCES: OSP applications and WSF's payment files.

Students who never used the OSP scholarship offered to them, or who did not use the scholarship consistently, could have found their way into other (non-OSP-participating) private schools, public charter schools, or traditional DC public schools. The same alternatives were available to students who applied to the OSP but were never offered a scholarship (the impact sample's control group). Both the treatment and control groups moved between public (both traditional and charter) and private schools or between SINI and non-SINI schools. As a result, over the 2 years after they applied to the OSP:

- Among the treatment group, 4 percent remained in the same school they were in when they applied to the Program; 71 percent switched schools once; and 25 percent switched schools twice.

- Among the control group, 22 percent remained in the same school they were in when they applied to the Program; 57 percent switched schools once; and 21 percent switched schools twice.

Figure 2. Most Common Reasons Given by Parents for Declining to Use the OSP Scholarship in Year 2



NOTES: Responses are unweighted. Respondents were able to select multiple responses, which generated a total of 180 responses provided by 153 parents. This equates to an average of 1.2 responses per parent. Responses that were not selected are unreported.

SOURCE: Impact Evaluation Parent Surveys.

Impact of the Program After 2 Years: Key Outcomes

The statute that authorized the OSP mandated that the Program be evaluated with regard to its impact on student test scores and school safety, as well as the "success" of the Program, which, in the design of this study, includes satisfaction with school choices. The impacts of the Program on these outcomes are presented in two ways: (1) the impact of the *offer* of an OSP scholarship, derived straight from comparing outcomes of the treatment and control groups, and (2) the impact of *using* an OSP scholarship, calculated from the unbiased treatment-control group comparison, but statistically netting out students who declined to use their scholarships.[5] The main focus of this study was on the overall group of

[5] This analysis uses straightforward statistical adjustments to account not only for the approximately 25 percent of impact sample respondents who received the offer of a scholarship but declined to use it (the "decliners"), but also the estimated 2.3 percent of the control group who never received a scholarship offer but who, by virtue of having a sibling with an OSP scholarship, ended up in a participating private school (we call this "program-enabled crossover"). These adjustments increase the size of the scholarship offer effect estimates, but cannot make a statistically insignificant result significant.

students, with a secondary interest in students who applied from SINI schools, followed by other subgroups of students (e.g., defined by their academic performance at application, their gender, or their grade level).

A previous report released in spring 2007 indicated that 1 year after application there were no statistically significant impacts on overall academic achievement or on student perceptions of school safety or satisfaction (Wolf et al. 2007). Parents were more satisfied if their child was in the Program and viewed their child's school as less dangerous. Among the secondary analyses of subgroups, there were impacts on math for students who applied from non-SINI schools and for those with relatively higher pre-Program test scores. Statistical adjustments for multiple comparisons suggested there is a possibility that the subgroup achievement impacts in year 1 were chance discoveries.

The analyses in this report were conducted using data collected on students 2 years after they applied to the OSP.

Impacts on Students and Parents Overall

- Across the full sample, there were no statistically significant impacts on reading achievement (effect size (ES) = .09)[6] or math achievement (ES = .01) from the offer of a scholarship (table 3) nor from the use of a scholarship.[7]

- Parents of students offered a scholarship were less likely to report serious concerns about school danger (ES = -.27) compared to parents of students not offered a scholarship (table 4); the same was true for parents of students who chose to use their scholarships (ES = -.34).

- On the other hand, students who were offered a scholarship reported similar levels of dangerous activities at school compared to those in the control group (ES = -.01; table 4); there was also no impact on student reports of school safety from using a scholarship (ES = -.01).

- The Program produced a positive impact on parent satisfaction with their child's school, for example regarding the likelihood of grading the school an "A" or "B," both for the impact of a scholarship offer (ES = .26; table 5) and the impact of scholarship use (ES = .33).

[6] An effect size (ES) is a standardized measure of the relative size of a program impact. In this report, effect sizes are expressed as a proportion of a standard deviation of the distribution of values observed for the study control group. One full standard deviation above and below the average value for a variable such as outcome test scores contains 64 percent of the observations in the distribution. Two full standard deviations above and below the average contain 95 percent of the observations.

[7] The magnitudes of these estimated achievement effects are below the threshold of .11 standard deviations, estimated by the power analysis to be the study's Minimum Detectable Effect size.

- Overall, there were no impacts of the OSP from being offered (ES = .05 to .13; table 5) or using a scholarship on students' satisfaction with his or her school.

Table 3. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Academic Achievement (Intent to Treat or ITT)

Student Achievement	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Reading	621.30	618.12	3.17	.09	.09
Math	614.09	613.85	.23	.01	.89

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for reading = 1,580; math = 1,585. Separate reading and math sample weights were used.

Table 4. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Parent and Student Reports of School Danger (ITT)

School Danger	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parents	2.06	3.00	-.94**	-.27	.00
Students	1.90	1.93	-.02	-.01	.87

**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for parent survey = 1,555. Valid *N* for student survey = 1,025. Parent and student survey weights were used. Survey given to students in grades 4-12.

Table 5. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Parent and Student Reports of Satisfaction with Their School (ITT)

Outcome	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parents who gave school a grade of A or B	.76	.63	.13**	.26	.00
Average grade parent gave school (5.0 scale)	4.02	3.73	.29**	.29	.00
School satisfaction scale	26.12	23.44	2.67**	.33	.00
Students who gave school a grade of A or B	.71	.68	.03	.05	.49
Average grade student gave school (5.0 scale)	3.97	3.84	.13	.12	.14
School satisfaction scale	34.12	33.24	.88	.13	.10

**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for parent measure of school grade = 1,549; parent satisfaction = 1,571. Parent survey weights were used. Parent school satisfaction scale was IRT scored and had a range of .96 to 35.43. Valid *N* for student measure of school grade = 974; student satisfaction = 1,042. Student survey weights were used. School satisfaction scale was IRT scored and had a range of 9.67 to 46.89. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Impacts on Subgroups

In addition to determining the general impacts of the OSP on all study participants, this evaluation also reports Programmatic impacts on policy-relevant subgroups of students. The subgroups were designated prior to data collection and include students who were attending SINI versus non-SINI schools at application, those relatively higher or lower performing at baseline, girls or boys, elementary versus high school students, and those from application cohort 1 or cohort 2. Since the subgroup analysis involves significance tests across multiple comparisons of treatment and control students, some of which may be statistically significant merely by chance, these subgroup-specific results should be interpreted with caution. Specifically:

Subgroup Achievement Impacts

- There were no statistically significant reading (ES = -.00) or math (ES = .05) achievement impacts for the high-priority subgroup of students who had attended a SINI public school under *No Child Left Behind* (*NCLB*) before applying to the Program.

- The Program may have had a positive impact on reading test scores in year 2 for three subgroups of students, although the statistical significance of the findings was not robust to adjustments for multiple comparisons:

  - Students who attended non-SINI public schools prior to application to the Program (56 percent of the impact sample) scored an average of 5.7 scale score points higher in reading (ES = .15) if they were offered the scholarship compared to not being offered a scholarship and 6.9 scale score points higher (ES = .18) if they used their scholarship compared to not being offered a scholarship.

  - Students who entered the Program in the higher two-thirds of the test-score performance distribution at baseline (66 percent of the impact sample) scored an average of 5.2 scale score points higher in reading (ES = .15) if they were offered a scholarship compared to not being offered a scholarship and 6.3 scale score points higher (ES = .18) if they used their scholarship compared to not being offered a scholarship.

  - Students from the first cohort of applicants (21 percent of the impact sample) scored an average of 8.7 scale score points higher in reading (ES = .27) if they were offered a scholarship compared to not being offered a scholarship and 12.2 scale score points higher (ES = .37) if they used their scholarship compared to not being offered a scholarship.

- The OSP had no statistically significant achievement impacts for other subgroups of participating students, including those in the lower third of the test-score performance distribution at baseline, boys, girls, elementary students, secondary students, and students from the second cohort of applicants (effect sizes ranging from -.14 to .11).

Subgroup Safety and Satisfaction Impacts

- Eight of the 10 subgroups analyzed, including parents of the high-priority subgroup of students who had attended SINI schools, reported viewing their child's school as less dangerous if the child was offered or using an OSP scholarship compared to not being offered a scholarship. Effect sizes for the impact of an offer of a scholarship on parent perceptions of school danger for the eight affected subgroups ranged from -.21 to -.35. Adjustments for multiple comparisons indicate that these eight subgroup impacts on parental perceptions of safety are not likely to be false discoveries. The parents of students who were relatively lower performing at baseline and those in high school were the exceptions, as they did *not* report lower or different levels of perceived school danger as a result of the treatment.

- Consistent with the finding for students overall, none of the subgroups of students reported experiencing differences in dangerous activities at school if they were in the Program. Thus, there was no impact on students' perceptions of school safety from either the offer or the use of a scholarship for any of the subgroups (effect sizes range from -.11 to .09).

- In addition to an overall impact on parental satisfaction with their child's school, the Program produced satisfaction impacts on 8 of the 10 subgroups analyzed, including the high-priority subgroup of parents of students who had attended SINI schools. Effect sizes for the impact of an offer of a scholarship on the likelihood of a parent grading their child's school "A" or "B" for the eight affected subgroups ranged from .18 to .34. Adjustments for multiple comparisons indicate that one of these eight subgroup impacts (for the parents of students who were relatively lower performing at baseline) may have been a false discovery. The statistical significance of the other seven subgroup impacts on parent satisfaction with their child's school was not affected by adjustments for multiple comparisons. The parents of high school students and those in the first cohort of applicants generally did *not* report higher levels of school satisfaction that were statistically significant as a result of the treatment (effect sizes range from .02 to .18).

- With one exception, there was no impact on school satisfaction if students were offered a scholarship, across subgroups. The high-priority subgroup of students who applied from a SINI school were more likely to give their school a grade of A or B (ES = .24) if they were offered a scholarship compared to not being offered a scholarship, although adjustments for multiple comparisons indicate that this finding may be a false discovery.

The Impact of the Program on Intermediate Outcomes

Understanding the mechanisms through which the OSP does or does not affect student outcomes requires examining the expectations, experiences, and educational environments made possible by Program participation. The analysis here estimates the impact of the Program on a set of "intermediate outcomes" that are influenced by parents' choice of whether to use an OSP scholarship and where to use it, but are not end outcomes themselves. The method used to estimate the impacts on intermediate outcomes is identical to that used to estimate impacts on the key Program outcomes, such as academic achievement.

Prior to data analysis, possible intermediate outcomes of the OSP were selected based on existing research and theory regarding scholarship programs and educational achievement. Because 24 intermediate outcome candidates were identified through this process, the variables were organized into four conceptual groups or clusters to aid in the analysis.[8]

There is no way to rigorously evaluate the linkages between the intermediate outcomes and achievement—students are not randomly assigned to the experience of various educational conditions and programs. That is why any findings from this element of the study do not suggest that we have learned what specific factors "caused" any observed test score impacts, only that certain factors emerge from the analysis as possible candidates for mediating influence. The analyses are exploratory, and, given the number of factors analyzed, some of the statistically significant findings may be "false discoveries" (due to chance).

Overall, 2 years after applying for a scholarship, the Program had an impact on 10 of the 24 intermediate outcomes, 8 of which remained statistically significant after adjustments for multiple comparisons:

- ***Home Educational Supports.*** The results suggest that the Program may have had an impact on two of four intermediate outcomes in this group. The Program appeared to produce a positive impact on parents' aspirations for how far in school their child would go (ES = .12); however, this result may be a false discovery. The Program led to students' experiencing more time spent commuting to school from their homes (ES = .25), a result that did not lose statistical significance after adjustments for multiple comparisons. There were no statistically significant differences between the treatment and control groups on the involvement in school reported by parents in year 2 (ES = -06) or on the use of a tutor outside of school (ES = -07).

- ***Student Motivation and Engagement.*** The Program had no statistically significant impacts on any of the six elements of this group of intermediate outcomes. Two years after they applied to the OSP, the treatment and control group students reported similar

[8] Intermediate Outcome Conceptual Grouping 1, *Home Educational Supports*, includes parent involvement, parent aspirations, out-of-school tutor usage, and school transit time. Intermediate Outcome Conceptual Grouping 2, *Student Motivation and Engagement*, includes student aspirations, attendance, tardiness, reading for fun, engagement in extracurricular activities, and frequency of homework. Intermediate Outcome Conceptual Grouping 3, *Instructional Characteristics*, includes student/teacher ratio, teacher attitude, challenge of classes, ability grouping, availability of tutors, in-school tutor usage, programs to assist students with learning disabilities or English language learners, programs for advanced learners, before-/after-school care programs, and enrichment programs. Intermediate Outcome Conceptual Grouping 4, *School Environment*, includes parent/ school communication, school size, percent non-white, and peer classroom behavior.

aspirations for future schooling (ES = -.11), frequency of doing homework (ES = -.10), time spent reading for fun (ES = .02), and engagement in extracurricular activities (ES = .08). There were no statistically significant differences in student attendance (ES = -.11) or tardiness rates (ES = -.11), as reported by parents.

- ***Instructional Characteristics.*** The offer of a scholarship appears to have had a statistically significant impact on 5 of the 10 intermediate outcomes in this group. Being offered a scholarship led to students' experiencing smaller classes, as measured by student/teacher ratios (ES = -.29). The Program also led to students' experiencing a lower likelihood that their school offered either tutoring (ES = -.32) or special programs for children who were English language learners or had learning problems (ES = -.66). At the same time, however, the Program had a positive impact on the use of an in-school tutor, presumably in schools that made them available (ES = .13). The OSP also led to students' experiencing a higher likelihood of being in a school that offered enrichment programs (ES = .19). The statistical significance of these five results was not affected by adjustments for multiple comparisons. There were no differences between the treatment and control groups in how students rated their teacher's attitude (ES = .02) or the challenge of their classes (ES = -.04), the school's use of ability grouping (ES = .13), the availability of programs for advanced learners (ES = .12), or before- and after-school programs (ES = .04).

- ***School Environment.*** The Program may have affected three of the four measures of school environment. Students in the treatment group experienced schools that were smaller (ES = -.43) and had a smaller percentage of non-white students (ES = -.39) than the schools of the control group, findings that were not affected by adjustments for multiple comparisons. Treatment group students also reported having better behaved peers in the classroom than did control group students (ES = .16), although adjustments for multiple comparisons suggest that this finding may be a false discovery. There were no differences in parents' reports of how their child's school communicates with them (ES = .01).

It is important to note that the findings regarding the impacts of the OSP reflect the particular Program elements that evolved from the law passed by Congress, and the characteristics of students, families, and schools—public and private—that exist in the Nation's capital. The same program implemented in another city could yield different results, and a different scholarship program in Washington, DC, might also produce different outcomes.

1. Introduction

The *District of Columbia School Choice Incentive Act of 2003,*[1] passed by the Congress in January 2004, established the first federally funded, private school voucher program in the United States. Since that time, more than 7,000 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 report from the ongoing evaluation describes the impacts of the Program 2 years after families who applied were given the option to move from a public school to a participating private school of their choice.

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 5 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.

- 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.

Following passage of the legislation, the Washington Scholarship Fund (WSF), a 501(c)3 organization in the District of Columbia, was selected in late March 2004 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. Since then the WSF has finalized the Program design, established protocols, recruited applicants and schools, awarded scholarships, and placed and monitored scholarship awardees in participating private schools. The funds appropriated for the OSP are sufficient to support approximately 1,700 to 2,000 students in a given year, 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 four rounds of applicants to the OSP (table 1-1):

- Applicants in spring 2004 (cohort 1) and spring 2005 (cohort 2), who represent the majority of Program applicants and from whom the evaluation sample was drawn,[2] and

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

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.[4] During fall of 2007, a total of 1,930 students were using Opportunity Scholarships to attend participating private schools.

[2] 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.

[3] 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 fourth year of implementation. WSF limited cohorts 3 and 4 applications to students entering grades K-6 because there were few slots available in participating junior high and 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 exits from the Program that enabled WSF to accommodate cohorts 3 and 4.

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

Table 1-1. OSP Applicants by Program Status, Cohorts 1 Through 4, Years 2004-2007

	Cohort 1 (Spring 2004)	Cohort 2 (Spring 2005)	Total Cohort 1 and Cohort 2	Cohort 3 (Spring 2006) and Cohort 4 (Spring 2007)	Total, All Cohorts
Applicants	2,692	3,126	5,818	1,308	7,126
Eligible applicants	1,848	2,199	4,047	846	4,893
Scholarship awardees	1,366	1,088	2,454	846	3,300
Scholarship users in initial year of receipt	1,027	797	1,824	712	2,536
Scholarship users fall 2005	919	797	1,716	NA	1,716
Scholarship users fall 2006	788	684	1,472	333	1,805
Scholarship users fall 2007	678	581	1,259	671	1,930

NOTES: 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 receipt was fall 2004 for cohort 1, fall 2005 for cohort 2, fall 2006 for cohort 3, and fall 2007 for cohort 4.

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

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 non-academic outcomes of concern to policymakers and use ". . . the strongest possible research design for determining the effectiveness" of the Program.[5]

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.[6] 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

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

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

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). 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 et al. 2002; Mayer et al. 2002).

Key Research Questions

The research priorities for the evaluation were shaped largely by the primary topics of interest specified in the statute.[7] 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 (e.g., school attendance and educational attainment)?* 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?

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.[8]

[7] 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)).

[8] The statute requests comparisons between "program participants" and non-participants. 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).

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.

Questions 1, 3, and 4 are central to this report. Questions 2 and 5 will be addressed in subsequent reports that are planned for the evaluation.[9] 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 (e.g., student motivation and engagement, school environment and instruction). These analyses will contribute to the literature on voucher programs.

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 in this report. 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 3 years during the evaluation period. During the first 2 years of recruitment, WSF received applications from 5,818 students. Of these, approximately 70 percent (4,047 of 5,818) were eligible to enter 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);

[9] We are deferring the analysis of education attainment (Question 2) until the spring 2009 report to allow a sufficient number of impact sample students (30 percent) to age into being potentially able to graduate from (or conversely drop out of) high school, in order to ensure the power to detect statistically significant differences (impact) between the treatment and control group if there are any. The analysis of how DC schools are responding to the OSP (Question 5) depends on changes over time and will also be examined in the spring 2009 report.

[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.

2. Non-SINI public school applicants (middle priority); and
3. Applicants already attending private schools (lowest priority).

However, not all applicants faced the conditions that necessitated scholarship award by lottery.[11,12] 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 represents 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).

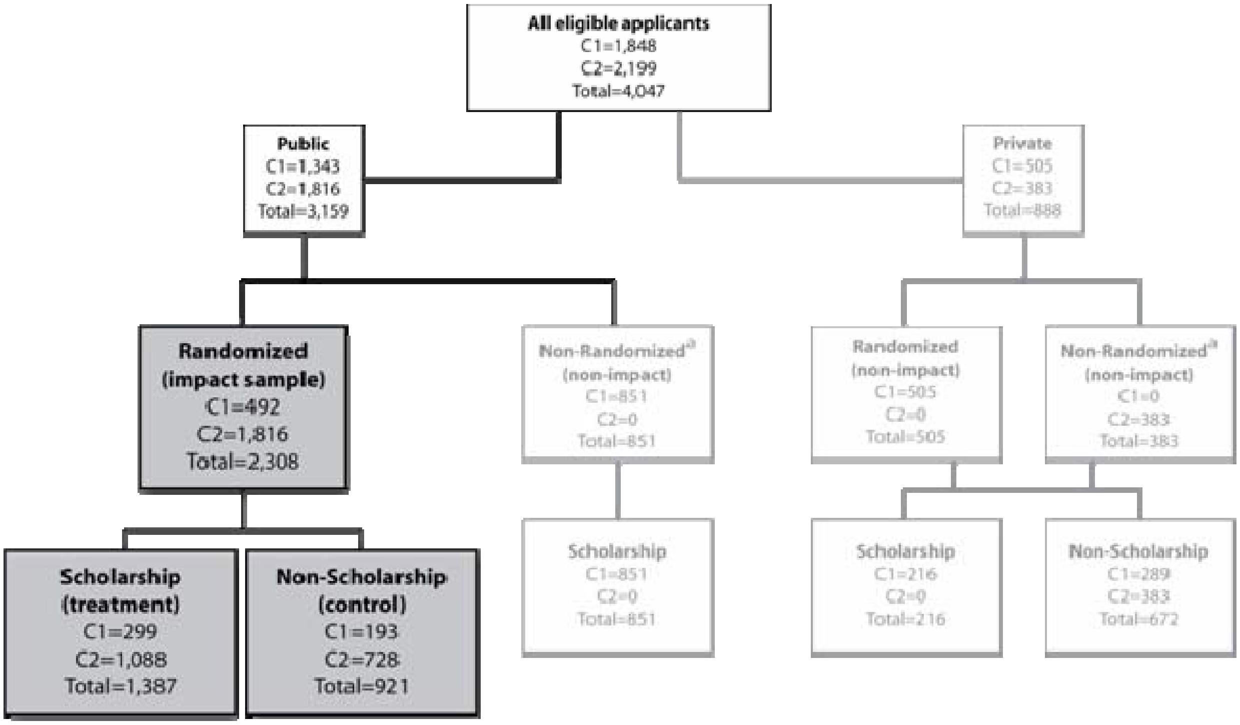
The total pool of eligible applicants comprised 1,848 applicants in cohort 1 (spring 2004) and 2,199 applicants in cohort 2 (spring 2005). Of those eligible applicants, 492 in cohort 1 and 1,816 in cohort 2 met the criteria to be randomly assigned by lottery to the evaluation's treatment and control 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. The impact sample comprised by 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).

[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 grades 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).

[12] For more information on the lotteries conducted in spring 2004 and spring 2005, see Wolf et al. 2006.

[13] A total of five members of the cohort 1 control group were awarded scholarships by lottery in the summer of 2005, and a total of seven members of the control group (cohorts 1 and 2) were awarded scholarships by lottery in the summer of 2006 as part of the control group follow-up lottery to reward control group members who cooperate with the evaluation's testing requirements. 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. They are treated as control group members for purposes of the Intent to Treat (ITT) and Bloom adjusted Impact on the Treated (IOT) analyses.

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).

Data Collection

The evaluation gathers annual information from students and families in the study, as well as 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 come 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 takes place primarily on Saturdays, during the spring, in

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

locations throughout DC arranged by the evaluators. The testing conditions are similar for members of the treatment and control groups.[15]

- **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 are being conducted at the Saturday testing events, while parents are waiting for their children to complete their outcome testing. The parent surveys provide the self-reported outcome measures for parental satisfaction and safety.[16]

- **Student surveys.** Each spring after the baseline year, surveys of students in grades 4 and above are being conducted at the outcome testing events. The student surveys provide the self-reported outcome measures for student satisfaction and safety.[17]

- **Principal surveys.** Each spring, surveys of principals of all public and private schools operating in the District of Columbia are being conducted. Topics include self-reports of school organization, safety, climate, principals' awareness of and response to the OSP, and, for private school principals, why they are or are not OSP participants.[18]

Several methods were used to encourage high levels of response to year 2 data collection in the spring of 2005 (cohort 1) and the spring of 2006 (cohort 2). Study participants were invited to at least three different data collection events if a member of the treatment group and at least five different data collection events if 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.

[15] For student assessments, the overall (effective) response rates were 74.6 percent for the treatment group and 69.3 percent for the control group. Actual response rates (before subsample weighting) for the control group were 46.1 percent (cohort 1) and 54.3 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 64.6 percent (cohort 1) and 70.6 percent (cohort 2). Actual and effective response rates for the treatment group were 71.2 percent (cohort 1) and 75.6 percent (cohort 2). See appendix A, figure A-1 and tables A-5 and A-7 for a detailed breakdown of the response rates and a discussion of the subsampling procedure.

[16] For the parent survey, the overall (effective) response rates were 74.8 percent for the treatment group and 68.8 percent for the control group. Actual response rates (before subsample weighting) for the control group were 45.6 percent (cohort 1) and 54.8 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 62.9 percent (cohort 1) and 70.3 percent (cohort 2). Actual and effective response rates for the treatment group were 69.6 percent (cohort 1) and 76.2 percent (cohort 2). See appendix A and table A-8 for a detailed breakdown of the response rates.

[17] For the student survey, the overall (effective) response rates were 71.8 percent for the treatment group and 61.8 percent for the control group. Actual response rates (before subsample weighting) for the control group were 37.0 percent (cohort 1) and 52.5 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 54.4 percent (cohort 1) and 64.8 percent (cohort 2). Actual and effective response rates for the treatment group were 60.5 percent (cohort 1) and 76.9 percent (cohort 2). See appendix A and table A-9 for a detailed breakdown of the response rates.

[18] For the principal survey, response rates for the 2006-07 school year were 53.2 percent (public schools) and 51.8 percent (private schools). For the 2005-06 school year, response rates were 67.0 percent (public schools) and 68.9 percent (private schools).

After these initial data collection activities were completed, the test score response rate for year 2 was 65.8 percent, with a response rate differential of 22 percentage points lower for the control group compared to the treatment group. To reduce this response rate differential, a random subsample of control non-respondents was drawn and subjected to intensive efforts at non-respondent conversion (see appendix A, section A.7). As a result of the subsample conversion process, the final effective test score response rate for year 2 was 72.5 percent, and the differential rate of response between the treatment and control groups was reduced to 5 percentage points (see appendix A, table A-7). Non-response weights that draw upon baseline information about participants were used to re-equate the treatment and control groups to reduce the threat of bias due to study attrition (see appendix A, section A.7). Sections A.3 and A.7 of appendix A provide additional details about data sources, collection methods, response rates, subsampling for non-response conversion, and final non-response sample weights.

Research Methodology

The evaluation of the OSP is 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 admission to the intervention or program and a control group that does not receive admission—with the control group's subsequent experiences indicating what probably 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 performed prior to data collection indicated that the evaluation is likely to be sufficiently powered to detect achievement impacts of .11 to .13 standard deviations for the entire study sample and .14 to .27 standard deviations for the subgroups of interest (see appendix A, section A.2). Observations were weighted after data collection, using baseline characteristics associated with study non-response, to re-establish the equivalence of the treatment and control groups in the face of differential rates of non-response (see appendix A, section A.7). A consistent set of 15 baseline student characteristics related to student achievement were included in the regression models that generated the estimates of Program impact (see appendix A, 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 any statistically significant impact estimates. 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 sub-sampling to convert control initial non-respondents 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 from the evaluation is the fourth in a series of required annual reports to Congress. It presents the impacts of the Program on students and families 2 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 2 years later as a result of the offer or the actual use of an Opportunity Scholarship. The final chapter (chapter 4) for the first time 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 exploratory 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 evaluation's final report, to be published in spring 2009, will examine impacts 3 years after application to the OSP and how DC schools have been changing in response to the Program.

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. Similarly, the patterns of scholarship use are part of a larger picture of school transfers, with both scholarship and non-scholarship students switching schools during the 2 years since they applied to the OSP. Research links elements of students' educational environments and their school mobility to later outcomes.[28] 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.

2.1 School Participation

The private schools participating in the OSP represent the choice set available to parents whose children received scholarships. For the 2006-07 school year, 66 of 104 private schools in the District of Columbia were participating in the Program.[29] Among the participating schools:[30]

- 55 percent (36) were faith-based, with most of them (23) being the parochial schools of the Catholic Archdiocese of Washington.

[28] For studies of the effects of school mobility on achievement see, for example, Hanushek, Kain, and Rivkin 2004; Temple and Reynolds 1999. For studies of the effects of elements of the school environment on achievement see, for example, Sander 1999; Nielsen and Wolf 2001; Hanushek, Kain, and Rivkin 2002; Card and Krueger 1992.

[29] This figure represents a loss of four schools since the prior year but a gain of two new schools to the Program. As reported by the WSF, the schools left the Program for various reasons. Two schools exclusively served students in first grade and below and for that reason were not attracting any OSP applicants. One school previously in the Program closed prior to the 2006-07 school year and another was excluded from the Program by WSF personnel due to school management concerns. The two private schools new to the OSP in 2006-07 include a K-12 school which first opened in 2006-07 and a school serving students in grades 6-8.

[30] Information was obtained for all 66 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 $N = 50$), percent minority students (valid $N = 45$), and student/teacher ratio (valid $N = 46$) were drawn from the National Center for Education Statistics' Private School Survey, last administered in 2003-04.

- 38 percent charged an average tuition above the OSP's scholarship cap of $7,500.[31]
- The average school had a total student population of 242 students.
- 23 percent served high school students.[32]
- The average percent minority among the student body was 73 percent.
- The average student/teacher ratio was 9.4.

These characteristics are similar to those presented in earlier evaluation reports because the group of participating private schools had mostly stabilized by the 2005-06 school year.[33]

Schools Attended by Scholarship Users in Year 2[34]

Not all of the schools that agreed to participate in the Program serve OSP students every year. [35] Two years after being awarded a scholarship, OSP students were enrolled in 57, and treatment students in 52, of the 69 schools available to them in that time period.[36] Since participating schools varied in how many slots they committed to the Program, OSP students tended to cluster in certain schools; this was also true of the students in the impact sample's treatment group (see figure 2-1).

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 average participating OSP school. In other words, the student-weighted average characteristics of schools attended by OSP students differed somewhat from the school-weighted average characteristics of the set of OSP schools. Only 11.2 percent of treatment group students were attending a school that charged tuition above the statutory cap of $7,500 during their second year in the Program (table 2-1), even though 39 percent and 38 percent of participating schools charged tuitions above that cap in 2005-06 and

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

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

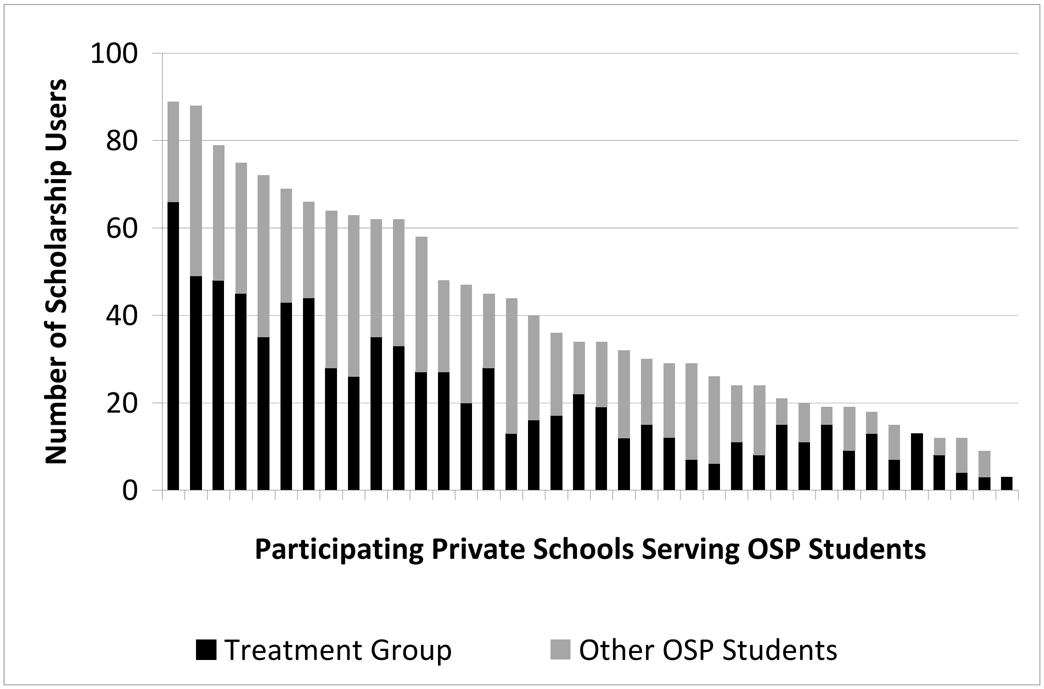
[33] See Wolf et al. 2007, pp. 15-17.

[34] "Year 2" is measured relative to the time each student applied to the Program. For cohort 1 students, who applied in spring 2004, year 2 is measured at spring 2006. For cohort 2 students who applied in spring 2005, their year 2 is spring 2007.

[35] The source for student enrollment in participating schools is the WSF OSP payment file for 2005-06 and 2006-07.

[36] The impact sample combines data from the experience of cohort 1 students in 2005-06 (their impact year 2) and cohort 2 students in 2006-07 (their impact year 2). Collectively, the total number of schools available for cohort 1 during 2005-06 and cohort 2 during 2006-07 was 69. While, technically, 72 individual campuses were available, the research team treats three of the schools with dual campuses as single entities because they have one principal for both campuses, following the classification practice used by the National Center for Education Statistics in the Annual School Survey.

Figure 2-1. Distribution of OSP Scholarship Users Across Participating Schools in Year 2, by Treatment Group vs. Other OSP Students



NOTES: Each bar represents a private school that enrolled OSP students during their second year in the Program. The dark area of each bar represents the number of students randomly assigned to the treatment group that used a scholarship (both partial and full users) and are included in the experimental evaluation of Program impact. The lighter area of each bar represents the number of other students that used OSP scholarships (both partial and full users) who are not a part of the evaluation. School *N* = 69. Student *N* = 1,599. Schools that did not enroll any OSP students have been omitted from this figure (*N* = 12). Additionally, data were suppressed for confidentiality purposes if a school enrolled only 1 or 2 treatment students or 1 or 2 other OSP students (*N* = 19).

SOURCE: WSF's payment files.

Table 2-1. Features of Participating OSP Private Schools Attended by the Treatment Group in Year 2

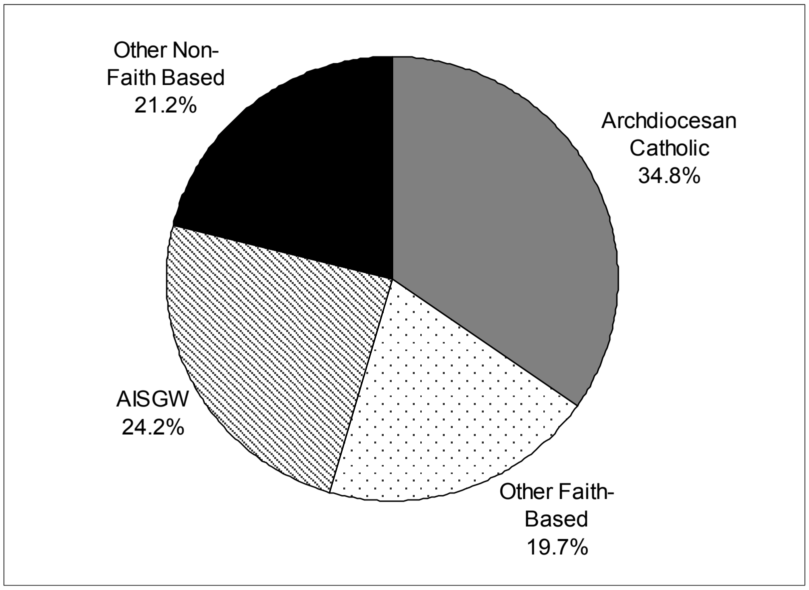
Characteristic	Weighted Mean	Highest	Lowest	Valid *N*
Schools charging over $7,500 tuition (percent of OSP students attending)	11.2	NA	NA	51
Archdiocesan Catholic schools	52.7%	NA	NA	51
Other faith-based schools	23.9%	NA	NA	51
Tuition	$5,928	$29,902	$3,500	51
Enrollment	196.4	1,056	20	50
Student *N*	841			

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.

SOURCE: OSP School Directory information, 2004-05, 2005-06, and 2006-07, WSF.

2006-07, respectively.[37] The average OSP student in this group attended a school with 196 students—somewhat smaller than the average of 236 (2005-06) and 242 (2006-07) students across the set of all participating schools. Although 55 percent of all participating schools were faith-based (35 percent part of the Catholic Archdiocese of Washington), 77 percent of the treatment group attended a faith-based school, with more than half of them (53 percent) attending the 23 participating Catholic parochial schools (figures 2-2 and 2-3).

Figure 2-2. Percent of Participating OSP Private Schools in Year 2 by Their Religious Affiliation

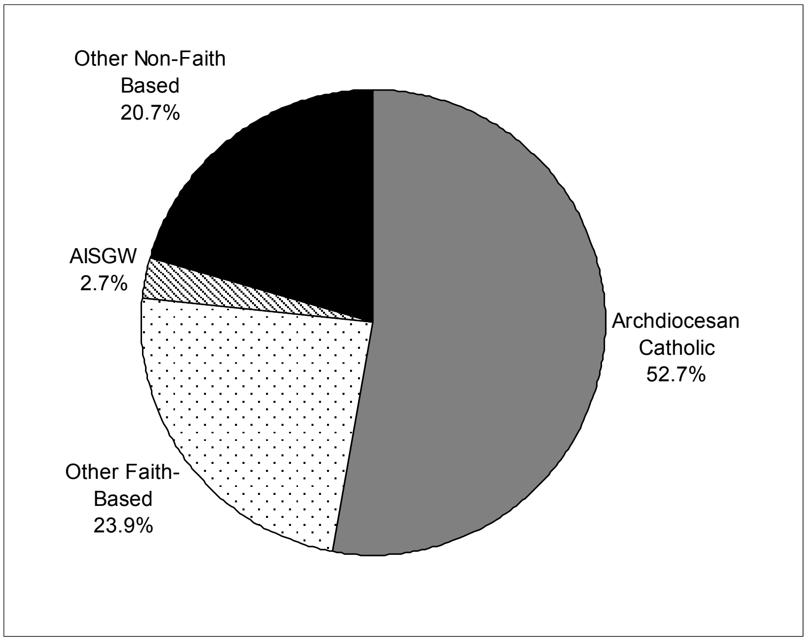


NOTES: *N* for schools = 66. AISGW is an abbreviation for the Association of Independent Schools of Greater Washington.

SOURCES: National Center for Education Statistics: Private School Universe Survey, 2003-04, supplemented by OSP School Directory information, 2004-05, 2005-06, WSF.

[37] The average tuition charged to these treatment group students who used their scholarships was $5,928 but varied between $3,500 and $29,902. The WSF reported that families in their second year of the Program were required to pay at least some money out-of-pocket for tuition in 164 cases where the tuition charged by the school exceeded the $7,500 cap.

Figure 2-3. Percent of Students Attending Participating OSP Private Schools in Year 2 by Their Religious Affiliation



NOTES: *N* for schools = 51. AISGW is an abbreviation for the Association of Independent Schools of Greater Washington.

SOURCES: National Center for Education Statistics: Private School Universe Survey, 2003-04, supplemented by OSP School Directory information, 2004-05, 2005-06, WSF.

Schools Attended by the Treatment Group in Relation to Those of the Control Group in Year 2

While the characteristics of the participating private schools are important considerations for parents, in many respects it is how they differ from the public school options available to them that matters most. How different are the school conditions? In the second year after applying to the OSP, students in the treatment and control groups did not differ significantly regarding the proportion attending schools that offered computer labs (93 and 92 percent), libraries (83 and 87 percent), gyms (70 and 66 percent), and art programs (90 and 86 percent) (table 2-2). Differences in school characteristics between the treatment and control groups 2 years after they applied to the OSP that were statistically significant at the .01 level included:

- Students in the treatment group were more likely to attend schools that offered a music program (92 percent), schools with an after-school program (97 percent), and schools with special programs for advanced learners (45 percent) compared to students in the control group (84 percent, 94 percent, and 33 percent, respectively).

- Students in the treatment group were less likely to attend a school with a separate cafeteria facility (72 percent) or a nurse's office (36 percent) compared to students in the control group (86 percent and 78 percent, respectively).

- Students in the treatment group were less likely to attend a school that offered counselors (74 percent), tutors (63 percent), programs for non-English speakers (19 percent), and programs for students with learning problems (55 percent) than were students in the control group (89 percent, 73 percent, 50 percent, and 79 percent, respectively, for each offering).

Table 2-2. Characteristics of School Attended by the Impact Sample, Year of Application and First 2 Years in the Program

Percentage of Students Attending a School with:	Baseline Year			Year 1			Year 2		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Separate facilities:									
Computer lab	73.53	72.90	0.63	95.51	89.13	6.38**	93.21	92.44	.77
Library	80.12	78.07	2.05	79.52	77.33	2.19	83.10	87.15	-4.05
Gym	63.67	66.16	-2.48	70.95	67.38	3.57	69.66	66.25	3.41
Cafeteria	87.39	88.68	-1.29	74.15	87.95	-13.80**	71.57	86.32	-14.75**
Nurse's office	87.43	88.51	-1.08	29.27	84.53	-55.26**	35.97	77.69	-41.72**
Percent missing	6.84	7.74	-0.89	35.42	42.38	-6.96	39.96	52.50	-12.54
Programs:									
Special program for non-English speakers	48.62	44.15	4.47	18.60	57.10	-38.50**	19.47	49.92	-30.44**
Special program for students with learning problems	64.35	65.58	-1.23	51.14	88.72	-37.58**	55.01	78.86	-23.85**
Special program for advanced learners	38.65	35.43	3.22	42.50	37.85	4.65	45.26	32.94	12.31**
Counselors	80.50	80.08	0.43	75.39	82.11	-6.72**	73.74	89.21	-15.47**
Individual tutors	36.58	39.10	-2.51	78.10	77.89	0.22	62.64	73.17	-10.53**
Music program	70.14	70.60	-0.47	93.57	74.82	18.75**	92.01	83.69	8.33**
Art program	69.18	66.66	2.52	84.23	81.45	2.78	90.03	86.43	3.60
After-school program	79.98	79.31	0.67	94.73	93.31	1.43	97.19	93.61	3.58**
Percent missing	7.16	7.89	-0.73	34.41	42.21	-7.80	39.82	52.50	-12.67
Sample size (unweighted)	1,387	921	466	1,387	921	466	1,387	921	466

**Statistically significant at the 99 percent confidence level.

NOTE: Data are weighted. For a description of the weights, see appendix A.

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

2.2 Student Participation

Whether the participating private schools are attractive to parents and students is reflected, to some degree, in the rates of students' scholarship use. A total of 2,454 students who applied to the OSP in the first 2 years of Program operation were offered scholarships, with 1,387 of them in the impact sample's treatment group. 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.

Patterns of Scholarship Use

According to rules determined by WSF, once a student was offered an OSP scholarship he or she could use it at any time. During the first 2 years of the Program (figure 2-4):

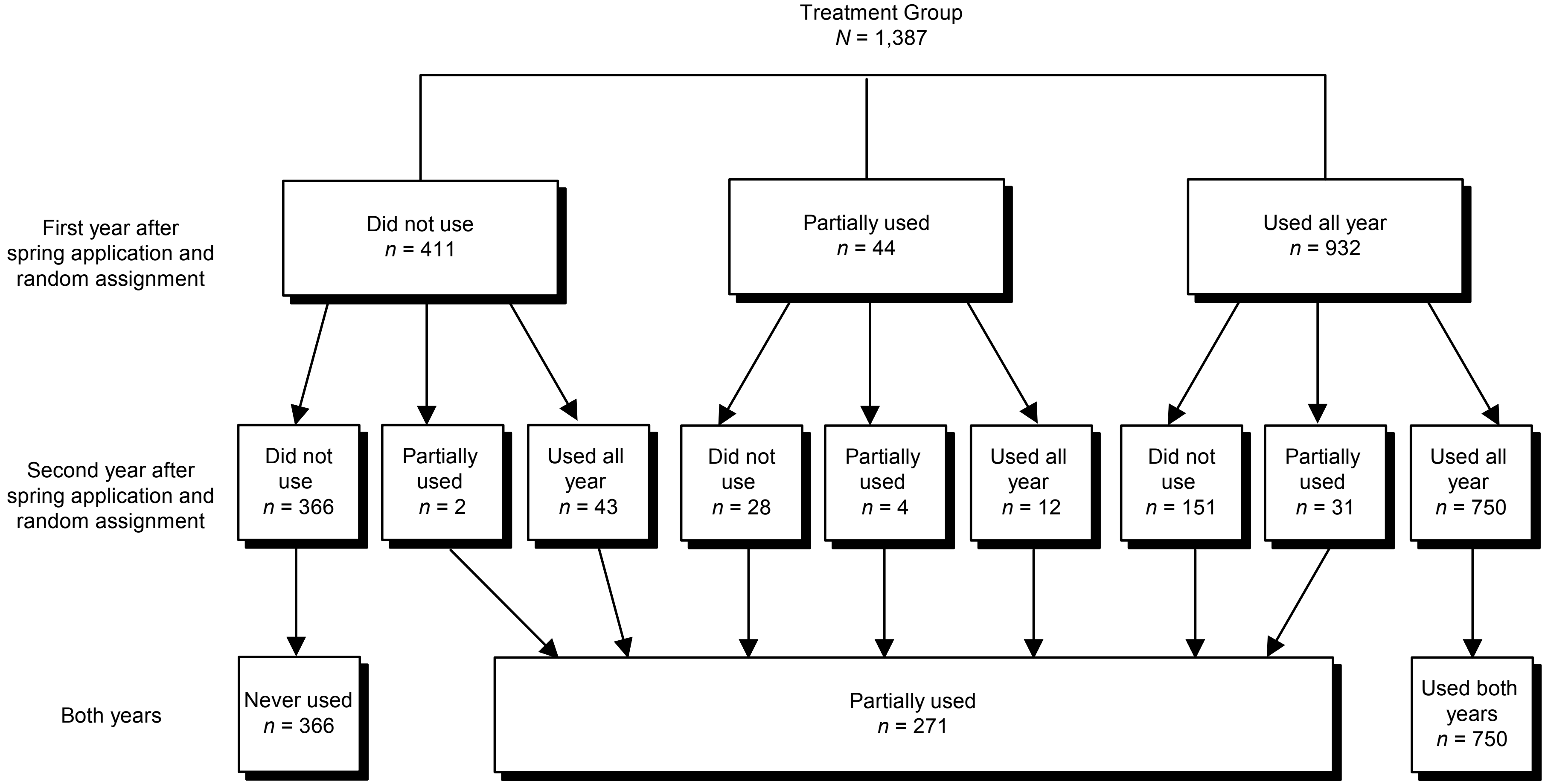
- 366 out of 1,387 (26 percent) treatment group students never used the OSP scholarships offered to them;

- 271 treatment students (20 percent) used their scholarships during some but not all of the first 2 years after the scholarship award. Among these students are 41 of the 179 students who either partially or fully used their scholarship in year 1, but were estimated to be "forced decliners" in year 2, meaning that they could not continue to use their scholarship because they "earned out" (their family income grew to exceed the Program's eligibility requirements) or because there was no space for them in a participating high school;[38] and

- The remaining 750 treatment group students (54 percent) used their scholarship during the entire 2 years after the scholarship lottery.

[38] The calculations regarding likely forced decliners were made using administrative data provided by the WSF. A total of 21 treatment students reported family income of over 200 percent of the poverty level after their first year in the Program, thereby "earning out" of subsequent Program eligibility. The estimate of the number of students forced to decline their scholarships due to the lack of high school slots was calculated by counting the number of treatment students who used a scholarship in 8th grade but declined to use it once they had advanced to 9th grade (*n* = 20). The transition from 8th to 9th grade was the focus of the analysis of slot constraints, since 9th grade is the primary intake grade for most OSP schools that serve high school students. It is impossible to know for certain if all 20 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. Therefore, the total estimate of 41 forced decliners for outcome year 2 is simply a rough estimate based on the limited data available. The actual number of forced decliners could be somewhat higher or lower than this estimate.

Figure 2-4. Scholarship Usage by Students Assigned to the Treatment Group in First 2 Years



NOTE: 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 non-users (no payments).

SOURCE: WSF's payment files.

Certain pre-program student characteristics were associated with the patterns of usage among students offered scholarships in the impact sample. Compared to treatment students who never used their scholarships, students who fully or partially used (i.e., "ever users") were significantly (table 2-3):

- More likely (58 percent compared to 43 percent) to be entering grades K-5 and less likely (6 percent compared to 20 percent) to be entering high school;
- Less likely (7 percent compared to 22 percent) to have special educational needs due to a disability;
- More likely (95 percent compared to 90 percent) to be African American; and
- Less likely (49 percent compared to 55 percent) to be male.

Compared to never users, ever users also tended to have fewer siblings and to have changed residence more recently. Ever users and never users were statistically similar regarding a number of baseline characteristics, including their test score performance, percentage having applied from SINI schools, percentage Hispanic, mother's average years of education and employment status, and family income.

Among the treatment students who ever used their scholarship, a somewhat different and smaller set of pre-program student characteristics were associated with full scholarship use. Compared to users who only partially used their scholarship, students who used their scholarship consistently for the 2-year period were significantly (table 2-4):

- Higher performing in reading and math but only if in high school (46 National Percentile Rank (NPR) points compared to 24 percent NPR points in math and 39 NPR points compared to 26 NPR points in reading); and
- More likely to be entering grades K-5 (61 percent compared to 49 percent) and less likely to be entering grades 6-8 (33 percent compared to 43 percent).

Full users and partial users were statistically similar regarding a number of baseline characteristics, including percentage having applied from SINI schools, race and ethnicity, gender, percentage with special needs, mother's average years of education and employment status, and various measures of family demographics.

Table 2-3. Baseline Characteristics of Treatment Group Students Who Ever Used Their OSP Scholarship Compared to Never Users in the First 2 Years

Characteristic	Ever User	Never User	Difference
Achievement:			
Reading percentile: Grade K-5	34.21	31.94	2.27
Reading percentile: Grade 6-8	34.95	31.00	3.95
Reading percentile: Grade 9-12	34.7	29.2	5.50
Percent missing	38.79	37.43	
Math percentile: Grade K-5	29.02	28.38	.64
Math percentile: Grade 6-8	36.88	34.52	2.37
Math percentile: Grade 9-12	38.64	38.92	-.28
Percent missing	16.75	29.51	
Student demographics:			
Percent SINI	30.75	31.42	-.67
Percent entering: Grade K-5	57.88	43.17	14.72**
Percent entering: Grade 6-8	35.85	37.16	-1.31
Percent entering: Grade 9-12	6.27	19.67	-13.40**
Percent missing	0	0	
Percent learning/physical disability	7.27	21.78	-14.51**
Percent missing	7.05	7.92	
Percent African American	95.44	89.91	5.53**
Percent missing	7.64	10.66	
Percent Hispanic	10.32	14.37	-4.05
Percent missing	6.07	6.83	
Percent Male	49.26	55.34	-6.08*
Percent missing	.20	.27	
Family demographics:			
Mother's average years of education	12.55	12.62	-.07
Percent missing	15.57	20.49	
Percent mother full-time job	44.01	42.66	1.35
Percent missing	16.55	19.55	
Average family income	$17,133.73	$17,033.79	$99.94
Percent missing	0	0	
Number of children	2.83	3.02	-.19*
Percent missing	0.20	1.09	
Months of residential stability	69.94	83.96	-14.01**
Percent missing	2.45	3.55	
Sample size (unweighted)	1,021	366	

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Data are not weighted. Ever users include full users and partial users. 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 non-users (no payments).

SOURCES: OSP applications and WSF's payment files.

Table 2-4. Baseline Characteristics of Treatment Group Students Who Fully Used Their OSP Scholarship Compared to Partial Users in First 2 Years

Characteristic	Full User	Partial User	Difference
Achievement:			
Reading percentile: Grade K-5	34.91	31.86	3.05
Reading percentile: Grade 6-8	35.22	34.36	.86
Reading percentile: Grade 9-12	38.95	26.20	12.75*
Percent missing	40.67	33.58	
Math percentile: Grade K-5	29.45	27.51	1.94
Math percentile: Grade 6-8	38.13	34.23	3.90
Math percentile: Grade 9-12	46.4	23.86	22.54**
Percent missing	17.33	15.13	
Student demographics (percent):			
Percent SINI	29.60	33.95	-4.35
Percent entering: Grade K-5	61.2	48.71	12.49**
Percent entering: Grade 6-8	33.20	43.17	-9.97**
Percent entering: Grade 9-12	5.6	8.12	-2.52
Percent missing	0	0	
Percent learning/physical disability	6.63	9.06	-2.43
Percent missing	6.87	7.56	
Percent African American	95.63	94.94	.69
Percent missing	8.53	5.17	
Percent Hispanic	10.59	9.56	1.03
Percent missing	5.60	7.38	
Percent male	47.59	53.87	-6.28
Percent missing	.27	0	
Family demographics:			
Mother's average years of education	12.52	12.64	-.12
Percent missing	15.60	15.50	
Percent mother full-time job	44.64	42.29	2.35
Percent missing	16.67	16.24	
Average family income	$17,341.74	$16,558.06	$783.68
Percent missing	0	0	
Number of children	2.83	2.84	-.01
Percent missing	0.13	.37	
Months of residential stability	69.96	69.89	.07
Percent missing	1.87	4.06	
Sample size (unweighted)	750	271	

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Data are not weighted. 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 non-users (no payments).

SOURCES: OSP applications and WSF's payment files.

Reasons for Not Participating Among the Treatment Group

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

- Lack of available space in the private school they wanted their child to attend, reported by 29 percent of these parents. For the parents of decliner students entering the high school grades, 50 percent listed "no space" as a reason for declining, compared to 26 percent of the parents of decliner students entering grades K-8;[39]

- Unable to find a participating school that offered services for their child's learning or physical disability or other special needs (17 percent of these parents);

- Child was accepted into a public charter school (16 percent of these parents);

- Moved outside of the DC area, and therefore no longer eligible for the Program (11 percent of these parents); and

- The location of the preferred private school made it difficult to use the scholarship (10 percent of these parents).

Parents whose children initially used a scholarship but subsequently decided to leave their chosen private school also were asked during year 2 data collection why they discontinued their scholarship use (table 2-6). The most common response given by these 48 treatment group parents was that the child did not get the academic support that the child needed (54 percent). Additionally, 21 percent of the parents of partial users said that their child did not like the private school, and 15 percent indicated that there was another private school their child liked better. Nineteen percent of parents who discontinued their child's scholarship use described the discipline at the private school as too strict. None of the remaining reasons offered to explain the partial use of a scholarship were reported by more than 8.3 percent of this small subgroup of parents whose children used a scholarship for less than the full 2-year impact period.

[39] Parents of cohort 2 decliners were about as likely as those of cohort 1 decliners to list a lack of available space as the reason for not using.

Table 2-5. Reasons Given by Parents of Treatment Group Students for Never Using an OSP Scholarship in Year 2

Reason Given by Parent for Child Not Using the Offer of the OSP Scholarship	Percent of Parents
There was no space at the participating private school that the child wanted to attend	29.4
The private school(s) did not have the services for the child's learning or physical disability or other special needs	17.0
Child got into a charter school	16.3
The child moved out of DC	10.5
The private school(s) the child was interested in were too far from home or too hard to get to	9.8
The private school the child wanted to attend was not participating	7.8
The child did not want to leave his/her friends in public school	6.5
Child did not want to be held back a grade	4.6
Child did not pass the private school's admission test	4.6
Child's public school teachers are better	3.3
Child thought the work might be too hard in the private school(s)	2.6
Child did not want to have religious instruction	2.0
Child's public school has sports that the private school(s) did not	2.0
Total respondents	153

NOTES: Responses are unweighted. Respondents were able to select multiple responses, which generated a total of 180 responses provided by 153 parents. This equates to an average of 1.2 responses per parent. Responses that were not selected are unreported, and categories with responses from fewer than three parents are not reported for confidentiality reasons.

SOURCE: Impact Evaluation Parent Surveys.

Table 2-6. Reasons Given by Parents of Treatment Group Students Who Left a Participating OSP Private School in Year 2

Reason Given by Parent for Child Not Staying in the Participating Private School Chosen with the Offer of the Scholarship	Percent of Parents
Child did not get the academic support he/she needed at the private school	54.2
Child did not like the private school	20.8
The discipline/rules were too strict at the private school	18.8
There was another private school the child liked better	14.6
The religious activities at the private school made the child uncomfortable	8.3
The work at the private school was too hard	6.3
It was too hard to get the child to the private school each day	6.3
Total respondents	48

NOTES: Responses are unweighted. Respondents were able to select multiple responses, which generated a total of 64 responses provided by 48 parents. This equates to an average of 1.3 responses per parent. Responses that were not selected are unreported, and categories with responses from fewer than three parents are not reported for confidentiality reasons.

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 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 non-participating private school. As indicated earlier, 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.[40]

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. In the subsequent 2 years, there was substantial movement across educational sectors (table 2-7).

Table 2-7. Percent of the Impact Sample by Type of School Attended: At Baseline, in Year 1, and in Year 2

	Baseline Year		Year 1			Year 2		
	Public		Public			Public		
	Traditional	Charter	Traditional	Charter	Private	Traditional	Charter	Private
Treatment	75.8	24.2	11.6	5.2	83.2	14.2	7.3	78.6
Control	73.7	26.3	60.5	26.8	12.6	51.0	35.0	14.0
Difference	2.1	-2.1	-48.9	-21.6	70.5	-36.8	-27.8	64.6

NOTES: 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* = 1,985. Percent missing baseline: Treatment = 5.4, Control = 9.9; percent missing year 1: Treatment = 18.5, Control = 42.9; percent missing year 2: Treatment = 23.9, Control = 47.6. Data are unweighted and represent actual responses. Given the high rates of missing data, readers are cautioned against drawing firm conclusions.

SOURCES: OSP applications and Impact Evaluation Parent Surveys.

[40] 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.

Based on data from survey respondents[41] in the first year:

- 83 percent of the treatment group and 13 percent of the control group moved from a public school to a private school;
- 12 percent of the treatment group and 61 percent of the control group attended a traditional public school; and
- The remaining 5 percent of the treatment group and 27 percent of the control group were enrolled in public charter schools.

Between the first and second year after they applied to the OSP:

- About 5 percent of the treatment group (83.2-78.6) who had been in a private school the first year moved back to public schools, dividing themselves about evenly between traditional public and public charter schools.
- The control group continued to exit traditional public schools in favor of private and charter schools. In year 2, a total of 51 percent were still in traditional public schools, while the share in private schools grew from 13 to 14 percent and the proportion in public charter schools increased from 27 percent to 35 percent.

These data show how assignment to treatment is not perfectly correlated with private school attendance and that assignment to the control group does not necessarily entail attendance at a traditional public school.[42] 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.

The enrollment patterns of students who attended schools designated as in need of improvement (SINI) is a special focus of this evaluation, given that Congress assigned that specific group of students to be the highest service priority of the OSP (Section 306). Among the applicant parents in the impact sample who provided the identity of their child's school (table 2-8):

[41] The subset of survey respondents in the treatment group are disproportionately 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 the WSF OSP payment file, which is used to calculate the Program-wide scholarship usage rates, does not contain any information on the types of schools attended by treatment decliners or control group members.

[42] These descriptive data regarding the types of school attended 1 and 2 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 (68 percent). Readers are cautioned not to draw conclusions about the impact of the OSP in causing these patterns of school-sector enrollments.

- 56 percent of the treatment and 52 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 ever).

- One year after random assignment, the number of treatment group students reportedly attending SINI-ever schools declined from 56 percent to 10 percent, while the number of control group students in such schools dropped from 52 percent to 43 percent.

- Two years after random assignment, 14 percent of treatment group students were reportedly attending SINI-ever public schools compared with 46 percent of control group students.

Table 2-8. Percentage of the Impact Sample Attending Schools Identified as in Need of Improvement (SINI): At Baseline, in Year 1, and in Year 2

	Baseline Year		Year 1			Year 2		
	SINI-ever Schools	SINI-never Schools	SINI-ever Schools	SINI-never Schools	Private	SINI-ever Schools	SINI-never Schools	Private
Treatment	55.7	44.3	10.1	6.8	83.2	13.7	7.7	78.6
Control	52.3	47.8	43.0	44.4	12.6	46.4	39.6	14.0
Difference	3.4	-3.4	-33.0	-37.6	70.5	-29.4	-32.0	64.6

NOTES: Schools were identified as SINI ever if they were officially designated as in need of improvement under the *Elementary and Secondary Education Act* between 2003 and 2005. 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* = 1,985. Percent missing baseline: Treatment = 5.4, Control = 9.9; percent missing year 1: Treatment = 18.5, Control = 42.9; percent missing year 2: Treatment = 23.9, Control = 47.6. Data are unweighted and represent actual responses. Given the high rates of missing data, readers are cautioned against drawing firm conclusions.

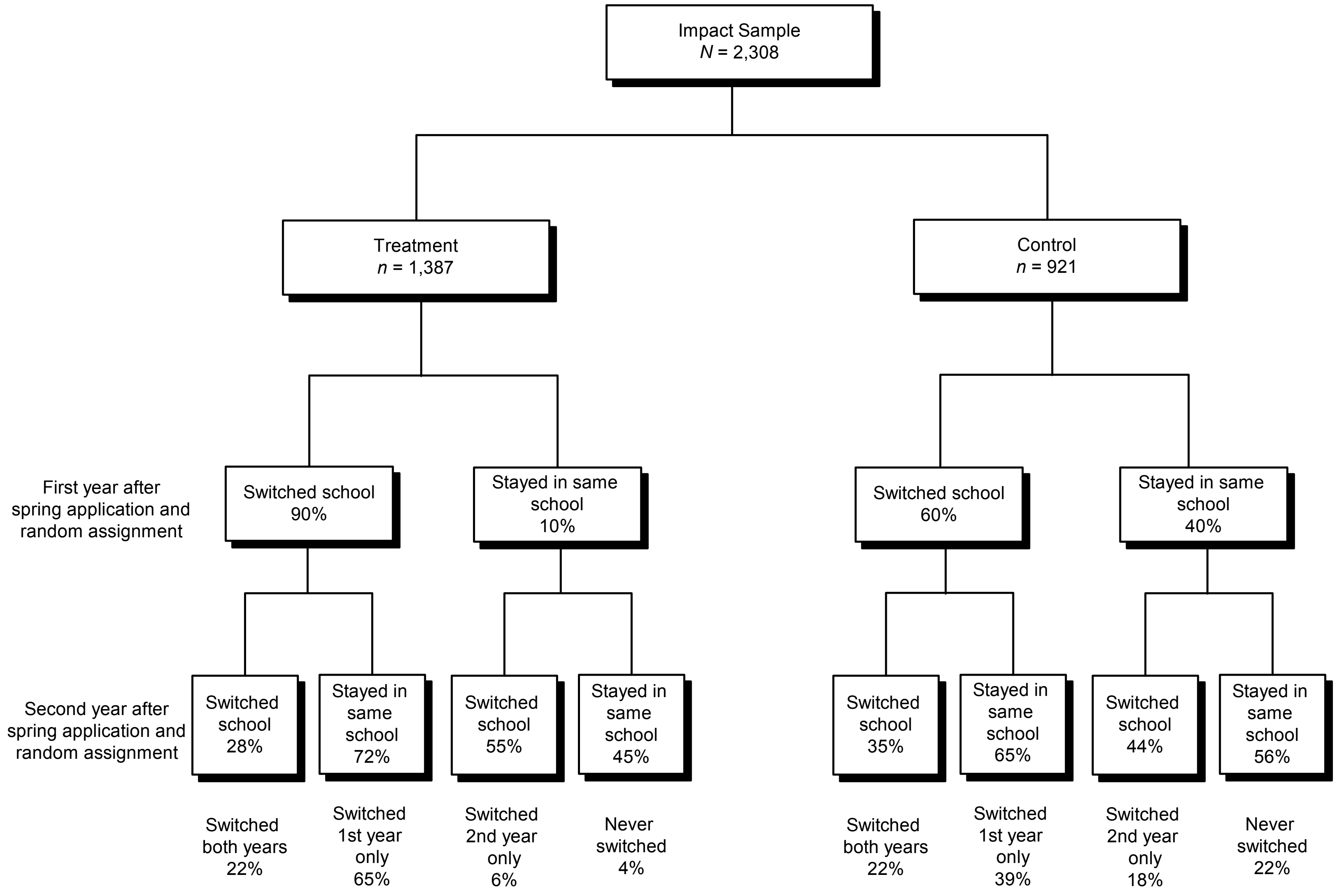
SOURCES: OSP applications and Impact Evaluation Parent Surveys.

The movement of impact sample students between public (both traditional and charter) and private schools or between SINI and non-SINI schools masks some additional transitions because students can change schools within the same sector. That is, some students moved from one charter school to another, or one private school to another. In terms of general student mobility, between the time students applied to the OSP and the next year (figure 2-5):

- 90 percent of the treatment group switched schools.
- 60 percent of the control group switched schools.[43]

[43] These represent an updating of findings reported in Wolf et al. 2007, p 6. The previous report estimated the year 1 school switching rates as 91 percent for the treatment group and 57 percent for the control group. The slight difference in the year 1 switching rates in that report and those presented here is the result of a reduction in missing data. The research team had access to WSF's payment files for the first time in late 2007 to provide additional information about school switching among the treatment group. The research team also made telephone calls to control parents who did not initially respond to the Impact Evaluation Parent Survey. Information from those two additional sources changed the previously reported rates slightly.

Figure 2-5. Movement of the Impact Sample Between Schools During the First 2 Years


Impact Sample
N = 2,308
Treatment
n = 1,387
Control
n = 921
First year after spring application and random assignment
Switched school 90%
Stayed in same school 10%
Switched school 60%
Stayed in same school 40%
Second year after spring application and random assignment
Switched school 28%
Stayed in same school 72%
Switched school 55%
Stayed in same school 45%
Switched school 35%
Stayed in same school 65%
Switched school 44%
Stayed in same school 56%
Switched both years 22%
Switched 1st year only 65%
Switched 2nd year only 6%
Never switched 4%
Switched both years 22%
Switched 1st year only 39%
Switched 2nd year only 18%
Never switched 22%

NOTES: Percent missing: after 1 year—Treatment = 18 percent and Control = 42 percent; after 2 years—Treatment = 28 percent and Control = 59 percent. Given the high rates of missing data, readers are cautioned against drawing firm conclusions.

SOURCES: OSP applications and Impact Evaluation Parent Surveys.

During the second year in the program:

- 28 percent of the treatment group students who switched schools during the first year switched schools again, while 55 percent of the treatment group who did not switch during the first year switched during the second year.

- 35 percent of the control group students who switched schools during the first year switched schools again, while 44 percent of the control group who did not switch during the first year switched during the second year.

Over the course of both years:

- Among the treatment group, 4 percent remained in the same school they were in when they applied to the Program, 71 percent switched schools once, and 25 percent switched schools twice during the 2-year period since application.

- Among the control group, 22 percent remained in the same school they were in when they applied to the Program, 57 percent switched schools once, and 21 percent switched schools twice during the 2-year period since application.

Both groups experienced higher rates of school mobility than the typical annual rate for urban students (22 to 28 percent).[44] However, the treatment group switched schools at a higher rate than the control group over the course of the first 2 years of the Program.[45]

[44] See Witte 2000, p. 144; and Wong, Dreeben, Lynn, and Sunderman 1997, p. 17.

[45] In an Ordered Logit estimation of the number of school switches experienced by students in the impact sample, the treatment variable was a statistically significant predictor of school switching ($Z = 3.69$, $p < .0001$).

3. Impacts on Key Outcomes, 2 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. This chapter presents the impacts of the Program on these outcomes 2 years after families and students applied to the OSP, or approximately 19 months after the start of their first possible school year in the Program. The first section provides an overview of the impacts 1 year after random assignment, as reported previously (Wolf et al. 2007). The second section summarizes the analytic methods used to determine the results and the techniques used to display them. Section 3 presents the impacts on student achievement. The fourth section discusses the safety impacts. Section 5 presents the satisfaction impacts. The sixth section provides a brief summary of the chapter findings.

3.1 Year 1 Impacts Reported Previously

The first year analysis reported the following findings regarding the impacts of a scholarship offer (Wolf et al. 2007, table ES-2):

- The main models indicated that the Program generated no statistically significant impacts, positive or negative, on student reading or math achievement for the entire impact sample in year 1. One of the two specifications that made up the sensitivity test indicated a positive and statistically significant math impact of 3.4 scale score points.

- No statistically significant achievement impacts were observed for the high-priority subgroup of students who had attended a SINI public school under *NCLB* before applying to the Program.

- The Program may have had an impact on math achievement for two subgroups of students with baseline characteristics associated with better academic preparation. The main models suggest that the OSP improved the math achievement of participating students who had not attended a SINI school by 4.7 scale score points and increased the math scores of those with relatively higher test score performance at baseline by 4.3 scale score points. However, these findings should be interpreted with caution, as adjustments for multiple comparisons suggested they may be false discoveries.

- No significant achievement impacts were observed for other subgroups of participating students, including those with lower test scores at baseline, girls, boys, elementary students, secondary students, or students within each of the individual cohorts that in combination made up the impact sample.

- The Program had a statistically significant positive impact on parents' views of school safety but not on students' actual school experiences with dangerous activities. Parents in the treatment group perceived their child's school to be less dangerous (an impact of -0.74 on a 10-point scale) than parents in the control group. Student reports of dangerous incidents in school did not differ systematically between the treatment and control groups.

- The Program also had an impact on parent satisfaction with their child's school. For example, an additional 19 percent of the parents of students in the treatment group graded their child's school "A" or "B" on a scale of A through F compared with the parents of control group students.

- For the most part, students' satisfaction with their school was unaffected by the Program. The main exception was for students with lower test score performance at baseline, who on average assigned their schools significantly lower grades. For example, 60 percent of this treatment subgroup graded their school "A" or "B" compared to 81 percent of the control subgroup.

These were the results of the analysis of data collected 1 year after random assignment and about 7 months into the students' new educational experiences if they were offered a scholarship. The results presented in the remainder of this report are based on data collected 2 years after random assignment and about 19 months into any new educational experiences that may have been induced by the scholarship offer.

3.2 Analytic and Presentation Approaches

For each key outcome that is a focus of the evaluation, we present the impacts of being awarded 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). The second set of results (the "Impact on the Treated" or IOT analysis) builds off of any statistically significant findings from the ITT analysis while adjusting for the rate of scholarship non-use. Appendix A (sections A.8 through A.10) provides a more detailed description of the analytic methods used for both types of analyses.

Table 3-1. Overview of the Analytic Approaches

Research Question	Approach
• What is the impact of being awarded (offered) an OSP scholarship?	Intent to Treat (ITT) Analysis 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?	Impact on the Treated (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) the approximately one-quarter who received a scholarship offer but declined to use it (the "decliners"); and (2) the hypothesized 2.3 percent 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").

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 subgroups of policy interest is also 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 or not students attended a school in need of improvement 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 (Mayer et al. 2002, appendix E; Barnard et al. 2003).

- **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; Sommers 2001) and therefore educational interventions might have differential effects on students based on their gender.

- **Whether students were in grades K-8 or 9-12 at the time of application.** Previous research found that elementary and high school education experiences differ in significant ways (e.g., Torgesen et al. 2007). Moreover, students entering the elementary or high school grades at the baseline of this study faced different sets of participating schools from which to choose, suggesting that the impact of the Program may differ for the two subgroups.

- **Whether students were in cohort 1 (applied in 2004) or cohort 2 (applied in 2005).** Cohort 1 students faced a different set of participating schools, and fewer slot constraints in those schools, than did cohort 2 students, conditions that could generate variance in Program impacts. Previous scholarship evaluations have examined whether achievement effects varied across study cohorts (Mayer et al. 2002, appendix D).

In presenting the results, 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 A, section A.8:

- The text and tables include "effect sizes" (ES) to translate each impact into a standard metric and to allow the reader to assess whether the size of the impact might be considered meaningful, whether or not it is statistically significant.[46]

- 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 evaluation research.

- A reliability test was administered to the results drawn from multiple comparisons of treatment and control group members across the 10 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). The estimates of the treatment impacts on parent and student perceptions of safety were not adjusted, since each was estimated using a single safety index. Although the treatment impact on perceptions of parent and student satisfaction is

[46] Specifically, the effect sizes are computed as a percentage of a standard deviation for the control group after 2 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 A, section A.2) forecasts that this year 2 evaluation will contain sufficient data to correctly identify an overall reading or math impact of the offer of a scholarship of .11-.12 standard deviations if such an impact actually exists. Subgroup ITT impacts are estimated to be detectable at various sizes, ranging from .14 to .38 standard deviations. Previous experimental evaluations of programs similar to the OSP have reported statistically significant overall achievement impacts only in math, of .16 to .24 standard deviations (Rouse 1998, p. 584), and in both reading and math of .25 standard deviations (Greene 2001, p. 57) after 2 years. Statistically significant achievement impacts for subgroups of African American participants of .28 standard deviations (Howell et al. 2006 p. 151) have been reported, also after 2 years. The effect sizes reported in previous experimental scholarship evaluations are based on estimations of the impact on the treated, whereas the Minimum Detectable Effects identified by the power analysis for this study are based on estimations of the impact of the scholarship offer.

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, reducing the danger of chance false discoveries in that specific outcome domain.

- The impact results from the primary analysis were subject to sensitivity tests involving a sample trimmed to exactly equalize the treatment and control response rates and the clustering of student observations on the school attended instead of family. These analyses were conducted to assess how robust the 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.

The final effective response rates for the year 2 analysis varied by data collection instrument, membership in the treatment or control group, and cohort. For the test score analysis, the overall effective response rate was 72.5 percent. Test scores were obtained from 74.6 percent of the treatment group and 69.3 percent of the control group. The overall test score response rate for cohort 1 was 68.6 percent and for cohort 2 was 73.6 percent. The effective response rate for the parent survey was 72.4 percent overall, with data provided for 74.8 percent of the treatment group and 68.8 percent of the control group. The parent survey effective response rate for cohort 1 was 67.0 percent and for cohort 2 was 73.8 percent. The administration of the student survey generated an overall effective response rate of 67.7 percent. A total of 71.8 percent of the treatment group students and 61.8 percent of the control group students provided survey data in year 2. The student survey effective response rate for cohort 1 was 58.1 and for cohort 2 was 71.9. Sample weights were used in all impact estimations to re-establish the equivalence of the treatment and control groups in the face of differential rates of assignment to treatment, non-random study attrition, and the statistical subsampling that was conducted in particular to increase the effective response rate for the control group (see appendix A, section A.7).

3.3 Impacts on Student Achievement

The statute clearly identifies students' academic achievement as the primary outcome to be measured as part of the evaluation. This emphasis is consistent with 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).

In summary, the analysis revealed:

- No significant impacts of the Program, either positive or negative, overall on student achievement after 2 years.

- No significant achievement impacts for students who came from SINI schools, the subgroup of students for whom the statute gave top priority.

- Among the other nine subgroups examined, there were no statistically significant test score differences between the treatment and control groups for students with lower performance at baseline, male students, female students, or elementary or high school students.

- Positive Programmatic impacts were observed in reading achievement for participants who applied from non-SINI schools, those who applied to the Program with relatively higher levels of academic performance, and students from the first cohort of applicants. This pattern of results holds for both the impact of being offered a scholarship and the impact of using a scholarship. However, these positive subgroup findings should be interpreted with caution, as reliability tests indicate that they could be false discoveries.

Impacts for the Full Sample of Students

Overall, the primary analysis indicated there were no statistically significant general impacts of the Program on reading or math achievement after 2 years. That is, the ITT analysis indicates that the outcome test scores of the treatment group as a whole, on average, were not significantly different from those of the control group as a whole in the second year (table 3-2).[47] Thus, neither the offer of a scholarship nor the use of a scholarship had an impact on achievement for students in general. In one of two robustness checks to the main analytic approach (equalizing the response rates), the impact estimate on reading achievement for the full sample of 3.17 (ES = .09) rises to 4.57 (ES = .12) and crosses the threshold to be statistically significant (appendix C). Otherwise, the sensitivity testing generates results consistent with the main analysis.

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

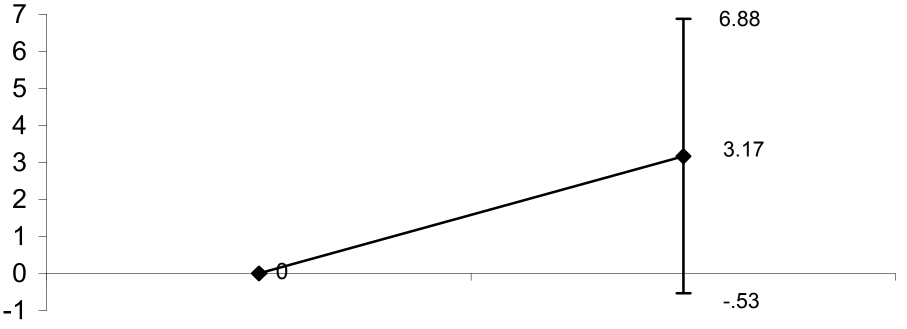
Table 3-2. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Academic Achievement (ITT)

Student Achievement	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Reading	621.30	618.12	3.17	.09	.09
Math	614.09	613.85	.23	.01	.89

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for reading = 1,580; math = 1,585. Separate reading and math sample weights were used.

While there were differences between the treatment and control groups in reading and math, none reached the 95 percent confidence level for statistical significance. This outcome can be viewed most clearly in figures 3-1 and 3-2. The confidence interval for the regression-adjusted difference between the treatment and control groups in reading outcomes ranges from a negative .53 to a positive 6.88, and includes the value zero.[48] Even though the estimate of the treatment impact on reading scale scores is a gain of about three points, it could plausibly lie anywhere within the interval; therefore, we are uncertain if the general reading impact is positive, zero, or negative. The same is true for the estimate of the general treatment impact on math scale scores. The statistical estimate of the Program's impact on math is a gain of .23 points; however, the actual impact could have been as high as 3.53 or as low as -3.07.

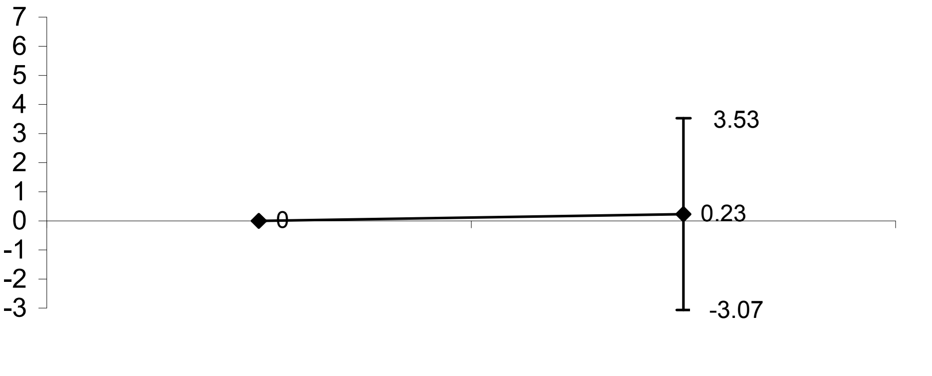
Figure 3-1. Regression-Adjusted Impact: Reading



NOTES: Valid *N* for reading = 1,580. The point on the vertical line (3.17) is the statistical estimate of the Program impact on reading gains in terms of scale score points. The high and low bounds of the vertical line illustrate the 95 percent confidence level associated with the estimate.

[48] 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 5th graders and 703.6 (SD = 36.5) for students in 12th grade.

Figure 3-2. Regression-Adjusted Impact: Math



NOTES: Valid *N* for math = 1,585. The point on the vertical line (.23) is the statistical estimate of the Program impact on math gains in terms of scale score points. The high and low bounds of the vertical line illustrate the 95 percent confidence level associated with the estimate.

Subgroup Impacts

The offer of a scholarship, and therefore also the use of a scholarship, did not appear to have an impact on academic achievement in the second year for most of the subgroups of students examined (table 3-3). That is, there were no statistically significant differences between the treatment and control groups in reading or math test scores for students defined in the following ways:

- Students who applied from a school designated SINI between 2003 and 2005—the highest service priority for the Program according to the statute;
- Students who entered the Program with relatively low academic achievement in reading and math;
- Males;
- Females;
- Students in either K-8 or in high school; and
- Students in application cohort 2.

Table 3-3. Year 2 Impact Estimates of the Offer of a Scholarship on Subgroups: Academic Achievement (ITT)

	Reading				
Student Achievement Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
SINI ever	640.47	640.48	-.01	-.00	1.00
SINI never	606.39	600.68	5.71*	.15	.04
Difference	34.09	39.80	-5.72	-.15	.12
Lower performance	597.68	599.27	-1.59	-.05	.65
Higher performance	631.66	626.43	5.23*	.15	.02
Difference	-33.98	-27.16	-6.81	-.18	.09
Male	616.89	613.00	3.90	.11	.17
Female	625.29	622.80	2.50	.07	.31
Difference	-8.40	-9.80	1.40	.04	.71
K-8	609.12	605.34	3.79	.10	.08
9-12	678.59	678.40	.19	.01	.96
Difference	-69.47	-73.06	3.59	.06	.38
Cohort 2	608.88	607.22	1.66	.04	.42
Cohort 1	664.96	656.23	8.74*	.27	.04
Difference	-56.08	-49.01	-7.07	-.19	.13

	Math				
Student Achievement Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
SINI ever	636.79	635.52	1.28	.05	.58
SINI never	596.46	597.05	-.59	-.02	.81
Difference	40.34	38.47	1.87	.06	.58
Lower performance	595.85	598.43	-2.58	-.09	.43
Higher performance	622.00	620.50	1.50	.05	.43
Difference	-26.15	-22.07	-4.08	-.12	.27
Male	612.30	611.78	.52	.02	.85
Female	615.69	615.72	-.03	-.00	.99
Difference	-3.39	-3.94	.55	.02	.88
K-8	601.35	600.44	.91	.03	.63
9-12	673.94	677.02	-3.08	-.14	.29
Difference	-72.59	-76.58	3.99	.12	.25
Cohort 2	600.33	600.25	.08	.00	.97
Cohort 1	662.37	661.58	.80	.03	.80
Difference	-62.05	-61.33	-.72	-.02	.84

*Statistically significant at the 95 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for reading = 1,580, including: SINI ever *N* = 687, SINI never *N* = 893, Lower performance *N* = 493, Higher performance *N* = 1,087, Male *N* = 782, Female *N* = 798, K-8 *N* = 1,354, 9-12 *N* = 226, Cohort 2 *N* = 1,262, Cohort 1 *N* = 318. Valid *N* for math = 1,585, including SINI ever *N* = 690, SINI never *N* = 895, Lower performance *N* = 492, Higher performance *N* = 1,093, Male *N* = 782, Female *N* = 803, K-8 *N* = 1,359, 9-12 *N* = 226, Cohort 2 *N* = 1,267, Cohort 1 *N* = 318. Separate reading and math sample weights were used.

However, based on estimates from the primary analysis, the Program did appear to have an impact on reading test scores in year 2 for certain subgroups of students, including at least two subgroups who applied with a relative advantage in academic preparation (table 3-3; table 3-4):

- Students who had attended non-SINI public schools prior to the Program scored an average of 5.7 scale score points higher in reading (ES = .15) if they were in the treatment group (the impact of the offer of a scholarship); the calculated impact of using a scholarship was 6.9 scale score points (ES = .18).
- Students who entered the Program in the higher two-thirds of the applicant test-score performance distribution—averaging a 43 National Percentile Rank (NPR) in reading at baseline—scored an average of 5.2 scale score points higher in reading (ES = .15) if they were in the treatment group; the impact of using a scholarship for this group was 6.3 scale score points (ES = .18).
- Students from the first cohort of applicants scored an average of 8.7 scale score points higher in reading (ES = .27) if they were in the treatment group; the impact of using a scholarship was 12.2 scale score points (ES = .37) for this group.

It is useful to place the estimated effect sizes for these subgroup impacts in context. For the SINI-never impacts on reading, the effect of .15 of a standard deviation (impact of scholarship offer) and .18 of a standard deviation (impact of scholarship use) equate to a NPR difference of 3.67 NPR points and 4.41 NPR points, respectively, on the standardized SAT-9 assessment.[49] Given the year 2 average scores of the control group, which provide the counterfactual for this experimental analysis, these figures indicate that the OSP raised non-SINI applicants' reading test scores from 29.45 NPRs to 33.12 NPRs (scholarship offer) and to 33.86 NPRs (scholarship users). For students who were higher performing at baseline, the scholarship offer raised their reading test scores from 36.52 NPRs to 40.12 NPRs and the scholarship use raised their reading test score to 40.84 NPRs. Finally, for students who were in the first cohort of applicants, the scholarship offer led to an increase in reading scores from 23.67 NPRs to 29.76 NPRs, and scholarship use raised them to 32.02 NPRs.

The three statistically significant subgroup impacts of the OSP on reading scores observed in this second year evaluation were the product of a subgroup analysis involving multiple comparisons of treatment and control group members. Statistical adjustments to account for the multiple comparisons suggest that the three significant subgroup achievement impacts in reading may be false discoveries and therefore should be interpreted with caution (see appendix B, table B-1).

[49] The standard deviations for the control group year 2 reading scores were 24.4957 NPRs for SINI-never students, 24.00 NPRs for higher baseline performers, and 22.57 NPRs for cohort 1 students.

Table 3-4. Year 2 Impact Estimates of Using a Scholarship on Subgroups: Academic Achievement (IOT)

Student Achievement Subgroups	Original ITT Estimates		Usage Rate	Single Bloom Adjustment	Program-Enabled Crossover	Double Bloom Adjustment
	Impact	*p*-value				
SINI never: Reading	5.71*	.04	85.6	6.67*	2.3	6.85*
(Effect Size)	.15			.18		.18
Higher performance: Reading	5.23*	.02	84.8	6.16*	2.3	6.33*
(Effect Size)	.15			.17		.18
Cohort 1: Reading	8.74*	.04	74.1	11.79*	2.3	12.17*
(Effect Size)	.27			.36		.37

*Statistically significant at the 95 percent confidence level.

NOTES: IOT estimates limited to impacts determined to be statistically significant in the ITT analysis, since non-significant impacts are understood to be zero. Valid *N* for reading = 1,580. SINI-never reading subgroup *N* = 893. Higher performance reading subgroup *N* = 1,087. Cohort 1 reading subgroup *N* = 318. Reading sample weights were used. Impacts are displayed in terms of scale scores.

3.4 Impacts on Reported School Safety/Danger

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).

In summary, the analysis suggests that:

- Overall, treatment group parents were less likely to report serious concerns about school danger compared to control group parents.
- Parents of students who applied to the Program from SINI schools reported being significantly less concerned about school danger if their child received a scholarship.
- Treatment group parents of non-SINI students, higher baseline performers, males, females, students in grades K-8, cohort 1 students, and cohort 2 students all reported school danger concerns that were significantly lower than their counterpart parents in the control group. Additional reliability tests indicated that none of these subgroup findings are likely to be false discoveries.

- Treatment and control group students, on the other hand, did not report experiencing differences in dangerous activities at school.

- This pattern of findings is consistent for the impact of a scholarship offer and the impact of scholarship use.

- The school danger impacts estimated for both parents and students, in general and for subgroups, were not affected by the sensitivity tests conducted.

Parent Self-Reports

Overall, the parents of students offered an Opportunity Scholarship in the lottery subsequently reported their child's school to be less dangerous than did the parents of students in the control group. The impact of the offer of a scholarship on parental perceptions of school danger was -0.94 on a 10-point index, an effect size of 0.27 standard deviations (see table 3-5). The impact of using a scholarship was -1.18 on the index, with an effect size of .34 standard deviations (tables 3-6 and 3-7). 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). The index of school danger items used here includes a variety of sources of possible parental concern, including school violence, weapons, teasing, truancy, etc (see appendix A, section A.3 for more information).

This impact of the offer of a scholarship on parental concerns about school danger and disorder was consistent across most subgroups of students (see table 3-5), including parents of students from SINI (ES = -.35) and non-SINI schools (ES = -21), parents of male (ES = -.27) and female students (ES = -.27), parents of students who entered the Program with higher levels of academic achievement (ES = -.32), parents of students in grades K-8 (ES = -.27), and parents of both cohort 1 (ES = -30) and cohort 2 (ES = -.27). All of these subgroup impacts on parental views of school safety remained statistically significant after adjustments to account for multiple comparisons (see appendix B, table B-2).

Because the impacts of the scholarship offer on perceptions of safety were statistically significant for these subgroups of parents, the Programmatic impacts on actual scholarship users also are statistically significant. For example, the impact of using a scholarship on parental reports of school danger for these affected subgroups ranged from -.86 for SINI-never parents to -1.62 for SINI-ever parents (table 3-6), which equates to subgroup effect sizes ranging from -.26 to -.46 standard deviations (table 3-7).

Table 3-5. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample and Subgroups: Parent Reports of School Danger (ITT)

School Danger: Parents	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Full sample	2.06	3.00	-.94**	-.27	.00
SINI ever	2.25	3.47	-1.22**	-.35	.00
SINI never	1.91	2.63	-.71**	-.21	.01
Difference	.33	.84	-.51	-.15	.22
Lower performance	2.37	2.91	-.53	-.16	.14
Higher performance	1.91	3.04	-1.12**	-.32	.00
Difference	.46	-.13	.59	.17	.16
Male	2.00	2.94	-.94**	-.27	.00
Female	2.11	3.04	-.94**	-.27	.00
Difference	-.10	-.10	.00	.00	1.00
K-8	1.91	2.84	-.92**	-.27	.00
9-12	2.74	3.75	-1.01	-.28	.06
Difference	-.83	-.92	.09	.02	.88
Cohort 2	1.92	2.83	-.91**	-.27	.00
Cohort 1	2.53	3.57	-1.04*	-.30	.04
Difference	-.60	-.73	.13	.04	.81

*Statistically significant at the 95 percent confidence level.
**Statistically significant at the 99 percent confidence level.
NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* = 1,555, including: SINI ever *N* = 674, SINI never *N* = 881, Lower performance *N* = 488, Higher performance *N* = 1,067, Male *N* = 772, Female *N* = 783, K-8 *N* = 1,336, 9-12 *N* = 219, Cohort 2 *N* = 1,251, Cohort 1 *N* = 304. Parent survey weights were used.

Table 3-6. Year 2 Impact Estimates of Using a Scholarship on the Full Sample and Subgroups: Parent Reports of School Danger (IOT)

	Original ITT Estimates					
School Danger: Parents	Impact	*p*-value	Usage Rate	Single Bloom Adjustment	Program-Enabled Crossover	Double Bloom Adjustment
Full sample	-.94**	.00	81.9	-1.15**	2.3	-1.18**
SINI ever	-1.22**	.00	78.0	-1.57**	2.3	-1.62**
SINI never	-.71**	.01	84.8	-.84**	2.3	-.86**
Higher performance	-1.12**	.00	83.9	-1.34**	2.3	-1.37**
Male	-.94**	.00	79.0	-1.19**	2.3	-1.22**
Female	-.94**	.00	84.8	-1.11**	2.3	-1.14**
K-8	-.92**	.00	85.1	-1.09**	2.3	-1.12**
Cohort 2	-.91**	.00	83.9	-1.09**	2.3	-1.12**
Cohort 1	-1.04*	.04	73.7	-1.41*	2.3	-1.46*

*Statistically significant at the 95 percent confidence level.
**Statistically significant at the 99 percent confidence level.
NOTES: Valid *N* = 1,555, including: SINI ever *N* = 674, SINI never *N* = 881, Higher performance *N* = 1,067, Male *N* = 772, Female *N* = 783, K-8 *N* = 1,336, Cohort 2 *N* = 1,251, Cohort 1 *N* = 304. Parent survey weights were used.

Table 3-7. Effect Sizes for Statistically Significant Impact Estimates of Using a Scholarship on Subgroups: Parent Reports of School Danger (IOT)

School Danger: Parents	Original ITT Estimates	Single Bloom Adjustment	Double Bloom Adjustment
Full sample	-.27	-.33	-.34
SINI ever	-.35	-.45	-.46
SINI never	-.21	-.25	-.26
Higher performance	-.32	-.38	-.39
Male	-.27	-.35	-.36
Female	-.27	-.32	-.33
K-8	-.27	-.32	-.33
Cohort 2	-.27	-.32	-.33
Cohort 1	-.30	-.40	-.42

NOTES: Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* = 1,555, including: SINI ever *N* = 674, SINI never *N* = 881, Higher performance *N* = 1,067, Male *N* = 772, Female *N* = 783, K-8 *N* = 1,336, Cohort 2 *N* = 1,251, Cohort 1 *N* = 304. Parent survey weights were used.

However, for high school students and those who applied to the Program with lower levels of academic achievement, there were no significant differences in their parents' perceptions of school danger.

Student Self-Reports

The students in grades 4-12 who completed surveys paint a different picture about dangerous activities at their school than do their parents. The student index of school danger 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 danger by students offered scholarships through the lottery were not statistically different from those of the control group (table 3-8). That is, there was no evidence of an impact from the offer of a scholarship or the use of a scholarship on students' reports of dangerous activities. No statistically significant findings were found across the subgroups analyzed. Nor did the sensitivity tests conducted lead to a different set of findings (see appendix C).

Table 3-8. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample and Subgroups: Student Perceptions of School Danger (ITT)

School Danger: Students	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Full sample	1.90	1.93	-.02	-.01	.87
SINI ever	1.96	1.78	-.17	.09	.40
SINI never	1.86	2.04	-.18	-.10	.36
Difference	.10	-.26	.35	.19	.22
Lower performance	1.93	1.86	.07	.03	.81
Higher performance	1.89	1.95	-.05	-.03	.73
Difference	.03	-.09	.12	.07	.70
Male	2.07	2.00	.07	.04	.74
Female	1.76	1.86	-.11	-.06	.57
Difference	.31	.13	.18	.09	.53
4-8	1.98	1.97	.01	.01	.94
9-12	1.53	1.73	-.20	-.11	.44
Difference	.45	.23	.21	.12	.50
Cohort 2	1.92	1.92	.00	-.00	.99
Cohort 1	1.85	1.95	-.10	-.06	.68
Difference	.07	-.03	.10	.05	.74

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* = 1,025, including: SINI ever *N* = 551, SINI never *N* = 474, Lower performance *N* = 317, Higher performance *N* = 708, Male *N* = 502, Female *N* = 523, K-8 *N* = 814, 9-12 *N* = 211, cohort 2 *N* = 760, cohort 1 *N* = 265. Student survey weights were used. Survey was 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 (see Section 309 of the *District of Columbia School Choice Incentive Act of 2003*). 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 with their school was measured in three ways—the percentage that assigned a grade of A or B, a standard grade-point average based on a 5-point A-F grade scale, and a satisfaction scale.[50] In summary, the analysis suggests that in year 2:

- Treatment group parents overall reported being more satisfied than parents of control group students across all three satisfaction measures.
- Parents of students who applied to the Program from SINI schools reported significantly higher levels of school satisfaction if their child had been awarded an Opportunity Scholarship, a finding that remained statistically significant after adjustments to guard against false discoveries (see appendix B, tables B-3, B-4, and B-5).
- Six of the other nine subgroups of parents reported significantly higher school satisfaction across all three of the measures if they were in the treatment group. Statistical adjustments for multiple comparisons indicated that only one of those 18 findings—for the parents of lower performing students assigning their school a grade of A or B—is at risk of being a false discovery (see appendix B, tables B-3, B-4, and B-5).
- Satisfaction impacts from the subgroup of parents whose students were lower performing at baseline were statistically significant for two of the three satisfaction measures (not the grade-point-average rating).
- No statistically significant satisfaction impacts were observed for the grades 9-12 and cohort 1 subgroups of parents.
- There were no treatment impacts overall on student satisfaction with school.
- Students who applied from SINI schools—the highest priority subgroup—were more satisfied with school if they were in the Program in year 2 on all three measures of satisfaction.
- The sensitivity testing we conducted did not alter the findings (see appendix C).

Parent Self-Reports

Nineteen months after the start of their experience with the OSP, parents overall are more satisfied with their child's school if they were offered a scholarship and if their child used a scholarship to attend a participating private school. The three different measures of parent satisfaction all show

[50] The parent satisfaction scale used in the analysis comprised 12 separate items asking how dissatisfied or satisfied they were with a variety of characteristics of their child's school, including location, academics, teachers, facilities, safety, communication, and parental support. Parents rated their degree of satisfaction with each of the items based on a 4-point scale, and a summary scale across all 12 items was constructed using Item Response Theory (IRT) techniques. Students answered similar questions, and a similar scale was constructed from their responses. For information about all three satisfaction measures used for parents and students see appendix A, sections A.3 and A.4.

statistically significant positive impacts of the Program on parental evaluations of their child's school (table 3-9 and table 3-11):

- A total of 76 percent of treatment parents assigned their child's school a grade of A or B compared with 63 percent of control parents—a difference of 13 percentage points (impact of the offer of a scholarship); the impact of using a scholarship was a difference of 16 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 .26 and .33, respectively.

- On a standard grade-point scale of A-F, the average grade assigned to the school by parents of treatment students was .29 (ES = .29) of a grade point higher than that of control parents; the impact of using a scholarship was .37 (ES = .36) of a grade point higher.

- Parents of students offered a scholarship scored an average of 2.67 points (ES = .33) higher than parents of students in the control group on the school satisfaction index; the impact of using a scholarship was 3.36 (ES = .42) points higher.

Table 3-9. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Parent Reports of Satisfaction with Their Child's School (ITT)

Outcome	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parents who gave school a grade of A or B	.76	.63	.13**	.26	.00
Average grade parent gave school (5.0 scale)	4.02	3.73	.29**	.29	.00
School satisfaction scale	26.12	23.44	2.67**	.33	.00

**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for school grade = 1,549; parent satisfaction = 1,571. Parent survey weights were used. School satisfaction scale was IRT scored and had a range of .96 to 35.43. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

The impact of the Program in year 2—both the offer and use of a scholarship—on parental satisfaction was positive and consistent across the various subgroups of participants (table 3-10, table 3-11, and table 3-12), with effect sizes ranging from .17 to .39 standard deviations for the offer of a scholarship and .24 to .50 for the use of a scholarship. This includes parents of SINI-ever students reporting higher levels of satisfaction with their child's school if they had been offered a scholarship, with effect sizes ranging from .26 to .38 for the offer of a scholarship, and .35 to .50 standard deviations for the use of a scholarship. There were some exceptions to the general tendency of the treatment to produce significant impacts on parental satisfaction. Specifically:

Table 3-10. Year 2 Impact Estimates of the Offer of a Scholarship on Subgroups: Parent Reports of Satisfaction with Their Child's School (ITT)

Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Parents Who Gave Their Child's School a Grade of A or B					
SINI ever	.70	.57	.13**	.26	.00
SINI never	.81	.69	.12**	.27	.00
Difference	-.11	-.12	.01	.01	.92
Lower performance	.68	.58	.11*	.22	.03
Higher performance	.79	.66	.14**	.29	.00
Difference	-.11	-.08	-.03	-.06	.62
Male	.73	.65	.09*	.18	.02
Female	.79	.62	.16**	.34	.00
Difference	-.05	.03	-.08	-.17	.17
K-8	.80	.64	.16**	.33	.00
9-12	.58	.59	-.01	.02	.89
Difference	.21	.05	.16*	.34	.02
Cohort 2	.79	.66	.14**	.29	.00
Cohort 1	.64	.56	.09	.18	.16
Difference	.15	.10	.05	.11	.44
Average Grade Parent Gave School (5.0 Scale)					
SINI ever	3.87	3.56	.31**	.29	.00
SINI never	4.14	3.86	.28**	.29	.00
Difference	-.26	3.66	.04	.04	.75
Lower performance	3.83	3.64	.18	.17	.10
Higher performance	4.11	3.77	.34**	.34	.00
Difference	-.28	-.12	-.16	-.16	.22
Male	3.96	3.79	.17*	.17	.03
Female	4.08	3.67	.41**	.39	.00
Difference	-.12	.12	-.24*	-.23	.03
K-8	4.10	3.74	.36**	.34	.00
9-12	3.65	3.66	-.01	-.01	.93
Difference	.45	.08	.37*	.36	.02
Cohort 2	4.08	3.78	.31**	.30	.00
Cohort 1	3.81	3.55	.25	.25	.06
Difference	.28	.22	.05	.05	.72

Table 3-10. Year 2 Impact Estimates of the Offer of a Scholarship on Subgroups: Parent Reports of Satisfaction withTheir Child's School (ITT)—(continued)

Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
School Satisfaction Scale					
SINI ever	25.07	21.86	3.21**	.38	.00
SINI never	26.92	24.68	2.25**	.30	.00
Difference	-1.85	-2.82	.97	.12	.29
Lower performance	24.85	22.80	2.05*	.24	.02
Higher performance	26.66	23.72	2.95**	.38	.00
Difference	-1.81	-.92	-.89	-.11	.39
Male	26.31	23.64	2.67**	.34	.00
Female	25.95	23.27	2.68**	.33	.00
Difference	.36	.36	-.00	-.00	1.00
K-8	26.52	23.68	2.84**	.35	.00
9-12	24.18	22.31	1.88	.25	.10
Difference	2.34	1.38	.96	.12	.43
Cohort 2	26.60	23.60	3.00**	.38	.00
Cohort 1	24.31	22.88	1.44	.18	.19
Difference	2.29	.72	1.57	.20	.19

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for school grade = 1,549, including: SINI ever *N* = 675, SINI never *N* = 874, Lower performance *N* = 479, Higher performance *N* = 1,070, Male *N* = 768, Female *N* = 781, K-8 *N* = 1,334, 9-12 *N* = 215, Cohort 2 *N* = 1,247, Cohort 1 *N* = 302. Valid N for parent satisfaction = 1,571, including: SINI ever *N* = 683, SINI never *N* = 888, Lower performance = 495, Higher performance *N* = 1,076, Male *N* = 776, Female *N* = 795, K-8 *N* = 1,350, 9-12 *N* = 221, Cohort 2 *N* = 1,266, Cohort 1 *N* = 305. Parent survey weights were used. School satisfaction scale was IRT scored and had a range of .96 to 35.43. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Table 3-11. Year 2 Impact Estimates of Using a Scholarship on the Full Sample and Subgroups: Parent Reports of Satisfaction with Their Child's School (IOT)

School Satisfaction: Parents	Original ITT Estimates		Usage Rate	Single Bloom Adjustment	Program-Enabled Crossover	Double Bloom Adjustment
	Impact	*p*-value				
School grade of A or B	.13**	.00	81.9	.15**	2.3	.16**
(Effect Size)	.26			.32		.33
School grade, 5.0 scale	.29**	.00	81.9	.36**	2.3	.37**
(Effect Size)	.29			.35		.36
School satisfaction scale	2.67**	.00	81.9	3.26**	2.3	3.36**
(Effect Size)	.33			.41		.42
Parents Who Gave Their Child's School a Grade of A or B						
SINI ever	.13**	.00	78.0	.17**	2.3	.17**
SINI never	.12**	.00	84.8	.15**	2.3	.15**
Lower performance	.11*	.03	77.8	.14*	2.3	.14*
Higher performance	.14**	.00	83.9	.16**	2.3	.17**
Male	.09*	.02	79.0	.11*	2.3	.12*
Female	.16**	.00	84.8	.19**	2.3	.20**
K-8	.16**	.00	85.1	.19**	2.3	.19**
Cohort 2	.14**	.00	83.9	.17**	2.3	.17**
Average Grade Parent Gave School (5.0 Scale)						
SINI ever	.31**	.00	78.0	.40**	2.3	.42**
SINI never	.28**	.00	84.8	.33**	2.3	.34**
Higher performance	.34**	.00	83.9	.41**	2.3	.42**
Male	.17*	.03	79.0	.22*	2.3	.23*
Female	.41**	.00	84.8	.48**	2.3	.49**
K-8	.36**	.00	85.1	.42**	2.3	.43**
Cohort 2	.31**	.00	83.9	.36**	2.3	.37**
School Satisfaction Scale						
SINI ever	3.21**	.00	78.0	4.12**	2.3	4.25**
SINI never	2.25**	.00	84.8	2.65**	2.3	2.72**
Lower performance	2.05*	.02	77.8	2.64*	2.3	2.72*
Higher performance	2.95**	.00	83.9	3.51**	2.3	3.61**
Male	2.67**	.00	79.0	3.38**	2.3	3.49**
Female	2.68**	.00	84.8	3.15**	2.3	3.24**
K-8	2.84**	.00	85.1	3.33**	2.3	3.43**
Cohort 2	3.00**	.00	83.9	3.58**	2.3	3.68**

* Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for school grade = 1,549, including: SINI ever *N* = 675, SINI never *N* = 874, Lower performance *N* = 479, Higher performance *N* = 1,070, Male *N* = 768, Female *N* = 781, K-8 *N* = 1,334, 9-12 *N* = 215, Cohort 2 *N* = 1,247, Cohort 1 *N* = 302. Valid N for parent satisfaction = 1,571, including: SINI ever *N* = 683, SINI never *N* = 888, Lower performance = 495, Higher performance *N* = 1,076, Male *N* = 776, Female *N* = 795, K-8 *N* = 1,350, Cohort 2 *N* = 1,266. Parent survey weights were used. School satisfaction scale was IRT scored and had a range of .96 to 35.43. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Table 3-12. Effect Sizes for Statistically Significant Impact Estimates of Using a Scholarship on Subgroups: Parent Reports of Satisfaction with Their Child's School (IOT)

School Satisfaction: Parents	Original ITT Estimates	Single Bloom Adjustment	Double Bloom Adjustment
Parents Who Gave Their Child's School a Grade of A or B			
SINI ever	.26	.34	.35
SINI never	.27	.31	.32
Lower performance	.22	.28	.29
Higher performance	.29	.34	.35
Male	.18	.23	.24
Female	.34	.40	.41
K-8	.33	.39	.40
Cohort 2	.29	.35	.36
Average Grade Parent Gave School (5.0 Scale)			
SINI ever	.29	.37	.39
SINI never	.29	.34	.35
Higher performance	.34	.41	.42
Male	.17	.22	.22
Female	.39	.47	.48
K-8	.34	.41	.42
Cohort 2	.30	.35	.36
School Satisfaction Scale			
SINI ever	.38	.49	.50
SINI never	.30	.36	.37
Lower performance	.24	.31	.32
Higher performance	.38	.45	.47
Male	.34	.43	.44
Female	.33	.39	.40
K-8	.35	.41	.42
Cohort 2	.38	.45	.46

NOTES: Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for school grade = 1,549, including: SINI ever *N* = 675, SINI never *N* = 874, Lower performance *N* = 479, Higher performance *N* = 1,070, Male *N* = 768, Female *N* = 781, K-8 *N* = 1,334, 9-12 *N* = 215, Cohort 2 *N* = 1,247, Cohort 1 *N* = 302. Valid N for parent satisfaction = 1,571, including: SINI ever *N* = 683, SINI never *N* = 888, Lower performance *N* = 495, Higher performance *N* = 1,076, Male *N* = 776, Female *N* = 795, K-8 *N* = 1,350, Cohort 2 *N* = 1,266. Parent survey weights were used. School satisfaction scale was IRT scored and had a range of .96 to 35.43. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

- There was no significant difference in the average grade assigned to a child's school as a result of the treatment among parents of students who entered the Program with lower levels of academic performance.

- Treatment and control group parents of students in cohort 1 were not statistically different on all three satisfaction measures.

- On all three measures of parent satisfaction, the reports of parents of high school students did not statistically differ if they were in the Program.

- Parents of both male and female students gave their school a higher average grade if their child had been offered a scholarship (ES = .17 and .39, respectively). The impact on the average grade given (A-F scale) was significantly higher for the parents of girls than for the parents of boys (.41 of a grade point compared to .17).

All but one of the parent satisfaction impacts across subgroups that were statistically significant initially remained significant after adjustments for the fact that they were the product of multiple comparisons (see appendix B, tables B-3, B-4, and B-5). The one exception was the finding regarding the likelihood of assigning an A or B for parents of lower baseline performers, which may have been a false discovery (see appendix B, table B-3).

Student Self-Reports

As was true with the dangerous activity measures, students had a different view of their satisfaction with their schools than did their parents. Nineteen months after applying to the OSP, the responses of members of the treatment group in general did not differ significantly from those of the control group regarding school satisfaction (table 3-13).[51] Specifically, there was no evidence of an impact of the offer of a scholarship or the use of a scholarship on the three measures: on students' likelihood of assigning their school a grade of A or B, the average grade they assigned their school, or their reports of satisfaction with their school.

There was one difference observed across the subgroups of students, however. Among students from SINI schools, the highest service priority of the Program, those awarded scholarships and those who used their scholarship were more likely to grade their school favorably than were those in the control group (table 3-14, table 3-15). Regarding the probability of grading their school A or B, SINI-ever students were 12 percentage points (ES = .24) more likely to do so if offered a scholarship and 15 percentage points (ES = .31) more likely if they used one. SINI-ever students assigned their schools an

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

Table 3-13. Year 2 Impact Estimates of the Offer of a Scholarship on the Full Sample: Student Reports of Satisfaction with Their School (ITT)

Outcome	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Students who gave school a grade of A or B	.71	.68	.03	.05	.49
Average grade student gave school (5.0 scale)	3.97	3.84	.13	.12	.14
School satisfaction scale	34.12	33.24	.88	.13	.10

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for school grade = 974; student satisfaction = 1,042. Student survey weights used. School satisfaction scale was IRT scored and had a range of 9.67 to 46.89. Impact estimates reported for the dichotomous variable "students who gave school a grade of A or B" are reported as marginal effects. Survey was given to students in grades 4-12.

Table 3-14. Year 2 Impact Estimates of the Offer of a Scholarship on Subgroups: Student Reports of Satisfaction with Their School (ITT)

Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Students Who Gave Their School a Grade of A or B					
SINI ever	.68	.58	.12*	.24	.02
SINI never	.70	.76	-.06	-.14	.27
Difference	-.03	-.18	.16**	.34	.01
Lower performance	.64	.63	.01	.03	.85
Higher performance	.73	.70	.03	.07	.46
Difference	-.09	-.07	-.02	-.05	.79
Male	.70	.65	.05	.12	.30
Female	.70	.71	-.00	-.00	.97
Difference	-.00	-.06	.06	.12	.40
4-8	.74	.71	.03	.07	.46
9-12	.54	.54	-.00	-.00	.99
Difference	.20	.17	.03	.07	.69
Cohort 2	.73	.72	.01	.02	.84
Cohort 1	.61	.54	.07	.15	.26
Difference	.12	.19	-.07	-.14	.41

Table 3-14. Year 2 Impact Estimates of the Offer of a Scholarship on Subgroups: Student Reports of Satisfaction with Their School (ITT)—(continued)

Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Average Grade Student Gave School (5.0 Scale)					
SINI ever	3.94	3.66	.28*	.25	.02
SINI never	3.99	3.98	.00	.00	.97
Difference	-.09	-.32	.28	.25	.10
Lower performance	3.90	3.64	.26	.20	.19
Higher performance	3.99	3.91	.08	.07	.42
Difference	-.09	-.27	.18	.16	.41
Male	3.96	3.76	.20	.17	.12
Female	3.97	3.90	.07	.06	.57
Difference	-.01	-.14	.13	.12	.43
4-8	4.06	3.89	.17	.15	.09
9-12	3.53	3.61	-.07	-.08	.66
Difference	.52	.28	.24	.22	.22
Cohort 2	4.04	3.91	.13	.12	.19
Cohort 1	3.71	3.59	.12	.11	.49
Difference	.33	.32	.01	.01	.95
School Satisfaction Scale					
SINI ever	33.74	32.09	1.65*	.24	.03
SINI never	34.44	33.40	.26	.04	.73
Difference	-.69	-.57	1.39	.20	.18
Lower performance	33.02	32.83	.19	.03	.85
Higher performance	34.54	33.40	1.14	.16	.07
Difference	-1.53	-.57	-.95	-.14	.44
Male	34.30	32.89	1.41	.20	.06
Female	33.94	33.55	.40	.06	.58
Difference	.36	-.66	1.01	.14	.32
4-8	34.38	33.50	.87	.12	.16
9-12	32.95	32.04	.91	.15	.31
Difference	1.43	1.46	-.04	-.01	.97
Cohort 2	34.27	33.50	.77	.11	.21
Cohort 1	33.67	32.38	1.29	.18	.21
Difference	.60	1.12	-.51	-.07	.66

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression-adjusted using a consistent set of baseline covariates. Effect sizes are displayed in terms of standard deviations of the study control group distribution. Valid *N* for school grade = 974, including: SINI ever *N* = 531, SINI never *N* = 443, Lower performance *N* = 306, Higher performance *N* = 668, Male *N* = 479, Female *N* = 495, K-8 *N* = 773, 9-12 *N* = 201, Cohort 2 *N* = 720, Cohort 1 *N* = 254. Valid *N* for student satisfaction = 1,042, including: SINI ever *N* = 562, SINI never *N* = 480, Lower performance *N* = 324, Higher performance *N* = 718, Male *N* = 513, Female *N* = 529, K-8 *N* = 828, 9-12 *N* = 214, Cohort 2 *N* = 775, Cohort 1 *N* = 267. Student survey weights were used. School satisfaction scale was IRT scored and had a range of 9.67 to 46.89. 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 3-15. Year 2 Statistically Significant Impact Estimates of Using a Scholarship on SINI-Ever Subgroup: Student Reports of Satisfaction with Their School (IOT)

	Original ITT Estimates					
Outcome	Impact	*p*-value	Usage Rate	Single Bloom Adjustment	Program-Enabled Crossover	Double Bloom Adjustment
SINI ever: School grade of A or B	.12*	.02	79.3	.15*	2.3	.15*
(Effect size)	.24			.30		.31
SINI ever: School grade, 5.0 scale	.28*	.02	79.3	.35*	2.3	.36*
(Effect size)	.25			.31		.32
SINI ever: School satisfaction scale	1.65*	.03	79.3	2.08*	2.3	2.14*
(Effect size)	.24			.30		.31

*Statistically significant at the 95 percent confidence level.

NOTES: Valid *N* for school grade = 974, including SINI ever *N* = 531. Student satisfaction *N* = 1,042, including SINI ever *N* = 562. Student survey weights were used. School satisfaction scale was IRT scored and had a range of 9.67 to 46.89. Impact estimates reported for the dichotomous variable “students who gave school a grade of A or B” are reported as marginal effects. Survey was given to students in grades 4-12.

average grade-point average that was .28 points higher (ES = .25) if offered a scholarship and .36 (ES = .32) points higher if they used one. SINI-ever students reported levels on the school satisfaction scale that were about 1.6 (ES = .24) points higher if offered a scholarship and 2.1 (ES = .31) points higher if they used it. These results should be interpreted with caution, as multiple comparison adjustments suggest that these three findings may be false discoveries (see appendix B, tables B-6, B-7, and B-8).

3.6 Chapter Summary

This chapter presented the results of an analysis of experimental data on the impacts of the Opportunity Scholarship Program 2 years after the initial random assignment of students to treatment or control groups. These second-year results are similar in most respects to the first-year results reported previously. In both years, no statistically significant achievement impacts were observed for the impact sample as a whole. Subgroups with certain relative advantages at baseline—students from non-SINI schools and those with higher test scores at baseline—exhibited achievement gains as a result of being offered the treatment in year 2 as in year 1; however, the apparent gains were in reading in year 2 but in math in year 1. In both years, adjustments for multiple comparisons suggested that the subgroup achievement impacts may be false discoveries. Overall, parents in the treatment group continue to perceive their child’s school to be safer and are more satisfied with it if they were offered a scholarship. Parents of students who applied to the Program from SINI schools, the top service priority of the OSP, report higher levels of school safety and satisfaction as a result of the treatment. Unlike in the first-year evaluation, when all parent subgroups demonstrated significant satisfaction impacts, the second-year analysis suggests that the school satisfaction impacts of the Program are not statistically significant across all subgroups, though they are significant for most of them. For information about the effects of attending private school, with or without an Opportunity Scholarship, see appendix E.

4. Exploratory Analysis of OSP Intermediate Outcomes

Whatever effect the OSP has on key outcomes (most important, achievement), researchers and policymakers have long been interested in understanding the *mechanisms* by which voucher programs may or may not benefit students (e.g., 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, there has so far been little research that empirically tests these relationships (Hess and Loveless 2005).

This chapter explores how these hypotheses may be playing out for the OSP, taking two possible analytic steps to link the Program to its intended outcomes. The first, most important and straightforward stage of the analysis answers 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 are influenced by the Program—parents’ choice to use or not use a scholarship and to select a specific participating private school for their child—but they themselves are not an end outcome as identified in the OSP statute. The method used to estimate the impacts on intermediate outcomes is identical to that used to estimate impacts on the key Program outcomes (see appendix A for more detail).

The second stage of this analysis is an exploration of whether any impacts in the educational experiences and behaviors of students and parents line up with any observed impacts on academic achievement. If there are impacts on the intermediate outcomes but not on achievement, we might hypothesize that there is no underlying relationship between those outcomes and achievement or that the impacts on education experiences and behaviors were not strong enough to affect achievement. We might offer a similar hypothesis if, conversely, there were impacts on achievement but not on the intermediate outcomes. In the case of the OSP, where achievement impacts were observed only for particular subgroups of students, this second part of the investigation focuses to a large extent on the subgroup analysis.

However, even these types of hypotheses must be quite tentative because there is no way to rigorously evaluate these linkages. Study participants are randomly assigned only to the offer of a scholarship; they are not randomly assigned to the experience of various educational conditions and programs. Parents of students offered scholarships select participating private schools and the environments that the schools offer. Thus, any connection between specific educational conditions and student test score outcomes could be partly or entirely a function of the types of scholarship students and families that sort themselves into the different school choices. Controlling for student background characteristics at baseline is unlikely to completely eliminate such self-selection effects. That is why any findings from this element of the study do not suggest that we have learned what specific factors “caused” any observed test score impacts, only that certain factors emerge from the analysis as possible candidates for mediating influence.

4.1 Impact of the OSP on Intermediate Outcomes

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 (appendix F). 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 sample of students overall and for each of the subgroups of students outlined earlier, using the same analytic model used to estimate the impacts reported in chapter 3. 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).

Impacts for the Full Sample

Overall, 2 years after applying for a scholarship, the Program appears to have had an impact on 10 of the 24 intermediate outcomes (table 4-1):

- ***Home Educational Supports.*** The results suggest that the Program may have had an impact on two of four intermediate outcomes in this group. The Program appeared to produce a positive impact of .28 additional years (ES = .12) on parents' aspirations for how far in school their child would go. The Program led to students' experiencing more time spent commuting to school from their homes[52] (ES = .25). The school transit impact remained significant after adjustments for multiple comparisons, whereas those adjustments indicated that the parent aspirations impact may be a false discovery (appendix B, table B-9). There were no statistically significant differences between the treatment and control groups on parents' reports of their involvement in school in year 2 (ES = -.06) or their child's use of a tutor outside of school (ES = -.07).

- ***Student Motivation and Engagement.*** There were no statistically significant impacts on this group of intermediate indicators. Two years after they applied to the OSP, students in the treatment and control groups reported similar aspirations for future schooling (ES = -.11), frequency of doing homework (ES = -.10), time spent reading for fun (ES = .02), and engagement in extracurricular activities (ES = .08). There were no statistically significant differences in student attendance (ES = -.11) or tardiness rates (ES = -.11), as reported by parents.

- ***Instructional Characteristics.*** The offer of a scholarship appears to have had a statistically significant impact on 5 of the 10 intermediate outcomes in this group of indicators. Being offered a scholarship led to students' experiencing classes that were smaller by an average of 1.6 students as measured by student/teacher ratios (ES = -.29). The Program also led to students' experiencing a lower likelihood that their school offered either tutoring (ES = -.32) or special programs for children who were English language learners or had learning problems (ES = -.66). At the same time, however, the Program had a positive impact on the use of an in-school tutor (ES = .13), presumably in schools that made them available. The OSP led to students' experiencing a higher likelihood of being in a school that offered enrichment programs (ES = .19). All five of these impacts remained statistically significant after adjustments for multiple comparisons (appendix B, table B-10). There were no differences between the treatment and control groups in how students rated their teacher's attitude (ES = .02) or the challenge of their classes (ES = -.04), the school's use of ability grouping (ES = .13), the availability of programs for advanced learners (ES = .12), or before- and after- school programs (ES = .04).

[52] Commuting time was selected as a possible intermediate outcome because students who exercise school choice tend to attend schools that are farther from their home than is their assigned public school. Commuting time also has been shown to be associated with student achievement (Dolton et al. 2003) (see appendix F).

Table 4-1. ITT Impacts on Intermediate Outcomes as Potential Mediators in Year 2

Mediators	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Section 1. Home Educational Supports					
Parental involvement	2.94	3.05	-.11	-.06	.32
Parent aspirations	17.39	17.12	.28*	.12	.04
Out-of-school tutor usage	.12	.14	-.02	-.07	.24
School transit time	N/A	N/A	.33**	.25	.00
Section 2. Student Motivation and Engagement					
Student aspirations	16.65	16.88	-.23	-.11	.16
Attendance	N/A	N/A	-.09	-.11	.42
Tardiness	N/A	N/A	-.09	-.11	.50
Reading for fun	.40	.39	.01	.02	.81
Engagement in extracurricular activities	2.35	2.24	.11	.08	.31
Frequency of homework (days)	3.73	3.87	-.14	-.10	.14
Section 3. Instructional Characteristics					
Student/teacher ratio	12.07	13.62	-1.55**	-.29	.00
Teacher attitude	2.85	2.81	.05	.02	.79
Challenge of classes	2.47	2.52	-.05	-.04	.59
Ability grouping	.65	.59	.06	.13	.18
Availability of tutors	.60	.74	-.14**	-.32	.00
In-school tutor usage	.27	.21	.06*	.13	.02
Programs for learning problems/ ELL	.77	1.28	-.51**	-.66	.00
Programs for advanced learners	.67	.58	.09	.12	.11
Before-/after-school programs	.96	.95	.01	.04	.33
Enrichment programs	2.49	2.34	.16*	.19	.02
Section 4. School Environment					
Parent/school communication	3.08	3.07	.01	.01	.93
School size	295.84	479.09	-172.32**	-.43	.00
Percent non-white	.93	.97	-.04**	-.39	.00
Peer classroom behavior	8.27	7.92	.34*	.16	.04

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

NOTES: Valid N for Parental involvement = 1,559; Parent aspirations *N* = 1,482; Out-of-school tutor usage *N* = 1,521; School transit time *N* = 1,571; Student aspirations *N* = 960; Attendance *N* = 1,520; Tardiness *N* = 1,499; Reading for fun *N* = 1,034; Engagement in extracurricular activities *N* = 986; Frequency of homework *N* = 1,009; Student/teacher ratio *N* = 1,241; Teacher attitude *N* = 1,027; Challenge of classes *N* = 1,018; Ability grouping *N* = 925; Availability of tutors *N* = 885; In-school tutor usage *N* = 1,522; Programs for learning problems/ELL *N* = 905; Programs for advanced learners *N* = 876; Before-/after-school programs *N* = 905; Enrichment programs *N* = 905; Parent/school communications *N* = 930; School size *N* = 1,312; Percent non-white *N* = 1,213; Peer classroom behavior *N* = 1,028. Separate weights were used for items from parent surveys, student surveys, and principal surveys. Impact estimates for the dichotomous variables "Out-of-school tutor usage," "Ability grouping," "Availability of tutors," "In-school tutor usage," and "Before-/after-school programs," are reported as marginal effects. Impact estimates for the ordered categorical variables "School transit time," "Attendance," and "Tardiness" were obtained by ordered logit; because of the categorical nature of the variables, treatment and control group means convey little useful information and thus were omitted as "Not Applicable" (N/A).

- ***School Environment.*** The Program may have affected three of four measures of school environment. Students in the treatment group experienced schools that were smaller by an average of 172 students (ES = -.43) and had a smaller percentage of non-white students (ES = -.39) than the schools of the control group, impacts that remained significant after adjustments for multiple comparisons (appendix B, table B-11). Treatment group students also reported having better behaved peers in the classroom than did control group students (ES = .16); however, statistical tests suggested that may be a false discovery. There were no differences in parents' reports of how their child's school communicates with them (ES = .01).

Impacts for Subgroups

The subgroup intermediate outcome impacts are important, not only because the Program might have differentially affected students' experiences and behaviors but also because statistically significant impacts on test scores were observed only for three subgroups of students—students from non-SINI schools, students who applied to the Program with higher academic performance, and students who were in cohort 1. A large number of statistical comparisons are involved in this analysis of Programmatic impacts on 10 different subgroups for 24 different intermediate outcomes. After adjustments for these multiple comparisons, only the subgroup findings with initial high levels of statistical significance remain significant and unlikely to be false discoveries.

This exploratory subgroup analysis suggests:

- Among the home educational supports (table 4-2), the impacts on parent aspirations overall were concentrated among two parental subgroups: the parents of students who entered the Program with higher academic performance (ES = .18) and the parents of students who were applying for high school placement (ES = .38). There were no statistically significant differences between the treatment and control students in other subgroups. The average increased transit time observed for participants overall did not apply to students from non-SINI schools, those who were lower performing at baseline, were female, or from cohort 2. Statistical adjustments indicated that any and all of the significant subgroup impacts on home educational supports could be false discoveries, so they should be interpreted with caution (appendix B, table B-13).

- In general, if the Program did not have an effect on an intermediate outcome for the full sample, it also did not have an effect for subgroups. The student motivation and engagement indicators had no statistically significant impacts overall, and this generally is duplicated in subgroup analyses (table 4-3). However, there were a few instances where an insignificant overall finding had significant subgroup effects. The Program reduced the future educational aspirations of students who entered the Program with relatively lower academic performance (ES = -.40). The Program reduced the frequency of doing homework for three subgroups of students: students from non-SINI schools (ES = -.22), female students (ES = -.20), and students in the elementary grades (ES = -.24). However, the OSP also increased the frequency of

Table 4-2. Year 2 Impact Estimates for Subgroups: Home Educational Supports (ITT)

Subgroup	Parental Involvement	Parent Aspirations	Out-of-School Tutor Usage	School Transit Time
Overall impact	**-.11**	**.28***	**-.02**	**.33****
SINI ever	.10	.23	-.03	.51**
SINI never	-.29	.32	-.01	.20
Difference	.39	-.09	-.02	.32
Lower performance	.12	-.02	.00	.28
Higher performance	-.22	.41**	-.04	.36**
Difference	.34	-.43	.04	-.07
Male	.07	.32	.00	.42**
Female	-.28	.24	-.05	.25
Difference	.35	.08	.06	.17
K-8	-.17	.13	-.02	.28*
9-12	.14	1.01**	-.03	.61*
Difference	-.31	-.89*	.00	-.33
Cohort 2	-.19	.24	-.03	.25
Cohort 1	.17	.41	-.01	.66*
Difference	-.36	-.17	-.01	-.42

* Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for Parental involvement = 1,559, including: SINI ever *N* = 675, SINI never *N* = 884, Lower performance *N* = 488, Higher performance *N* = 1,071, Male *N* = 771, Female *N* = 788, K-8 *N* = 1,340, 9-12 *N* = 219, Cohort 2 *N* = 1,257, Cohort 1 *N* = 302. Valid *N* for Parent aspirations = 1,482, including: SINI ever *N* = 631, SINI never *N* = 851, Lower performance *N* = 458, Higher performance *N* = 1,024, Male *N* = 729, Female *N* = 753, K-8 *N* = 1,275, 9-12 *N* = 207, Cohort 2 *N* = 1,191, Cohort 1 *N* = 291. Valid *N* for Out-of-school tutor usage = 1,521, including: SINI ever *N* = 662, SINI never *N* = 859, Lower performance *N* = 476, Higher performance *N* = 1,045, Male *N* = 750, Female *N* = 771, K-8 *N* = 1,303, 9-12 *N* = 218, Cohort 2 *N* = 1,222, Cohort 1 *N* = 299. Valid *N* for School transit time = 1,571, including: SINI ever *N* = 678, SINI never *N* = 893, Lower performance *N* = 497, Higher performance *N* = 1,074, Male *N* = 778, Female *N* = 793, K-8 *N* = 1,349, 9-12 *N* = 222, Cohort 2 *N* = 1,268, Cohort 1 *N* = 303. 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. The shaded rows indicate subgroups that demonstrated statistically significant achievement gains in reading and that, therefore, are the focus of the analysis. Impact estimates for the dichotomous variable "Out-of-school tutor usage" are reported as marginal effects.

doing homework for students in high school (ES = .36). Statistical adjustments indicated that any and all of the significant subgroup impacts on student motivation and engagement could be false discoveries, so they should be interpreted with caution (appendix B, table B-14).

- The observed impacts on instructional characteristics showed a variable pattern across subgroups (table 4-4). For two intermediate outcomes—student-teacher ratio (ES range from -.22 to -.75) and the availability of programs for students with special learning needs (ES range from -.55 to -.97)—the negative impacts for the full sample were consistent across all subgroups of students. The Program reduced the likelihood that tutors were available at school for all subgroups (ES range from -.24 to -.85) except students who applied from non-SINI schools and those in the first cohort. In contrast, the offer of a scholarship led to the greater use of an in-school tutor for two subgroups: students who were lower performing academically at baseline (ES = .28) and students

Table 4-3. Year 2 Impact Estimates for Subgroups: Student Motivation and Engagement (ITT)

Subgroup	Student Aspirations	Attendance	Tardiness	Reading for Fun	Engagement in Extra-curricular Activities	Frequency of Homework
Overall impact	**-.23**	**-.09**	**-.09**	**.01**	**.11**	**-.14**
SINI ever	-.25	-.15	-.20	.04	.06	.02
SINI never	-.21	-.04	.01	-.01	.14	-.27*
Difference	-.03	-.11	-.22	.06	-.09	.29
Lower performance	-.83*	-.07	-.22	-.02	.27	-.14
Higher performance	-.00	-.10	-.03	.02	.04	-.15
Difference	-.83*	.03	-.19	-.05	.23	.01
Male	-.25	-.13	-.27	.02	.19	-.02
Female	-.20	-.06	.08	.00	.03	-.26*
Difference	-.05	-.07	-.36	.01	.16	.24
K-8	-.26	-.04	-.00	-.00	.12	-.30**
9-12	-.09	-.38	-.51	.07	.06	.58**
Difference	-.17	.35	.51	-.08	.06	-.88**
Cohort 2	-.22	-.15	-.10	-.01	.05	-.15
Cohort 1	-.27	-.26	-.03	.09	.31	-.11
Difference	.05	.11	-.07	-.09	-.25	-.05

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for Student Aspirations = 960, including: SINI ever *N* = 527, SINI never *N* = 433, Lower performance *N* = 297, Higher performance *N* = 663, Male *N* = 473, Female *N* = 487, K-8 *N* = 760, 9-12 *N* = 200, Cohort 2 *N* = 707, Cohort 1 *N* = 253. Valid *N* for Attendance = 1,520, including: SINI ever *N* = 649, SINI never *N* = 871, Lower performance *N* = 474, Higher performance *N* = 1,046, Male *N* = 749, Female *N* = 771, K-8 *N* = 1,311, 9-12 *N* = 209, Cohort 2 *N* = 1,227, Cohort 1 *N* = 293. Valid *N* for Tardiness = 1,499, including: SINI ever *N* = 646, SINI never *N* = 853, Lower performance *N* = 466, Higher performance *N* = 1,033, Male *N* = 741, Female *N* = 758, K-8 *N* = 1,293, 9-12 *N* = 206, Cohort 2 *N* = 1,209, Cohort 1 *N* = 290. Valid *N* for Reading for fun = 1,034, including: SINI ever *N* = 555, SINI never *N* = 479, Lower performance *N* = 319, Higher performance *N* = 715, Male *N* = 509, Female *N* = 525, K-8 *N* = 822, 9-12 *N* = 212, Cohort 2 *N* = 770, Cohort 1 *N* = 264. Valid *N* for Engagement in extracurricular activities = 986, including: SINI ever *N* = 535, SINI never *N* = 451, Lower performance *N* = 308, Higher performance *N* = 678, Male *N* = 486, Female *N* = 500, K-8 *N* = 783, 9-12 *N* = 203, Cohort 2 *N* = 731, Cohort 1 *N* = 255. Valid *N* for Frequency of homework = 1,009, including: SINI ever *N* = 542, SINI never *N* = 467, Lower performance *N* = 310, Higher performance *N* = 699, Male *N* = 499, Female *N* = 510, K-8 *N* = 798, 9-12 *N* = 211, Cohort 2 *N* = 746, Cohort 1 *N* = 263. 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. The shaded rows indicate subgroups that demonstrated statistically significant achievement gains in reading and that, therefore, are the focus of the analysis. Impact estimates for "Attendance" and "Tardiness" are derived from ordered logistic regression. Impact estimates for the dichotomous variable "Reading for fun" are reported as marginal effects. Data regarding student aspirations, reading for fun, engagement in extracurricular activities, and frequency of homework were drawn from the student survey and therefore limited to students in grades 4-12.

Table 4-4. Year 2 Impact Estimates by Subgroup: Instructional Characteristics (ITT)

Subgroup	Student/ Teacher Ratio	Teacher Attitude	Challenge of Classes	Ability Grouping	Avail-ability of Tutors	In School Tutor Usage	Programs for Learning Problems/ ELL	Programs for Advanced Learners	Before-/ After-School Programs	Enrichment Programs
Overall Impact	**-1.55****	**.05**	**-.05**	**.06**	**-.14****	**.06***	**-.51****	**.09**	**.01**	**.16***
SINI ever	-1.78**	-.07	-.14	.12	-.29**	.06	-.60**	.20*	-.02	.19
SINI never	-1.36**	.14	.02	.01	-.05	.05	-.45**	.01	.03	.14
Difference	-.42	-.20	-.17	.10	-.27**	.01	-.15	.19	-.08	.04
Lower performance	-2.28**	.12	-.08	.00	-.18**	.12**	-.59**	.13	-.00	.17
Higher performance	-1.23**	.03	-.04	.08	-.12*	.03	-.48**	.08	.02	.16*
Difference	-1.05	.09	-.04	-.07	-.06	.10	-.10	.05	-.02	.00
Male	-1.07*	-.15	-.20	.03	-.15**	.08*	-.45**	.10	-.00	.22*
Female	-2.03**	.23	.08	.08	-.11*	.03	-.58**	.09	.03	.10
Difference	.96	-.38	-.28	-.05	-.04	.05	.12	.01	-.03	.12
K-8	-1.40**	.10	.01	.03	-.11**	.07	-.50**	.09	.01	.20**
9-12	-2.31**	-.21	-.34**	.20	-.28**	.00	-.57**	.12	.01	.01
Difference	.90	.31	.35*	-.18	.17	.07	.07	-.04	.00	.21
Cohort 2	-1.71**	.11	-.07	.04	-.19**	.08	-.47**	.13*	.02	.14
Cohort 1	-1.00	-.19	.02	.10	.13	-.03	-.67**	-.03	-.00	.23
Difference	-.71	.30	-.09	-.06	-.33**	.11	.20	.16	.02	-.09

* Statistically significant at the 95 percent confidence level.
** Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for Student/Teacher Ratio = 1,241, including: SINI ever *N* = 539, SINI never *N* = 702, Lower performance *N* = 376, Higher performance *N* = 865, Male *N* = 618, Female *N* = 623, K-8 *N* = 1,077, 9-12 *N* = 164, Cohort 2 *N* = 975, Cohort 1 *N* = 266. Valid *N* for Teacher attitude = 1,027, including: SINI ever *N* = 552, SINI never *N* = 475, Lower performance *N* = 318, Higher performance *N* = 709, Male *N* = 505, Female *N* = 522, K-8 *N* = 815, 9-12 *N* = 212, Cohort 2 *N* = 762, Cohort 1 *N* = 265. Valid *N* for Challenge of classes = 1,018, including: SINI ever *N* = 546, SINI never *N* = 472, Lower performance *N* = 316, Higher performance *N* = 702, Male *N* = 504, Female *N* = 514, K-8 *N* = 806, 9-12 *N* = 212, Cohort 2 *N* = 754, Cohort 1 *N* = 264. Valid *N* for Ability grouping = 980, including: SINI ever *N* = 416, SINI never *N* = 564, Lower performance *N* = 281, Higher performance *N* = 699, Male *N* = 500, Female *N* = 480, K-8 *N* = 866, 9-12 *N* = 114, Cohort 2 *N* = 730, Cohort 1 *N* = 250. Valid *N* for Availability of tutors = 940, including: SINI ever *N* = 395, SINI never *N* = 545, Lower performance *N* = 267, Higher performance *N* = 673, Male *N* = 480, Female *N* = 460, K-8 *N* = 825, 9-12 *N* = 115, Cohort 2 *N* = 697, Cohort 1 *N* = 243. Valid *N* for In-school tutor usage = 1,522, including: SINI ever *N* = 660, SINI never *N* = 862, Lower performance *N* = 476, Higher performance *N* = 1,046, Male *N* = 753, Female *N* = 769, K-8 *N* = 1,304, 9-12 *N* = 218, Cohort 2 *N* = 1,224, Cohort 1 *N* = 298. Valid *N* for Programs for learning problems/ELL = 961, including: SINI ever *N* = 407, SINI never *N* = 554, Lower performance *N* = 274, Higher performance *N* = 687, Male *N* = 489, Female *N* = 472, K-8 *N* = 846, 9-12 *N* = 115, Cohort 2 *N* = 718, Cohort 1 *N* = 243. Valid *N* for Programs for advanced learners = 930, including: SINI ever *N* = 393, SINI never *N* = 537, Lower performance *N* = 263, Higher performance *N* = 667, Male *N* = 476, Female *N* = 454, K-8 *N* = 815, 9-12 *N* = 115, Cohort 2 *N* = 687, Cohort 1 *N* = 243. Valid *N* for Before-/after-school programs = 960, including: SINI ever *N* = 407, SINI never *N* = 553, Lower performance *N* = 274, Higher performance *N* = 686, Male *N* = 488, Female *N* = 472, K-8 *N* = 846, 9-12 *N* = 114, Cohort 2 *N* = 717, Cohort 1 *N* = 243. Valid *N* for Enrichment programs = 961, including: SINI ever *N* = 407, SINI never *N* = 554, Lower performance *N* = 274, Higher performance *N* = 687, Male *N* = 489, Female *N* = 472, K-8 *N* = 846, 9-12 *N* = 115, Cohort 2 *N* = 718, Cohort 1 *N* = 243. 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. The shaded rows indicate subgroups that demonstrated statistically significant achievement gains in reading and that, therefore, are the focus of the analysis. 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 the student survey and therefore limited to students in grades 4-12.

who were male (ES = .20). The greater availability of enrichment programs for students offered scholarships was concentrated among three subgroups: higher performing students at baseline (ES = .19), males (ES = .25), and students in grades K-8 (ES = .23). Twenty-eight of the 36 subgroup findings of impacts on intermediate outcomes remained statistically significant after adjustments for multiple comparisons (appendix B, table B-14).

- Among the school environment indicators, the reduced school size (ES range from -.33 to -.49) and smaller percentage of non-white classmates (ES range from -.36 to -1.10) for students in the Program were consistent across subgroups, except in the case of high school students, who experienced an increase in the percentage of non-white classmates (ES = .25) (table 4-5). However, the positive peer behavior impacts of the treatment, observed in the entire sample of participants were observed only for two subgroups: students from SINI schools (ES = .27) and those who were higher performing at baseline (ES = .22). There were no other subgroups of students for whom we observed statistically significant impacts on peer behavior. All but one of these statistically significant subgroup impacts on school environment factors remained significant after adjustments for multiple comparisons (appendix B, table B-15). The exception was the impact of the Program on reducing the percentage of non-white students for high school students, which may be a false discovery.

4.2 Association Between Intermediate Outcomes and Student Achievement

In order to hypothesize even a tentative link between specific intermediate outcomes and test scores, we should see a consistent pattern of impacts on both sets of measures. However, the impacts on the educational conditions and behaviors of treatment students indicated by this exploratory analysis do not, at this point, align closely with the pattern of subgroup test score impacts reported in chapter 3. Students in the three subgroups for whom there were positive reading impacts as a result of the treatment—students not from SINI schools, students with higher baseline performance, and students from cohort 1—were about as likely to experience most of these intermediate outcomes of the OSP as were members of subgroups that did not report reading impacts. As such, the year 2 impacts on intermediate outcomes reported here do not provide a solid basis for identifying possible mediators of the treatment impact on student test scores.

Table 4-5. Year 2 Impact Estimates for Subgroups: School Environment (ITT)

Subgroup	Parent/ School Communication	School Size	Percent Non-White	Peer Classroom Behavior
Overall impact	**.01**	**-172.32****	**-.04****	**.34***
SINI ever	.04	-182.82**	-.05**	.58*
SINI never	-.02	-183.51**	-.04**	.15
Difference	.06	.69	-.01	.44
Lower performance	-.00	-236.91**	-.06**	-.05
Higher performance	.01	-159.84**	-.04**	.50*
Difference	-.01	-77.07	-.02	-.54
Male	.04	-146.86**	-.07**	.25
Female	-.03	-218.41**	-.02	.43
Difference	.07	71.55	-.05**	-.18
K-8	-.05	-197.99**	-.06**	.33
9-12	.29	-107.84	.02*	.43
Difference	-.34	-90.15	-.08**	-.10
Cohort 2	-.09	-207.86**	-.04**	.30
Cohort 1	.35*	-96.79*	-.04*	.50
Difference	-.44**	-111.08*	-.00	-.20

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for Parent/school communication = 985, including: SINI ever *N* = 420, SINI never *N* = 565, Lower performance *N* = 282, Higher performance *N* = 703, Male *N* = 504, Female *N* = 481, K-8 *N* = 868, 9-12 *N* = 117, Cohort 2 *N* = 734, Cohort 1 *N* = 251. Valid *N* for School size = 1,312, including: SINI ever *N* = 574, SINI never *N* = 738, Lower performance *N* = 403, Higher performance *N* = 909, Male *N* = 655, Female *N* = 657, K-8 *N* = 1,144, 9-12 *N* = 168, Cohort 2 *N* = 1,031, Cohort 1 *N* = 281. Valid *N* for Percent non-white = 1,213, including: SINI ever *N* = 539, SINI never *N* = 674, Lower performance *N* = 371, Higher performance *N* = 842, Male *N* = 605, Female *N* = 608, K-8 *N* = 1,047, 9-12 *N* = 166, Cohort 2 *N* = 949, Cohort 1 *N* = 264. Valid *N* for Peer classroom behavior = 1,028, including: SINI ever *N* = 553, SINI never *N* = 475, Lower performance *N* = 319, Higher performance *N* = 709, Male *N* = 507, Female *N* = 521, K-8 *N* = 816, 9-12 *N* = 212, Cohort 2 *N* = 765, Cohort 1 *N* = 263. 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. The shaded rows indicate subgroups that demonstrated statistically significant achievement gains in reading and that, therefore, are the focus of the analysis. Data regarding Peer classroom behavior were drawn from the student survey and therefore limited to students in grades 4-12.

References

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.

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.

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.

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

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.

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

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.

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

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

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

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.

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

Harcourt Assessment, Inc. *Stanford-9 Technical Data Report.* San Antonio TX: Harcourt Educational Measurement, 1997.

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.

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.

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.

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.

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.

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, 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.

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.

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.

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.

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.law.harvard.edu/civilrights/].

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.

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, 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.

Sheehan, Eugene P. and Tara DuPrey. "Student Evaluations of University Teaching." *Journal of Instructional Psychology* 1999, 26(3): 188-193.

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.

Sommers, Christina Hoff. *The War Against Boys: How Misguided Feminism is Harming Our Young Men*. New York: Simon and Schuster, 2001.

Spector, Paul E. *Summated Rating Scale Construction: An Introduction.* Newbury Park, CA: Sage Publications, 1992.

Stewart, Thomas, Patrick J. Wolf, and Stephen Q. Cornman. *Parent and Student Voices on the First Year of the DC Opportunity Scholarship Program*. SCDP Report 05-01. Washington, DC: School Choice Demonstration Project, Georgetown University, 2005. Available online at [http://www.georgetown.edu/research/scdp/PSV-FirstYear.html].

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

Temple, Judy A., and Arthur J. Reynolds. "School Mobility and Achievement: Longitudinal Findings from an Urban Cohort." *Journal of School Psychology* 1999, 37: 355-377.

Torgesen, Joseph K., Greg Roberts, Sharon Vaught, Jade Wexler, David J. Francis, Mabel O. Rivera, and Nonie Lesaux. *Academic Literacy Instruction for Adolescents: A Guidance Document of the Center on Instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction, 2007.

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].

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, 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 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].

Wong, Kenneth K., Robert Dreeben, Laurence E. Lynn, Jr., and Gail L. Sunderman. *Integrated Governance as a Reform Strategy in the Chicago Public Schools*. Department of Education and Irving B. Harris Graduate School of Public Policy Studies, University of Chicago, January 1997.

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 non-awardees, except for the fact that they were offered scholarships. More precisely, there may be some non-programmatic 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). In addition, the evaluation estimates the relationship between attending a private school, regardless of whether an OSP scholarship is used, and key outcomes. These different analyses are described below in separate sections of this appendix.

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, for simplicity, we present the power analysis numbers for two representative years—the first and third outcome years.

- Central to analytic power is the sample size of study participants *who actually provide outcome information in a given year*. Thus, this power analysis factors in expected study attrition and non-response over time.

- The analysis also takes account of the correlation between baseline test scores and outcome test scores. 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. (We have, as discussed below, imputed missing baseline data, thereby producing an analysis sample with complete baseline data.)[1]

- 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.

- If all else is equal, power is greatest when the treatment and control groups are the same size. This condition, however, is not met in this evaluation. Instead, the OSP evaluation is based on the actual number of applicants, private school slots available, and the ratio of those two in the first 2 years of the Program. Because the treatment group is about 50 percent larger than the control group, our analysis will have slightly less power than a study with a comparable number of participants equally distributed across the treatment condition.

- 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" 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 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.[2]

- 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;

[1] We also estimate (but do not report) MDEs assuming no baseline characteristics. Our analysis suggests the inclusion of baseline characteristics seems to improve our MDEs slightly, reducing them by about 6 to 7 percent.

[2] 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 OSP 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.

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

α the correlation between a given student's test scores at baseline and outcome year 1, set at 0.57; and

ζ the correlation between sibling test scores (set at 0.50).

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 fairly representative of the degree to which sibling test scores are correlated.

Table A-1 presents our basic estimates of MDEs for the combined cohort evaluation sample. We present estimates for the SINI priority group and its converse (non-SINI school designation at time of application) and for the overall sample, broken down by grade band and adjusted for attrition in the first and third year of evaluation. The grade-level groupings described in the column headings are both substantively meaningful and hold some prospect of generating detectible effects. For example, K-8 is included as a grouping because it spans the entire set of elementary grades—a common educational category—and combines the smaller set of grade 6-8 applicants with the larger set of K-5 applicants. We do not present separate MDEs for applicants in grades 6-8 because there were too few of them to generate meaningful MDE estimates.

We also show the sample size and MDEs in the third evaluation year, adjusted for forecasted cumulative study attrition. We assume attrition of 20 percent of the treatment sample and 30 percent of the control sample in the first evaluation year, and 35 percent of the treatment sample and 45 percent of the control sample by the third evaluation year. These assumptions generate sample sizes that are consistent with observed baseline testing outcomes and follow-up data from the first year of the OSP.

To place these estimated effect sizes in context, an effect of 0.13 to 0.15 of a standard deviation equates to a Normal Curve Equivalent (NCE) difference of 2.73 to 3.15 NCE points.[3] 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 3.15 NCEs would bring it up to about the 24th percentile.

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

Table A-1. Minimum Detectable Effects, Combined Cohorts, By SINI Status and Grade Level

	Subtotal			Total	
Impact Sample	K-5	K-8	9-12	K-12 (First Evaluation Year)	K-12 (Third Evaluation Year)
SINI					
Treatment	174	284	59	343	278
Control	42	74	84	158	122
Subtotal SINI	216	358	143	501	400
MDE SINI sample	**0.39**	**0.29**	**0.38**	**0.21**	**0.24**
Non-SINI					
Treatment	423	713	50	763	620
Control	257	371	109	480	377
Subtotal Non-SINI	679	1,085	159	1243	997
MDE Non-SINI sample	**0.18**	**0.14**	**0.38**	**0.13**	**0.14**
All (SINI & non-SINI)					
Treatment	596	996	109	1105	898
Control	299	445	190	635	499
Total	895	1,441	299	1,740	1,397
MDE Total sample	**0.16**	**0.13**	**0.27**	**0.11**	**0.12**
Total treatment/control ratio	**2.0**	**2.2**	**0.6**	**1.7**	**1.8**

NOTES: Estimates at 80 percent power using a two-tailed hypothesis test at the 0.05 level of statistical significance.

Finally, we examined the feasibility of estimating Program effects of reasonable magnitude for other subgroups of interest to policymakers, in addition to the separate cohort and grade-level groupings discussed above, and determined that we will be able to report on the following subgroups. As shown in table A-2, the following subgroups provide sufficient power for separate impact analysis:

- **SINI designation.** Because the lotteries had to be conducted in the spring, before DCPS reports its SINI designations each August, the lottery priority group categories were always based on SINI designations that are a year behind. For the purposes of examining SINI applicants, however, it is more accurate to consider the designation for the school year in which a student applies to the DC OSP, even if that designation was not announced until the fall after the student had applied. (We refer to this as SINI ever.)

- **Gender.** Boys or girls.

- **Baseline test performance.** We are interested in the magnitude of the Program's impact on students who, at the time of random assignment, were "lower academic performers." We considered several possible "cut points" for determining the composition of the lower performing subgroup and determined that we have adequate statistical power for a group defined as at or below the bottom one-third of the baseline test score distribution.

Table A-2. Minimum Detectable Effects, Combined Cohorts by Subgroups

	Total	
Impact Subgroup	K-12 (First Evaluation Year)	K-12 (Third Evaluation Year)
SINI (Ever)		
Treatment	633	514
Control	377	296
Subtotal SINI	1,010	811
MDE SINI sample	**0.15**	**0.17**
Total treatment/control ratio	1.68	1.73
Gender: Boys		
Treatment	704	572
Control	445	350
Subtotal non-SINI	1,149	922
MDE non-SINI sample	**0.14**	**0.16**
Total treatment/control ratio	1.58	1.64
Gender: Girls		
Treatment	680	553
Control	472	371
Subtotal non-SINI	1,152	923
MDE non-SINI sample	**0.14**	**0.16**
Total treatment/control ratio	1.44	1.49
Lower baseline performers (bottom third)		
Treatment	489	397
Control	280	220
Total	769	617
MDE total sample	**0.17**	**0.19**
Total treatment/control ratio	1.74	1.81

NOTE: Estimates at 80 percent power using a two-tailed hypothesis test at the 0.05 level of statistical significance.

In summary, the analysis shows that we are able to estimate treatment effects of reasonable magnitudes in year 1 and year 3 for the overall combined-cohort impact sample, the non-SINI impact sample in year 1, and several grade-band subsamples within these two larger populations in year 1. 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 outcomes for all randomly assigned participants as well as several 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 shown below in table A-3. As shown, the impact analysis samples are defined on the basis of the elapsed time after random assignment (1, 2, and 3 years after random assignment), which for the two cohorts actually occurred in different years.

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)
Year 1 impact	Spring 2005 (1st follow-up)	Spring 2006 (1st follow-up
Year 2 impact	Spring 2006 (2nd follow-up)	Spring 2007 (2nd follow-up)
Year 3 impact	Spring 2007 (3rd follow-up)	Spring 2008 (3rd follow-up)

The full data collection activity includes the following separate sources of information:

- **Student assessments.** Baseline measures of student achievement in reading and math for public school applicants came from the SAT-9 standardized assessment administered by the 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.[4] Each spring after the baseline year, the evaluation team administers the SAT-9 to all cohort 1 and 2 students who were offered a scholarship, as well as to all

[4] 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.

members of the control group who did not receive a scholarship.[5] The testing takes place primarily on Saturdays, during the spring, in locations throughout DC arranged by the evaluators. The testing conditions are similar for members of the treatment and control groups, and the test administrators hired and trained by the evaluation team do not know whether specific students are members of the treatment or control groups. The standardized testing in reading and math provides the outcome measures for student achievement. The sample-wide response rates for these data collection instruments were 83 percent for the baseline year and effectively 73 percent for the second year follow-up assessments.[6]

- **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.[7] Each spring after the baseline year, surveys of parents of all applicants are being conducted at the Saturday testing events, while parents are waiting for their children to complete their outcome testing. The parent surveys provide the self-reported outcome measures for parental satisfaction and safety. Other topics include reasons for applying, school involvement, educational climate, and curricular offerings at the school. The response rate for this data collection instrument was 100 percent for the baseline year and effectively 72 percent for the second year follow-up.

- **Student surveys.** Each spring after the baseline year, surveys of students in grades 4 and above are being conducted at the outcome testing events. The student surveys provide the self-reported outcome measures for student satisfaction and safety. Additional topics include attitude toward school, school environment, friends and classmates, and individual activities. In the second year follow-up data collection, effectively 68 percent of students in grade 4 or higher completed surveys.

- **Principal surveys.** Each spring, surveys of principals of all public and private schools operating in the District of Columbia are being conducted. Topics include self-reports of school organization, safety, climate, principals' awareness of and response to the OSP, and, for private school principals, why they are or are not participating in the OSP. Information from the principal surveys will be analyzed in future reports to describe what is happening within the public and private schools in DC, possibly as a result of the operation of the OSP. In addition, information from principals of impact sample members (treatment and control group) is being used to assess the relationship between school characteristics and impacts. The response rate for these surveys was 52 percent in the second year follow-up data.

[5] 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).

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

[7] 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.

Outcome Measures

Congress specified in the Program statute that the rigorous evaluation study possible impacts regarding academic achievement, school safety, and satisfaction. For this second year impact report, impact estimates were produced for all three of these outcome domains: (1) academic achievement in reading and math (two measures); (2) parent self-reports of school safety (one measure) and student self-reports of school safety (one measure); and (3) parental self-reports of satisfaction (three measures) and student self-reports of satisfaction (three measures). As in this report, previous studies of scholarship program impacts have used multiple measures of the outcomes of interest because achievement, safety, and satisfaction are constructs that often cannot be measured completely or well using any single indicator (see Mayer et al. 2002; Witte 2000).

All outcome data were obtained from impact sample respondents in the spring of their first and second years after random assignment 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.[8] 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 problem solving and math procedures. This norm-referenced test is designed to measure how a student's performance compares with the scores of other students who took the test for norming purposes.[9] 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.

- **Parent self-reports of school safety.** 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:

[8] 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.

[9] 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.)

- 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.[10]

- **Student self-reports of school safety.** 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;
  - Being offered drugs;
  - Physical assault;
  - Threats of physical harm;
  - Observations of weapons being carried by other students; and
  - Bullying.

  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 6 (see Spector 1992).[11]

- **Parental self-reports of satisfaction.** Parent satisfaction with their child's school was measured three ways. First, parents were asked "What overall grade would you give this child's current school?" Two outcomes were created from this question: (1) a 5-point grading scale ranging from 1 (an F) to 5 (an A) and a dichotomous variable equal to 1 if the parent assigned an A or B, and equal to zero otherwise.

[10] 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.

[11] 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.

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;
- 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 which 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.5 below for a more detailed description of IRT.) In situations such as exist here, when individual questions each capture some piece of a more general construct (e.g., satisfaction) and the response categories capture the degree as well as the direction of the response, the IRT method is superior to count-based indices in measuring subjective conditions or traits. Two specific advantages of IRT scoring are that: (1) it allows scores to be assigned in the event that a respondent missed one of the scale items in his/her response, and (2) it identifies specific items that are highly effective in distinguishing respondents and assigns more weight to those items in the scale. For example, the IRT method is commonly used to score standardized tests. It will identify the questions that most clearly separate the better performing students from the worse performing students and count those items more heavily in generating the final test scores.

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. [12]

- **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. Students were similarly asked to rate various specific aspects of their current school on a 4-point scale. The individual items covered the following topics:
  - Behavior and discipline;
  - Academic quality;

[12] 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).

- 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. The student scale also exhibited a high level of reliability with a Cronbach's Alpha of .85.

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 assignment via the application form (including a parent survey) and administration of the SAT-9 in reading and math.[13] 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 non-response 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.[14] 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).[15] 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:[16]

- 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,

[13] Cohort 1 baseline test scores were obtained from the DCPS accountability testing database. Because DCPS administered the SAT-9 in fewer grades in 2005, baseline test scores for cohort 2 were obtained through SAT-9 administration by Westat.

[14] 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).

[15] 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 non-response weights.

[16] 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).

- 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 4-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.

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 their 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 Item Response Theory (IRT) (see Hambleton et al. 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.

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 2-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, i.e., 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, i.e., 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, e.g., 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.

The SAT-9 is designed to provide relatively intensive testing of the various aspects of reading ability for first and second graders. During the baseline test administrations, this posed a special problem for first graders, many of whom struggled to complete both of the reading subtests, and their parents, who were required to remain at the testing event longer than the parents of students in other grades. As a result, the decision was made to only administer the extensive Reading Comprehension subtest to all first graders at outcome testing and to use that subtest score as the measure of reading outcomes for those students. Other students provided some, but not all, outcome subtest scores within the two domains because they either missed or skipped entire subtests. This included 77 students of various grades (besides first) in reading, and 59 students in grades K-12 in math.[17]

Before deciding whether to include or exclude respondents who contributed only subtest scores during outcome data collection, an analysis was conducted to determine how closely subtest

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

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.[18] Given such high levels of correlations, and consistent with the principle of bringing as many observations as possible to the test score impact analysis, a decision was made to substitute subtest scores for the composite scores in the 136 cases where only the subtest scores were available. Those cases were considered respondents for the purposes of calculating the test score non-response 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.[19] 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.[20] 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.[21]

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

[19] 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.

[20] 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 was picked up by the stepwise regression procedures, while providing a set of imputations with distributional patterns similar to those of the real values.

[21] 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.

A.7 Sampling and Non-Response Weights

Sampling weights were used in the impact analyses to account for the fact that the study sample was selected differently in the 2 years of 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 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 non-response to data collection.

Base Weights

The base weight is the inverse of the probability of being assigned to either the treatment or control groups. 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

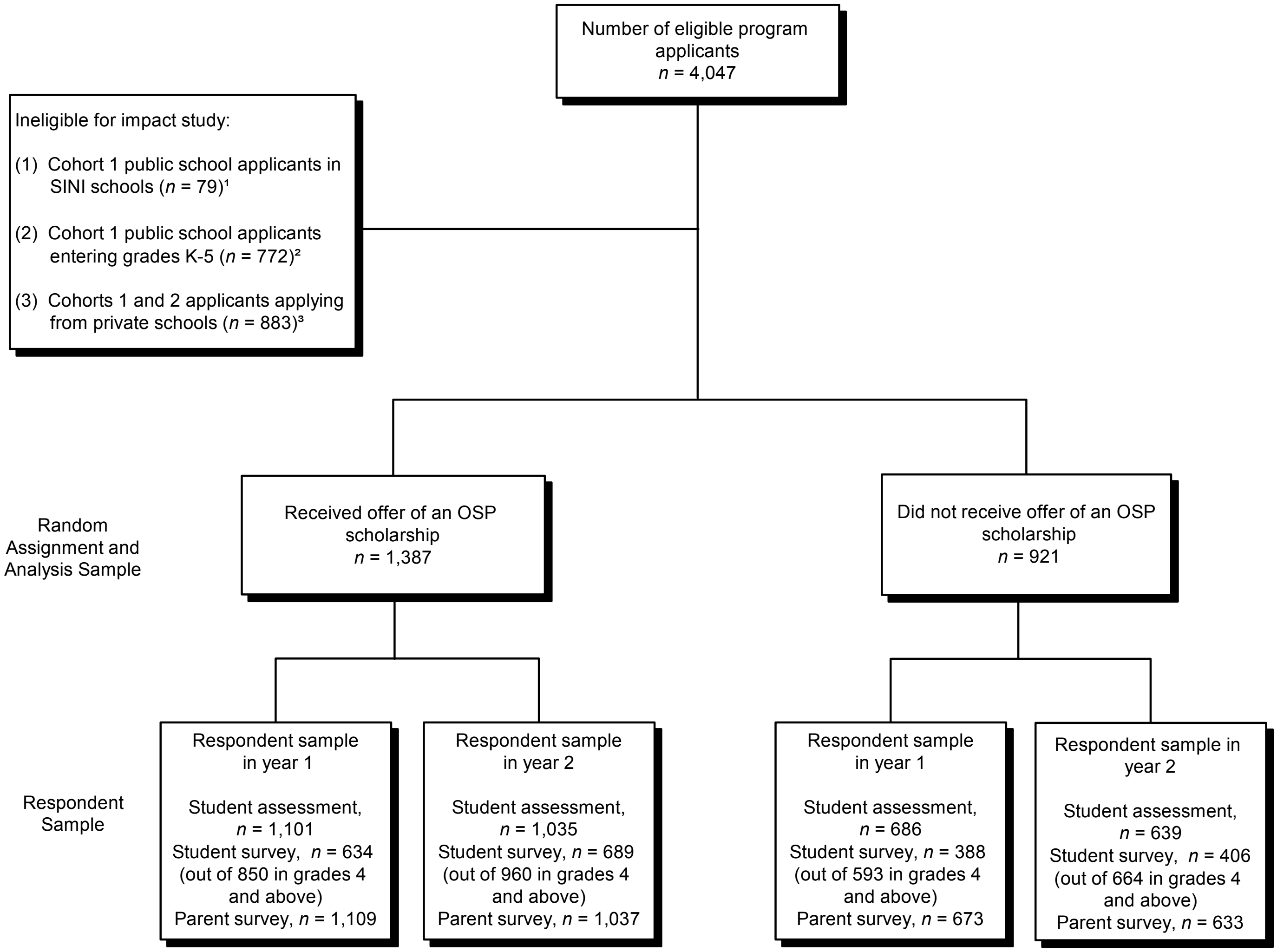
Stratum	Cohort	SINI Status	Grade Band	Treatment Sampling Rate (%)	Base Weight for Control Group	Base Weight for Treatment Group
1	Cohort 1	Non-SINI	6th to 8th	75.89	4.15	1.32
2	Cohort 1	Non-SINI	9th to 12th	28.21	1.39	3.54
3	Cohort 2	SINI	K to 5th	78.34	4.62	1.28
4	Cohort 2	SINI	6th to 8th	75.00	4.00	1.33
5	Cohort 2	SINI	9th to 12th	38.14	1.62	2.62
6	Cohort 2	Non-SINI	K to 5th	59.05	2.44	1.69
7	Cohort 2	Non-SINI	6th to 8th	55.33	2.24	1.81
8	Cohort 2	Non-SINI	9th to 12th	28.57	1.40	3.50

Adjustments for Non-Response

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 non-response, where a "respondent" was considered a student with reading or mathematics test data in year 2 (figure A-1).[22] 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

[22] 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: 2 Years After Application and Random Assignment


Number of eligible program applicants
n = 4,047
Ineligible for impact study:
(1) Cohort 1 public school applicants in SINI schools (*n* = 79)[1]
(2) Cohort 1 public school applicants entering grades K-5 (*n* = 772)[2]
(3) Cohorts 1 and 2 applicants applying from private schools (*n* = 883)[3]
Random Assignment and Analysis Sample
Received offer of an OSP scholarship
n = 1,387
Did not receive offer of an OSP scholarship
n = 921
Respondent Sample
Respondent sample in year 1
Student assessment, *n* = 1,101
Student survey, *n* = 634 (out of 850 in grades 4 and above)
Parent survey, *n* = 1,109
Respondent sample in year 2
Student assessment, *n* = 1,035
Student survey, *n* = 689 (out of 960 in grades 4 and above)
Parent survey, *n* = 1,037
Respondent sample in year 1
Student assessment, *n* = 686
Student survey, *n* = 388 (out of 593 in grades 4 and above)
Parent survey, *n* = 673
Respondent sample in year 2
Student assessment, *n* = 639
Student survey, *n* = 406 (out of 664 in grades 4 and above)
Parent survey, *n* = 633

[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.

non-response bias by compensating for different data collection response rates across various demographic groups of students organized within classification "cells." In effect, the non-response adjustment factor "spreads the weight" of the non-responding students over the 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.[23] This maintains the same mix of the impact sample across classification cells as would have been present had there been no non-response (see Howell et al. 2006, pp. 209-216; U.S. Department of Health and Human Services 2005). As a last step, the non-response-adjusted base weights were trimmed. This is done to prevent extremely large weights from unduly inflating the estimated variances and thus reducing the precision of the impact estimates.[24]

Even with the weighting protocol to adjust for non-response described above, there was a large differential between the response rates of the two experimental groups, which could undermine their comparability and therefore bias the impact analysis. After four invitations to attend data collection events, the evaluation team had obtained responses from nearly 75 percent of the treatment group but only about 53 percent of the control group (table A-5).

Table A-5. Test Score Response Rates for Year 2 Before Drawing Subsample

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)
Cohort 1 C	193	89	46.1
Cohort 1 T	299	213	71.2
Cohort 2 C	728	395	54.3
Cohort 2 T	1,088	822	75.6
Cohort 1 total	492	302	61.4
Cohort 2 total	1,816	1,217	67.0
C total	921	484	52.6
T total	1,387	1,035	74.6
Combined total	2,308	1,519	65.8

[23] To determine the factors used to create the non-response 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 non-response cells nested within the sampling strata and within treatment status. The non-response adjustment for each respondent in the cell was equal to the reciprocal of the base-weighted response rate within the cell.

[24] 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.

Recently, a new technique was developed to help reduce non-response bias in longitudinal impact analyses. Non-response subsampling is a strategy to reduce the differences between the characteristics of baseline and outcome samples by way of random sampling and non-response conversion. After the regular period of outcome data collection is over, a subsample of non-respondents is drawn and subjected to intensive efforts at non-response conversion. If initial non-response was significantly higher in one experimental group compared with the other, as was the case in this evaluation, then the subsample can be drawn exclusively from the underresponded group (e.g., controls). Each initial non-respondent 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 non-respondent 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 non-respondents that converts represents themselves as well as the same proportion of nonsampled non-respondents.

This technique was applied for the spring 2007 data collection, as it had been in 2006, to increase the outcome response rates for the control group and reduce the response rate differential across the experimental subgroups. The initial data gathering effort was followed by a targeted intensive recruitment of control group initial non-responders. A random sample of 203 of the 413 control group non-respondents was drawn (49 percent),[25] 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 non-respondent to respondent status. A total of 76 initial non-respondents were converted to respondents as a result of this effort (37 percent, with 16 from cohort 1 and 60 from cohort 2) (table A-6**.** These "converted" control group cases were more heavily weighted than the other observations in the outcome sample, by a factor of 2, to account for the complementary set of initial non-respondents 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, Kling, Duncan, and Brooks-Gunn 2006).[26] The weights ensure that each converted member of the subsample represents him or herself as well as

[25] There were 96 control group non-responders from cohort 1 and 317 from cohort 2. The random sample of 203 consisted of 43 from cohort 1 and 160 from cohort 2.

[26] For example, the Moving to Opportunity Section 8 housing voucher experimental evaluation obtained an initial year 1 response rate of 78 percent. Evaluators then drew a random sample of 30 percent of the initial non-responders 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 non-respondent 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.

another study participant: a nonrespondent like them who would have converted had they been included in the subsample. As a result of implementing this approach, the combined cohort control group response rate increased to an effective rate of 69 percent for outcome testing in math and reading, and the treatment-control response differential decreased to 5 percentage points. For other outcome measures, the differential decreased to 6 percentage points for parent surveys and 10 percentage points for student surveys (tables A-7 through A-9).

Table A-6. Subsample Conversion Response Rates for Year 2

	Subsample Members	Actual Response Conversions	Actual Conversion Rate (%)
Cohort 1	43	16	37.21
Cohort 2	160	60	37.50
Total	203	76	37.44

Table A-7. Final Test Score Response Rates for Year 2, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	193	105	54.4	125	64.6
Cohort 1 T	299	213	71.2	213	71.2
Cohort 2 C	728	455	62.5	514	70.6
Cohort 2 T	1,088	822	75.6	822	75.6
Cohort 1 total	492	318	64.6	338	68.6
Cohort 2 total	1,816	1,277	70.3	1,336	73.6
C total	921	560	60.8	639	69.3
T total	1,387	1,035	74.6	1,035	74.6
Combined total	2,308	1,595	69.1	1,674	72.5

Table A-8. Final Parent Survey Response Rates for Year 2, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	193	103	53.4	122	62.9
Cohort 1 T	299	208	69.6	208	69.6
Cohort 2 C	728	456	62.6	512	70.3
Cohort 2 T	1088	829	76.2	829	76.2
Cohort 1 total	492	311	63.2	330	67.0
Cohort 2 total	1816	1285	70.8	1341	73.8
C total	921	559	60.7	633	68.8
T total	1387	1037	74.8	1037	74.8
Combined total	2308	1596	69.2	1670	72.4

Table A-9. Final Student Survey Response Rates for Year 2, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	192	86	44.8	105	54.4
Cohort 1 T	299	181	60.5	181	60.5
Cohort 2 C	465	273	58.7	302	64.8
Cohort 2 T	661	508	76.9	508	76.9
Cohort 1 total	491	267	54.4	286	58.1
Cohort 2 total	1126	781	69.4	810	71.9
C total	657	359	54.6	406	61.8
T total	960	689	71.8	689	71.8
Combined total	1617	1048	64.8	1095	67.7

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 or 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 evidence standards without reservation if the study demonstrates that the treatment and control group have remained approximately equivalent in spite of 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 percentiles, as constituting a possible attrition problem. The test score effective response rates obtained in year 2 met the overall standard of 70 percent or higher but exceeded by three-tenths-of-a-percentile the response differential standard of 5 percentiles or less. Similarly, the parent survey effective response rates that informed this analysis met the overall response rate standard but exceeded by 1 percentile the differential response rate standard. The effective response rates regarding the student survey did not meet either standard. In this study, the non-response weights that are generated from student test score performance and demographic data collected at baseline re-established the equivalence of the treatment and control groups in the wake of the year 2 sample attrition experienced here. Thus, the evaluation continues to meet the WWC evidence standards.

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 control initial non-respondents (with X_i equal to 2.233 for cohort 1 (96 divided by 43) and 1.981 for cohort 2 (317 divided by 160) for this set, and equal to 1 otherwise), NR_j is the non-response 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 pre-defined set of participant subgroups, study response rates and subsequent analytic sample sizes are presented for each of those subgroups and for all three primary data collection instruments (student tests, parent surveys, and student surveys). The year 2 subgroup-level effective response rates for student test scores ranged from a low of 62 percent for participants entering the high school grades at baseline to a high of 75 percent for their counterparts entering grades K-8 at baseline (table A-10). The subgroup of students entering a high school grade at baseline were comprised of the smallest subgroup sample size for the analysis, of 253 observations, compared to 1,421 observations in the K-8 subgroup. Besides the grade-level subgroups, the only other subgroup pair with year 2 response rates that differed by more than 5 percentiles was the subgroup of higher baseline performers (75 percent response) compared to the subgroup of lower baseline performers (68 percent response).

Table A-10. Effective Test Score Response Rates for Year 2 Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	1,010	721	71.4
SINI never	1,298	953	73.4
Lower performance	788	535	67.8
Higher performance	1,520	1,139	74.9
Male	1,149	826	71.9
Female	1,159	847	73.1
K-8	1,900	1,421	74.8
9-12	408	253	61.9
Cohort 2	1,816	1,336	73.6
Cohort 1	492	338	68.6

The year 2 subgroup-level effective response rates for parent surveys ranged from a low of 61 percent for participants entering the high school grades at baseline to a high of 75 percent for their counterparts entering grades K-8 at baseline (table A-11). The subgroup pairs based on baseline test score performance and cohort exhibited response rate differentials of 8 percentiles and 7 percentiles, respectively, regarding the parent surveys.

Table A-11. Effective Parent Survey Response Rates for Year 2 Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	1,010	715	70.8
SINI never	1,298	956	73.6
Lower performance	788	531	67.4
Higher performance	1,520	1,139	75.0
Male	1,149	826	71.9
Female	1,159	844	72.8
K-8	1,900	1,421	74.8
9-12	408	249	61.0
Cohort 2	1,816	1,341	73.8
Cohort 1	492	329	67.0

The year 2 subgroup-level effective response rates for student surveys ranged from a low of 58 percent for participants in cohort 1 to a high of 71 percent for their counterparts in cohort 2 (table A-12). The subgroup pairs based on grade level and baseline test score performance exhibited response rate differentials of 11 percentiles and 10 percentiles, respectively, regarding the student surveys.

Table A-12. Effective Student Survey Response Rates for Year 2 Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	861	715	68.4
SINI never	763	956	66.3
Lower performance	557	531	61.5
Higher performance	1,067	1,139	70.5
Male	807	826	66.7
Female	817	844	68.1
K-8	1,216	1,421	70.3
9-12	408	249	58.8
Cohort 2	1,133	1,341	71.4
Cohort 1	491	329	58.1

NOTE: Student surveys administered to students in grades 4-12.

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}):

$$(1)\ Y_{it} = \alpha + \tau\, T_{it} + \varepsilon_{it} \qquad \text{if } t>k \text{ (period after Program takes effect),}$$

where T_{it} is equal to 1 if the student *has the opportunity to participate* in the scholarship Program (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.[27] 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:

[27] 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 1 year, which could be biased by the higher average baseline test scores for the control group, to one on comparative achievement gains after 1 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.

$$(2) Y_{it} = \alpha + \tau T_{it} + X_i \gamma + \delta_R R_{it} + \delta_M M_{it} + \varepsilon_{it}.$$

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.

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.[28] This "days until test" variable, signified by DT in the equation below, controls for the fact that test scores were obtained over a 4-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.[29] Thus, the models that produced the regression-adjusted impact estimates for this analysis took the general form:[30]

$$(3) Y_{it} = \alpha + \tau T_{it} + X_i \gamma + \delta_R R_{it} + \delta_M M_{it} + \delta_{DT} DT_{it} + \varepsilon_{it}.$$

[28] 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.

[29] 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.

[30] 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.

The same set of baseline covariates and the DT variable were used in all impact regression models, regardless of whether the outcomes being estimated were student achievement, school satisfaction, school safety, or any of the intermediate outcomes.[31]

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:

$$(4)\ Y_{ikt} = \mu + \tau T_{ikt} + \tau_B P_i{*}T_{ikt} + \sum\nolimits^{b}_{j=2} \varphi^{j}_{is} + X_{ik}\,\gamma + \delta_R R_{it} + \delta_M M_{it} + \delta_{DT} DT_{it} + \varepsilon_{ik,t}$$

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). The coefficient τ_P indicates the marginal treatment effect for students in the designated subgroup. 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. 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. Consequently, subgroup impacts were estimated for the following groups:

- Applied from a school ever designated SINI—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);[32]
- Gender—male and female;
- Grade band—K-8 and high school; and
- Cohort—1 and 2.

[31] After the initial impacts were obtained, 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 were found.

[32] 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.

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 non-response adjustments. As a consequence, all of the impact analyses were completed using sampling weights in STATA.[33] The effects of family clustering, which is not part of the sample design, but which may be having 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).[34]

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 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 year 2 outcome test scores, 82 percent had used a scholarship for all or part of the 2 years after random assignment. The 18 percent of the treatment students 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

[33] 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.

[34] We also examined the effect on the standard errors of the estimates of clustering on the school students attended at baseline. 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. They may appear to be counter-intuitive, 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.

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 18 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 by 16-35 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 74-86 percent.[35]

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 Program 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

[35] The Bloom adjustment is generated by dividing the ITT estimate by the usage rate for that outcome. Any number that is divided by .74 will generate a dividend that is 35 percent larger. Any number that is divided by .86 will generate a dividend that is 16 percent larger.

"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 who provided outcome tests in math, 17.9 percent reported attending a private school. 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).[36] The crossover rate also is higher for control group students with siblings in the treatment group (20.3 percent) compared to those without treatment siblings (15.6 percent),[37] a difference that is statistically significant beyond the 99 percent confidence level. At outcome data collection events, some parents of control group students commented to evaluation staff that their control-group child was 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 2.3 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 15.6 percent of the control group students without treatment siblings are attending private schools, whereas 17.9 percent of the control group overall is in private schools. Since the 15.6 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 17.9 percent to arrive at the hypothesized Program-enabled crossover rate of 2.3 percent. To adjust for the fact that this small component of the control group

[36] 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.

[37] 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.

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. We rescale the pure ITT impacts that are statistically significant by an amount equal to the treatment decliner rate (~18 percent), as described above and, in addition, rescale in the same manner for the possible Program-enabled crossover rate (~2.3 percent). This strategy provides upper and lower bounds for the IOT estimates.

Appendix B
Benjamini-Hochberg Adjustments for Multiple Comparisons

Below is a series of tables (tables B-1 through B-15) that 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 source of the multiple comparisons was either various subgroups of the impact sample (chapters 3 and 4), or the multiple comparisons made within the conceptual groupings of intermediate outcomes (chapter 4 only). 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 five domains that comprise the primary analysis: student achievement, parent perceptions of safety, student perceptions of safety, parent satisfaction with school, and student satisfaction with school. These five 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, math, student perceptions of safety, or student satisfaction for the entire sample of study participants in year 2, so there could not have been false discoveries in those domains. Significant impacts for the entire sample were observed regarding parental perceptions of safety, but they were not the result of multiple comparisons. Finally, significant impacts were observed for the entire sample across three measures of parental satisfaction with their child's school, but two of those measures were alternative classifications of the exact same outcome data ("percent of parents assigning a grade of A or B" and "average grade parents assigned to school"), reducing the likelihood that mere chance produced the parental satisfaction impacts reported for the entire sample.

Table B-1. Multiple Comparisons Adjustments, Reading

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	1.00	1.00
SINI never	.04*	.14
Lower performance	.65	.81
Higher performance	.02*	.14
Male	.17	.34
Female	.31	.52
K-8	.08	.20
9-12	.96	1.00
Cohort 2	.42	.61
Cohort 1	.04*	.14

*Statistically significant at the 95 percent confidence level.

Table B-2. Multiple Comparisons Adjustments, Parental School Danger

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.00**	.00**
SINI never	.01**	.01**
Lower performance	.14	.14
Higher performance	.00**	.00**
Male	.00**	.00**
Female	.00**	.00**
K-8	.00**	.00**
9-12	.06	.07
Cohort 2	.00**	.00**
Cohort 1	.04*	.05*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-3. Multiple Comparisons Adjustments, Parents Gave Their Child's School a Grade of A or B

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.00**	.00**
SINI never	.00**	.00**
Lower performance	.03*	.09
Higher performance	.00**	.00**
Male	.02*	.03*
Female	.00**	.00**
K-8	.00**	.00**
9-12	.89	.89
Cohort 2	.00**	.00**
Cohort 1	.16	.17

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-4. Multiple Comparisons Adjustments, Average Grade Parent Gave Their Child's School

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.00**	.00**
SINI never	.00**	.00**
Lower performance	.10	.11
Higher performance	.00**	.00**
Male	.03*	.05*
Female	.00**	.00**
K-8	.00**	.00**
9-12	.93	.93
Cohort 2	.00**	.00**
Cohort 1	.06	.08

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-5. Multiple Comparisons Adjustments, Parental Satisfaction

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.00**	.00**
SINI never	.00**	.00**
Lower performance	.02*	.03*
Higher performance	.00**	.00**
Male	.00**	.00**
Female	.00**	.00**
K-8	.00**	.00**
9-12	.10	.11
Cohort 2	.00**	.00**
Cohort 1	.19	.19

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-6. Multiple Comparisons Adjustments, Students Gave Their School a Grade of A or B

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.02*	.16
SINI never	.28	.74
Lower performance	.85	.99
Higher performance	.45	.76
Male	.30	.74
Female	.97	.99
4-8	.45	.77
9-12	.99	.99
Cohort 2	.83	.99
Cohort 1	.26	.74

*Statistically significant at the 95 percent confidence level.

Table B-7. Multiple Comparisons Adjustments, Average Grade Student Gave Their School

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.02*	.22
SINI never	.97	.97
Lower performance	.19	.38
Higher performance	.42	.97
Male	.12	.38
Female	.56	.70
4-8	.09	.38
9-12	.66	.97
Cohort 2	.19	.97
Cohort 1	.49	.70

*Statistically significant at the 95 percent confidence level.

Table B-8. Multiple Comparisons Adjustments, Student Satisfaction Scale

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.03*	.25
SINI never	.73	.81
Lower performance	.85	.85
Higher performance	.07	.25
Male	.06	.25
Female	.58	.72
4-8	.16	.35
9-12	.31	.44
Cohort 2	.21	.35
Cohort 1	.21	.85

*Statistically significant at the 95 percent confidence level.

Table B-9. Multiple Comparisons Adjustments, Home Educational Supports

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental involvement	.32	.32
Parent aspirations	.04*	.07
Out-of-school tutor usage	.22	.30
School transit time	.00**	.01*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-10. Multiple Comparisons Adjustments, Instructional Characteristics

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/teacher ratio	.00**	.00**
Teacher attitude	.79	.79
Challenge of classes	.59	.65
Ability grouping	.18	.26
Availability of tutors	.00**	.00**
In-school tutor usage	.02*	.04*
Programs for learning problems/ELL	.00**	.00**
Programs for advanced learners	.11	.18
Before-/after-school programs	.28	.35
Enrichment programs	.02*	.04*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-11. Multiple Comparisons Adjustments, School Environment

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/school communication	.89	.89
School size	.00**	.00**
Percent non-white	.00**	.00**
Peer classroom behavior	.04*	.06

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-12. Multiple Comparisons Adjustments, Subgroup Impacts on Home Educational Supports

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Parental involvement	SINI ever	.53	.68
Parental involvement	SINI never	.07	.23
Parental involvement	Lower performance	.59	.69
Parental involvement	Higher performance	.10	.24
Parental involvement	Male	.70	.77
Parental involvement	Female	.06	.23
Parental involvement	K-8	.17	.31
Parental involvement	9-12	.65	.74
Parental involvement	Cohort 2	.13	.28
Parental involvement	Cohort 1	.53	.68
Parent aspirations	SINI ever	.29	.41
Parent aspirations	SINI never	.07	.23
Parent aspirations	Lower performance	.94	.96
Parent aspirations	Higher performance	.01**	.07
Parent aspirations	Male	.11	.24
Parent aspirations	Female	.19	.33
Parent aspirations	K-8	.37	.51
Parent aspirations	9-12	.01**	.07
Parent aspirations	Cohort 2	.09	.24
Parent aspirations	Cohort 1	.23	.37
Outside tutor usage	SINI ever	.25	.38
Outside tutor usage	SINI never	.55	.68
Outside tutor usage	Lower performance	.99	.99
Outside tutor usage	Higher performance	.11	.24
Outside tutor usage	Male	.91	.96
Outside tutor usage	Female	.05	.21
Outside tutor usage	K-8	.29	.41
Outside tutor usage	9-12	.56	.68
Outside tutor usage	Cohort 2	.22	.37
Outside tutor usage	Cohort 1	.72	.78
School transit time	SINI ever	.00**	.07
School transit time	SINI never	.15	.29
School transit time	Lower performance	.15	.29
School transit time	Higher performance	.01**	.07
School transit time	Male	.01**	.07
School transit time	Female	.09	.24
School transit time	K-8	.01*	.09
School transit time	9-12	.05*	.21
School transit time	Cohort 2	.05*	.21
School transit time	Cohort 1	.02*	.10

Table B-13. Multiple Comparisons Adjustments, Subgroup Impacts on Student Motivation and Engagement

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Student aspirations	SINI ever	.33	.75
Student aspirations	SINI never	.28	.75
Student aspirations	Lower performance	.02*	.38
Student aspirations	Higher performance	.00	.00
Student aspirations	Male	.31	.75
Student aspirations	Female	.30	.75
Student aspirations	4-8	.17	.75
Student aspirations	9-12	.75	.97
Student aspirations	Cohort 2	.27	.75
Student aspirations	Cohort 1	.28	.75
Attendance	SINI ever	.38	.81
Attendance	SINI never	.77	.97
Attendance	Lower performance	.74	.97
Attendance	Higher performance	.46	.89
Attendance	Male	.43	.89
Attendance	Female	.72	.97
Attendance	K-8	.77	.97
Attendance	9-12	.21	.75
Attendance	Cohort 2	.24	.75
Attendance	Cohort 1	.69	.97
Tardiness	SINI ever	.31	.75
Tardiness	SINI never	.95	.98
Tardiness	Lower performance	.33	.75
Tardiness	Higher performance	.86	.98
Tardiness	Male	.12	.75
Tardiness	Female	.64	.97
Tardiness	K-8	.00	.00
Tardiness	9-12	.14	.75
Tardiness	Cohort 2	.46	.89
Tardiness	Cohort 1	.92	.98
Reading for fun	SINI ever	.48	.89
Reading for fun	SINI never	.80	.98
Reading for fun	Lower performance	.74	.97
Reading for fun	Higher performance	.65	.97
Reading for fun	Male	.78	.97
Reading for fun	Female	.94	.98
Reading for fun	4-8	.95	.98
Reading for fun	9-12	.34	.75
Reading for fun	Cohort 2	.83	.98
Reading for fun	Cohort 1	.30	.75
Engagement in extracurricular activities	SINI ever	.70	.97
Engagement in extracurricular activities	SINI never	.31	.75
Engagement in extracurricular activities	Lower performance	.13	.75
Engagement in extracurricular activities	Higher performance	.73	.97
Engagement in extracurricular activities	Male	.22	.75

Table B-13. Multiple Comparisons Adjustments, Subgroup Impacts on Student Motivation and Engagement—(continued)

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Engagement in extracurricular activities	Female	.84	.98
Engagement in extracurricular activities	4-8	.33	.75
Engagement in extracurricular activities	9-12	.76	.97
Engagement in extracurricular activities	Cohort 2	.67	.97
Engagement in extracurricular activities	Cohort 1	.09	.75
Frequency of homework	SINI ever	.91	.98
Frequency of homework	SINI never	.04*	.50
Frequency of homework	Lower performance	.51	.91
Frequency of homework	Higher performance	.18	.75
Frequency of homework	Male	.92	.98
Frequency of homework	Female	.04*	.50
Frequency of homework	4-8	.01**	.21
Frequency of homework	9-12	.01**	.21
Frequency of homework	Cohort 2	.19	.75
Frequency of homework	Cohort 1	.49	.89

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-14. Multiple Comparisons Adjustments, Subgroup Impacts on Instructional Characteristics

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Student/teacher ratio	SINI ever	0.00**	0.02*
Student/teacher ratio	SINI never	0.00**	0.01**
Student/teacher ratio	Lower performance	0.01**	0.02*
Student/teacher ratio	Higher performance	0.00**	0.01*
Student/teacher ratio	Male	0.03*	0.09
Student/teacher ratio	Female	0.00**	0.00**
Student/teacher ratio	K-8	0.00**	0.01**
Student/teacher ratio	9-12	0.00**	0.00**
Student/teacher ratio	Cohort 2	0.00**	0.00**
Student/teacher ratio	Cohort 1	0.06	0.15
Teacher attitude	SINI ever	0.80	0.91
Teacher attitude	SINI never	0.56	0.73
Teacher attitude	Lower performance	0.71	0.82
Teacher attitude	Higher performance	0.90	0.95
Teacher attitude	Male	0.53	0.71
Teacher attitude	Female	0.35	0.51
Teacher attitude	4-8	0.63	0.74
Teacher attitude	9-12	0.50	0.69
Teacher attitude	Cohort 2	0.58	0.74
Teacher attitude	Cohort 1	0.62	0.77
Challenge of classes	SINI ever	0.27	0.45
Challenge of classes	SINI never	0.87	0.95
Challenge of classes	Lower performance	0.64	0.78
Challenge of classes	Higher performance	0.71	0.82
Challenge of classes	Male	0.16	0.29
Challenge of classes	Female	0.53	0.71
Challenge of classes	4-8	0.94	0.97
Challenge of classes	9-12	0.01**	0.03*
Challenge of classes	Cohort 2	0.53	0.71
Challenge of classes	Cohort 1	0.89	0.95
Ability grouping	SINI ever	0.08	0.18
Ability grouping	SINI never	0.83	0.92
Ability grouping	Lower performance	0.96	0.97
Ability grouping	Higher performance	0.11	0.24
Ability grouping	Male	0.60	0.76
Ability grouping	Female	0.16	0.29
Ability grouping	K-8	0.54	0.71
Ability grouping	9-12	0.09	0.20
Ability grouping	Cohort 2	0.33	0.50
Ability grouping	Cohort 1	0.29	0.46
Availability of tutors	SINI ever	0.00**	0.00**
Availability of tutors	SINI never	0.34	0.51
Availability of tutors	Lower performance	0.03*	0.07
Availability of tutors	Higher performance	0.01**	0.04*
Availability of tutors	Male	0.01**	0.03*

Table B-14. Multiple Comparisons Adjustments, Subgroup Impacts on Instructional Characteristics—(continued)

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Availability of tutors	Female	0.05*	0.12
Availability of tutors	K-8	0.01**	0.04*
Availability of tutors	9-12	0.01**	0.03*
Availability of tutors	Cohort 2	0.00**	0.00**
Availability of tutors	Cohort 1	0.14	0.27
In-school tutor usage	SINI ever	0.12	0.24
In-school tutor usage	SINI never	0.09	0.20
In-school tutor usage	Lower performance	0.01**	0.03
In-school tutor usage	Higher performance	0.39	0.56
In-school tutor usage	Male	0.03*	0.07
In-school tutor usage	Female	0.33	0.50
In-school tutor usage	K-8	0.01*	0.04*
In-school tutor usage	9-12	0.98	0.98
In-school tutor usage	Cohort 2	0.00**	0.02*
In-school tutor usage	Cohort 1	0.67	0.81
Programs for learning problems/ ELL	SINI ever	0.00**	0.00**
Programs for learning problems/ ELL	SINI never	0.00**	0.00**
Programs for learning problems/ ELL	Lower performance	0.00**	0.00**
Programs for learning problems/ ELL	Higher performance	0.00**	0.00**
Programs for learning problems/ ELL	Male	0.00**	0.00**
Programs for learning problems/ ELL	Female	0.00**	0.00**
Programs for learning problems/ ELL	K-8	0.00**	0.00**
Programs for learning problems/ ELL	9-12	0.00**	0.00**
Programs for learning problems/ ELL	Cohort 2	0.00**	0.00**
Programs for learning problems/ ELL	Cohort 1	0.00**	0.00**
Programs for advanced learners	SINI ever	0.02*	0.05
Programs for advanced learners	SINI never	0.90	0.95
Programs for advanced learners	Lower performance	0.17	0.31
Programs for advanced learners	Higher performance	0.26	0.43
Programs for advanced learners	Male	0.19	0.32
Programs for advanced learners	Female	0.30	0.47
Programs for advanced learners	K-8	0.15	0.29
Programs for advanced learners	9-12	0.44	0.62
Programs for advanced learners	Cohort 2	0.05*	0.12
Programs for advanced learners	Cohort 1	0.83	0.92

Table B-14. Multiple Comparisons Adjustments, Subgroup Impacts on Instructional Characteristics—(continued)

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Before-/after-school programs	SINI ever	0.43	0.62
Before-/after-school programs	SINI never	0.02*	0.05*
Before-/after-school programs	Lower performance	0.96	0.97
Before-/after-school programs	Higher performance	0.18	0.32
Before-/after-school programs	Male	0.70	0.82
Before-/after-school programs	Female	0.12	0.24
Before-/after-school programs	K-8	0.08	0.18
Before-/after-school programs	9-12	0.73	0.84
Before-/after-school programs	Cohort 2	0.20	0.34
Before-/after-school programs	Cohort 1	0.93	0.97
Enrichment programs	SINI ever	0.08	0.18
Enrichment programs	SINI never	0.10	0.22
Enrichment programs	Lower performance	0.14	0.27
Enrichment programs	Higher performance	0.04	0.12
Enrichment programs	Male	0.02	0.07
Enrichment programs	Female	0.28	0.45
Enrichment programs	K-8	0.01*	0.03*
Enrichment programs	9-12	0.90	0.95
Enrichment programs	Cohort 2	0.06	0.16
Enrichment programs	Cohort 1	0.12	0.24

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-15. Multiple Comparisons Adjustments, Subgroup Impacts on School Environment

Intermediate Outcome	Subgroup	Original *p*-value	False Discovery Rate *p*-value
Parent/school communication	SINI ever	0.70	0.79
Parent/school communication	SINI never	0.80	0.86
Parent/school communication	Lower performance	0.98	0.98
Parent/school communication	Higher performance	0.91	0.93
Parent/school communication	Male	0.67	0.79
Parent/school communication	Female	0.77	0.85
Parent/school communication	K-8	0.51	0.63
Parent/school communication	9-12	0.10	0.15
Parent/school communication	Cohort 2	0.27	0.36
Parent/school communication	Cohort 1	0.02*	0.04*
School size	SINI ever	0.00**	0.00**
School size	SINI never	0.00**	0.00**
School size	Lower performance	0.00**	0.00**
School size	Higher performance	0.00**	0.00**
School size	Male	0.00**	0.00**
School size	Female	0.00**	0.00**
School size	K-8	0.00**	0.00**
School size	9-12	0.06	0.10
School size	Cohort 2	0.00**	0.00**
School size	Cohort 1	0.02*	0.04*
Percent non-white	SINI ever	0.00**	0.00**
Percent non-white	SINI never	0.00**	0.01**
Percent non-white	Lower performance	0.00**	0.00**
Percent non-white	Higher performance	0.00**	0.00**
Percent non-white	Male	0.00**	0.00**
Percent non-white	Female	0.19	0.26
Percent non-white	K-8	0.00**	0.00**
Percent non-white	9-12	0.03*	0.06
Percent non-white	Cohort 2	0.00**	0.00**
Percent non-white	Cohort 1	0.02*	0.03*
Peer classroom behavior	SINI ever	0.01*	0.03*
Peer classroom behavior	SINI never	0.54	0.65
Peer classroom behavior	Lower performance	0.88	0.93
Peer classroom behavior	Higher performance	0.01*	0.03*
Peer classroom behavior	Male	0.30	0.39
Peer classroom behavior	Female	0.07	0.11
Peer classroom behavior	4-8	0.09	0.15
Peer classroom behavior	9-12	0.16	0.23
Peer classroom behavior	Cohort 2	0.13	0.19
Peer classroom behavior	Cohort 1	0.10	0.15

*Statistically significant at the 95 percent confidence level.

**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., non-response 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*: The sample of students was trimmed back to equalize the actual response rates of the treatment and control groups prior to any subsampling of control non-respondents. Since the actual response rate of the treatment group was higher (75 percent), in effect the "latest 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 53 percent. This approach differs from the primary analysis, where all observations were used even though a higher percentage of the treatment than the control group actually responded to outcome data collection. This sensitivity testing is designed to address whether the difference in response rates is adequately controlled for by non-response weighting of subsampled initial non-respondents.

- *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.

C.1 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, 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 any statistically significant impacts from the exploratory subgroup analyses using these same sensitivity tests.

Sensitivity Checks for the ITT Impacts on Reading and Math Achievement

The sensitivity test produced only two changes in the findings for reading and math impacts. First, the overall estimate of a positive impact in reading crosses the threshold to be statistically significant at the 95 percent confidence level when the analysis is limited to only the trimmed sample of respondents (table C-1). Second, for higher performing students in reading, the trimmed sample *p*-value is .052, thus dropping below the .05 threshold and changing to not statistically significant. The other two statistically significant subgroup findings from the primary analysis—for non-SINI and cohort 1 students in reading—remain significant under models run with only the trimmed sample. The SINI-never subgroup impact estimate under the trimmed sample analysis grows in magnitude from 5.7 to 8.5 scale score points with a more significant *p*-value of .00. The cohort 1 subgroup's impact estimate is smaller under the trimmed sample than it was based on the primary analysis, and it maintains a .04 *p*-value.

Table C-1. Year 2 Test Score ITT Impact Estimates and *P*-Values with Different Specifications

Student Achievement Groups	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample: reading	3.17	.09	4.57*	.02	3.17	.12
Full sample: math	.23	.89	1.66	.33	.23	.89
SINI never: reading	5.71*	.04	8.47**	.00	5.71	.06
Higher performing: reading	5.23*	.02	4.33	.05	5.23*	.02
Cohort 1: reading	8.74*	.04	4.35*	.04	8.74*	.03

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Impacts are displayed in terms of scale scores. Original estimates valid *N* for reading = 1,580; math = 1,585. Trimmed sample valid *N* for reading = 1,468; math = 1,471. Separate reading and math sample weights were used.

The other sensitivity specification that involves the use of robust regression analysis that clusters on students' current school in place of the clustering by family that was employed in the primary analysis does not change the overall impact estimates but does affect the *p*-values and thus significance levels of those estimates. For the statistically significant ITT subgroup reading impacts, this specification does not yield different results from the main analysis for higher performing students (both *p* values are at .02). However, the *p*-value for the reading impact on the cohort 1 subgroup changes from .04 to .03, while the *p*-value for the reading impact of the SINI-never subgroup increases from .04 and .06, thereby crossing the .05 level and losing its statistical significance.

In sum, the finding from the primary analysis of no significant Programmatic impact overall on math achievement was consistent across the analysis approaches. The finding from the primary analysis of no significant impact overall on reading achievement was consistent with the results when the standard errors were clustered by school, but not consistent with the finding of a statistically significant overall impact in reading when the treatment group was trimmed back to the control group response rate. The finding from the primary analysis of a significant Programmatic impact in reading for cohort 1 students was consistent across specifications. The findings of significant reading impacts on students who did not apply from a SINI school or who were relatively higher performing at baseline were each consistent with the results from one sensitivity specification and inconsistent with the results from the other.

Sensitivity Checks for ITT Impacts on Parent Perceptions of School Danger

The Programmatic impacts on parental reports of school danger discussed in chapter 3 were consistent across analytic approaches (table C-2). Regardless of how the data were analyzed, parents' perception of school danger was significantly lower 2 years later if their child had been offered a scholarship.

Table C-2. Year 2 Parent Perceptions of School Danger ITT Impact Estimates and *P*-Values with Different Specifications

Outcome	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
School danger: parents	-.94**	.00	-.84**	.00	-.94**	.00

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Original estimates valid *N* = 1,555. Trimmed Sample valid *N* = 1,418. Parent survey weights were used.

Sensitivity Checks for ITT Impacts on Student Reports of School Danger

The primary analysis discussed in chapter 3 found no treatment effect on students' perceptions of school danger. This result is consistent across different analytic approaches (table C-3). Regardless of how the data were analyzed, responses of those offered a scholarship did not differ significantly from control group students' perception of school danger.

Table C-3. Year 2 Student Perceptions of School Danger ITT Impact Estimates and *P*-Values with Different Specifications

	Original Estimates		Trimmed Sample		Clustering on Current School	
Outcome	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
School danger: students	-.02	.87	-.13	.42	-.02	.87

*Statistically significant at the 95 percent confidence level.
**Statistically significant at the 99 percent confidence level.
NOTES: Original estimates valid *N* = 1,025. Trimmed Sample valid *N* = 941 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 was not sensitive to different analytic approaches (table C-4). Across the different specifications of the parent satisfaction impacts, parents self-reported significantly higher levels of school satisfaction if their child had been awarded a scholarship.

Table C-4. Year 2 Parent Satisfaction ITT Impact Estimates and *P*-Values with Different Specifications

	Original Estimates		Trimmed Sample		Clustering on Current School	
Outcome	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Graded school A or B	.13**	.00	.15**	.00	.13**	.00
Average grade given school (5.0 scale)	.29**	.00	.33**	.00	.29**	.00
School satisfaction scale	2.67**	.00	2.31**	.00	2.67**	.00

*Statistically significant at the 95 percent confidence level.
**Statistically significant at the 99 percent confidence level.
NOTES: Original estimates valid *N* for school grade = 1,549; parent satisfaction = 1,571. Trimmed sample valid *N* for school grade = 1,444; parent satisfaction = 1,464. 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 Satisfaction

The results of the primary analysis found no Programmatic impact on overall student self-reports of satisfaction on the three outcomes examined. Two of those three main findings are consistent across the different methodological approaches (table C-5). In every specification, there are no differences in the likelihood of a student grading his or her school A or B or in the average grade a child gave his or her school. However, the trimmed sample analysis did yield a statistically significant positive treatment impact on the student satisfaction scale.

Table C-5. Year 2 Student Satisfaction ITT Impact Estimates and *P*-Values with Different Specifications

	Original Estimates		Trimmed Sample		Clustering on Current School	
Outcome	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Graded school A or B	.03	.49	.03	.38	.03	.55
Average grade given school (5.0 scale)	.13	.14	.13	.14	.13	.19
School satisfaction scale	.88	.10	1.24*	.02	.88	.10

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Original estimates valid *N* for school grade = 974; student satisfaction = 1,042. Trimmed sample valid *N* for school grade = 910; student satisfaction = 975. Student survey weights were 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.

Sensitivity Checks for ITT Impacts on Student Reports of Satisfaction—SINI Ever

For the subgroup of students who attended schools designated as SINI, the primary analysis found a statistically significant Programmatic impact on student self-reports of satisfaction on all three outcomes examined**.** These three findings are consistent across the different methodological approaches (table C-6).

Table C-6. Year 2 SINI-Ever Student Satisfaction ITT Regression-Based Impact Estimates and *P*-Values with Different Specifications

	Original Estimates		Trimmed Sample		Clustering on Current School	
Outcome	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
SINI ever: School grade of A or B	.12*	.02	.13**	.01	.12*	.02
SINI ever: School grade, 5.0 scale	.28*	.02	.30*	.02	.28*	.03
SINI ever: School satisfaction scale	1.65*	.03	1.66*	.03	1.65*	.02

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Original estimates valid *N* for school grade = 974; student satisfaction = 1,042. Trimmed sample valid *N* for school grade = 910; student satisfaction = 975. Student survey weights were 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.

Appendix D
Detailed ITT Tables

Table D-1. Year 2 Test Score ITT Impacts: Reading

	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	620.88	618.72	2.16	.57	3.17	.09	.09
	(35.64)	(37.05)	(3.79)		(1.89)		(37.05)
Subgroups							
SINI ever	645.50	640.80	4.70	.34	-.01	1.00	-.00
	(32.81)	(34.29)	(4.90)		(2.41)		(34.29)
SINI never	601.59	601.40	.19	.97	-5.71	.04	.15
	(37.07)	(37.75)	(5.30)		(2.80)		(37.75)
Difference	43.91	39.40	4.51	.54	-5.71	.12	-.15
	(4.30)	(5.90)	(7.26)		(3.69)		(37.05)
Lower performance	599.90	600.84	-.94	.89	-1.59	.65	-.05
	(28.78)	(30.75)	(6.47)		(3.45)		(30.75)
Higher performance	630.88	626.43	4.44	.32	5.23*	.02	.15
	(33.70)	(35.96)	(4.48)		(2.16)		(35.96)
Difference	-30.97	-25.59	-5.38	.49	-6.81	.09	-.18
	(4.69)	(6.35)	(7.86)		(3.96)		(37.05)
Male	618.15	613.90	4.25	.47	3.90	.17	.11
	(35.99)	(36.94)	(5.89)		(2.83)		(36.94)
Female	623.64	623.09	.55	.91	2.50	.31	.07
	(34.53)	(34.82)	(4.96)		(2.48)		(34.82)
Difference	-5.49	-9.19	3.70	.63	1.40	.71	.04
	(4.64)	(6.16)	(7.76)		(3.74)		(37.05)
K-8	608.09	605.97	2.11	.60	3.79	.08	.10
	(36.71)	(38.07)	(4.04)		(2.15)		(38.07)
9-12	679.41	678.40	1.01	.82	.19	.96	.01
	(26.18)	(32.27)	(4.36)		(3.48)		(32.27)
Difference	-71.32	-72.43	1.10	.85	3.59	.38	.06
	(3.96)	(4.50)	(5.98)		(4.06)		(37.05)
Cohort 2	608.47	607.90	.58	.89	1.66	.42	.04
	(36.59)	(37.61)	(4.20)		(2.08)		(37.61)
Cohort 1	666.52	656.23	10.30	.10	8.74*	.04	.27
	(32.58)	(32.50)	(6.34)		(4.26)		(32.50)
Difference	-58.05	-48.33	-9.72	.20	-7.07	.13	-.19
	(4.03)	(6.44)	(7.60)		(4.70)		(37.05)

NOTE: Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for reading = 1,580. Reading sample weights were used.

Table D-2. Year 2 Test Score ITT Impacts: Math

	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	613.43	614.09	-.66	.87	.23	.89	.01
	(31.18)	(32.65)	(4.01)		(1.68)		(32.65)
Subgroups							
SINI ever	641.89	635.52	6.37	.19	1.28	.58	.05
	(26.94)	(26.77)	(4.88)		(2.27)		(26.77)
SINI never	590.96	597.39	-6.43	.26	-.59	.81	-.02
	(33.11)	(35.89)	(5.73)		(2.46)		(35.89)
Difference	50.93	38.13	12.80	.09	1.86	.58	.06
	(4.32)	(6.23)	(7.54)		(3.38)		(32.65)
Lower performance	596.92	599.06	-2.14	.77	-2.58	.43	-.09
	(24.91)	(27.64)	(7.25)		(3.23)		(27.64)
Higher performance	621.53	620.50	1.04	.83	1.50	.43	.05
	(30.23)	(31.85)	(4.75)		(1.91)		(31.85)
Difference	-24.61	-21.43	-3.18	.72	-4.08	.27	-.12
	(4.97)	(7.04)	(8.72)		(3.70)		(32.65)
Male	614.04	611.88	2.15	.73	.52	.85	.02
	(32.11)	(32.64)	(6.32)		(2.69)		(32.64)
Female	612.83	616.08	-3.25	.52	-.03	.99	-.00
	(29.23)	(30.56)	(5.09)		(2.19)		(30.56)
Difference	1.20	-4.19	5.40	.51	.55	.88	.02
	(4.76)	(6.47)	(8.12)		(3.55)		(32.65)
K-8	599.94	600.69	-.75	.86	.91	.63	.03
	(30.82)	(34.76)	(4.32)		(1.91)		(34.76)
9-12	675.23	677.02	-1.79	.64	-3.08	.29	-.14
	(24.89)	(22.75)	(3.80)		(2.93)		(22.75)
Difference	-75.29	-76.34	1.04	.85	3.99	.25	.12
	(3.88)	(4.20)	(5.65)		(3.43)		(32.65)
Cohort 2	600.19	600.50	-.31	.94	.08	.97	.00
	(32.23)	(35.00)	(4.47)		(1.92)		(35.00)
Cohort 1	662.21	661.58	.64	.90	.80	.80	.03
	(26.93)	(23.54)	(4.92)		(3.13)		(23.54)
Difference	-62.02	-61.07	-.95	.89	-.72	.84	-.02
	(3.80)	(5.48)	(6.65)		(3.59)		(32.65)

NOTE: Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for math = 1,585. Math sample weights were used.

Table D-3. Year 2 Parental Perceptions of School Danger: ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Parental Perceptions of School Danger	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	2.09	2.99	-.90**	.00	-.94**	.00	-.27
	(3.26)	(3.45)	(.20)		(.20)		(3.45)
Subgroups							
SINI ever	2.37	3.47	-1.10**	.00	-1.22**	.00	-.35
	(3.23)	(3.49)	(.32)		(.31)		(3.49)
SINI never	1.86	2.62	-.75**	.01	-.71**	.01	-.21
	(3.27)	(3.38)	(.26)		(.26)		(3.38)
Difference	.50	.85	-.35	.41	-.51	.22	-.15
	(.26)	(.35)	(.42)		(.41)		(3.45)
Lower performance	2.33	2.91	-.58	.11	-.53	.14	-.16
	(3.43)	(3.37)	(.36)		(.36)		(3.37)
Higher performance	1.97	3.03	-1.05**	.00	-1.12**	.00	-.32
	(3.17)	(3.49)	(.24)		(.24)		(3.49)
Difference	.35	-.12	.47	.28	.59	.16	.17
	(.26)	(.36)	(.43)		(.42)		(3.45)
Male	2.09	2.94	-.85**	.00	-.94**	.00	-.27
	(3.18)	(3.43)	(.29)		(.29)		(3.43)
Female	2.08	3.03	-.95**	.00	-.94**	.00	-.27
	(3.35)	(3.48)	(.28)		(.27)		(3.48)
Difference	.01	-.09	.10	.81	.00	1.00	.00
	(.24)	(.33)	(.41)		(.40)		(3.45)
K-8	1.91	2.83	-.92**	.00	-.92**	.00	-.27
	(3.22)	(3.41)	(.22)		(.22)		(3.41)
9-12	2.90	3.75	-.85	.10	-1.01	.06	-.28
	(3.31)	(3.56)	(.51)		(.54)		(3.56)
Difference	-.99	-.92	-.07	.91	.09	.88	.02
	(.43)	(.37)	(.56)		(.59)		(3.45)
Cohort 2	1.94	2.83	-.89**	.00	-.91**	.00	-.27
	(3.23)	(3.42)	(.21)		(.21)		(3.42)
Cohort 1	2.63	3.57	-.94	.06	-1.04*	.04	-.30
	(3.33)	(3.50)	(.50)		(.49)		(3.50)
Difference	-.69	-.75	.06	.92	.13	.81	.04
	(.36)	(.46)	(.54)		(.53)		(3.45)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTE: Effect sizes are in terms of standard deviations. Valid *N* = 1,555. Parent survey weights were used.

Table D-4. Year 2 Student Perceptions of School Danger: ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Student Perceptions of School Danger	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	1.91	1.94	-.03	.82	-.02	.87	-.01
	(1.89)	(1.85)	(.15)		(.14)		(1.85)
Subgroups							
SINI ever	1.95	1.79	.16	.45	-.17	.40	.09
	(1.99)	(1.85)	(.22)		(.21)		(1.85)
SINI never	1.87	2.06	-.19	.33	-.18	.36	-.10
	(1.81)	(1.85)	(.20)		(.20)		(1.85)
Difference	.08	-.27	.35	.23	.35	.22	.19
	(.17)	(.24)	(.30)		(.29)		(1.85)
Lower performance	1.90	1.87	.03	.91	.07	.81	.03
	(2.04)	(2.06)	(.29)		(.28)		(2.06)
Higher performance	1.91	1.96	-.06	.73	-.05	.73	-.03
	(1.83)	(1.77)	(.17)		(.16)		(1.77)
Difference	-.00	-.10	.09	.79	.12	.70	.07
	(.19)	(.28)	(.34)		(.32)		(1.85)
Male	2.05	2.00	.05	.82	.07	.74	.04
	(1.94)	(1.92)	(.22)		(.21)		(1.92)
Female	1.76	1.88	-.12	.53	-.11	.57	-.06
	(1.83)	(1.79)	(.19)		(.19)		(1.79)
Difference	.29	.12	.17	.55	.18	.53	.09
	(.17)	(.24)	(.29)		(.28)		(1.85)
4-8	2.01	1.98	.03	.88	.01	.94	.01
	(1.94)	(1.86)	(.17)		(.16)		(1.86)
9-12	1.44	1.74	-.29	.22	-.20	.44	-.11
	(1.56)	(1.81)	(.24)		(.26)		(1.81)
Difference	.56	.24	.32	.28	.21	.50	.12
	(.20)	(.22)	(.30)		(.31)		(1.85)
Cohort 2	1.94	1.93	.01	.98	-.00	.99	-.00
	(1.96)	(1.88)	(.17)		(.17)		(1.88)
Cohort 1	1.78	1.95	-.17	.55	-.10	.68	-.06
	(1.64)	(1.74)	(.29)		(.24)		(1.74)
Difference	.16	-.02	.17	.60	.10	.74	.05
	(.17)	(.30)	(.33)		(.29)		(1.85)

NOTE: Effect sizes are in terms of standard deviations. Valid *N* = 1025. Student survey weights were used. Survey given to students in grades 4-12.

Table D-5. Year 2 Parental Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Parents Who Gave Child's School a Grade of A or B	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	.63	.12**	.00	.13**	.00	.26
	(.43)	(.48)	(.03)		(.03)		(.48)
Subgroups							
SINI ever	.70	.57	.12**	.00	.13**	.00	.26
	(.46)	(.50)	(.04)		(.04)		(.50)
SINI never	.80	.69	.13**	.00	.12**	.00	.27
	(.40)	(.46)	(.04)		(.04)		(.46)
Difference	-.11	-.11	-.00	.97	.01	.92	.01
	(.04)	(.04)	(.06)		(.06)		(.48)
Lower performance	.70	.58	.12*	.01	.11*	.03	.22
	(.46)	(.49)	(.05)		(.05)		(.49)
Higher performance	.78	.66	.13**	.00	.14**	.00	.29
	(.41)	(.47)	(.03)		(.03)		(.47)
Difference	-.09	-.08	-.01	.87	-.03	.62	-.06
	(.04)	(.05)	(.06)		(.06)		(.48)
Male	.73	.65	.08*	.04	.09*	.02	.18
	(.44)	(.48)	(.04)		(.07)		(.48)
Female	.78	.62	.16**	.00	.16**	.00	.34
	(.41)	(.49)	(.04)		(.04)		(.49)
Difference	.06	.02	-.08	.15	-.08	.17	-.17
	(.04)	(.04)	(.06)		(.06)		(.48)
K-8	.79	.64	.15**	.00	.16**	.00	.33
	(.41)	(.48)	(.03)		(.03)		(.48)
9-12	.60	.59	.01	.91	-.01	.89	.27
	(.49)	(.49)	(.07)		(.07)		(.49)
Difference	.21	.05	.14*	.04	.16*	.02	.01
	(.07)	(.05)	(.07)		(.07)		(.48)
Cohort 2	.78	.66	.13**	.00	.14**	.00	.29
	(.41)	(.47)	(.03)		(.03)		(.47)
Cohort 1	.66	.55	.09	.16	.09	.16	.18
	(.48)	(.50)	(.06)		(.06)		(.50)
Difference	.14	.10	.04	.54	.05	.44	.11
	(.05)	(.06)	(.07)		(.07)		(.48)

**Statistically significant at the 99 percent confidence level.

NOTE: Valid *N* for school grade = 1,549. 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.

Table D-6. Year 2 Parental Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Average Grade Parent Gave Child's School (5.0 Scale)	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	4.02	3.73	.29**	.00	.29**	.00	.29
	(.94)	(1.02)	(.06)		(.06)		(1.02)
Subgroups							
SINI ever	3.89	3.56	.33**	.00	.31**	.00	.29
	(1.00)	(1.08)	(.10)		(.09)		(1.08)
SINI never	4.13	3.86	.27**	.00	.28**	.00	.29
	(.88)	(.96)	(.07)		(.07)		(.96)
Difference	-.25	-.30	.06	.64	.04	.75	.04
	(.07)	(.10)	(.12)		(.12)		(1.02)
Lower performance	3.85	3.64	.21	.05	.18	.10	.17
	(1.02)	(1.06)	(.11)		(.11)		(1.06)
Higher performance	4.10	3.77	.33**	.00	.34**	.00	.34
	(.90)	(1.01)	(.07)		(.07)		(1.01)
Difference	-.25	-.12	-.12	.34	-.16	.22	-.16
	(.07)	(.11)	(.13)		(.13)		(1.02)
Male	3.96	3.79	.17*	.05	.17*	.03	.17
	(.98)	(1.01)	(.08)		(.08)		(1.01)
Female	4.08	3.67	.41**	.00	.41**	.00	.39
	(.90)	(1.03)	(.08)		(.08)		(1.03)
Difference	-.12	.12	-.24*	.04	-.24*	.03	-.23
	(.07)	(.10)	(.12)		(.11)		(1.02)
K-8	4.10	3.75	.35**	.00	.36**	.00	.34
	(.93)	(1.03)	(.06)		(.06)		(1.03)
9-12	3.67	3.66	.02	.89	-.01	.93	-.01
	(.93)	(.97)	(.14)		(.14)		(.97)
Difference	.42	.09	.33*	.04	.37*	.02	.36
	(.12)	(.10)	(.16)		(.16)		(1.02)
Cohort 2	4.08	3.78	.30**	.00	.31**	.00	.30
	(.93)	(1.03)	(.06)		(.06)		(1.03)
Cohort 1	3.81	3.55	.26	.07	.25	.06	.25
	(.97)	(.99)	(.14)		(.14)		(.99)
Difference	.28	.24	.04	.80	.05	.72	.05
	(.09)	(.13)	(.16)		(.15)		(1.02)

**Statistically significant at the 99 percent confidence level.

NOTE: Valid *N* for school grade = 1,549. Parent survey weights were used.

Table D-7. Year 2 Parental Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
School Satisfaction Scale	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	26.03	23.44	2.60**	.00	2.67**	.00	.33
	(7.65)	(8.01)	(.47)		(.45)		(8.01)
Subgroups							
SINI ever	25.10	21.85	3.25**	.00	3.21**	.00	.38
	(8.05)	(8.43)	(.78)		(.73)		(8.43)
SINI never	26.78	24.68	2.10**	.00	2.25**	.00	.30
	(7.23)	(7.43)	(.56)		(.56)		(7.43)
Difference	-1.68	-2.83	1.15	.24	.97	.29	.12
	(.57)	(.84)	(.98)		(.92)		(8.01)
Lower performance	25.02	22.80	2.22*	.02	2.05*	.02	.24
	(8.25)	(8.58)	(.93)		(.90)		(8.58)
Higher performance	26.52	23.72	2.80**	.00	2.95**	.00	.38
	(7.30)	(7.72)	(.53)		(.51)		(7.72)
Difference	-1.50	-.92	-.58	.59	-.89	.39	-.11
	(.56)	(.92)	(1.06)		(1.03)		(8.01)
Male	26.24	23.64	2.60**	.00	2.67**	.00	.34
	(7.34)	(7.93)	(.64)		(.63)		(7.93)
Female	25.82	23.26	2.56**	.00	2.68**	.00	.33
	(7.95)	(8.07)	(.68)		(.64)		(8.07)
Difference	.42	.38	.04	.97	-.00	1.00	-.00
	(.53)	(.76)	(.94)		(.89)		(8.01)
K-8	26.44	23.68	2.77**	.00	2.84**	.00	.35
	(7.59)	(8.07)	(.51)		(.48)		(8.07)
9-12	24.11	22.30	1.81	.11	1.88	.10	.25
	(7.62)	(7.61)	(1.12)		(1.14)		(7.61)
Difference	2.34	1.38	.96	.43	.96	.43	.12
	(.94)	(.83)	(1.22)		(1.23)		(8.01)
Cohort 2	26.50	23.60	2.90**	.00	3.00**	.00	.38
	(7.70)	(8.00)	(.50)		(.47)		(8.00)
Cohort 1	24.29	22.88	1.41	.21	1.44	.19	.18
	(7.21)	(7.99)	(1.12)		(1.10)		(7.99)
Difference	2.21	.72	1.49	.22	1.57	.19	.20
	(.71)	(1.07)	(1.22)		(1.18)		(8.01)

**Statistically significant at the 99 percent confidence level.

NOTE: Valid *N* for parent satisfaction = 1,571. Parent survey weights were used.

Table D-8. Year 2 Student Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Students Who Gave Their School a Grade of A or B	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	.70	.68	.02	.57	.03	.49	.05
	(.46)	(.47)	(.04)		(.04)		(.47)
Subgroups							
SINI ever	.68	.58	.09	.08	.12*	.02	.24
	(.47)	(.49)	(.05)		(.05)		(.49)
SINI never	.72	.76	-.05	.34	-.06	.27	-.14
	(.45)	(.43)	(.05)		(.05)		(.43)
Difference	-.04	-.18	.13*	.04	.16*	.01	.34
	(.04)	(.06)	(.06)		(.06)		(.47)
Lower performance	.67	.62	.04	.50	.01	.85	.03
	(.47)	(.48)	(.07)		(.06)		(.48)
Higher performance	.71	.70	.01	.77	.03	.46	.07
	(.45)	(.46)	(.04)		(.04)		(.46)
Difference	-.04	-.07	.03	.68	-.02	.79	-.05
	(.05)	(.07)	(.08)		(.08)		(.47)
Male	.69	.65	.04	.49	.05	.30	.12
	(.46)	(.48)	(.05)		(.05)		(.48)
Female	.71	.70	.01	.86	-.00	.97	-.00
	(.45)	(.46)	(.05)		(.05)		(.46)
Difference	-.02	-.05	.03	.71	.06	.40	.12
	(.04)	(.06)	(.07)		(.07)		(.47)
4-8	.73	.71	.03	.53	.03	.46	.07
	(.44)	(.46)	(.04)		(.04)		(.46)
9-12	.54	.54	-.00	1.00	-.00	.99	-.00
	(.50)	(.50)	(.07)		(.07)		(.50)
Difference	.19	.16	.03	.73	.03	.69	.07
	(.07)	(.06)	(.08)		(.08)		(.47)
Cohort 2	.73	.72	.01	.80	.01	.84	.02
	(.44)	(.45)	(.04)		(.04)		(.45)
Cohort 1	.58	.52	.05	.46	.07	.26	.15
	(.49)	(.50)	(.07)		(.07)		(.50)
Difference	.15	.19	-.04	.63	-.07	.41	-.14
	(.05)	(.08)	(.08)		(.08)		(.47)

NOTES: Valid *N* for school grade = 974. Student survey weights were 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 D-9. Year 2 Student Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
Average Grade Student Gave Their School (5.0 Scale)	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	3.95	3.84	.12	.21	.13	.14	.12
	(1.06)	(1.10)	(.09)		(.09)		(1.10)
Subgroups							
SINI ever	3.92	3.66	.26	.07	.28*	.02	.20
	(1.05)	(1.13)	(.14)		(.12)		(1.13)
SINI never	3.98	3.99	-.00	1.00	.00	.97	-.14
	(1.07)	(1.05)	(.12)		(.12)		(1.05)
Difference	-.07	-.32	.26	.17	.28	.10	.34
	(.10)	(.16)	(.19)		(.17)		(1.10)
Lower performance	3.92	3.63	.29	.16	.26	.19	.02
	(1.10)	(1.27)	(.20)		(.19)		(1.27)
Higher performance	3.97	3.91	.06	.58	.08	.42	.07
	(1.04)	(1.02)	(.10)		(.09)		(1.02)
Difference	-.05	-.28	.23	.31	.18	.41	-.05
	(.11)	(.20)	(.23)		(.22)		(1.10)
Male	3.93	3.76	.17	.22	.20	.12	.11
	(1.05)	(1.15)	(.14)		(.13)		(1.15)
Female	3.98	3.90	.08	.53	.07	.57	-.00
	(1.07)	(1.06)	(.12)		(.11)		(1.06)
Difference	-.04	-.14	.09	.61	.13	.43	.12
	(.10)	(.16)	(.18)		(.17)		(1.10)
4-8	4.05	3.89	.16	.13	.17	.09	.07
	(1.04)	(1.13)	(.11)		(.10)		(1.13)
9-12	3.54	3.61	-.07	.65	-.07	.66	-.00
	(1.04)	(.92)	(.16)		(.16)		(.92)
Difference	.51	.27	.23	.23	.24	.22	.07
	(.14)	(.13)	(.19)		(.20)		(1.10)
Cohort 2	4.04	3.91	.13	.21	.13	.19	.02
	(1.03)	(1.09)	(.10)		(.10)		(1.09)
Cohort 1	3.64	3.57	.07	.71	.12	.49	.15
	(1.11)	(1.10)	(.19)		(.17)		(1.10)
Difference	.40	.34	.06	.79	.01	.95	-.14
	(.12)	(.18)	(.22)		(.20)		(1.10)

NOTES: Valid *N* for school grade = 974. Student survey weights were used. Survey given to students in grades 4-12.

Table D-10. Year 2 Student Satisfaction ITT Impacts

	Mean Differences				Regression-Based Impact Estimates		
School Satisfaction Scale	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	34.04	33.25	.78	.15	.88	.10	.13
	(6.44)	(7.01)	(.54)		(.53)		(7.01)
Subgroups							
SINI ever	33.80	32.09	1.70*	.03	1.65*	.03	.24
	(6.31)	(6.96)	(.78)		(.74)		(6.96)
SINI never	34.24	34.20	.03	.96	.26	.73	.04
	(6.54)	(6.91)	(.73)		(.75)		(6.91)
Difference	-.44	-2.11	1.67	.12	1.39	.18	.20
	(.56)	(.92)	(1.07)		(1.05)		(7.01)
Lower performance	33.16	32.82	.34	.73	.19	.85	.03
	(6.59)	(7.01)	(.98)		(1.02)		(7.01)
Higher performance	34.41	33.41	.99	.13	1.14	.07	.16
	(6.34)	(7.00)	(.65)		(.64)		(7.00)
Difference	-1.25	-.59	-.66	.58	-.95	.44	-.14
	(.59)	(1.02)	(1.19)		(1.22)		(7.01)
Male	34.18	32.89	1.29	.09	1.41	.06	.20
	(6.64)	(7.12)	(.76)		(.75)		(7.12)
Female	33.89	33.56	.32	.66	.40	.58	.06
	(6.22)	(6.90)	(.74)		(.71)		(6.90)
Difference	.30	-.67	.97	.36	1.01	.32	.14
	(.53)	(.94)	(1.05)		(1.01)		(7.01)
4-8	34.23	33.51	.71	.26	.87	.16	.12
	(6.68)	(7.16)	(.63)		(.62)		(7.16)
9-12	33.16	32.04	1.12	.16	.91	.31	.15
	(5.08)	(6.13)	(.80)		(.89)		(6.13)
Difference	1.07	1.48	-.41	.69	-.04	.97	-.01
	(.64)	(.80)	(1.02)		(1.10)		(7.01)
Cohort 2	34.18	33.51	.67	.28	.77	.21	.11
	(6.42)	(6.99)	(.62)		(.61)		(6.99)
Cohort 1	33.50	32.38	1.12	.29	1.29	.21	.18
	(6.47)	(7.00)	(1.06)		(1.02)		(7.00)
Difference	.68	1.13	-.45	.71	-.51	.66	-.07
	(.60)	(1.10)	(1.23)		(1.18)		(7.01)

NOTES: Valid *N* for student satisfaction = 1,042. Student survey weights were used. Survey given to students in grades 4-12.

Appendix E
Relationship Between Attending a Private School and Key Outcomes

Scholarship programs such as the 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 less rigorous findings with some caution.

E.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 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. As with the estimation of the IOT, however, 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. 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 school in year 2. 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.[1] 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 places greater demands upon the data, 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.

Applying IV analytic methods to the experimental data from the evaluation, we find a statistically significant relationship between enrollment in a private school in year 2 and the following outcomes for groups of students and parents (table E-1):

- Reading achievement for students who applied from non-SINI schools; that is, among students from non-SINI schools, those who were enrolled in private school in year 2 scored 10.73 scale score points higher (ES = .30)[2] than those who were not in private school in year 2.

- Reading achievement for students who applied with relatively higher academic performance; the difference between those who were and were not attending private schools in year 2 was 8.36 scale score points (ES = .24).

- Parents' perceptions of danger at their child's school, with those whose children were enrolled in private schools in year 2 reporting 1.53 fewer areas of concern (ES = -.45) than those with children in the public schools.

- Parental satisfaction with schooling, such that, for example, parents are 20 percentage points more likely to give their child's school a grade of A or B if the child was in a private school in year 2.

- Satisfaction with school for students who applied to the OSP from a SINI school; for example, they were 23 percentage points more likely to give their current school a grade of A or B if it was a private school.

[1] 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 (82.4 percent) and to each member of the control group (17.9 percent), 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 only distinguishes between treatments and controls (who were randomly assigned) and cannot distinguish between private school enrollees and non-private school enrollees (who were self-selected), the use of the lottery as the instrumental variable in this analysis does generate unbiased estimates of the effects of private schooling.

[2] ES stands for Effect Size and is measured as a fraction of a standard deviation of the distribution of control group values of the outcome variable.

Table E-1. Private Schooling Effect Estimates for Statistically Significant ITT Results

Outcomes	IV Regression Estimate	*p*-value	Effect Size
Student Achievement Subgroups			
SINI never: Reading	10.73*	.03	.30
Higher performance: Reading	8.36*	.03	.24
Cohort 1: Reading	12.47	.15	.41
School Danger: Parents			
School danger	-1.53**	.00	-.45
School Satisfaction: Parents			
School grade of A or B	.20**	.00	.41
School grade, 5.0 scale	.50**	.00	.46
School satisfaction scale	4.19**	.00	.53
School Satisfaction: Student Subgroups			
SINI ever: School grade of A or B	.23*	.02	.46
SINI ever: School grade, 5.0 scale	.55*	.02	.50
SINI ever: School satisfaction scale	2.90*	.05	.43

*Statistically significant at the 95 percent confidence level.
**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for reading = 1,451. Reading sample weights used. Difference displayed in terms of scale scores.

Valid *N* for school danger = 1,440. Parent survey weights used.

Valid *N* for school grade = 1,435; parent satisfaction = 1,454. Parent survey weights used. School satisfaction scale was IRT scored and had a range of .96 to 35.43.

Valid *N* for school grade = 891; student satisfaction = 950. Student survey weights used. School satisfaction scale was IRT scored and had a range of 9.67 to 46.89.

E.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 E-2).

- The effect of private school attendance on reading for the cohort 1 subgroup was not statistically significant in the original IV analysis or in either of the different specifications.

- The sensitivity testing suggests that one of two different approaches to the estimation (using the trimmed sample), increased the size and statistical significance of the IV results for reading achievement of students from non-SINI schools (a change in *p*-values from .03 to .01) but led to smaller and non-significant results (a change in *p*-values from .03 to .05) for the reading achievement of students who were higher performing at baseline.

- Similarly, the approach of equalizing the response rates resulted in a higher estimate and statistical significance of the relationship between private schooling and SINI students' likelihood of giving their school a grade of A or B (a change in *p*-values from

.02 to .00) compared with the original analysis, and a lower and a non-significant estimate on the overall satisfaction scale for those students if they attended a private school (a change in *p*-values from .05 to .06).

- The finding that parental perceptions of school danger are lower for those who enrolled their child in private school is not sensitive to different analytic methods.
- The finding that parental satisfaction is higher for those who enrolled their child in a private school is not sensitive to different analytic methods.
- The second of the two different specifications (clustering on the school attended by the student), did not lead to any changes in the overall statistical significance of any of the findings from the original model.

Table E-2. Private Schooling Achievement Effects and *P*-Values with Different Specifications

	Original IV Estimate		Trimmed Sample		Clustering on Current School	
Outcomes	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Student Achievement Subgroups						
SINI never: reading	10.73*	.03	14.45**	.01	10.73*	.05
Higher performing: reading	8.36*	.03	7.96	.05	8.36*	.03
Cohort 1: reading	12.47	.15	3.27	.73	12.47	.13
School Danger: Parents						
School danger	-1.53**	.00	-1.36**	.00	-1.53**	.00
School Satisfaction: Parents						
Graded school A or B	.20**	.00	.25**	.00	.20**	.00
Grade given school (5.0 scale)	.50**	.00	.56**	.00	.50**	.00
School satisfaction scale	4.19**	.00	4.26**	.00	4.19**	.00
School Satisfaction Student Subgroups						
SINI ever: School grade of A or B	.23*	.02	.28**	.00	.23*	.03
SINI ever: School grade, 5.0 scale	.55*	.02	.51*	.02	.55*	.04
SINI ever: School satisfaction scale	2.90*	.05	2.32	.06	2.90*	.04

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Reading sample weights were used. Impact displayed in terms of scale scores. Parent survey weights were used for parent survey items. Student survey weights were used for student survey items. Impact estimates reported for the dichotomous variable “parents who gave school a grade of A or B” are reported as marginal effects.

Appendix F
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 as mediators of the impact of the treatment on student achievement. Previous research regarding the 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 within each of these four categories, as described below.

F.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 3 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) and to have a relationship to student achievement (Henderson and Berla 1994; Sui-Chu and Willms 1996).

The parental involvement variable ranges from .7 to 7.66 with a mean of 2.99 and a standard deviation of 2.03. The Cronbach's Alpha for the parental involvement scale is .71.[1]

[1] Cronbach's Alpha is a measure of the consistency and reliability of a scale (Spector 1992). 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 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.

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 (Natriello and McDill 1986; Singh et al. 1995). The measure of parent aspirations ranges from 11 to 19. The mean of parent aspirations is 17.25, and the standard deviation is 2.33.

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 school work 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 et al. 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 .13, and the standard deviation is .33.

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= 0
b. 11-20 minutes=1
c. 21-30 minutes=2
d. 31-45 minutes=3
e. 46 minutes to an hour=4
f. More than 1 hour=5.

This variable was chosen because it has been shown to be associated with student achievement (Dolton et al. 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 0 to 5 with a mean of 2.68 and a standard deviation of 1.34.

F.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, Reading for Fun, Engagement in Extracurricular Activities, and Frequency of Homework (measured in days).

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 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 16.74, and the standard deviation is 1.98.

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 days =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 .75 and a standard deviation of .84.

3. Tardiness

Tardiness is a measure of how often the student has missed school. Taken from the parent survey and evaluating 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 et al. 1992). The tardiness variable ranges from 0 to 3. Tardiness has a mean of .47 and a standard deviation of .76.

4. Reading for Fun

Reading 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 responds that he/she reads 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 et al. 2003). Reading for fun has a mean of .40 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 5 items including 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 2.29 and a standard deviation of 1.30.

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 0 to 5, 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 (Rutter et al. 1979; Natriello and McDill 1986; Rumberger and Palardy 2005). The mean of frequency of homework is 3.90, and the standard deviation is 1.45.

F.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, Challenge of Classes, Ability Grouping, Availability of Tutors, In-School Tutor Usage, Programs for Learning Problems or English Language Learners, Programs for Advanced Learners, Before-/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). Student/teacher ratio ranges from 1.30 to 42.50. The mean of student/teacher ratio is 12.81, and the standard deviation is 5.17.

2. Teacher Attitude

Teacher attitude measures how students report being treated by their classroom teachers. Taken from the student survey, the variable is an IRT scale that combines student evaluations of 4 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 (Hanushek 1971; Card and Krueger 1992; Wayne and Youngs 2003; Wolf and Hoople 2006). Teacher attitude ranges from .39 to 10.74 with a mean of 2.82 and a standard deviation of 2.23. The Cronbach's Alpha for teacher attitude is .75.

3. Challenge of Classes

Challenge of classes measures how difficult the student finds the classes in which he or she is enrolled. Taken from the student survey, the variable is an IRT scale that combines student evaluations of 4 items involving how hard class work was to learn, how difficult it was to keep up with homework, if he or she needed additional help from teachers, and if he or she understood what the teachers explained. Challenge of classes was chosen because it has been shown to be related to student achievement (Lee and Bryk 1988; Sheehan and DuPrey 1999). Challenge of classes ranges from .66 to 5.15 with a mean of 2.48 and a standard deviation of 1.22. The Cronbach's Alpha for challenge of classes is .72.

4. Ability Grouping

Ability grouping is a measure of the ways in which a school differentiates instruction based on student ability level. Taken from the school (i.e., principal's) 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 .62 and a standard deviation of .48.

5. Availability of Tutors

Availability of tutors measures whether the school a student attends has tutors available for its students. Taken from the school (i.e., principal's) 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: tutors have 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 .67 and a standard deviation of .47.

6. 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 .24 and a standard deviation of .43.

7. Programs to Assist Students with Learning Disabilities or English Language Learners

Programs to assist students with learning disabilities or English language learners is a count of how many programs a school reports offering out of a list of 3 items in the principal survey that includes special instruction for non-English speakers, special instruction for students with learning problems, and special instruction approaches along the lines of Success for All and Reading Recovery. 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). This variable ranges from 0 to 2. The mean of this variable is 1.03 and the standard deviation is .81.

8. Programs for Advanced Learners

Programs for advanced learners is a count of how many programs a school reports offering out of a list of 3 items in the principal survey that includes 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). The programs for advanced learners variable ranges from 0 to 3 with a mean of .64 and a standard deviation of .78.

9. Before-/After-School Programs

Before-/after-school programs was taken from the school (i.e., principal's) survey and is a dichotomous variable that equals 1 if the school offers a program to care 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 .97, indicating that almost every student in the impact sample attended a school with a before- or after-school program, and the standard deviation is .16.

10. Enrichment Programs

Enrichment programs is a count of how many programs a school reports offering out of a list of 3 items that includes 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 2.42 and a standard deviation of .78.

F.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. Parent/School Communication

Parent/school communication measures the amount of communication a school attempts to conduct with its students' parents. Taken from the school (i.e., principal's) survey, the variable is a count of the number of activities that a school reports implementing out of a list of 4 items that includes informing parents of their students' 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. Parent/school communication 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 3.10 and a standard deviation of .90.

2. School Size

School size is the total reported student enrollment in the attended school and is taken from the NCES CCD and NCES PSS. The variable was included in the analysis as a possible mediator because it has been associated with student achievement (Sander 1999). School size ranges from 12 to 3,017. The mean of school size is 379.47 and the standard deviation is 358.82.

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 NCES PSS. The variable was included in the analysis as a possible mediator because it has been shown to vary across public and private schools (Reardon and Yun 2002; Schneider and Buckley 2002) and to be associated with student achievement (Coleman et al. 1966; Coleman 1990; Hanushek et al. 2002; Nielsen and Wolf 2002). Percent non-white ranges from .01 to 1.00 with a mean of .95 and a standard deviation of .16.

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 5 statements about their peers including whether or not students behave well with teachers, students neglect their homework, students are made fun of by other students, other students often disrupt class, and students who misbehave often get away with it. Peer classroom behavior was chosen for the analysis as a possible mediator because it has been shown to vary across public and private schools (Lee et al. 1991) and to be associated with student achievement (Card and Krueger 1992). Peer classroom behavior ranges from 2.73 to 13.06 with a mean of 8.08 and a standard deviation of 2.23. The Cronbach's Alpha for peer classroom behavior is .68.[2]

[2] This Alpha rating falls short of the standard critical value of .70 for scale reliability. Thus, the results involving the peer classroom behavior variable in the mediator analysis should be treated with caution.