NCEE 2009-4050

U.S. DEPARTMENT OF EDUCATION

Evaluation of the DC Opportunity Scholarship Program

Impacts After Three Years

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

Evaluation of the DC Opportunity Scholarship Program

Impacts After Three Years

March 2009

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 2009-4050
U.S. Department of Education

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

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
Sue Betka
Acting Director

National Center for Education Evaluation and Regional Assistance
Phoebe Cottingham
Commissioner

March 2009

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

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

To order copies of this report,

- Write to ED Pubs, 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

Contents (continued)

Contents (continued)

List of Tables

Contents (continued)

List of Tables (continued)

Contents (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

List of Tables (continued)

Contents (continued)

List of Figures

Acknowledgments

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

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

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

The challenging task of assembling the analysis files was capably undertaken by Yong Lee, Quinn Yang, and Yu Cao at Westat. The management and conduct of the data collection was performed by Juanita Lucas-McLean and Sabria Hardy of Westat. Expert editorial and production assistance was provided by Evarilla Cover and Saunders Freeland of Westat. Jeffery Dean of the University of Arkansas ably assisted with the intermediate outcomes analysis and the drafting of chapter 4 and appendix F.

Disclosure of Potential Conflicts of Interests[1]

The research team for this evaluation consists of a prime contractor, Westat, and two subcontractors, Patrick Wolf (formerly at Georgetown University) and his team at the University of Arkansas Department of Education Reform and Michael Puma of Chesapeake Research Associates (CRA). None of these organizations or their key staff has financial interests that could be affected by findings from the evaluation of the DC Opportunity Scholarship Program (OSP). No one on the seven-member 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 Congress in January 2004, established the first federally funded, private school voucher program in the United States. As part of this legislation, Congress mandated a rigorous evaluation of the impacts of the Program, now called the DC Opportunity Scholarship Program (OSP). This report presents findings from the evaluation of the impacts 3 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 3 years, there was a statistically significant positive impact on reading test scores, but not math test scores.** Overall, those offered a scholarship were performing at statistically higher levels in reading—equivalent to 3.1 months of additional learning—but at similar levels in math compared to students not offered a scholarship (table 3). Analysis in prior years indicated no significant impacts overall on either reading or math achievement.

- **The OSP had a positive impact overall on parents' reports of school satisfaction and safety** (figures 3 and 4)**, but not on students' reports** (figures 3 and 4). Parents were more satisfied with their child's school (as measured by the percentage giving the school a grade of A or B) and viewed their child's school as safer and more orderly if the child was offered a scholarship. Students had a different view of their schools than did their parents. Reports of safety and school climate 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.** Fourteen percent of students in our impact sample who were randomly assigned by lottery to receive a scholarship and who responded to year 3 data collection chose not to use their scholarship at any point over the 3-year period after applying to the Program.[1] 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 reading achievement, parent views of school safety and climate, and parent views of

[1] This 14 percent "never user" rate among year 3 respondents in the impact sample differs from the 25 percent "never user" rate for the impact sample as a whole (Figure 1) because scholarship "never users" in the impact sample responded to year 3 data collection events at lower rates than did scholarship "ever users."

satisfaction all increase in size, and there remains no impact on math achievement and no overall impact on students' perceptions of school safety and climate or satisfaction from using an OSP scholarship.

- **The OSP improved reading achievement for 5 of the 10 subgroups examined.**[2] Being offered or using a scholarship led to higher reading test scores for participants who applied from schools that were not classified as "schools in need of improvement" (non-SINI). There were also positive impacts for students who applied to the Program with relatively higher levels of academic performance, female students, students entering grades K-8 at the time of application, and students from the first cohort of applicants. These impacts translate into 1/3 to 2 years of additional learning growth. However, the positive subgroup reading impacts for female students and the first cohort of applicants should be interpreted with caution, as reliability tests suggest that they could be false discoveries.

- **No achievement impacts were observed for five other subgroups of students, including those who entered the Program with relative academic disadvantage.** Subgroups of students who applied from SINI schools (designated by Congress as the highest priority group for the Program) or were in the lower third of the test score distribution among applicants did not demonstrate significant impacts on reading test scores if they were offered or used a scholarship. In addition, male students, those entering high school grades upon application, and those in application cohort 2 showed no significant impacts in either reading or math after 3 years.

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.

[2] The subgroups that are analyzed in this study were designated prior to the collection and analysis of data and are of particular policy interest based on the Program statute and education policy literature. The subgroups are: (1) whether students attended a school designated as in need of improvement (SINI) under the *No Child Left Behind Act* prior to application to the Program—students were either attending a SINI-ever or SINI-never school; (2) whether students were relatively lower performing or relatively higher performing at baseline—students were either in the bottom one-third or the top two-thirds of the test score distribution; (3) student gender; (4) whether students were entering grades K-8 or 9-12 at the time of application; and (5) whether students were in application cohort 1 (applied in 2004) or application cohort 2 (applied in 2005).

The Program is operated by the Washington Scholarship Fund (WSF). To date, there have been five 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.[3] A smaller number of applicants in spring 2006 (cohort 3), spring 2007 (cohort 4), and spring 2008 (cohort 5) were recruited and enrolled by WSF in order to keep the Program operating at capacity each year.

Table 1. OSP Applicants by Program Status, Cohorts 1 Through 5, Years 2004-2008

	Cohort 1 (Spring 2004)	Cohort 2 (Spring 2005)	Total Cohort 1 and Cohort 2	Cohort 3 (Spring 2006), Cohort 4 (Spring 2007), and Cohort 5 (Spring 2008)	Total, All Cohorts
Applicants	2,692	3,126	5,818	2,034	7,852
Eligible applicants	1,848	2,199	4,047	1,284	5,331
Scholarship awardees	1,366	1,088	2,454	1,284	3,738
Scholarship users in initial year of receipt	1,027	797	1,824	1,057	2,881
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
Scholarship users fall 2008	496	411	909	807	1,714

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, fall 2007 for cohort 4, and fall 2008 for cohort 5.

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 nonacademic outcomes of concern to policymakers and use ". . . the strongest possible research design for determining the effectiveness" of the

[3] 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.

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 attending public schools or were rising kindergarteners 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),[4] surveys of parents, and surveys of students in grade 4 and above—administered by the evaluation team in central District of Columbia (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.

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

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 3 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 sample's treatment group were clustered, have characteristics that differed somewhat from the average participating OSP school. Although 56 percent of all participating schools were faith-based (39 percent were part of the Catholic Archdiocese of Washington), 82 percent of the treatment group attended a faith-based school, with 59 percent of them attending the 22 participating Catholic parochial schools (table 2). Twenty-two percent of treatment group students were attending a school that charged tuition above the statutory cap of $7,500 during their third year in the Program (table 2) even though 38 percent and 46 percent of participating schools charged tuitions above that cap in 2006-07 and 2007-08, respectively.

Table 2. Features of Participating OSP Private Schools Attended by the Treatment Group in Year 3

Characteristic	Weighted Mean	Highest	Lowest	Valid *N*
Archdiocesan Catholic schools (percent of treatment students attending)	59.2	NA	NA	66
Other faith-based schools (percent of treatment students attending)	22.5	NA	NA	66
Charging over $7,500 tuition (percent of treatment students attending)	22.3	NA	NA	48
Tuition	$6,620	$29,902	$3,600	48
Enrollment	260.5	1,072	10	43
Student *N*	701			

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, 2006-07, and 2007-08, Washington Scholarship Fund.

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 third year after applying to the OSP, students in the treatment and control groups did

not differ significantly regarding the proportion attending schools that offered a separate library (88 vs. 91 percent), gyms (71 and 72 percent), and art programs (89 and 87 percent). There were the following statistically significant differences (at the .01 level):

- Students in the treatment group were more likely than those in the control group to attend schools with a computer lab (96 vs. 87 percent), with special programs for advanced learners (48 vs. 32 percent), and that offered a music program (89 vs. 82 percent).
- Students in the treatment group were less likely than the control group to attend a school with a cafeteria facility (79 vs. 88 percent) or a nurse's office (30 vs. 81 percent).
- Students in the treatment group were also less likely than those in the control group to attend a school that offered special programs for non-English speakers (26 vs. 57 percent), special programs for students with learning problems (71 vs. 88 percent), counselors (69 vs. 82 percent), tutors (50 vs. 67 percent), and after-school programs (86 vs. 92 percent).

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 3 years of the Program (figure 1):

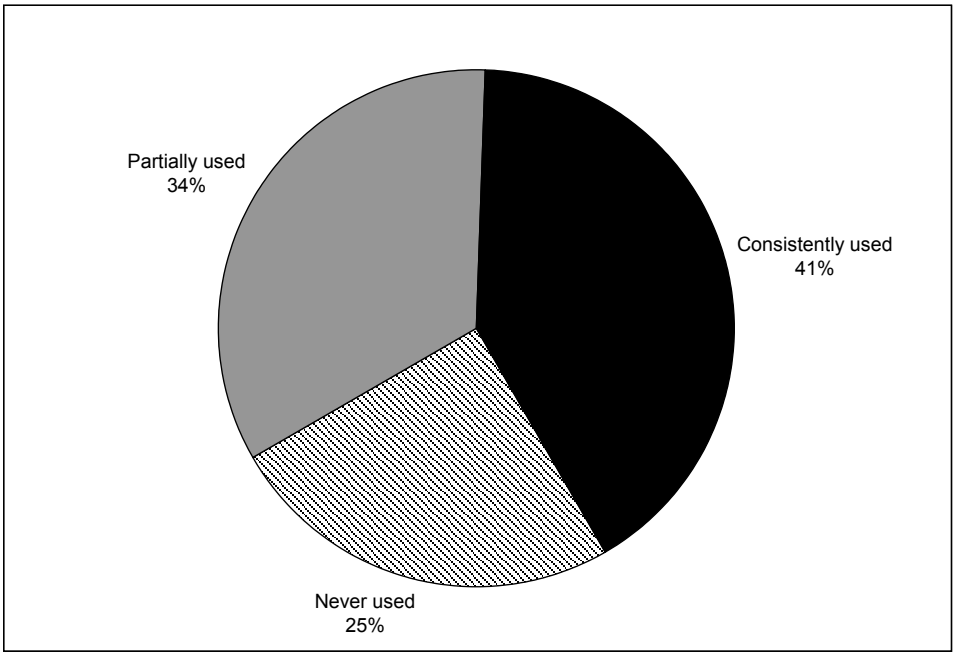
- 25 percent (346 out of 1,387 students) of those offered an OSP scholarship never used it;
- 34 percent (473 students) used their scholarship during some but not all of the first 3 years after the award; and
- The remaining 41 percent (568 students) used their scholarship consistently for the entire 3 years after the lottery.

The reasons for not using the scholarship—either initially or consistently—varied. The most common reasons cited by parents whose child never used their scholarship at anytime in year 3 and who completed surveys were (figure 2):

- Lack of available space in the private school they wanted their child to attend (22 percent of these parents);
- Child moved out of DC (21 percent of these parents);
- Child was accepted into a public charter school (19 percent of these parents); and

- Participating schools did not offer services for their child's learning or physical disability or other special needs (16 percent of these parents).

Figure 1. Proportions of Treatment Group Students Who Experienced Various Categories of Usage in First 3 Years



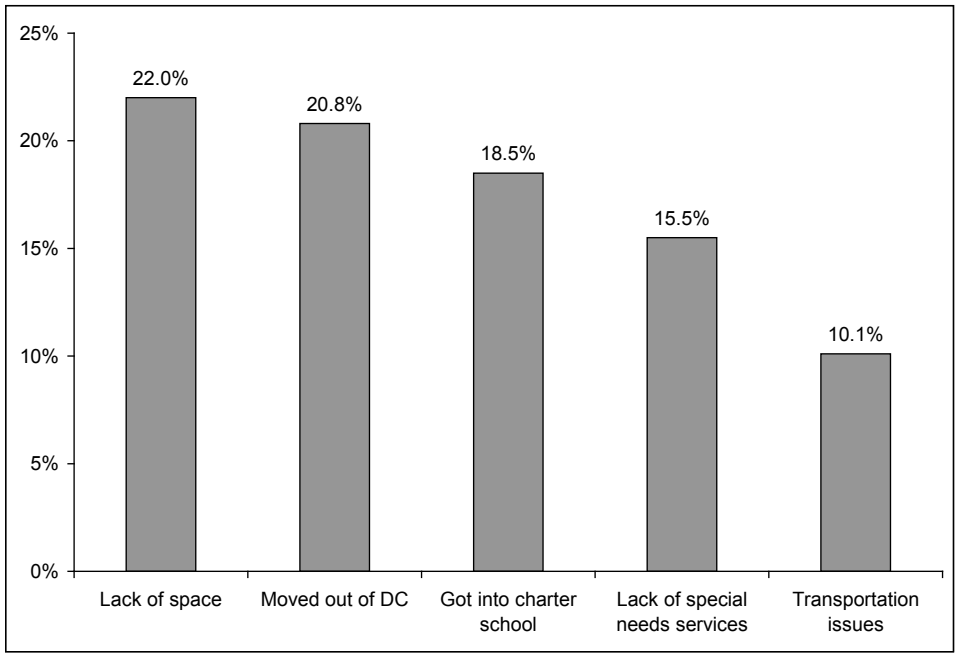
NOTES: Data are not weighted. Valid *N* = 1,387. Students were identified as scholarship users based upon information from WSF's payment files. Because some schools use a range of tuitions and some students had alternative sources of funding, students were classified as full users if WSF made payments 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 nonusers (no payments).

SOURCES: OSP applications and WSF's payment files.

The most common responses given by parents whose child initially used a scholarship in year 3 but dropped out of the OSP include:

- Lack of academic support that the child needed (39 percent of these parents);
- "Child did not like the private school" (25 percent);
- There was another private school the child liked better (13 percent);
- Work at the private school was too hard (11 percent);
- It was too difficult to get the child to the private school each day (11 percent); and
- The discipline or rules at the private school were too strict (7 percent).

Figure 2. Most Common Reasons Given by Parents for Declining to Use the OSP Scholarship in Year 3



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.

SOURCE: Impact Evaluation Parent Surveys.

Students who were partial users were more likely to have special needs and those entering the higher grades averaged lower baseline test scores than students who participated consistently across the 3 years.[5]

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 3 years after they applied to the OSP:

- Among the treatment group, 3 percent remained in the same school they were in when they applied to the Program; 46 percent switched schools once; 40 percent switched schools twice; and 11 percent switched three times.

[5] At baseline, partial users in grades 9-12 were lower performing in reading (27 National Percentile Ranks (NPRs) vs. 40 NPRs for full users, statistically significant at the .05 level) and in math (29 NPRs vs. 49 NPRs for full users, statistically significant at the .01 level); partial users in grades 6-8 were lower performing in math (34 NPRs vs. 41 NPRs for full users, statistically significant at the .01 level); and partial users were more likely to have special needs (5 percent vs. 10 percent for full users, statistically significant at the .05 level).

- Among the control group, 15 percent remained in the same school they were in when they applied to the Program; 40 percent switched schools once; 37 percent switched schools twice; and 8 percent switched three times.

These patterns of student mobility are important because previous studies suggest that switching schools has an initial short-term negative effect on student achievement (Hanushek, Kain, and Rivkin 2004).

Impact of the Program After 3 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 adjusting for students who declined to use their scholarships.[6] The main focus of this study was on the overall group of 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).

Previous reports released in spring 2007 and spring 2008 indicated that 1 and 2 years 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; Wolf et al. 2008). Parents were more satisfied if their child was in the Program and viewed their child's school as safer and more orderly. Among the secondary analyses of subgroups, there were impacts on math test scores in year 1 for students who applied from non-SINI schools and those with relatively higher pre-Program test scores, and impacts in reading test scores (but not math) in year 2 for those same two subgroups plus students who applied in the first year of Program implementation. 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. Throughout this report, the phrases "appears to have an impact" and "may have

[6] This analysis uses straightforward statistical adjustments to account not only for the approximately 14 percent of impact sample year 3 respondents who received the offer of a scholarship but declined to use it over the 3-year period after application (the "never users"), but also the estimated 1.6 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 do not alter the statistical significance of the impact estimate.

had an impact" are used to caution readers regarding statistically significant impacts that may have been false discoveries.

The analyses in this report were conducted using data collected on students 3 years after they applied to the OSP.[7]

Impacts on Students and Parents Overall

- Across the full sample, there was a statistically significant impact on reading achievement of 4.5 scale score points (effect size (ES) = .13)[8] from the offer of a scholarship and 5.3 scale score points (ES = .15) from the use of a scholarship (table 3). These impacts are equivalent to 3.1 and 3.7 months of additional learning, respectively.[9]

- There was no statistically significant impact on math achievement, overall (ES = .03) from the offer of a scholarship nor from the use of a scholarship (table 3).[10]

- Parents of students offered a scholarship were more likely to report their child's school to be safer and have a more orderly school climate (ES = .29) compared to parents of students not offered a scholarship (figure 3); 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 school safety and an orderly climate compared to those in the control group (ES = .06; figure 3); there was also no significant impact on student reports of school safety and an orderly climate from using a scholarship (ES = .07).

- The Program produced a positive impact on parent satisfaction with their child's school as measured by the likelihood of grading the school an "A" or "B," both for the impact of a scholarship offer (ES = .22; figure 4) and the impact of scholarship use (ES = .26).

---

[7] Specifically, year 3 test scores were obtained from 69 percent of study participants, whereas parent survey data were gathered from 68 percent of participants and student survey data from 67 percent of participants. Response rates to the principal survey varied between 51.8 percent and 57.3 percent, depending on academic year and school sector. Missing outcome data create the potential for nonresponse bias in a longitudinal evaluation such as this one, if the nonrespondent portions of the sample are different between the treatment and control groups. Response rates differed by less than 2 percent between the treatment and control groups for the tests and parent and student surveys, meaning that similar proportions of the treatment and control groups provided outcome data. In addition, nonresponse weights were used to equate the two groups on important baseline characteristics, thereby reducing the threat of nonresponse bias in this case.

[8] 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.

[9] Scale score impacts were converted to approximate months of learning first by dividing the impact ES by the ES of the weighted (by grade) average annual increase in reading scale scores for the control group. The result was the proportion of a typical year of achievement gain represented by the programmatic impact. That number was further divided by nine to convert the magnitude of the gain to months, since the official school year in the District of Columbia comprises 9 months of instruction.

[10] The magnitudes of these estimated achievement effects are below the threshold of .12 standard deviations, estimated by the power analysis to be the study's Minimum Detectable Effect (MDE) size.

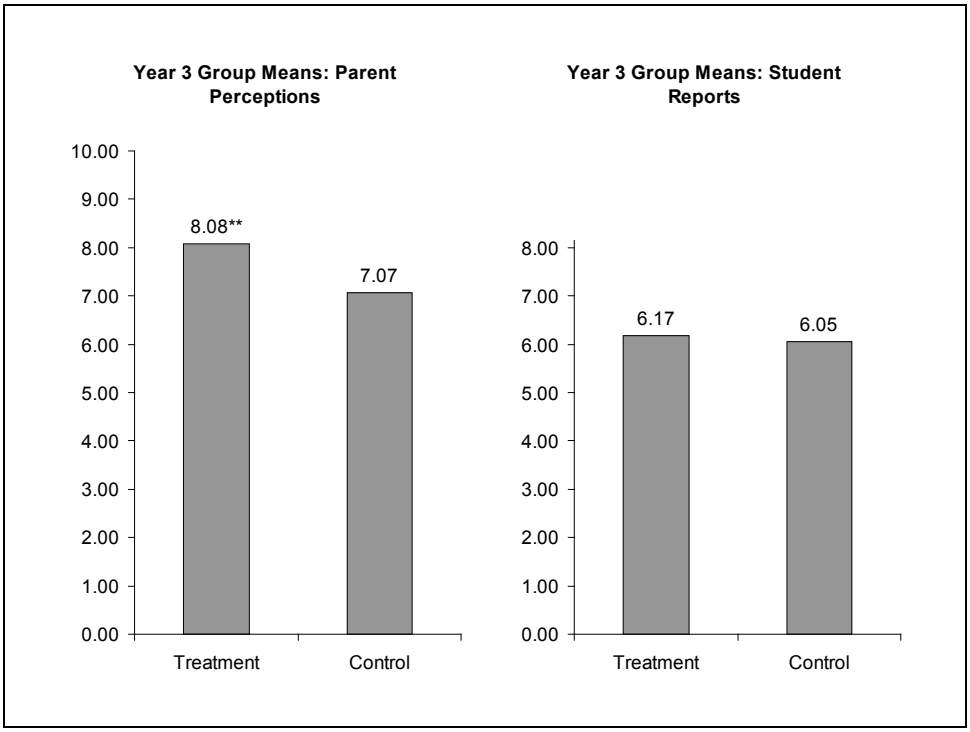
Table 3. Year 3 Impact Estimates of the Offer and Use of a Scholarship on the Full Sample: Academic Achievement

	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Reading	635.44	630.98	4.46*	.13	5.27*	.15	.01
Math	630.15	629.35	.81	.03	.95	.03	.62

*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 in terms of standard deviations. Valid *N* for reading = 1,460; math = 1,468. Separate reading and math sample weights used.

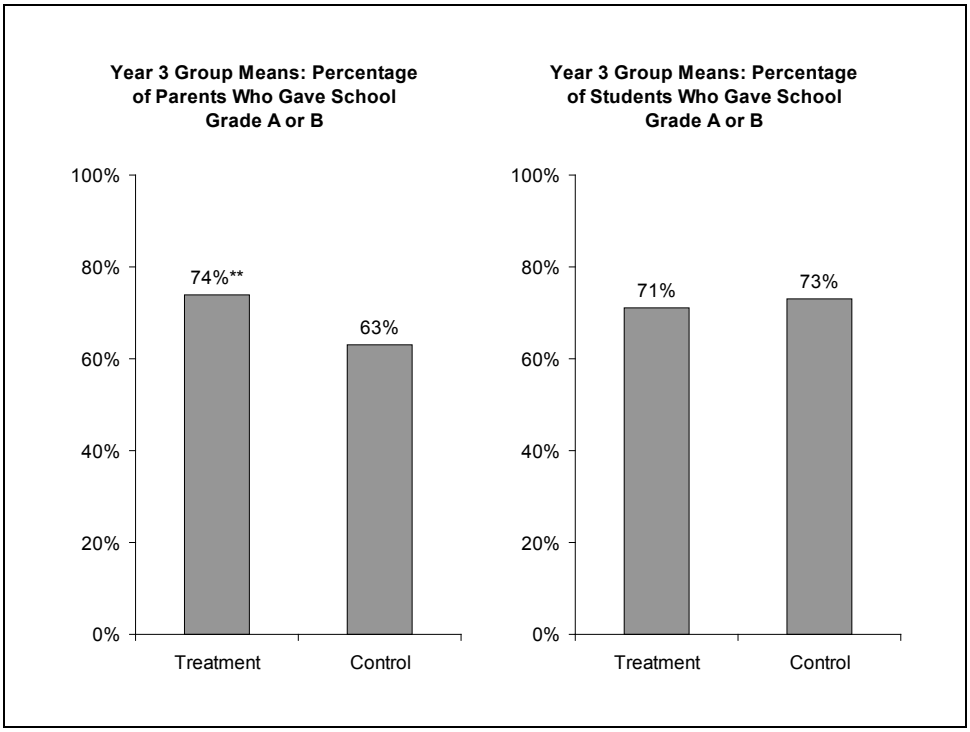
Figure 3. Parent Perceptions and Student Reports of Safety and an Orderly School Climate



**Statistically significant at the 99 percent confidence level.

NOTES: Parent perceptions are based on a ten-point scale; student reports are based on an eight-point scale. For parent perceptions, valid *N* = 1,423; parent survey weights were used; the ten-point index of indicators of school safety and an orderly environment includes the absence of property destruction, tardiness, truancy, fighting, cheating, racial conflict, weapons, drug distribution, drug/alcohol use, and teacher absenteeism. For student reports, valid *N* = 1,098; student survey weights were used; the survey was given to students in grades 4-12; the means represent the absence of incidents on an eight-item index for student reports of students being a victim of theft, drug-dealing, assaults, threats, bullying or taunting, or had observed weapons at school. Means are regression adjusted using a consistent set of baseline covariates.

Figure 4. Parent and Student Reports of School Satisfaction



**Statistically significant at the 99 percent confidence level.

NOTES: For parent reports, valid N = 1,410; parent survey weights were used. For student reports, valid N = 1,014; student survey weights were used; the survey was given to students in grades 4-12. Means are regression adjusted using a consistent set of baseline covariates.

- Overall, there were no impacts of the OSP from being offered (ES = -.06; figure 4) or using a scholarship (ES = -.07) on students' satisfaction with their schools as measured by the likelihood of assigning their school a grade of "A" or "B."

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 = .05) or math (ES = .01) 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.

- There were statistically significant impacts on reading test scores in year 3 for five subgroups of students, although the statistical significance of two of the subgroup 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 6.6 scale score points higher in reading (ES = .19) if they were offered the scholarship compared to not being offered a scholarship and 7.7 scale score points higher (ES = .22) if they used their scholarship compared to not being offered a scholarship. These scale score differences between the treatment and control groups translate into 4.1 and 4.9 additional months of learning, or half a year of schooling based on a typical 9-month school year.

    - 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 5.5 scale score points higher in reading (ES = .17) if they were offered a scholarship and 6.2 scale score points higher (ES = .19) if they used their scholarship, impacts equivalent to 4.0 and 4.6 months of learning gains.

    - Female students scored an average of 5.1 scale score points higher in reading (ES = .15) if they were offered a scholarship and 5.8 scale score points higher (ES = .17) if they used their scholarship. These impacts represent 3.1 and 3.6 months of additional learning, respectively. The statistical significance of this finding was not robust to adjustments for multiple comparisons.

    - Students who entered the Program in grades K-8 (81 percent of the impact sample) scored an average of 5.2 scale score points higher in reading (ES = .15) or 2.9 months of additional learning if they were offered a scholarship compared to not being offered a scholarship and 6.0 scale score points higher (ES = .17) or 3.3 months of additional learning 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 = .31) if they were offered a scholarship compared to not being offered a scholarship and 11.7 scale score points higher (ES = .42) if they used their scholarship compared to not being offered a scholarship. These impacts translate into 14.1 and 18.9 months of additional learning (1.5 to 2 years of typical schooling). The statistical significance of this finding was not robust to adjustments for multiple comparisons.

- The OSP had no statistically significant reading impacts for other subgroups of participating students, including those in the lower third of the test-score performance distribution at baseline, boys, secondary students, and students from the second cohort of applicants (ES ranging from -.00 to .11).

- The OSP had no statistically significant math impacts for any of the 10 subgroups (ES ranging from -.16 to .23).

Subgroup Safety and Satisfaction Impacts

- All of the 10 subgroups analyzed, including parents of the high-priority subgroup of students who had attended SINI schools at baseline, reported viewing their child's school as safer and more orderly 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 safety and an orderly school climate for the 10 subgroups ranged from .27 to .40. Adjustments for multiple comparisons indicate that these 10 subgroup impacts on parental perceptions of safety and school climate are not likely to be false discoveries.

- Consistent with the finding for students overall, none of the subgroups of students reported experiencing differences in safety and an orderly school climate if they were offered (ES range from -.03 to .08) or using an OSP scholarship.

- In addition to an overall impact on parental satisfaction with their child's school, the Program produced satisfaction impacts on 7 of the 10 subgroups analyzed. Effect sizes for the impact of an offer of a scholarship on the likelihood of a parent grading his/her child's school "A" or "B" for these seven subgroups ranged from .16 to .41. Adjustments for multiple comparisons indicated that none of these parent satisfaction subgroup impacts may have been a false discovery. The parents of students who had attended SINI schools, parents of students in the lower one-third of the test score distribution, and parents of high school students generally did not report higher levels of school satisfaction that were statistically significant as a result of the treatment (ES ranged from -.03 to .13).

- There were no statistically significant differences between the treatment group and the control group for all 10 subgroups in the likelihood that students gave their school a grade of A or B (ES ranged from -.18 to .05).

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 may be 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, as described below, to aid in the analysis.

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 because the Program affected students' experience of these factors. 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, 3 years after applying to the Program, the offer of an Opportunity Scholarship appears to have had an impact on 8 of the 24 intermediate outcomes examined, 7 of which remained statistically significant after adjustments for multiple comparisons:

- ***Home Educational Supports.*** Of the four intermediate outcomes in this category, the offer of a scholarship had an impact on one of them. There was a significant negative impact on tutor usage outside of school (ES = -.14), and this impact remained statistically significant after adjustments for multiple comparisons. There were no statistically significant differences between the treatment and control groups on parents' reports of their involvement in school in year 3 (ES = -.11), parents' aspirations for how far in school their children would go (ES = .02), or time required for the student to get to school (odds ratio = 1.13).[11]

- ***Student Motivation and Engagement.*** Of the six intermediate outcomes in this category, the offer of a scholarship may have had an impact on one of them. Based on student surveys, the offer of a scholarship seems to have had a significant negative impact on whether students read for fun (ES = -.16). Adjustments for multiple comparisons, however, indicate that this result could be a false discovery, so it should be interpreted with caution. There were no statistically significant differences between the treatment and control groups in their reported aspirations for future schooling (ES = -.14), engagement in extracurricular activities (ES = .04), and frequency of doing homework (ES = .08), or in their parents' reports of student attendance (odds ratio = 1.11) or tardiness rates (odds ratio = 1.19).

[11] The effect size for this categorical variable is expressed as an odds ratio, which describes the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response.

- ***Instructional Characteristics.*** The offer of a scholarship had a statistically significant impact on 5 of the 10 intermediate outcomes in this group of indicators. Students offered a scholarship experienced a lower likelihood that their school offered tutoring (ES = -.38), special programs for children who were English language learners (ES = -.61), or special programs for students with learning problems (ES = -.36) compared to control group students; these impacts remained statistically significant after adjustments for multiple comparisons. Students offered a scholarship experienced a higher likelihood that their school offered programs for advanced learners (ES = .27) and such enrichment programs as art, music, and foreign language (ES = .23); these two impact estimates also remained statistically significant after adjustments for multiple comparisons. There were no significant differences between the treatment and control groups in student/teacher ratio (ES = .01), how students rated their teacher's attitude (ES = -.04), the school's use of ability grouping (ES = .02), in-school tutor usage (ES = .04), or the availability of before- and after-school programs (ES = -.11).

- ***School Environment.*** The offer of a scholarship affected one of four measures of school environment. Students offered a scholarship experienced schools that were smaller by an average of 182 students (ES = -.29) than the schools attended by students in the control group; this impact remained statistically significant after adjustments for multiple comparisons. There were no statistically significant differences between the treatment and control groups, on average, in school reports of parent/school communication practices (ES = -.06), the percentage of minority students at the school (ES = -.10), or the classroom behavior of peers (ES = .09) based on student reports.

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 scholarship program in Washington, DC, with different design features than the OSP might also produce different outcomes.

1. Introduction

The *District of Columbia School Choice Incentive Act of 2003*,[1] passed by Congress in January 2004, established the first federally funded, private school voucher program in the United States. Since that time, more than 7,800 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 3 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 five 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), spring 2007 (cohort 4), and spring 2008 (cohort 5) 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 2008, a total of 1,714 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 fifth year of implementation. WSF limited applications to students entering grades K-6 for cohort 3 and grades K-7 for cohorts 4 and 5 because there were few slots available in participating high schools, as large numbers of students from cohorts 1 and 2 advanced to those grades. Applications also were limited to students previously attending public schools or rising kindergarteners, since public school students are a higher service priority of the Program than are otherwise eligible private school students. See chapter 2 for more detail on the cohort 1 and 2 exits from the Program that enabled WSF to accommodate cohorts 3 through 5.

[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 5, Years 2004-2008

	Cohort 1 (Spring 2004)	Cohort 2 (Spring 2005)	Total Cohort 1 and Cohort 2	Cohort 3 (Spring 2006), Cohort 4 (Spring 2007), and Cohort 5 (Spring 2008)	Total, All Cohorts
Applicants	2,692	3,126	5,818	2,034	7,852
Eligible applicants	1,848	2,199	4,047	1,284	5,331
Scholarship awardees	1,366	1,088	2,454	1,284	3,738
Scholarship users in initial year of receipt	1,027	797	1,824	1,057	2,881
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
Scholarship users fall 2008	496	411	909	807	1,714

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, fall 2007 for cohort 4, and fall 2008 for cohort 5.

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 nonacademic 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

[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)).

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).

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., 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 nonparticipants. Since the central purpose of the Program is to provide students with the option of attending a private school, the evaluation team has understood this provision as consistent with the examination of the effects of actual attendance at a private school. Previous experimental evaluations of scholarship programs have examined the effects of actual private school attendance on study participants (Howell et al. 2006, pp. 144-167; Greene 2001; Rouse 1998).

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 a subsequent, final report in 2010.[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, included in the current report, will contribute to the literature on voucher programs. Finally, the pattern of impact results raises a number of questions regarding, for example, the length of scholarship use or how the OSP affects the frequency or timing of school switching and what these mean for effects on achievement. Exploratory analyses to address these questions will be presented in the final evaluation report, benefiting from an additional year of data collection and examination.

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:

---

[9] We are deferring the analysis of education attainment (Question 2) until the final 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 appropriate 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 final evaluation 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.

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

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

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 represented a new opportunity. Thus, the impact sample for the evaluation comprised all eligible applicants who were previously attending public schools (or were rising kindergarteners) AND were subject to a lottery to determine whether they would receive an Opportunity Scholarship (figure 1-1, shaded area).

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).

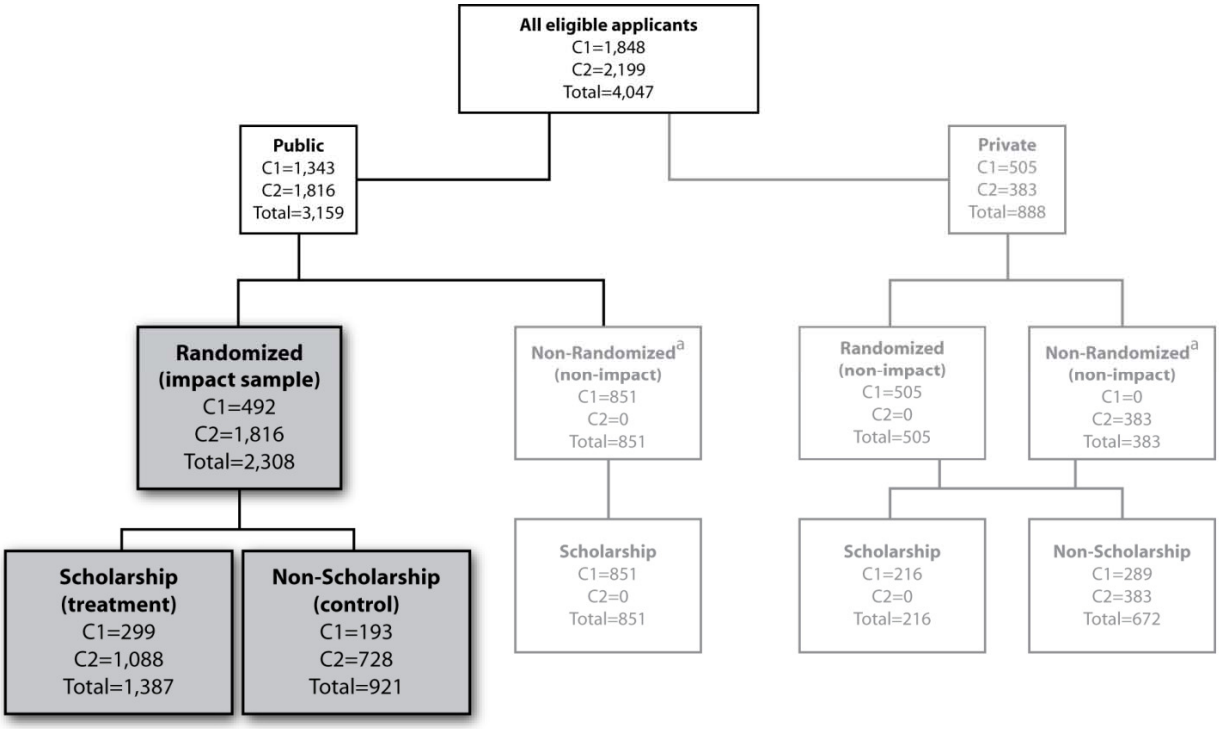
---

[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] As part of the control group follow-up lottery to reward control group members who cooperate with the evaluation's testing requirements, five members of the control group (cohort 1) were awarded scholarships by lottery in the summer of 2005, seven members of the control group (cohorts 1 and 2) were awarded scholarships by lottery in the summer of 2006, seven members of the control group (cohorts 1 and 2) were awarded scholarships by lottery in the summer of 2007, and four members of the control group (cohort 2) were awarded scholarships by lottery in the summer of 2008. Control group students who win a follow-up incentive lottery remain in the analysis as control group members, even though they have been awarded scholarships, to preserve the integrity of the original random assignment. They are treated as control group members for purposes of the intent-to-treat (ITT) and Bloom adjusted impact-on-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 information annually from students and families in the study, as well as from their schools, in order to address the key research questions. These data include:

- **Student assessments.** Measures of student achievement in reading and math for public school applicants 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 locations throughout DC arranged by the evaluators. The testing conditions are similar for members of the treatment and control groups.

- **Parent surveys.** The OSP application included baseline surveys for parents applying to the Program. These surveys were appended to the OSP application form and

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

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.

- **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.

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

Several methods were used to encourage high levels of response to year 3 data collection in spring 2007 (cohort 1) and spring 2008 (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.

After these initial data collection activities were completed, the test score response rate for year 3 was 63.9 percent—the treatment group response rate was 67.8 percent and the control group response rate was 57.8 percent, a response rate differential of 10 percentage points lower for the control group compared to the treatment group. To reduce this response rate differential, a random subsample of half of the control nonrespondents was drawn and subjected to intensive efforts at nonrespondent conversion (see appendix A, section A.7). Since these initial nonrespondents were selected at random, each one that was successfully converted to a respondent counts double, as he or she "stands in" for an approximately similar control nonrespondent that was not subsampled (see Kling, Ludwig, and Katz 2005; Sanbonmatsu et al. 2006). The "effective" response rate after subsample conversion is the number of actual respondents prior to the subsample plus two times the number of subsampled respondents, all divided by the total number of students in the impact sample.

As a result of the subsample conversion process, the final effective test score response rate for year 3 was 68.5 percent, and the differential rate of response between the treatment and control groups

was reduced to 1.7 percentage points higher for the control group.[15] The effective parent survey response rate was 67.9 percent.[16] The effective student survey response rate was 67.0 percent.[17] The principal survey was the data collection instrument with the lowest response rate in year 3, ranging from 51.8 percent to 57.3 percent depending on school sector and calendar year of administration.[18]

Missing outcome data create the potential for nonresponse bias in a longitudinal evaluation such as this one, if the nonrespondent portions of the sample are different between the treatment and control groups. Response rates for the various data collection instruments differed by less than 2 percent between the treatment and control groups, meaning that similar proportions of the treatment and control groups provided outcome data. In addition, nonresponse weights were used to equate the two groups on important baseline characteristics, thereby reducing the threat of nonresponse bias in this case (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 nonresponse conversion, and final nonresponse sample weights.

The test score response rate of 69 percent for this year 3 analysis of the OSP is higher than the response rates obtained in any of the three previous experimental evaluations of privately-funded K-12 scholarship programs 3 years after random assignment. The previous evaluations of such programs in New York City and Washington, DC reported year 3 test score response rates of 67 percent and 60 percent, respectively (Howell et al. 2006, p. 47). A previous experimental evaluation of the publicly-

---

[15] Specifically the overall effective response rates were 67.8 percent for the treatment group and 69.5 percent for the control group. Prior to drawing the subsample, response rates for the control group were 41.4 percent (cohort 1) and 61.9 percent (cohort 2). Response rates (after drawing the subsample) for the control group were 51.1 percent (cohort 1) and 66.8 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 60.9 percent (cohort 1) and 71.7 percent (cohort 2). Actual and effective response rates for the treatment group were 63.8 percent (cohort 1) and 68.9 percent (cohort 2). Eighty-five impact sample students awarded scholarships were entering grades 10 or higher at baseline and, therefore, were no longer grade-eligible for the OSP by year 3; these students are excluded from response rate calculations. See appendix A, figure A-1 and tables A-4 and A-6 for a detailed breakdown of the response rates and a further discussion of the subsampling procedure.

[16] Specifically the overall effective response rates were 67.4 percent for the treatment group and 68.6 percent for the control group. Response rates (after drawing the subsample) for the control group were 51.1 percent (cohort 1) and 66.2 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 60.9 percent (cohort 1) and 70.5 percent (cohort 2). Actual and effective response rates for the treatment group were 65.2 percent (cohort 1) and 68.0 percent (cohort 2). See appendix A and table A-7 for a detailed breakdown of the response rates.

[17] Specifically the overall effective response rates were 67.0 percent for the treatment group and 67.1 percent for the control group. Response rates (after drawing the subsample) for the control group were 50.9 percent (cohort 1) and 65.8 percent (cohort 2). After subsample weights were applied, the effective response rates for the control group were 60.7 percent (cohort 1) and 69.3 percent (cohort 2). Actual and effective response rates for the treatment group were 65.9 percent (cohort 1) and 67.4 percent (cohort 2). See appendix A and table A-8 for a detailed breakdown of the response rates.

[18] Since the principal survey is designed to gather information about all public and private schools in D.C., as opposed to a defined set of students in the impact sample, the response rates for this instrument are broken down by school sector (public or private) and by academic year (since cohort 1 students in the study experienced year 3 in these schools in 2006-07 but cohort 2 students experienced year 3 in 2007-08). For the principal survey, response rates for the 2007-08 school year were 56.4 percent (public schools) and 57.3 percent (private schools). For the 2006-07 school year, response rates were 53.2 percent (public schools) and 51.8 percent (private schools).

funded Milwaukee Parental Choice Program reported test score response rates in year 3 of 47 percent for the treatment group and 23 percent for the control group (Rouse 1998, p. 555).

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 .12 standard deviations for the entire study sample and .14 to .38 standard deviations for the subgroups of interest (see appendix A, section A.2).[19] Observations were weighted after data collection, using baseline characteristics associated with study nonresponse, to re-establish the equivalence of the treatment and control groups in the face of differential rates of nonresponse (see appendix 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 subsampling to convert control initial nonrespondents and (2) clustering the standard errors of the observations on school attended instead of family (see appendix C).

[19] This year's power analyses combined observation numbers (i.e., *N*s) from the actual year 3 respondent sample with assumptions regarding the strength of relationships in the data drawn from an earlier experimental analysis of a privately funded K-12 scholarship program in DC.

1.3 Contents of This Report

This report from the evaluation is the fifth in a series of required annual reports to Congress. It presents the impacts of the Program on students and families 3 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 3 years later as a result of the offer or the actual use of an Opportunity Scholarship. The final chapter (chapter 4) assesses the impacts of the Program on intermediate outcomes—such as parent aspirations and supports, student motivation and engagement, school instructional characteristics, and the school environment. This analysis is an attempt to develop hypotheses about the mechanisms through which private school vouchers may or may not lead to higher student achievement or better outcomes for students. The evaluation's final report will examine impacts at least 4 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 nonscholarship students switching schools during the 3 years since they applied to the OSP. Research links elements of students' educational environments and their school mobility to later outcomes.[20] 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. A total of 57[21] of 102 private schools in the District of Columbia were participating in the program at the start of the 2007-08 school year.[22] Among the participating schools:[23]

---

[20] 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, Cain, and Rivkin 2002; Card and Krueger 1992.

[21] While, technically, 61 individual campuses were participating in the OSP from the start of the school year, the research team treats four of the schools with dual campuses as single entities because they have one financial office that serves both campuses, following the classification practice used by the National Center for Education Statistics in their Private School Survey.

[22] This figure represents a net loss of 9 schools since the prior year. Because the data on participating schools was obtained from the WSF DC Opportunity Scholarship Program School Directory, prepared and distributed in the summer of 2007, the count does not include 3 private schools that confirmed their continuation in the OSP after the start of the 2007-08 school year. Thus, 60 schools were participating in the OSP during the 2007-08 school year, a net loss of six schools from 2006-07. According to WSF, 7 schools left the Program prior to the 2007-08 school year. Five of the 7 schools either closed or merged with other schools due to financial difficulties. The other two schools left the OSP because they were unable to attract OSP students (both served grades pre-K through 1). One new private school joined the Program for the first time prior to the 2007-08 school year. It served students in grades pre-K through K in that year but plans to add another grade in each subsequent year. Additionally, 3 schools that did not fully participate in 2007-08 are not included in this report. One school joined the Program mid-year, after the typical student enrollment period. Two schools that initially expressed the intent to participate did not fully complete the required documentation to be considered full participants. These three schools did not enroll any OSP students in 2007-08.

[23] Information was obtained for all 57 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 $N = 47$), percent minority students (valid $N = 49$) and student/teacher ratio (valid $N = 44$) were drawn from the National Center for Education Statistics' Private School Survey, last administered in 2005-06.

- Fifty-six percent (32) were faith-based, with most of them (22) the parochial schools of the Catholic Archdiocese of Washington;
- Forty-six percent charged an average tuition above the OSP's scholarship cap of $7,500;[24]
- The average school had a total student population of 265 students;
- Twenty-five percent served high school students;[25]
- The average minority percentage among the student body was 77 percent; and
- The average student/teacher ratio was 10.3.

These characteristics are similar to those presented in earlier reports from this evaluation.[26]

Schools Attended by Scholarship Users in Year 3

Not all of the schools that agreed to participate in the Program serve OSP students every year.[27] Three years after being awarded a scholarship, OSP students were enrolled in 54, and the impact sample's treatment students in 48, of the 66 schools available to them in that time period.[28] 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 typical 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. Twenty-two percent of treatment group students were attending a school that charged tuition above the statutory cap of $7,500 during their third year in the Program (table 2-1), even though 38 percent and 46 percent of participating schools charged tuitions above that cap in 2006-07 and 2007-08, respectively. Although 56 percent of all participating schools were faith-based (39 percent part of the

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

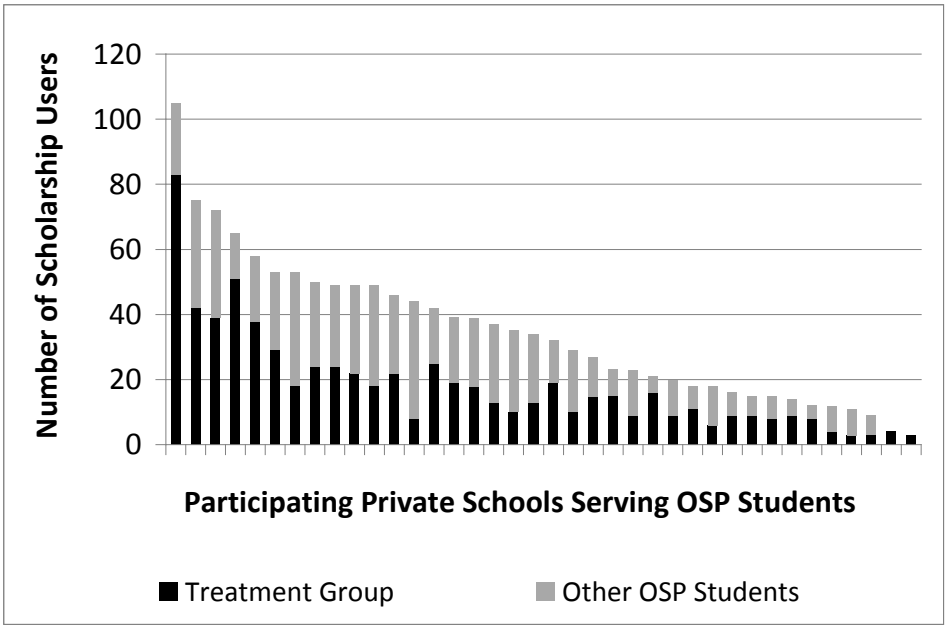
[25] Schools were classified as serving high school students if they enrolled students in any grade 9-12.

[26] See Wolf et al. 2007, pp. 15-17; Wolf et al. 2008, pp. 11-13.

[27] The source for student enrollment in participating schools is the WSF OSP payment file for 2006-07 and 2007-08.

[28] The impact sample combines data from the experience of cohort 1 students in 2006-07 (their impact year 3) and cohort 2 students in 2007-08 (their impact year 3). Collectively, the total number of schools available for cohort 1 during 2006-07 and cohort 2 during 2007-08 was 66.

Figure 2-1. Distribution of OSP Scholarship Users Across Participating Schools, by Impact Sample Treatment Group vs. Other OSP Students, Year 3



NOTES: Each bar represents a private school that enrolled OSP students during their third 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 $N = 66$. Student $N = 1{,}393$. Schools that did not enroll any OSP students in year 3 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 ($N = 16$).

SOURCE: WSF's payment files.

Table 2-1. Features of Participating OSP Private Schools Attended by the Treatment Group in Year 3

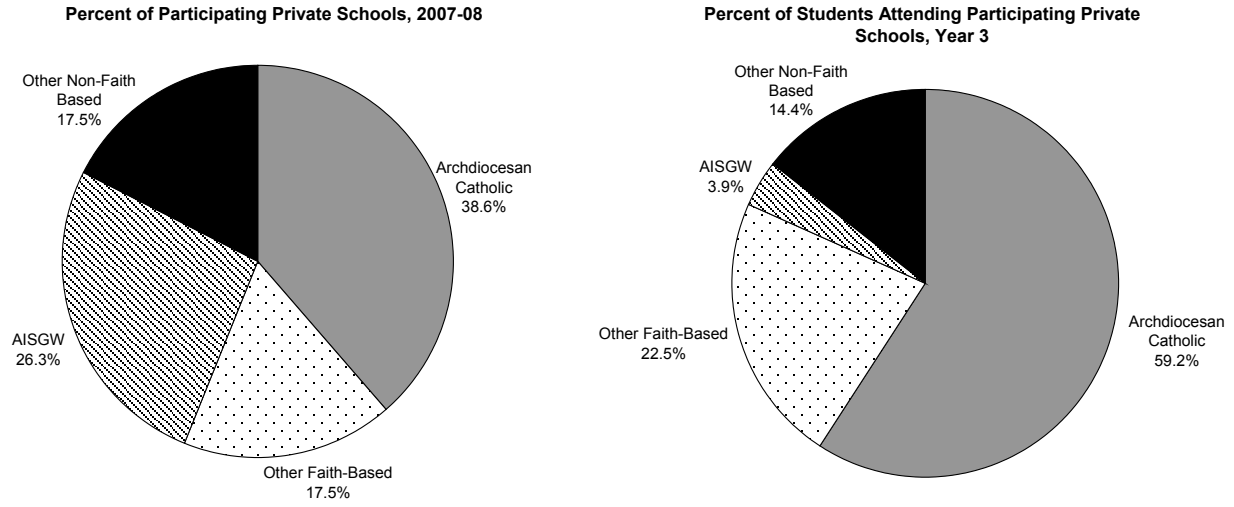
Characteristic	Weighted Mean	Highest	Lowest	Valid *N*
Charging over $7,500 tuition (percent of treatment students attending)	22.3	NA	NA	48
Tuition	$6,620	$29,902	$3,600	48
Enrollment	260.5	1,072	10	43
Student *N*	701			

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, 2006-07, and 2007-08, Washington Scholarship Fund.

Catholic Archdiocese of Washington), 82 percent of the treatment group attended a faith-based school, with most of them (59 percent) attending the 22 participating Catholic parochial schools (figure 2-2).

Figure 2-2. Religious Affiliation of Participating Schools



NOTES: School *N* = 57 for percent of participating private schools in 2007-08. School *N* = 66 and Student *N* = 701 for percent of students in the treatment group attending participating private schools in year 3 (which includes schools participating in 2006-07 and 2007-08). AISGW is an acronym for the Association of Independent Schools of Greater Washington.

SOURCES: National Center for Education Statistics: Private School Universe Survey, 2003-2004, supplemented by OSP School Directory information, 2004-05, 2005-06, 2006-07, 2007-08, Washington Scholarship Fund.

Schools Attended by the Treatment Group in Relation to Those of the Control Group in Year 3

While the characteristics of the participating private schools are important considerations for parents, how those characteristics differ from the public school options available to parents matter more. How different are the school conditions? Students in the treatment and control groups did not differ significantly regarding the proportion attending schools that offered a separate library (88 vs. 91 percent), gyms (71 and 72 percent), and art programs (89 and 87 percent) (table 2-2). Three years after they applied to the OSP, there were the following statistically significant differences (at the .01 level) between students in the treatment and control groups:[29]

- Students in the treatment group were more likely than those in the control group to attend schools with a computer lab (96 vs. 87 percent), with special programs for advanced learners (48 vs. 32 percent), and that offered a music program (89 vs. 82 percent).

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

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

Percentage of Students Attending a School with:	Baseline Year			Third Follow-up Year		
	Treatment	Control	Difference	Treatment	Control	Difference
Separate Facilities:						
Computer lab	73.53	73.88	-.35	95.51	86.68	8.83**
Library	80.12	79.33	.79	87.52	91.43	-3.90
Gym	63.67	65.26	-1.59	70.57	71.58	1.01
Cafeteria	87.39	88.08	-.69	79.02	87.57	-8.55**
Nurse's office	87.43	88.81	-1.38	30.22	80.65	-50.43**
Percent missing	6.84	7.86	-1.02	38.13	56.51	-18.38
Programs:						
Special program for non-English speakers	48.62	43.08	5.55	25.57	57.07	-31.50**
Special program for students with learning problems	64.35	66.28	-1.93	70.70	88.13	-17.43**
Special program for advanced learners	38.65	35.70	2.95	48.34	31.64	16.70**
Counselors	80.50	80.58	-.07	69.01	81.86	-12.85**
Individual tutors	36.58	38.68	-2.10	49.60	66.65	-17.05**
Music program	70.14	70.83	-0.70	89.01	82.28	6.73**
Art program	69.18	67.51	1.67	88.77	87.07	1.69
After-school program	79.98	79.37	0.60	86.19	91.89	-5.69**
Percent missing	7.16	8.13	-0.97	38.13	56.51	-18.38
Sample size (unweighted)	1,387	921	466	1,387	921	466

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Data are weighted. For a description of the weights, see appendix A. Some of the missing data rates for year 3 are a product of students naturally grading out of the Program's eligibility requirements. Of the 2,308 students in the impact sample, 85 are considered to have graded out during the third year, including 54 control group members and 31 treatment group members. Baseline year means presented here for the control group differ slightly from those presented in previous reports due to a minor correction in the weights applied to these observations.

SOURCES: The DC Opportunity Scholarship Program Application, the Impact Evaluation Parent Survey (for school attended), and the Impact Evaluation Principal Survey.

- Students in the treatment group were less likely than those in the control group to attend a school with a cafeteria facility (79 vs. 88 percent) or a nurse's office (30 vs. 81 percent).

- Students in the treatment group were also less likely than those in the control group to attend a school that offered special programs for non-English speakers (26 vs. 57 percent), special programs for students with learning problems (71 vs. 88 percent), counselors (69 vs. 82 percent), tutors (50 vs. 67 percent), and after-school programs (86 vs. 92 percent).

2.2 Student Participation

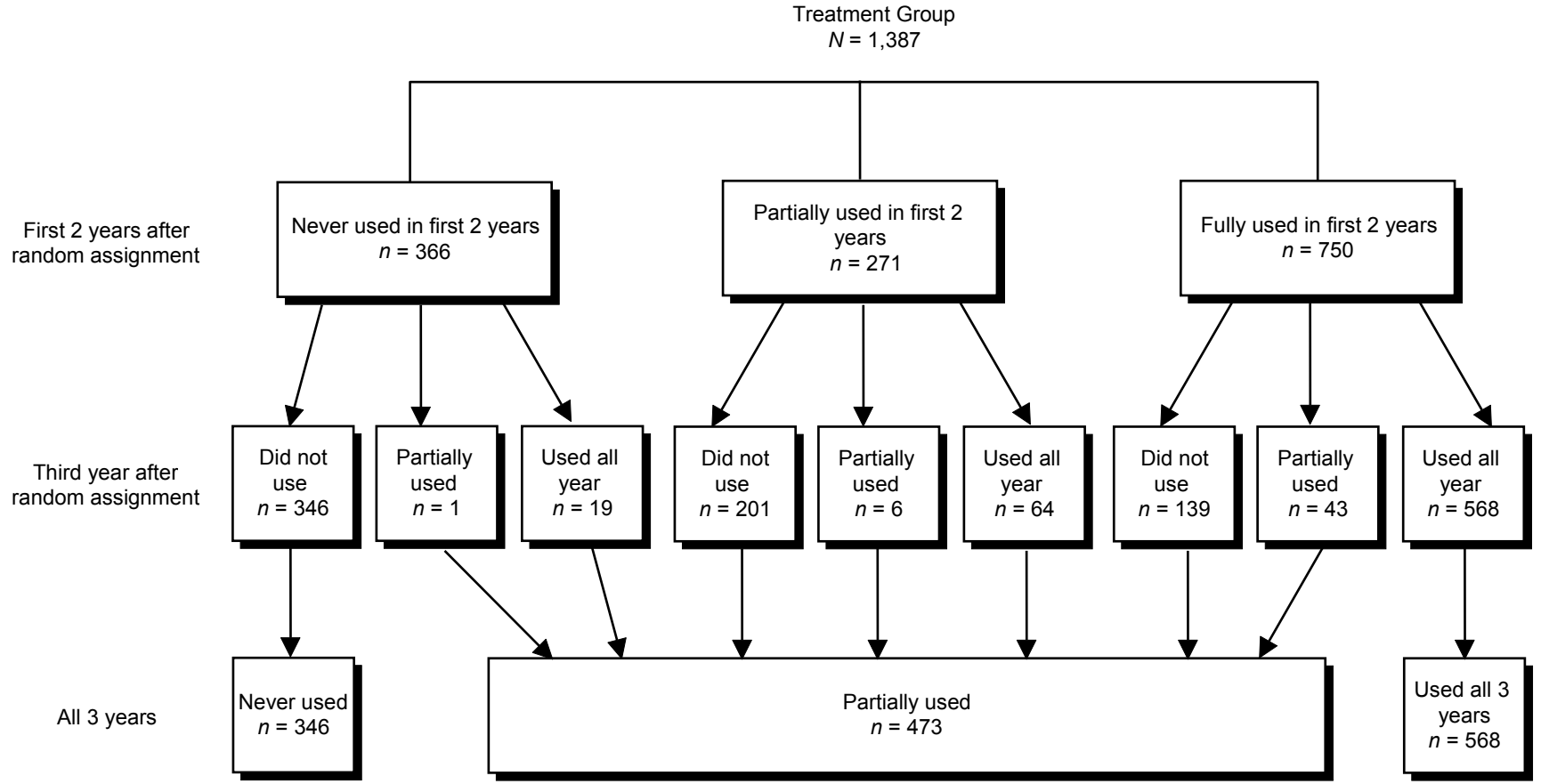
The degree to which students initially and consistently used their scholarships provides some signal of the attractiveness of the OSP to parents and the ability of the Program and its participating schools to accommodate their needs. A total of 2,454 students who applied to the OSP in the first two years of Program operation were offered scholarships, with 1,387 of them in the impact sample's treatment group. 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 the Program operator (the WSF), once a student was offered an OSP scholarship, he or she could use it at any time. During the first 3 years of the Program (figure 2-3):

- 346 out of 1,387 (25 percent) treatment group students never used the OSP scholarships offered to them;

Figure 2-3. Scholarship Usage by Students Assigned to the Treatment Group in First 3 Years



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

By year 3, 31 treatment group students had "graded out" and 7 had "earned out" of Program eligibility. They are included in this diagram as nonusers from the point at which they lost eligibility.

SOURCE: WSF's payment files.

- 473 treatment students (34 percent) used their scholarships during some but not all of the first 3 years after the scholarship award. Among these students are 142 students estimated to be "forced decliners," meaning that they could not continue to use their scholarship because they "graded out" (graduated high school), "earned out" (their family income grew to exceed the Program's eligibility requirements), or there was no space for them in a participating high school;[30] and

- The remaining 568 treatment group students (41 percent) used their scholarship during the entire 3 years after the scholarship lottery.

Certain preprogram 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 (57 vs. 44 percent) to be entering grades K-5 and less likely (6 vs. 20 percent) to be entering high school;

- Less likely (7 vs. 22 percent) to have special educational needs due to a disability;

- More likely (95 vs. 90 percent) to be African American and less likely (10 vs. 15 percent) to be Hispanic; and

- Less likely (49 vs. 56 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, percent having applied from SINI schools, mother's average years of education and employment status, and family income.

Among the subgroup of treatment students who ever used their scholarship, a somewhat different set of preprogram student characteristics were associated with full scholarship use. Compared to users who only partially used their scholarship, students who used their scholarship consistently (full users) for the 3-year period were significantly (table 2-4):

[30] The calculations regarding likely forced decliners were made using information from the baseline application/survey and administrative data provided by the WSF. A total of 85 students awarded scholarships in cohort 1 or cohort 2 were entering grades 10 or higher at baseline and therefore were no longer grade-eligible for the scholarship by the third year. A total of seven treatment students initially qualified for the Program but later reported family income of over 300 percent of the poverty level, 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 comparing the higher rate of scholarship continuation for 7th graders moving to 8th grade with the lower rate of scholarship continuation for 8th graders moving to 9th grade. The difference between those two continuation rates, applied to the number of OSP students moving from 8th to 9th grade generates the estimate of forced decliners due to high school slot constraints of 50 (20 in year 2 plus 30 new ones in year 3). It is impossible to know for certain if all 50 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 142 forced decliners for outcome year 3 is simply an estimate based on the limited data available.

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

Characteristic	Ever User	Never User	Difference
Achievement:			
Reading percentile: Grade K-5	34.00	32.77	1.22
Reading percentile: Grade 6-8	35.11	30.14	4.97
Reading percentile: Grade 9-12	33.62	30.05	3.57
Percent missing	38.42	38.44	-.01
Math percentile: Grade K-5	28.99	28.54	.45
Math percentile: Grade 6-8	37.10	33.56	3.54
Math percentile: Grade 9-12	38.86	38.70	.15
Percent missing	16.81	30.06	-13.25
Student demographics:			
Percent SINI	30.74	31.50	-.76
Percent entering: K-5	57.35	43.93	13.42**
Percent entering: 6-8	36.22	36.13	.09
Percent entering: 9-12	6.44	19.94	-13.51**
Percent missing	0	0	
Percent learning/physical disability	7.45	21.99	-14.54**
Percent missing	7.25	7.37	-.12
Percent African American	95.42	89.64	5.78**
Percent missing	7.68	10.69	-3.01
Percent Hispanic	10.34	14.55	-4.21*
Percent missing	6.15	6.65	-.50
Percent male	49.18	55.94	-6.76*
Percent missing	.19	.29	-.10
Family demographics:			
Mother's average years of education	12.54	12.63	-.09
Percent missing	15.47	21.10	-5.63
Percent mother full-time job	43.84	43.12	.73
Percent missing	16.52	20.23	-3.71
Average family income	$17,121.00	$17,067.00	$53.66
Percent missing	0	0	
Number of children	2.83	3.03	-.20*
Percent missing	.19	1.16	-.96
Months of residential stability	69.78	85.29	-15.50**
Percent missing	2.40	3.76	-1.36
Sample size (unweighted)	1,041	346	695

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

SOURCES: The DC Opportunity Scholarship Program Application and the WSF's payment files.

Table 2-4. Baseline Characteristics of Treatment Group Students Who Fully Used Their OSP Scholarship Compared to Partial Users in the First 3 Years

Characteristic	Full User	Partial User	Difference
Achievement:			
Reading percentile: Grade K-5	36.56	30.09	6.46
Reading percentile: Grade 6-8	37.38	33.10	4.28
Reading percentile: Grade 9-12	39.66	27.39	12.27*
Percent missing	41.55	34.67	6.88
Math percentile: Grade K-5	30.55	26.46	4.09
Math percentile: Grade 6-8	40.61	34.04	6.57*
Math percentile: Grade 9-12	49.19	28.53	20.66**
Percent missing	16.90	16.70	.20
Student demographics:			
Percent SINI	29.75	31.92	-2.17
Percent entering: K-5	63.56	49.89	13.66**
Percent entering: 6-8	30.63	42.92	-12.28**
Percent entering: 9-12	5.81	7.19	-1.38
Percent missing	0	0	
Percent learning/physical disability	5.34	10.07	-4.73**
Percent missing	5.37	9.51	-4.14
Percent African American	95.38	95.48	-.10
Percent missing	8.63	6.55	2.07
Percent Hispanic	10.78	9.79	.99
Percent missing	5.28	7.19	-1.91
Percent male	46.03	52.97	-6.93*
Percent missing	.18	.21	-.04
Family demographics:			
Mother's average years of education	12.59	12.49	.10
Percent missing	15.85	15.01	.83
Percent mother full-time job	45.24	42.17	3.07
Percent missing	16.73	16.28	.45
Average family income	$17,501	$16,664	$837.31
Percent missing	0	0	0
Number of children	2.71	2.99	-.28**
Percent missing	.00	.42	-.42
Months of residential stability	70.09	69.41	.68
Percent missing	2.11	2.75	-.64
Sample size (unweighted)	568	473	95

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

SOURCES: The DC Opportunity Scholarship Program Application and the WSF's payment files.

- Higher performing in reading but only if in high school at baseline (40 vs. 27 National Percentile Rank (NPR) points);
- Higher performing in math if in middle school at baseline (41 vs. 34 NPR points) and if in high school at baseline (49 vs. 29 NPR points);
- More likely to be entering grades K-5 (64 vs. 50 percent) and less likely to be entering grades 6-8 (31 vs. 43 percent); and
- Less likely to be students with special needs (5 vs. 10 percent).

Compared to partial users, full users were also more likely to be female and tended to have fewer siblings. Full users and partial users were statistically similar regarding a number of baseline characteristics, including percentage having applied from SINI schools, race and ethnicity, mother's average years of education and employment status, and various measures of family demographics.

Reasons for Not Participating Among the Treatment Group

Students who were offered a scholarship could decline to participate in the OSP either initially or at any point during the 3-year follow-up period observed so far. Among those who completed surveys, the subgroup of parents of treatment group students who never used their scholarships in a given year 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 not using their scholarship in a given year were:

- Lack of available space in the private school they wanted their child to attend, cited by 44 percent of nonusers in year 1, 29 percent of nonusers in year 2, and 22 percent of nonusers in year 3;
- Unable to find a participating school that offered services for their child's special needs, listed by 21 percent of nonusers in year 1, 17 percent of nonusers in year 2, and 16 percent of nonusers in year 3;
- Child was accepted into a public charter school (15 percent in year 1, 16 percent in year 2, 19 percent in year 3);
- Moved outside of the DC area, and therefore no longer eligible for the Program (4 percent in year 1, 11 percent in year 2, and 21 percent in year 3); and
- Child did not want to leave public school friends (15 percent in year 1, 7 percent in year 2, and 9 percent in year 3).

Table 2-5. Reasons Given by Parents of Treatment Students for Not Using an OSP Scholarship in Year 1, Year 2, and Year 3

Reason given by parent for child not using the offer of the scholarship	Year 1	Year 2	Year 3
There was no space at the participating private school that the child wanted to attend	43.6	29.4	22.0
The private school(s) did not have the services for the child's learning or physical disability or other special needs	21.0	17.0	15.5
Child got into a charter school	14.5	16.3	18.5
The child moved out of DC	4.0	10.5	20.8
The child did not want to leave his/her friends in public school	14.5	6.5	8.9
The private school(s) the child was interested in were too far from home or too hard to get to	8.1	9.8	10.1
The private school the child wanted to attend was not participating	12.1	7.8	6.6
Child did not pass the private school's admission test	6.5	4.6	2.4
Child did not want to be held back a grade	4.8	4.6	2.4
Child's public school has sports that the private school(s) did not	4.0	2.0	1.8
Other reasons	6.5	9.2	7.7
Total respondents	124	153	168

NOTES: Responses are unweighted. Respondents were able to select multiple responses. Categories with responses from fewer than three parents in any year are collapsed into the "Other reasons" category for confidentiality reasons. Each year column reports responses from parents who did not use an OSP scholarship for the given year.

SOURCE: Evaluation Parent Surveys.

Parents whose children initially used a scholarship but subsequently decided to leave their chosen private school also were asked during year 1, year 2, and year 3 data collection why they discontinued their scholarship use (table 2-6). Those who completed surveys cited a number of reasons for doing so. The most common responses were:

- Lack of academic support that the child needed, cited by 45 percent of parents who stopped using their scholarship and responded to data collection efforts in year 1, 54 percent in year 2, and 39 percent in year 3.
- "Child did not like the private school" was cited by 31 percent of these parents in year 1, 21 percent in year 2, and 25 percent in year 3.
- The discipline or rules at the private school being too strict was cited by 29 percent of parents in year 1, 19 percent of parents in year 2, and 7 percent of parents in year 3.
- There was another private school the child liked better, reported by 8 percent, 15 percent, and 13 percent of parents in year 1, 2, and 3, respectively.

- Work at the private school was too hard (8 percent, 6 percent, and 11 percent in year 1, year 2, and year 3, respectively), or it was too difficult to get the child to the private school each day (6 percent, 6 percent, and 11 percent in year 1, 2, and 3, respectively).

Table 2-6. Reasons Given by Parents of Treatment Group Students Who Left a Participating OSP Private School in Year 1, Year 2, and Year 3

Reason given by parent for child not staying in the private school chosen with the offer of the scholarship	Year 1	Year 2	Year 3
Child did not get the academic support he/she needed at the private school	45.1	54.2	39.3
Child did not like the private school	31.4	20.8	25.0
The discipline/rules were too strict at the private school	29.4	18.8	7.1
There was another private school the child liked better	7.8	14.6	12.5
The work at the private school was too hard	7.8	6.3	10.7
It was too hard to get the child to the private school each day	5.9	6.3	10.7
Other reasons	19.6	12.5	21.4
Total respondents	51	48	56

NOTES: Responses are unweighted. Respondents were able to select multiple responses, which generated a total of 210 responses provided by 155 parents. Categories with responses from fewer than three parents in any year are collapsed into the "Other reasons" category for confidentiality reasons.

SOURCE: Evaluation Parent Surveys.

Overall Movement Into and Out of Private and Public Schools

Where did students who declined to participate in the OSP attend school instead? Children in the treatment group who never used the OSP scholarship offered to them, or who did not use the scholarship consistently, could have remained in or transferred to a public charter school or a traditional DC public school, or enrolled in a non-OSP-participating private school. The same alternatives were available to students who applied to the OSP, were entered into the lottery, but were never offered a scholarship (the impact sample's control group); they could remain in their current DC public school (traditional or charter), enroll in a different public school, or try to find a way to attend a participating or nonparticipating private school. 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 (Hanushek, Kain, and Rivkin 2004).

The members of the impact sample were all attending DC public schools or were rising kindergarteners in the year they applied to the OSP. Of the students who were not entering kindergarten, approximately three-fourths were attending traditional DC public schools, while the remaining one-fourth were attending public charter schools. Three years after random assignment, there was substantial variation across educational sectors (table 2-7).

Table 2-7. Percentage of the Impact Sample by Type of School Attended: At Baseline and in Year 3

	Baseline		3 Years After Random Assignment		
	Public		Public		
	Traditional	Charter	Traditional	Charter	Private
Treatment	75.8	24.2	19.1	9.3	71.6
Control	73.7	26.3	53.9	33.8	12.3
Difference	2.1	-2.1	-34.7	-24.6	59.3

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 3: Treatment = 31.3, Control = 47.6. Some of the missing data rates for year 3 are a product of students naturally grading-out of the Program's eligibility requirements: 85 students are considered to have graded out during the third year, including 54 control group members and 31 treatment group members. Data are unweighted and represent actual responses. Given the rates of missing data, readers are cautioned against drawing firm conclusions.

SOURCES: Program applications and Evaluation Parent Surveys.

Based on data from survey respondents,[31] in the third year:

- Nineteen percent of the treatment group and 54 percent of the control group attended a traditional public school;
- Nine percent of the treatment group and 34 percent of the control group were enrolled in public charter schools; and
- Seventy-two percent of the treatment group and 12 percent of the control group attended a private school.

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.[32] 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 SINI schools is a special focus of this evaluation, given that Congress assigned that specific group of students to be the highest service priority

[31] The subset of survey respondents in the treatment group comprises 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 WSF's payment files, which are used to calculate the Program-wide scholarship usage rates, do not contain any information on the types of schools attended by treatment nonusers or control group members.

[32] These descriptive data regarding the types of school attended at baseline and 3 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 (62 percent). Readers are cautioned not to draw conclusions about the impact of the OSP in causing these patterns of school-sector enrollments.

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):

- Fifty-six 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).

- Three years after random assignment, the number of treatment group students reported to be attending SINI-ever schools was 15 percent, while the number of control group students in such schools was 42 percent.

Table 2-8. Percentage of the Impact Sample Attending Schools Identified as in Need of Improvement (SINI): Baseline and Year 3

	Baseline		3 Years After Random Assignment		
	SINI-Ever Schools	SINI-Never Schools	SINI-Ever Schools	SINI-Never Schools	Private
Treatment	55.7	44.3	14.6	13.8	71.6
Control	52.3	47.8	42.0	45.7	12.3
Difference	3.4	-3.4	-27.4	-31.9	59.3

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 3: Treatment = 31.3, Control = 47.6. Some of the missing data rates for year 3 are a product of students naturally grading-out of the Program's eligibility requirements: 85 students are considered to have graded out during the third year, including 54 control group members and 31 treatment group members. Data are unweighted and represent actual responses. Given the rates of missing data, readers are cautioned against drawing firm conclusions.

SOURCES: Program applications and Evaluation Parent Surveys.

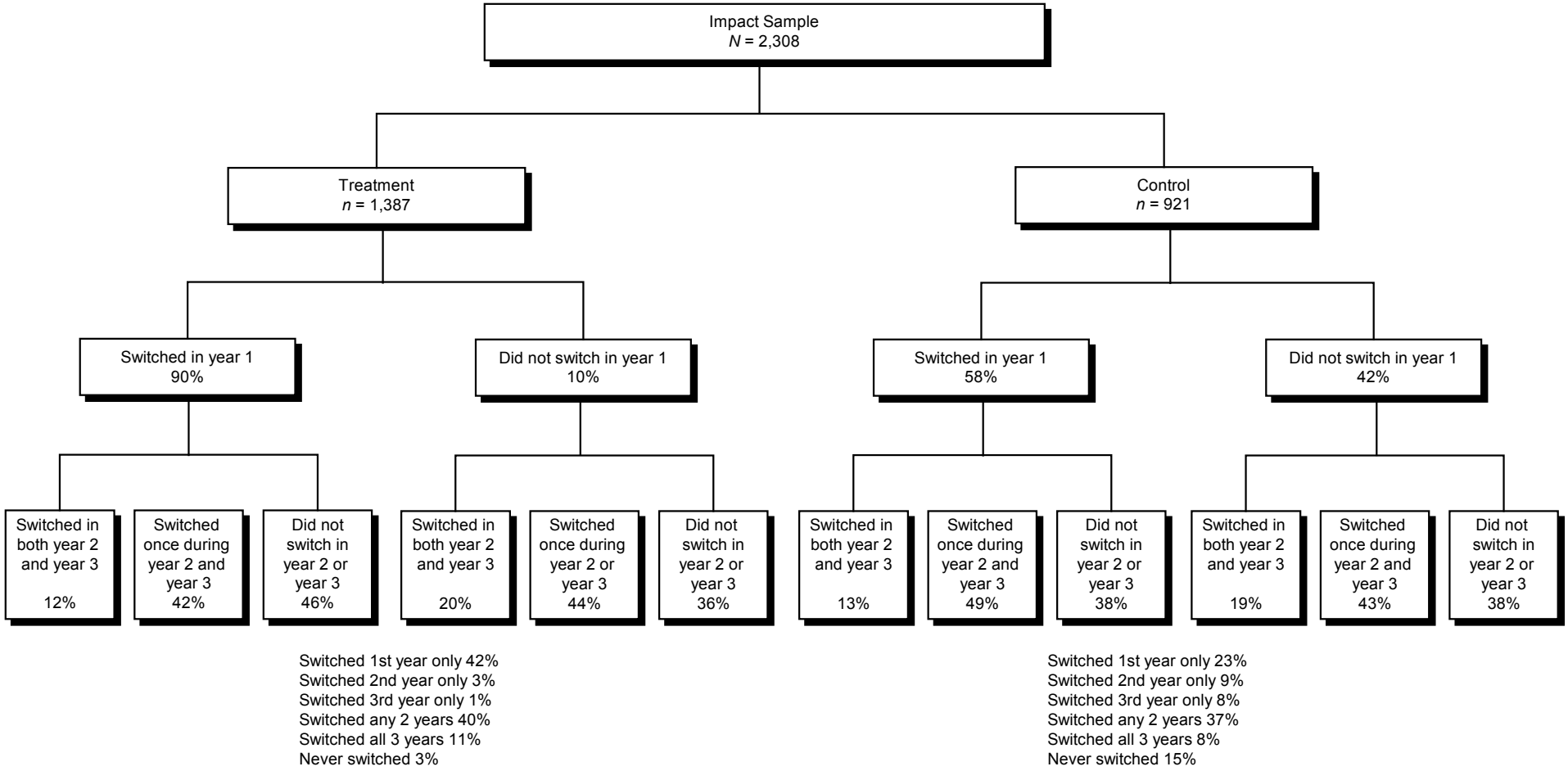
The movement of impact sample students between public (both traditional and charter) and private schools or between SINI and non-SINI schools masks some additional transitions because students can change schools within the same sector. That is, some students moved from one charter school to another charter school, or one private school to another one. In terms of general student mobility, between the time students applied to the OSP and the next year (figure 2-4):

- 90 percent of the treatment group switched schools; and

- 58 percent of the control group switched schools.[33]

[33] 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 the 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 survey. Information from those two additional sources changed the previously reported rates slightly.

Figure 2-4. Movement of the Impact Sample Between Schools During the First 3 Years



NOTES: School switching includes switching in and out of private schools and switching to different schools within the same sector, as well as the natural transitions from elementary to secondary school. Percent missing year 1: treatment = 21.6, control = 46.1; percent missing year 1 and year 2: treatment = 30.6, control = 60.8; percent missing all 3 years: treatment = 41.0, control = 68.8. 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 and third years since the random assignment:

- 54 percent of the treatment group students who switched schools during the first year switched schools at least once more, while 64 percent of the treatment group who did not switch during the first year switched during the second year, the third year, or both years; and
- 62 percent of the control group students who switched schools during the first year switched schools again during year 2, year 3, or both years, while 62 percent of the control group who did not switch during the first year switched during years 2 or 3 or both.

Over the course of all 3 years since random assignment to the treatment or control group:

- Among the treatment group, 3 percent remained in the same school they were in when they applied to the Program, 46 percent switched schools once, 40 percent switched twice, and 11 percent switched three times; and
- Among the control group, 15 percent remained in the same school they were in when they applied to the Program, 40 percent switched schools once, 37 percent switched schools twice, and 8 percent switched three times.

Both groups experienced higher rates of school mobility than the typical annual rate for urban students. The treatment group students switched schools at an annual rate of 53 percent and the control group at an annual rate of 46 percent from the baseline year to year 3.[34] In contrast, other studies of urban school populations report annual school-switching rates of 22 to 28 percent.[35] The higher school-switching rate of the treatment group compared to the control group of 7 percentiles annually (21 percentiles cumulatively over 3 years) is itself statistically significant.[36] The treatment group students have switched schools more frequently than the control group students, but both groups have switched schools more often than is the norm for inner-city K-12 students.

[34] Annual school-switching rates were calculated by totaling the number of switches since baseline (2,205 for the treatment and 1,271 for the control students), dividing by 3 years to generate the average annual number of switches, and further dividing by the number of students in each group to generate the average annual rate.

[35] See Witte 2000, p. 144; Wong et al. 1997, p. 17.

[36] 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=2.15, p=.03).

3. Impacts on Key Outcomes, 3 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 3 years after families and students applied to the OSP, or approximately 30 months after the start of their first school year in the Program. The first section summarizes the analytic methods used to determine the results and the techniques used to display them. The second section provides a review of the impacts 1 year and 2 years after random assignment, as reported previously (Wolf et al. 2007, Wolf et al. 2008). Section 3 presents the year 3 impacts on student achievement. The fourth section discusses the year 3 safety impacts. Section 5 presents the year 3 satisfaction impacts. The sixth section provides a brief summary of the chapter findings.

3.1 Analytic and Presentation Approaches

For each key year 3 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) takes the impacts from the ITT analysis but adjusts them by the rate of scholarship nonuse, effectively re-scaling the ITT impacts across only the treatment students who actually used their scholarship. Appendices A.8 and A.9 provides a more detailed description of the analytic methods used for both types of analyses. For information about the relationship between attending private school, with or without an Opportunity Scholarship, and outcomes see appendix E.

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; includes never users, partial users, and full users as members of the treatment group. We compare the outcomes of students randomly assigned to receive the offer of a scholarship (treatment group) with the outcomes of students randomly assigned to not receive the offer (control group). The difference in outcomes is the impact of being offered a scholarship.
• What is the impact of using an OSP scholarship to attend a participating private school?	"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 "never users") and (2) the hypothesized 1.7 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 student subgroups of policy interest also is discussed. The participant subgroups that are analyzed in this study were designated prior to the collection and analysis of Program impacts, with the designation based on their use in previous evaluations of scholarship programs or importance to contemporary policy discussions about educational improvement. They are:

- **Whether students attended a SINI school prior to application to the Program.** The Program statute designates such students as the highest service priority for the OSP, making the question of whether or not 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; Summers 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 finds that the elementary and high school education experiences differ in significant ways (e.g., Torgesen 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 application cohort 1 (applied in 2004) or application 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.8:[37]

- The text and tables include effect sizes (ES) to translate each impact into a standard metric and allow the reader to assess whether the size of the impact might be considered meaningful, whether or not it is statistically significant.[38]

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

- Whenever the results from different subgroup pairs (e.g., males and females) are presented in tables, the difference between the average impact on each subgroup is also presented, in a row labeled "Difference," and the difference is subjected to the same statistical test for significance (i.e., T-Test) as the impacts on the subgroups themselves.

- 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" (Schochet 2007, p. 5; Benjamini and Hochberg 1995)

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

[38] Specifically, the effect sizes are computed as a percentage of a standard deviation for the control group after 3 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.2) forecasts that this year 3 evaluation will contain sufficient data to correctly identify an overall reading or math impact of the offer of a scholarship of .12 standard deviations or higher if such an impact actually exists. Subgroup ITT impacts are estimated to be detectable at various sizes, ranging from .14 to .38 standard deviations.

(Appendix B).[39] Throughout this report, the phrases "appears to have an impact" and "may have had an impact" are used to caution readers regarding statistically significant impacts that may have been false discoveries.

- The impact results from the primary analysis were subject to sensitivity tests involving a sample trimmed to exactly equalize the treatment and control response rates at the level before response conversion efforts were applied to the control group (trimmed sample) and the clustering of student observations on the school attended instead of family (clustering on current school). 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.

3.2 Impacts Reported Previously

The evaluation of the impact of the OSP is a longitudinal study, in that it tracks the outcomes of students over multiple years of their potential participation in the scholarship program.[40] Two earlier reports described impacts 1 and 2 years after students applied to the OSP and were randomly assigned by lottery to either the treatment or control group.[41] The results from those analyses indicated:

- In neither year were there statistically significant impacts on academic achievement overall or for students from SINI schools, the key student subgroup defined in the law.[42]

- In both year 1 and year 2, initial estimates suggested there were positive achievement impacts for several other subgroups. In year 1, the subgroup impacts were in math achievement, while in year 2 the impacts were in reading achievement.[43] In both years, those impact estimates lost their statistical significance when adjustments for multiple comparisons were made, and thus the subgroup findings may have been "false discoveries."

---

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

[40] Each year of analysis is cumulative: for example, year 2 impacts were estimated based on outcomes at the end of the second year students could be in the Program, incorporating any outcomes achieved after 1 year as well as after 2 years.

[41] See Wolf et al. 2007, table ES-2 and Wolf et al. 2008, xxii-xxiv.

[42] Overall: in year 1, ES = .03 in reading and .08 in math; in year 2, ES = .09 in reading and .01 in math. Students from SINI schools: in year 1, ES = -.01 in reading and .01 in math; in year 2, ES = -.00 in reading and .05 in math.

[43] Students from non-SINI schools: ES = .12 in math in year 1, ES = .15 in reading in year 2; students who were higher performing when they applied to the OSP: ES = .12 in math in year 1, ES = .15 in reading in year 2; and students from cohort 1: ES = .27 in reading in year 2.

- There were no impacts in year 1 or year 2 on either reading or math achievement for students in other subgroups, including those who were lower performing at application, males or females separately, students entering grades K-8 or high school, or students who applied to the Program the second year (cohort 2).

- In both years, the Program had a positive impact on parents' satisfaction with their child's school and parents' perceptions of school safety.[44]

- In year 1 and year 2, students with OSP scholarships did not report being more satisfied with school or feeling safer than those without access to scholarships; in year 2 the Program seemed to have a negative impact on school satisfaction for students with lower academic performance when they applied to the OSP,[45] though adjustments for multiple comparisons suggest that could have been a false discovery.

- This same pattern of findings holds for the impact of *using* a scholarship as well as being *offered* a scholarship.

These were the results of the analysis of data collected 1 and 2 years after random assignment. The results presented in the remainder of this report are based on data collected 3 years after random assignment and about 30 months into any new educational experiences that may have been induced by the scholarship offer.

3.3 Year 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:

- Positive and statistically significant impacts of the Program on overall student achievement in reading after 3 years.

- No significant impacts on overall student achievement in math after 3 years.

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

---

[44] Parents' satisfaction with their child's school (as measured by giving the school a grade of A or B): ES = .38 in year 1 and ES = .26 in year 2. Parent's perceptions of the safety of their child's school: ES = -.22 in year 1 and ES = -.27 in year 2.

[45] Students' satisfaction with their school (as measured by giving the school a grade of A or B): ES = -.03 in year 1 and ES = .05 in year 2. Students' reports of school safety: ES = -.11 in year 1 and ES = -.01 in year 2.

- Positive programmatic impacts in reading achievement for 5 of the 10 subgroups examined: participants who applied from non-SINI schools, those who applied to the Program with relatively higher levels of academic performance, female students, students entering grades K-8 at the time of application, and students from the first cohort of applicants. However, the positive subgroup reading impacts for female students and the first cohort of applicants should be interpreted with caution, as reliability tests for multiple comparison adjustments suggest that they could be false discoveries.

- No statistically significant test score differences in math between the treatment and control groups for any of the 10 subgroups of students.

Impacts for the Full Sample of Students

The primary analysis indicates there was a statistically significant overall impact of the Program on reading achievement after 3 years (ES = .13, p-value = .01). That is, the average reading test scores of the treatment group as a whole were significantly higher than those of the control group as a whole in the third year (table 3-2).[46] The impact of the offer of a scholarship was 4.46 scale score points, and the impact of the use of a scholarship was 5.27 scale score points on reading achievement. The magnitude and statistical significance of the overall impact on reading achievement is consistent in the sensitivity tests conducted (appendix C, table C-1).

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

	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Reading	635.44	630.98	4.46*	.13	5.27*	.15	.01
Math	630.15	629.35	.81	.03	.95	.03	.62

*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 in terms of standard deviations. Valid *N* for reading = 1,460; math = 1,468. Separate reading and math sample weights used.

The primary analysis found no statistically significant general impact of the Program on math achievement after 3 years. The two sensitivity tests confirmed this finding of no significant math effects (appendix C, table C-1).

[46] 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.

These findings can be viewed most clearly in figures 3-1 and 3-2. The 95 percent confidence interval for the regression-adjusted difference of 4.46 scale score points between the treatment and control group in reading outcomes ranges from a positive .89 to a positive 8.03.[47] Because both ends of the confidence interval are greater than the value for a zero impact (the horizontal axis), we can be confident that the overall reading impact is positive. In contrast, while the estimate of the treatment impact on math scale scores is a positive gain of .81 points, the confidence interval ranges from a positive 4.03 to a negative 2.42. Therefore, we are uncertain if the general math impact is positive, zero, or negative.

These year 3 achievement impacts can be placed in the context of impacts estimated in the prior years (figure 3-3). The reading impact was 1.03 (nonsignificant) in year 1, 3.17 (nonsignificant) in year 2, and 4.46 (significant) in year 3.[48] The math impact estimates were 2.74 (nonsignificant) in year 1, 0.23 (nonsignificant) in year 2, and 0.81 (nonsignificant) in year 3.

Subgroup Impacts

The statistical significance of impacts for particular subgroups of students in year 3 are consistent with those for students overall in math but not in reading. There were no impacts on math achievement for any of the 10 subgroups examined, as was true for the full impact sample. The offer of a scholarship, and therefore the use of a scholarship, had a statistically significant positive impact on reading achievement in the third year for half of the student subgroups, including at least two subgroups who applied with a relative advantage in academic preparation (table 3-3). The subgroups with positive reading impacts include:[49]

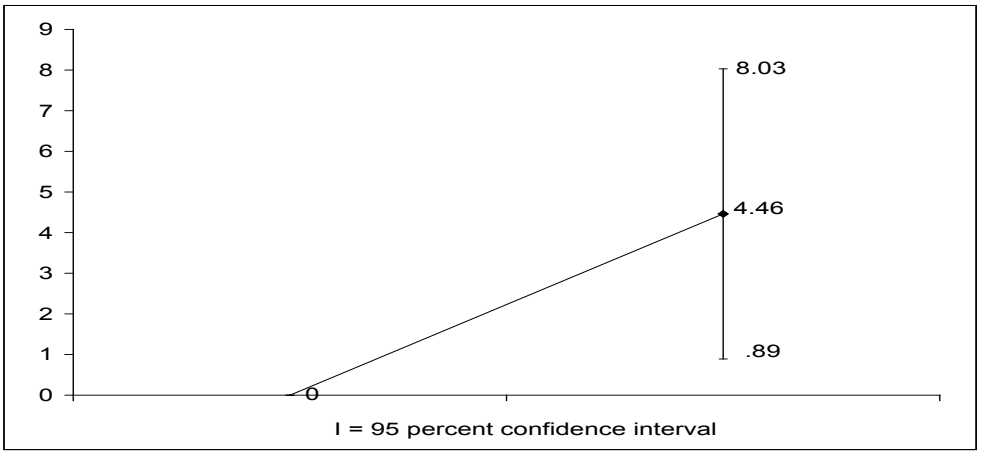
- Students in the treatment group who had attended non-SINI public schools prior to the Program, who scored an average of 6.6 scale score points higher in reading (ES = .19) than students in the control group from non-SINI schools (the impact of the offer of a scholarship); the calculated impact of using a scholarship was 7.7 scale score points (ES = .22).

[47] The scale score mean and standard deviation (SD) for the SAT-9 norming population varies by grade and is 463.8 (SD = 38.5) for kindergarteners tested in the spring, compared to 652.1 (SD = 39.1) for fifth graders and 703.6 (SD = 36.5) for students in 12th grade.

[48] It is not known whether the three annual impact estimates in reading or math are significantly different from each other. Since the estimates are not derived from independent samples (i.e., most of the same students tested in all 3 years), the covariance of the estimates across the years would need to be known in order to determine if the three estimates generate a statistically significant trend over time in achievement. The research team plans to pool the annual analysis data across years and estimate the magnitude and statistical significance of any achievement trends (or "slopes") for the next impact report, when a fourth annual estimate will be available to inform those calculations.

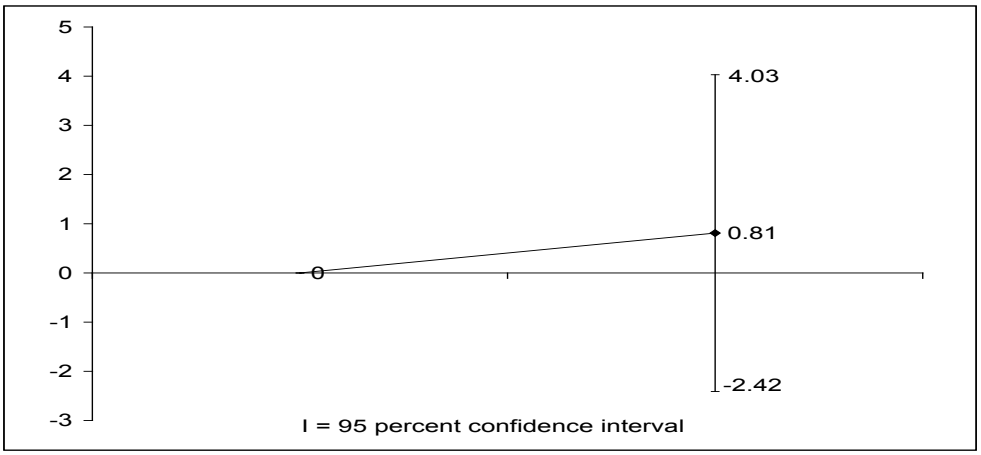
[49] Each of these findings refers to one subgroup and not to the paired subgroup categories, so that a significant finding refers to a treatment vs. control difference, not a difference between, for example, males and females.

Figure 3-1. Regression-Adjusted Impact and Confidence Interval in Year 3: Reading



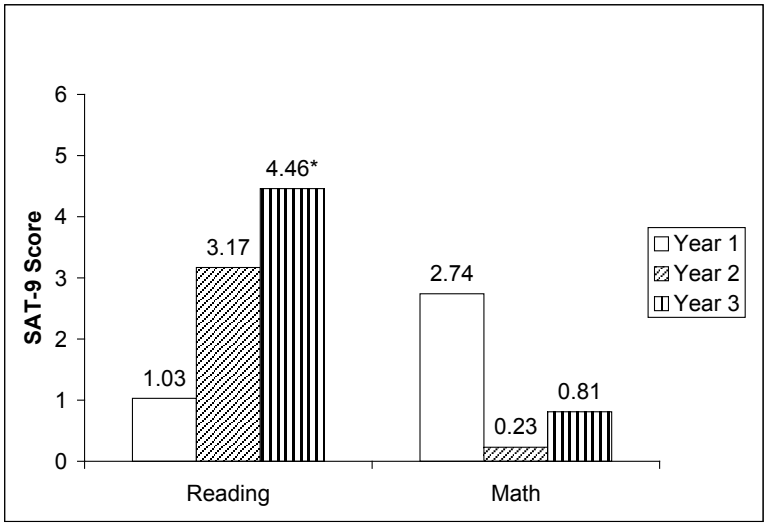
NOTES: Valid *N* for reading = 1,460. The point on the vertical line (4.46) is the statistical estimate of the Program impact on reading gains in terms of scale score points. The high and low bounds of the vertical line illustrate the 95 percent confidence level associated with the estimate.

Figure 3-2. Regression-Adjusted Impact and Confidence Interval in Year 3: Math



NOTES: Valid *N* for math = 1,468. The point on the vertical line (.81) 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.

Figure 3-3. Impact of OSP on Reading and Math Achievement Overall, in Years 1 Through 3



*The individual year's impact is statistically significant at the 95 percent confidence level.

NOTES: Displayed numbers are regression-adjusted mean impacts using a consistent set of baseline covariates. Impacts are displayed in terms of scale scores. Valid *N* for reading = 1,649 in year 1; 1,580 in year 2; and 1,460 in year 3. Valid *N* for math = 1,715 in year 1; 1,585 in year 2; and 1,468 in year 3. Separate reading and math sample weights were used each year.

- Students in the treatment group who entered the Program in the higher two-thirds of the applicant test-score performance distribution—averaging a 43 National Percentile Rank in reading at baseline—who scored an average of 5.5 scale score points higher in reading (ES = .17) than students in the control group who applied to the OSP in the higher two-thirds of the test-score distribution; the impact of using a scholarship for this group was 6.2 scale score points (ES = .19).

- Female students in the treatment group, who scored an average of 5.1 scale score points higher in reading (ES = .15) than females in the control group; the impact of using a scholarship was 5.8 scale score points (ES = .17).

- Students in the treatment group who entered the Program in grades K-8, who scored an average of 5.2 scale score points higher in reading (ES = .15) than students in the control group who applied to the OSP entering grades K-8; the impact of using a scholarship was 6.0 points (ES = .17).

- Students in the treatment group from the first cohort of applicants, who scored an average of 8.7 scale score points higher in reading (ES = .31) than students in the control group from cohort 1; the impact of using a scholarship was 11.7 scale score points (ES = .42).

Table 3-3. Year 3 Impact Estimates of the Offer and Use of a Scholarship on Subgroups: Academic Achievement

	Reading						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement: Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
SINI ever	649.77	648.25	1.52	.05	1.81	.06	.59
SINI never	625.29	618.72	6.57**	.19	7.72**	.22	.01
Difference	24.48	29.53	-5.05	-.15			.17
Lower performance	614.48	612.38	2.10	.07	2.68	.10	.47
Higher performance	644.74	639.29	5.45*	.17	6.21*	.19	.02
Difference	-30.26	-26.90	-3.35	-.10			.35
Male	631.32	627.48	3.83	.11	4.67	.14	.15
Female	639.30	634.24	5.07*	.15	5.81*	.17	.04
Difference	-7.99	-6.75	-1.23	-.04			.73
K-8	627.30	622.07	5.23**	.15	6.04**	.17	.01
9-12	682.41	682.50	-.10	-.00	-.14	.04	.98
Difference	-55.11	-60.43	5.33	.15			.21
Cohort 2	625.64	622.27	3.37	.09	3.87	.11	.09
Cohort 1	672.87	664.17	8.70*	.31	11.67*	.42	.04
Difference	-47.23	-41.90	-5.34	-.15			.25

	Math						
	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Student Achievement: Subgroups	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
SINI ever	646.73	646.56	.17	.01	.20	.01	.95
SINI never	618.39	617.12	1.27	.04	1.49	.04	.58
Difference	28.34	29.44	-1.10	-.03			.75
Lower performance	615.43	615.08	.35	.01	.45	.02	.90
Higher performance	636.64	635.65	.98	.03	1.12	.04	.64
Difference	-21.20	-20.57	-.63	-.02			.86
Male	629.35	629.31	.04	.00	4.67	.00	.99
Female	630.92	629.38	1.54	.05	1.76	.06	.50
Difference	-1.56	-.07	-1.50	-.05			.65
K-8	621.74	620.73	1.01	.03	1.16	.03	.57
9-12	678.77	679.18	-.41	-.02	-.60	-.03	.92
Difference	-57.04	-58.45	1.42	.04			.74
Cohort 2	619.32	619.53	-.21	-.01	-.24	-.01	.91
Cohort 1	671.48	666.74	4.74	.23	6.37	.31	.19
Difference	-52.16	-47.21	-4.95	-.16			.23

*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. Impacts are displayed in terms of scale scores. Effect sizes are in terms of standard deviations. Total valid *N* for Reading = 1,460, including: SINI ever *N* = 610, SINI never *N* = 850, Lower performance *N* = 465, Higher performance *N* = 995, Male *N* = 714, Female *N* = 746, K-8 *N* = 1,287, 9-12 *N* = 173, Cohort 2 *N* = 1,186, Cohort 1 *N* = 274. Total Valid N for Math = 1,468, including SINI ever *N* = 613, SINI never *N* = 855, Lower performance *N* = 469, Higher performance *N* = 999, Male *N* = 717, Female *N* = 751, K-8 *N* = 1,295, 9-12 *N* = 173, Cohort 2 *N* = 1,195, Cohort 1 *N* = 273.

There was no statistically significant subgroup impact in reading for students who applied from a school designated SINI between 2003 and 2005—the highest service priority for the Program according to the statute. The analysis also did not show subgroup impacts for students who entered the program in the lower one-third of the applicant test-score performance distribution, male students, students who entered the Program from high school, and cohort 2 students.[50]

It is useful to place the estimated effect sizes for these overall and subgroup impacts in context (table 3-4). The overall reading impact of the scholarship offer (ITT) of 4.5 scale score points is equivalent to 3.1 months of additional learning for members of the treatment group. The overall impact of the actual use of a scholarship (IOT) of 5.3 scale score points is equivalent to 3.7 additional months of learning. The SINI-never impacts on reading are equivalent to 4.1 additional months of learning for the offer and 4.9 additional months of learning for the use of a scholarship. With the exception of cohort 1, the positive reading achievement impacts on the other subgroups ranged from 3 to 5 additional months of learning, or one-third to one-half of a typical 9-month school year. The reading impacts for cohort 1 are equivalent to 1.5 or 2 years of extra learning (14 to 19 months).

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

Student Achievement: Reading	Impact of the Scholarship Offer (ITT)		Impact of Scholarship Use (IOT)	
	Effect Size	Months of Schooling	Effect Size	Months of Schooling
Full sample	.13	3.11	.15	3.68
SINI never	.19	4.13	.22	4.86
Higher performance	.17	4.00	.19	4.56
Female	.15	3.11	.17	3.56
K-8	.15	2.87	.17	3.32
Cohort 1	.31	14.06	.42	18.86

NOTES: Scale score impacts were converted to approximate months of learning first by dividing the impact effect size by the effect size of the weighted (by grade) average annual increase in reading scale scores for the control group. The result was the proportion of a typical year of achievement gain represented by the programmatic impact. That number was further divided by nine to convert the magnitude of the gain to months, since the official school year in the District of Columbia comprises 9 months of instruction.

The five statistically significant subgroup impacts of the OSP on reading test scores in year 3 were the product of an analysis involving multiple comparisons of treatment and control group members. Under such conditions, statistically significant findings can emerge by chance. Statistical adjustments to

[50] Since the analysis of impacts on subgroups draws on smaller samples of students, such analyses inevitably have less power to detect statistically significant impacts than do analyses on the entire impact sample (Appendix A, tables A-1 and A-2). The five subgroup impacts on reading discussed here were detectable, even with lower power, because they were larger in magnitude than the overall reading impact.

account for the multiple comparisons suggest that two of the five significant subgroup achievement impacts in reading—the impact on female students and the impact on cohort 1 students—may be false discoveries and therefore should be interpreted with caution (see appendix B, table B-1).

As with the achievement impacts for the overall sample, the statistically significant impacts on reading for the SINI-never, higher baseline performing, female, K-8, and cohort 1 subgroups of students were subjected to sensitivity tests. All five impact estimates were larger and retained statistical significance when drawn from the trimmed sample; the estimates for females and cohort 1 students were no longer statistically significant when the statistical model was modified to control for similarities among students who attended the same school as opposed to controlling for similarities among students who are from the same family (appendix C, table C-1).

3.4 Impacts on Reported Safety and an Orderly School Climate

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

There are no specific tests to evaluate the safety of a school as there are for evaluating student achievement. There are various indicators of the relative orderliness of the school environment, such as the presence or absence of property destruction, cheating, bullying, and drug distribution to name a few (see appendix A.3 for more information). Students and parents can be surveyed regarding the extent to which such indicators of disorder are or are not a problem at their or their child's school. The responses then can be consolidated into an index of safety and an orderly school climate and analyzed, as we do here.

In summary, the analysis suggests that:

- Overall, treatment group parents rated their child's school significantly higher regarding safety and an orderly climate than did control group parents (figure 3-4 and table 3-5).

- Parents of students who applied to the Program from SINI schools reported significantly higher perceptions of school safety if their child had been awarded a scholarship (table 3-5).

- Treatment group parents of non-SINI students, higher and lower baseline performers, males, females, students in grades K-8 and 9-12, and cohort 1 and cohort 2 students all produced ratings of safety and an orderly school climate that were significantly higher than their counterpart parents in the control group (table 3-5). Additional reliability tests indicated that none of the subgroup impacts are likely to be a false discovery.

- Treatment and control group students, on the other hand, did not report differences on the evaluation's measure of safety and an orderly school climate (figure 3-4 and table 3-6).

- The school safety impacts estimated for both parents and students, in the overall sample and for all subgroups, were confirmed by the sensitivity tests in all but one case (cohort 1 parents' views of school safety).

Parent Self-Reports

Overall, the parents of students offered an Opportunity Scholarship in the lottery subsequently reported their child's school to be safer and more orderly than did the parents of students in the control group. The impact of the offer of a scholarship on parental perceptions of safety and an orderly school climate was 1.01 on a 10-point index of indicators of school safety and orderliness, an effect size of 0.29 standard deviations (table 3-5; figure 3-4 for a visual display). The impact of using a scholarship was 1.20 on the index, with an effect size of .34 standard deviations. These findings persisted through the sensitivity tests: that is, the statistical significance of the findings did not change as a result of the different models (see appendix C, table C-2).

This impact of the offer of a scholarship on parental perceptions of safety and an orderly school climate was consistent across all subgroups of students examined, including parents of students from SINI (ES = .32) and non-SINI schools (ES = .27), parents of students who entered the program with relatively higher (ES = .29) and lower (ES = .28) levels of academic achievement, parents of male (ES = .30) and female students (ES = .28), parents of students in grades K-8 (ES = .27) and 9-12 (ES = .40), and parents of both cohort 1 (ES = .31) and cohort 2 (ES = .29) (table 3-5). 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). The overall impact of the Program, as well as the impacts for 9 of the 10 subgroups, was robust to both sensitivity tests. The exception was the impact of the Program on parents of cohort 1 students, which lost statistical significance in the sensitivity test using the trimmed sample (appendix C, table C-2).

Table 3-5. Year 3 Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Parent Perceptions of Safety and an Orderly School Climate

	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		
Safety and an Orderly School Climate: Parents	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	***p*-value of estimates**
Full sample	**8.08**	**7.07**	**1.01****	**.29**	**1.20****	**.34**	**.00**
SINI ever	7.91	6.74	1.16**	.32	1.38**	.38	.00
SINI never	8.20	7.30	.90**	.27	1.06**	.32	.00
Difference	-.30	-.56	.26	.07			.53
Lower performance	7.81	6.80	1.02**	.28	1.30**	.36	.01
Higher performance	8.20	7.19	1.01**	.29	1.15**	.33	.00
Difference	-.38	-.39	.01	.00			.99
Male	8.12	7.06	1.06**	.30	1.29**	.37	.00
Female	8.04	7.08	.96**	.28	1.10**	.32	.00
Difference	.08	-.02	.10	.03			.78
K-8	8.22	7.29	.93**	.27	1.08**	.32	.00
9-12	7.31	5.80	1.51*	.40	2.15*	.56	.02
Difference	.91	1.49	-.58	-.17			.37
Cohort 2	8.29	7.33	.97**	.29	1.11**	.33	.00
Cohort 1	7.28	6.08	1.20*	.31	1.61*	.42	.03
Difference	1.02	1.25	-.23	-.07			.70

*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 in terms of standard deviations. Valid N = 1,423, including: SINI ever N = 592, SINI never N = 831, Lower performance N = 455, Higher performance N = 968, Male N = 699, Female N = 724, K-8 N = 1,265, 9-12 N = 158, Cohort 2 N = 1,152, Cohort 1 N = 271. Parent survey weights were used.

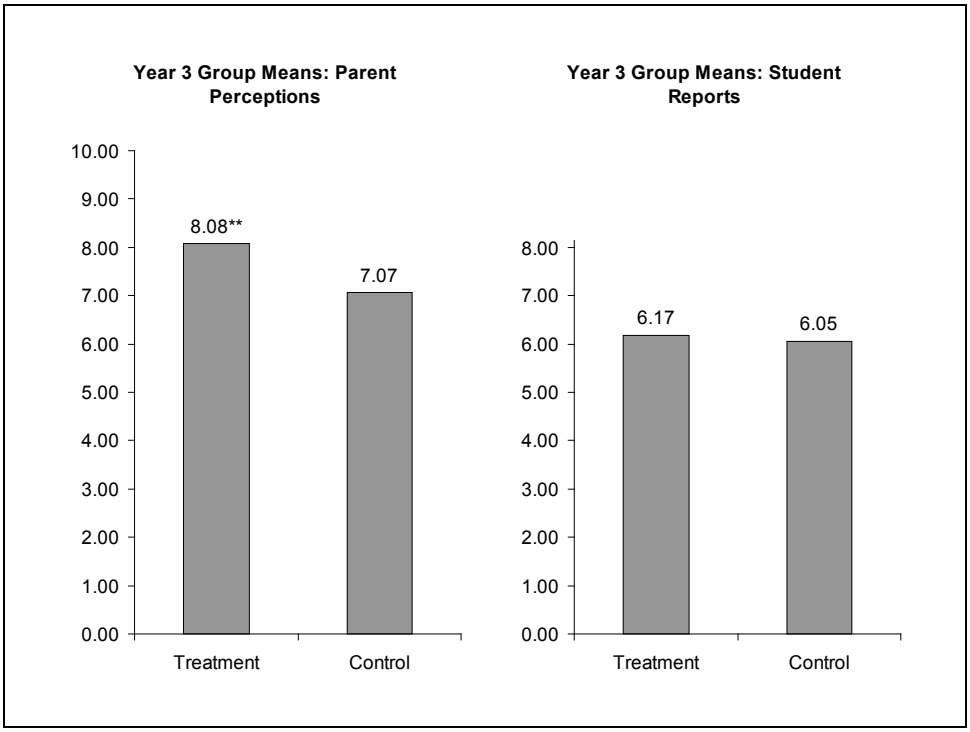
Because the impacts of the scholarship offer on perceptions of safety and an orderly school climate were statistically significant for these subgroups of parents, the programmatic impacts on actual scholarship users were also statistically significant. For example, the impact of using a scholarship on parental perceptions of school safety for these affected subgroups ranged from 1.06 for parents whose students were in SINI-never schools to 2.15 for parents of students entering grades 9-12 at baseline, which equates to subgroup effect sizes ranging from .32 to .56 standard deviations (table 3-5).

Student Self-Reports

The students in grades 4-12 who completed surveys paint a different picture about school safety at their school than do their parents. The student index of school climate and safety asked students if they personally had been a victim of theft, drug-dealing, assaults, threats, bullying, or taunting or had observed weapons at school. On average, reports of school climate and safety by students offered scholarships through the lottery were not statistically different from those of the control group (table 3-6;

figure 3-4 for a visual display). That is, there was no evidence of an impact from the offer of a scholarship or the use of a scholarship on students' reports. No statistically significant findings were evident across the subgroups analyzed. Nor did the sensitivity tests conducted lead to a different set of overall findings (see appendix C, table C-3).

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



**Statistically significant at the 99 percent confidence level.

NOTES: Parent perceptions are based on a ten-point scale; student reports are based on an eight-point scale. For parent perceptions, valid N = 1,423; parent survey weights were used; the ten-point index of indicators of school safety and an orderly environment includes the absence of property destruction, tardiness, truancy, fighting, cheating, racial conflict, weapons, drug distribution, drug/alcohol use, and teacher absenteeism. For student reports, valid N = 1,098; student survey weights were used; the survey was given to students in grades 4-12; the means represent the absence of incidents on an eight-item index for student reports of students being a victim of theft, drug-dealing, assaults, threats, bullying or taunting, or had observed weapons at school. Means are regression adjusted using a consistent set of baseline covariates.

Table 3-6. Year 3 Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Student Reports of Safety and an Orderly School Climate

Safety and an Orderly School Climate: Students	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		***p*-value of estimates**
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**6.17**	**6.05**	**.12**	**.06**	**.14**	**.07**	**.36**
SINI ever	6.06	5.98	.08	.04	.10	.05	.66
SINI never	6.24	6.10	.14	.08	.17	.10	.41
Difference	-.18	-.12	-.06	-.03			.83
Lower performance	6.04	5.89	.14	.07	.19	.09	.55
Higher performance	6.22	6.12	.10	.06	.12	.07	.50
Difference	-.19	-.22	.04	.02			.90
Male	5.96	5.86	.10	.05	.12	.06	.59
Female	6.37	6.23	.13	.07	.16	.09	.44
Difference	-.41	-.37	-.04	-.02			.88
4-8	6.12	6.00	.12	.06	.14	.07	.39
9-12	6.42	6.31	.11	.06	.15	.08	.74
Difference	-.30	.31	.01	.01			.97
Cohort 2	6.16	6.01	.16	.08	.18	.09	.29
Cohort 1	6.16	6.21	-.04	-.02	-.06	-.03	.88
Difference	.00	-.20	.20	.10			.54

NOTES: Means are regression adjusted using a consistent set of baseline covariates. Effect sizes are in terms of standard deviations. Valid N = 1,098, including: SINI ever N = 552, SINI never N = 546, Lower performance N = 347, Higher performance N = 751, Male N = 542, Female N = 556, K-8 N = 928, 9-12 N = 170, Cohort 2 N = 823, Cohort 1 N = 275. Student survey weights were used. Survey given to students in grades 4-12.

3.5 Impacts on School Satisfaction

Economists have long used customer satisfaction as a proxy measure for product or service quality (see Johnson and Fornell 1991). While not specifically identified as an outcome to be studied, it is an indicator of the "success of the Program in expanding options for parents," which Congress asked the evaluation to consider.[51] 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).

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

Satisfaction of both parents and students was measured by the percentage that assigned a grade of A or B to their child's or their school.[52] In summary, the analysis suggests that in year 3:

- Treatment group parents overall were more likely to assign their child's school a high grade than were control group parents (figure 3-5 and table 3-7).
- Parents of students who applied to the Program from SINI schools were not significantly more likely to grade their child's school highly if their child had been awarded an Opportunity Scholarship, nor were the lower performing at baseline or grade 9-12 subgroups of parents (table 3-7).
- The remaining seven subgroups of parents were significantly more likely to report a high grade for their child's school if they were in the treatment group (table 3-7). Statistical adjustments for multiple comparisons indicated that none of these parent satisfaction subgroup impact estimates are at risk of being false discoveries (see appendix B, table B-3).
- There were no treatment impacts overall, or for any subgroup, on student satisfaction with school (figure 3-5 and table 3-8).
- The school satisfaction impacts estimated for both parents and students, in the overall sample and for subgroups, were confirmed by the sensitivity tests in all but one case (satisfaction of parents of female students).

Parent Self-Reports

About 30 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. A total of 74 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 11 percentage points (impact of the offer of a scholarship) (table 3-7; figure 3-5 for a visual display); the impact of using a scholarship was a difference of 12 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 .22 and .26, respectively (table 3-7).

There were also statistically significant positive impacts for 7 of 10 subgroups, including parents of students from non-SINI schools, and students who had higher test-score performance at baseline, were male or female, were entering grades K-8, or were in cohort 1 or cohort 2. Parents of these students were significantly more likely to give their child's school a grade of A or B if they were in the treatment group. The effect sizes ranged from .16 to .41 standard deviations for the offer of, and from .19 to .55 standard deviations for the use of, a scholarship.

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

Table 3-7. Year 3 Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Parent Reports of Satisfaction with Their Child's School

Parents Who Gave Their School a Grade of A or B	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (ITT)		*p*-value of estimates
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**.74**	**.63**	**.11****	**.22**	**.12****	**.26**	**.00**
SINI ever	.69	.63	.06	.13	.07	.15	.16
SINI never	.78	.64	.14**	.29	.17**	.35	.00
Difference	-.09	-.01	-.09	-.18			.17
Lower performance	.63	.57	.06	.12	.08	.16	.21
Higher performance	.79	.67	.13**	.27	.15**	.31	.00
Difference	-.17	-.10	-.07	-.15			.26
Male	.71	.60	.11**	.22	.13**	.27	.01
Female	.77	.67	.10**	.22	.12**	.25	.01
Difference	-.07	-.07	.01	.01			.93
K-8	.77	.64	.13**	.27	.15**	.31	.00
9-12	.59	.60	-.01	-.03	-.02	-.04	.88
Difference	.18	.04	.14	.28			.11
Cohort 2	.75	.68	.08*	.16	.09*	.19	.02
Cohort 1	.68	.48	.20**	.41	.27**	.55	.00
Difference	.07	.20	-.13	-.27			.06

*Statistically significant at the 95 percent confidence level.

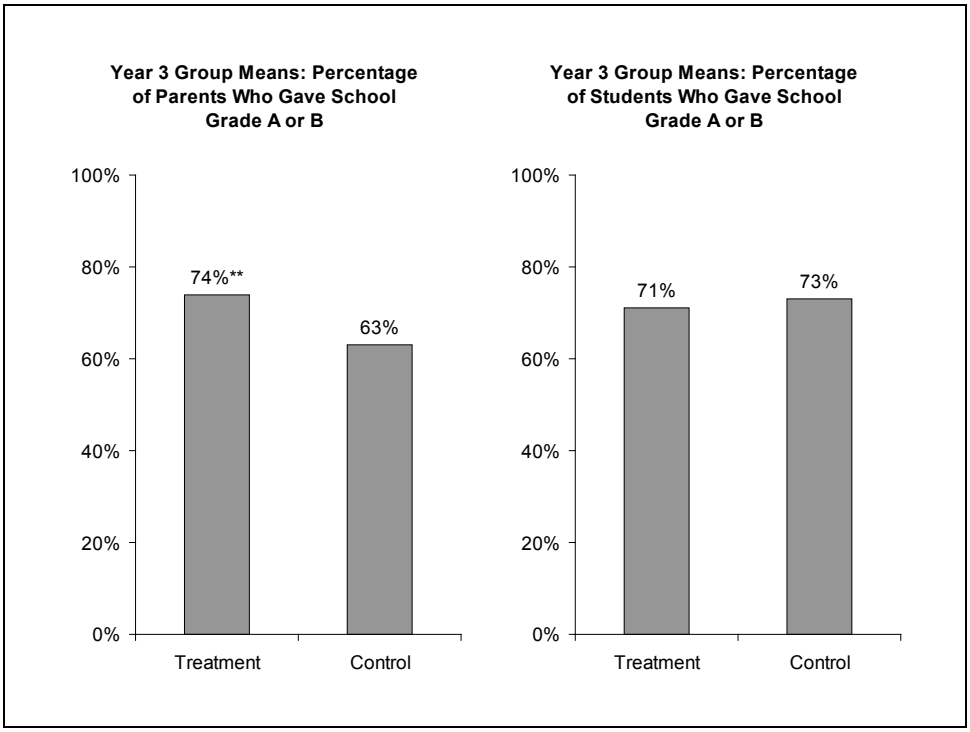
**Statistically significant at the 99 percent confidence level.

NOTES: Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid N = 1,410, including: SINI ever N = 586, SINI never N = 824, Lower performance N = 451, Higher performance N = 959, Male N = 694, Female N = 716, K-8 N = 1,254, 9-12 N = 156, Cohort 2 N = 1,141, Cohort 1 N = 269. Parent survey weights were used.

Three groups of parents were no more satisfied with their child's school if they had been offered a scholarship. Parents of students who entered the program from SINI schools, with lower levels of academic performance at baseline, or in grades 9-12 were just as likely to grade their child's school A or B if they were in the treatment as the control group. As described in section 3.3, those subgroups of students did not experience the positive impacts on reading achievement that many other students gained from the Program.

All seven of the parent satisfaction subgroup impacts that initially were statistically significant remained significant after adjustments for the fact that they were the product of multiple comparisons (see appendix B, table B-3). Sensitivity tests confirmed the parent satisfaction results from the primary analysis with one exception. The positive impact of the Program on school satisfaction for parents of female students lost statistical significance in the estimation using the trimmed sample (appendix C, table C-4).

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



**Statistically significant at the 99 percent confidence level.

NOTES: For parent reports, valid N = 1,410; parent survey weights were used. For student reports, valid N = 1,014; student survey weights were used; the survey was given to students in grades 4-12. Means are regression adjusted using a consistent set of baseline covariates.

Student Self-Reports

As was true with the school safety and climate measures, students had a different view of their schools than did their parents. Three years after random assignment, there were no significant differences between the treatment group and the control group in their likelihood of assigning their schools a grade of A or B (table 3-8; figure 3-5 for a visual display).[53] Student reports of school satisfaction were statistically similar between the treatment and control groups for all 10 subgroups examined. These results were confirmed by both sensitivity tests.

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

Table 3-8. Year 3 Impact Estimates of the Offer and Use of a Scholarship on the Full Sample and Subgroups: Student Reports of Satisfaction with Their School

Students Who Gave Their School a Grade of A or B	Impact of the Scholarship Offer (ITT)				Impact of Scholarship Use (IOT)		*p*-value of estimates
	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	Adjusted Impact Estimate	Effect Size	
Full sample	**.71**	**.73**	**-.03**	**-.06**	**-.03**	**-.07**	**.41**
SINI ever	.64	.72	-.08	-.18	-.09	-.21	.07
SINI never	.76	.74	.02	.05	.03	.07	.60
Difference	-.12	-.02	-.11	-.25			.10
Lower performance	.63	.69	-.06	-.13	-.08	-.17	.29
Higher performance	.74	.75	-.01	-.02	-.01	-.02	.81
Difference	-.11	-.06	-.05	-.12			.46
Male	.67	.72	-.05	-.11	-.06	-.14	.27
Female	.74	.74	-.00	-.01	-.00	-.01	.95
Difference	-.06	-.02	-.05	-.11			.46
4-8	.73	.76	-.03	-.06	-.03	-.07	.44
9-12	.55	.57	-.02	-.04	-.03	-.06	.76
Difference	.19	.19	-.01	-.01			.94
Cohort 2	.73	.78	-.05	-.11	-.05	-.13	.24
Cohort 1	.59	.56	.03	.05	.04	.07	.63
Difference	.15	.22	-.07	-.16			.28

NOTES: Means are regression adjusted using a consistent set of baseline covariates. Impact estimates are reported as marginal effects. Effect sizes are in terms of standard deviations. Valid $N = 1{,}014$, including: SINI ever $N = 513$, SINI never $N = 501$, Lower performance $N = 320$, Higher performance $N = 694$, Male $N = 494$, Female $N = 520$, K-8 $N = 851$, 9-12 $N = 163$, Cohort 2 $N = 749$, Cohort 1 $N = 265$. Student survey weights were used. Survey given to students in grades 4-12.

3.6 Chapter Summary

This chapter presents the estimated impacts of the OSP 3 years after the initial random assignment of students to treatment or control groups. The evidence indicates that the treatment—*the offer* of a scholarship—generated a positive and statistically significant impact on the average reading test scores of the students in the study. The size of the overall reading impact is .13 standard deviations or 3.1 months of additional schooling for the offer of the scholarship and .15 standard deviations or 3.7 months of additional schooling for the use of a scholarship. Five of the 10 student subgroups examined demonstrated statistically significant reading impacts, with the positive impacts on the reading scores of SINI-never, higher baseline performing, and K-8 applicants retaining significance after adjustments for multiple comparisons. The highest priority subgroup—SINI-ever students—did not experience statistically significant achievement impacts from the offer or use of an OSP scholarship. No statistically significant treatment impacts were observed in math, overall or for any of the 10 student subgroups.

Overall, parents in the treatment group continued to perceive their child's school to be safer and were more likely to assign it a high grade compared to parents of the control group. The subgroup impacts were consistent with the overall estimates for school satisfaction, but 3 of the 10 parent groups—those of students who were more academically challenged (that is, students from SINI-ever schools or students in the lower one-third of the test-score distribution) or entering the slot-constrained high school grades when they applied to the Program—did not report being more satisfied with their child's school if they were offered an Opportunity Scholarship.

As in the year 1 and year 2 impact reports released previously, student perceptions of school safety and satisfaction differed significantly from those of their parents in year 3. Student reports of school safety and climate and their likelihood of grading their school A or B were statistically indistinguishable between the treatment and control groups—overall and for all 10 student subgroups.

4. Exploratory Analysis of OSP Intermediate Outcomes

Whatever effect the OSP has on key outcomes (most importantly, 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; Howell and Peterson 2006, pp. 158-166). 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; Nielsen and Wolf 2001; Hoxby 2000); (2) school organization or instruction is different (Chubb and Moe 1990); (3) parents and students develop different expectations for their success (Akerlof and Kranton 2002; Bryk, Lee, and Holland 1993); (4) the school community surrounding students is more comprehensive and nurturing (Brandl 1998; Coleman and Hoffer 1987); and (5) parents become more involved in that school community (Coulson 1999). The conceptual basis for these hypotheses depends on two important linkages: (1) access to a voucher alters the educational experiences or behaviors mentioned above, and (2) those differences lead to better student outcomes. However, so far there has been little research that empirically tests these relationships (Hess and Loveless 2005).

This chapter examines how the first set of linkages in the hypothesized causal chain may be playing out for the OSP. The investigation explores the question, "Did the OSP change the daily educational life or experiences of participating students?" While part of this question was examined descriptively in Chapter 2, the analysis here estimates the actual impact of the Program on a set of variables that we call "intermediate outcomes" because they may be influenced by the Program but they themselves are not an "end outcome" as identified in the OSP statute. The method used to estimate the impacts on intermediate outcomes is identical to that used to estimate impacts on the key programmatic end outcomes (see appendix A for more detail).

The first section of this chapter examines the impact of the OSP on the educational conditions experienced by the treatment group overall. The second section explores those impacts across the 10 policy-relevant subgroups that were the subject of the subgroup analyses in chapter 3. The final section briefly discusses the pattern of results.

4.1 Impact of the OSP on Intermediate Outcomes Overall

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 using the same analytic model used to estimate the impacts reported in chapter 3.[54] 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).

Overall, 3 years after applying for a scholarship, the Program appears to have had an impact on 8 of the 24 intermediate outcomes examined (table 4-1).

Home Educational Supports

The offer of an Opportunity Scholarship had an impact on one of the four intermediate outcomes in this conceptual grouping (table 4-1).

- Students offered a scholarship experienced a lower likelihood of tutor usage outside of the school environment (ES = -.14). According to parents, 9 percent of students in the treatment group, compared to 14 percent of students in the control group, received some tutoring services outside of school during year 3. This impact remained statistically significant after adjustments for multiple comparisons.

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

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

Mediators:	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)	Effect Size	*p*-value
Section 1. Home Educational Supports					
Parental Involvement	2.68	2.91	-.23	-.11	.06
Parent Aspirations	17.24	17.18	.06	.02	.69
Out-of-school Tutor Usage	.09	.14	-.05**	-.14	.00
School Transit Time	2.78	2.65	.12	1.13#	.29
Section 2. Student Motivation and Engagement					
Student Aspirations	16.60	16.87	-.27	-.14	.06
Attendance	.90	.80	.11	1.11#	.36
Tardiness	.64	.46	.17	1.19#	.19
Reads for Fun	.38	.46	-.08*	-.16	.03
Engagement in Extracurricular Activities	2.25	2.20	.05	.04	.62
Frequency of Homework (days)	3.49	3.36	.13	.08	.21
Section 3. Instructional Characteristics					
Student/Teacher Ratio	12.95	12.89	.06	.01	.83
Teacher Attitude	2.58	2.67	-.09	-.04	.54
Ability Grouping	.72	.71	.01	.02	.83
Availability of Tutors	.49	.67	-.18**	-.38	.00
In-school Tutor Usage	.25	.23	.02	.04	.50
Programs for Learning Problems	.76	.88	-.12**	-.36	.00
Programs for English Language Learners	.27	.57	-.30**	-.61	.00
Programs for Advanced Learners	.52	.38	.13*	.27	.01
Before-/After-School Programs	.88	.91	-.03	-.11	.09
Enrichment Programs	2.49	2.30	.19**	.23	.00
Section 4. School Environment					
School Communication Policies	3.01	3.06	-.05	-.06	.46
School Size	379.74	561.72	-181.98**	-.29	.00
Percent Non-White	.96	.97	-.01	-.10	.13
Peer Classroom Behavior	8.32	8.10	.22	.09	.17

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Effect sizes for categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for these variables see appendix F.5.

NOTES: Valid *N* for Parental Involvement = 1,425; Parent Aspirations = 1,373; Out-of-school Tutor Usage = 1,373; School Transit Time = 1,433; Student Aspirations = 1,027; Attendance = 1,403; Tardiness = 1,380; Reads for Fun = 1,104; Engagement in Extracurricular Activities = 1,037; Frequency of Homework = 1,072; Student/Teacher Ratio = 1,069; Teacher Attitude = 1,096; Ability Grouping = 848; Availability of Tutors = 854; In-school Tutor Usage = 1,384; Programs for Learning Problems = 819; Programs for English Language Learners = 867; Programs for Advanced Learners = 866; Before-/After-School Programs = 867; Enrichment Programs = 867; School Communication Policies = 873; School Size = 1,186; Percent Non-White = 1,188; Peer Classroom Behavior = 1,099. 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. Regression models for Before-/After-School Programs omit the grade level 4 dummy variable because it predicted success perfectly.

- There were no statistically significant differences between the treatment and control groups on parents' reports of their involvement in school in year 3 (ES = -.11), parents' aspirations for how far in school their children would go (ES = .02), or time required for the student to get to school (odds ratio = 1.13).[55]

Student Motivation and Engagement

Of the six intermediate outcomes in this category, the offer of a scholarship may have had an impact on one of them (table 4-1).

- Based on student surveys, the offer of a scholarship seems to have had a significant negative impact on students' likelihood of reading for fun (ES = -.16). Adjustments for multiple comparisons, however, indicate that this result could be a false discovery, so it should be interpreted with caution.

- Three years after they applied to the OSP, there were no statistically significant differences between students in the treatment and control groups on student reports of their aspirations for future schooling (ES = -.14), engagement in extracurricular activities (ES = .04), and frequency of doing homework (ES = .08). There were no statistically significant differences in student attendance at school (odds ratio = 1.11) or tardiness rates (odds ratio = 1.19), as reported by parents.[56]

Instructional Characteristics

The offer of a scholarship had a statistically significant impact on 5 of the 10 intermediate outcomes in this group of indicators (table 4-1).

- Students offered a scholarship experienced a lower likelihood that their school offered tutoring (ES = -.38), special programs for students with learning problems (ES = -.36), or special programs for children who were English language learners (ES = -.61) compared to control group students.[57] These impacts remained statistically significant after adjustments for multiple comparisons.

---

[55] The effect size for this categorical variable is expressed as an odds ratio, which describes the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for these variables see appendix F.5. Commuting time was selected as a possible intermediate outcome because students who exercise school choice tend to attend schools that are further 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).

[56] The effect sizes for these categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for these variables see appendix F.5.

[57] As reported in chapter 2, 50 percent of schools attended by the treatment group and 67 percent of schools attended by the control group made tutors available at school; 26 percent of schools attended by the treatment group and 57 percent of schools attended by the control group offered special programs for non-English speakers; and 71 percent of schools attended by the treatment group and 88 percent of schools attended by the control group offered special programs for students with learning problems.

- Students offered a scholarship experienced a higher likelihood that their school offered programs for advanced learners (ES = .27) and enrichment programs, such as art, music, and foreign language (ES = .23) compared to control group students. These two impact estimates also remained statistically significant after adjustments for multiple comparisons.

- There were no significant differences between the treatment and control groups in how students rated their teacher's attitude (ES = -.04) or their likelihood of using a tutor in school (ES = .04). The school attended by the two groups were statistically similar in their student/teacher ratio (ES = .01), use of ability grouping (ES = .02), and the availability of before- and after-school programs (ES = -.11).

School Environment

The Program affected one of four measures of the school environment (table 4-1).

- Students offered a scholarship experienced schools that were smaller by an average of 182 students (ES = -.29) than the schools attended by students in the control group. This impact remained statistically significant after adjustments for multiple comparisons.

- There were no statistically significant differences between the treatment and control groups in school reports of communication policies (ES = -.06), the percentage of minority students at the school (ES = -.10), and the classroom behavior of peers (ES = .09) based on student reports.

4.2 Impacts on Intermediate Outcomes for Student Subgroups and Their Association with Achievement

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 for five specific subgroups of students—students from non-SINI schools, students who applied to the Program with higher academic performance, female students, students entering grades K-8 at baseline, and students who were in cohort 1. If the educational experiences of the subgroups of students who most clearly benefited academically from the treatment appear to be systematically different from those of the subgroups of students who do not appear to have so benefited, it might provide some support for the hypotheses about the linkages between those school conditions or behaviors and achievement.

However, any effort to examine a relationship between the intermediate outcomes and achievement can only be suggestive because there is no way to rigorously evaluate these linkages. Study participants are only randomly assigned 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 subgroups of participants that are demonstrating clear achievement impacts could be partly or entirely a function of the types of scholarship students and families that sort themselves into the different school choices. Moreover, the order of causality between intermediate and end outcomes is not certain in this case. Specific educational conditions may be associated with achievement gains because the conditions induced the gains or because the fact that a student was or was not achieving at a satisfactory level led to the application of a specific educational condition. For example, the OSP both increased reading achievement and decreased the use of tutors outside of school, possibly because OSP parents perceived less of a need for their children to be tutored in reading as their reading ability approved. Similarly, parents may be less motivated to be involved in their child's school if the student is improving academically, which could explain the combination of a positive impact of the OSP on reading achievement and a negative impact of the Program on parental involvement for SINI-never, higher baseline performance, and female students. 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.

To explore the possibility of treatment mediators, we estimated the difference between the experiences of the various subgroup pairs (males versus females, SINI versus non-SINI) by interacting the treatment variable with the subgroup indicator variable.[58] Much of the discussion that follows focuses on the cases where the analysis confirms a statistically significant difference in the experience of an intermediate outcome between two subgroups – one of which demonstrated a year 3 reading impact and one of which did not.[59]

[58] Details regarding how the standard statistical model was modified to estimate the differences in outcomes between subgroup pairs are provided in the appendix A.8 subsection on "Subgroup ITT Impacts."

[59] A large number of statistical comparisons are involved in this analysis of programmatic impacts on 10 different subgroups for 24 different intermediate outcomes. To ease the burden on the reader, only the effect sizes of the impact estimates and differences between the subgroup impacts, along with statistical significance indicators, are presented in the tables that follow. The large number of statistical estimates also means that, after adjustments for these multiple comparisons, only the subgroup findings with initial high levels of statistical significance in this exploratory analysis remain significant and unlikely to be false discoveries.

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 students' educational conditions and behaviors indicated by the exploratory analysis presented below do not, at this point, align closely with the pattern of test score impacts reported in chapter 3. There are distinctive patterns of impacts on intermediate outcomes for different subgroups of students. However, subgroups that experienced significant changes in elements of their educational experience as a result of the OSP so far appear to be no more or less likely to have experienced reading impacts as a result of the Program. Therefore, none of the intermediate outcomes explored here seem to be likely mediators of the impact of the scholarship treatment on year 3 reading outcomes.

Home Educational Supports

There were subgroup impacts on two of the four outcomes in this group (table 4-2).

- While there was no impact overall, there may have been a negative impact on parental involvement in school for three of the five student subgroups that had statistically significant reading impacts from the scholarship offer. Lower levels of parental involvement in the treatment group were reported for SINI-never (ES = -.16), higher baseline performance (ES = -.19), and female (ES = -.16) students compared with similar control group students. Only the impact on the higher baseline performance subgroup of students remained statistically significant after adjustment for multiple comparisons.

- Students who entered the Program in grades 9-12 also showed a statistically significant negative impact on parental involvement (ES = -.34) though that subgroup did not demonstrate reading impacts. This impact was no longer statistically significant after adjustment for multiple comparisons and so may be a false discovery.

- Four of the five student subgroups with positive reading impacts at the same time experienced a negative impact on usage of tutors outside their school. Lower levels of tutor usage were an intermediate outcome of the Program for the SINI-never (ES = -.14), higher baseline performance (ES = -.14), female (ES = -.15) and K-8 (ES = -.14) subgroups of students. Adjustments for multiple comparisons indicate that the impacts on the SINI-never and female subgroups of students could be false discoveries.

- The male (ES = -.14) and cohort 2 (ES = -.17) subgroups of students also appear to have used outside tutors less frequently as a result of being offered a scholarship, although the impacts for males could be due to chance. Neither of these groups experienced an impact on achievement, positive or negative.

Table 4-2. Year 3 Effect Sizes for Subgroups: Home Educational Supports (ITT)

Subgroup:	Parental Involvement	Parent Aspirations	Out-of-school Tutor Usage	School Transit Time #
Overall Impact	**-.11**	**.02**	**-.14****	**1.13**
SINI ever	-.04	.04	-.14	1.16
SINI never	-.16*	.01	-.14*	1.11
Difference	.12	.03	-.00	1.05
Lower performance	.08	.06	-.15	.87
Higher performance	-.19**	.00	-.14*	1.28
Difference	.27*	.06	-.01	.68
Male	-.06	.09	-.14*	1.11
Female	-.16*	-.05	-.15*	1.15
Difference	.10	.14	.01	.97
K-8	-.08	.02	-.14**	1.21
9-12	-.34*	.07	-.18	.77
Difference	.26	-.06	.05	1.56
Cohort 2	-.10	-.01	-.17**	1.11
Cohort 1	-.14	.13	-.00	1.20
Difference	.04	-.14	-.17	.93

* Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Effect sizes for categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for these variables see appendix F.5.

NOTES: Valid *N* for Parental Involvement = 1,425, including: SINI ever *N* = 592, SINI never *N* = 833, Lower performance *N* = 45, Higher performance *N* = 972, Male *N* = 696, Female *N* = 729, K-8 *N* = 1,265, 9-12 *N* = 160, Cohort 2 *N* = 1,154, Cohort 1 *N* = 271. Valid *N* for Parent Aspirations = 1,373, including: SINI ever *N* = 571, SINI never *N* = 802, Lower performance *N* = 431, Higher performance *N* = 942, Male *N* = 674, Female *N* = 699, K-8 *N* = 1,215, 9-12 *N* = 158, Cohort 2 *N* = 1,106, Cohort 1 *N* = 267. Out-of-school Tutor Usage *N* = 1,387, including: SINI ever *N* = 575, SINI never *N* = 812, Lower performance *N* = 444, Higher performance *N* = 943, Male *N* = 686, Female *N* = 701, K-8 *N* = 1,228, 9-12 *N* = 159, Cohort 2 *N* = 1,122, Cohort 1 *N* = 265. Valid *N* for School Transit Time = 1,434, including: SINI ever *N* = 598, SINI never *N* = 836, Lower performance N = 458, Higher performance N = 976, Male N = 708, Female N = 726, K-8 N = 1,271, 9-12 N = 163, Cohort 2 N = 1,159, Cohort 1 *N* = 275. 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.

Student Motivation and Engagement

With two exceptions, for this group of measures, 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 (table 4-3). Thus, there were no statistically significant impacts for any of the 10 subgroups on attendance, engagement in extracurricular activities, or frequency of homework. There are impacts for subgroups on three outcomes in this category.

Table 4-3. Year 3 Effect Sizes for Subgroups: Student Motivation and Engagement (ITT)

Subgroup:	Student Aspirations	Attendance#	Tardiness#	Reads for Fun	Engagement in Extra-curricular Activities	Frequency of Homework
Overall Impact	**-.14**	**1.11**	**1.19**	**-.16***	**.04**	**.08**
SINI ever	-.09	1.14	1.07	-.07	-.03	.12
SINI never	-.17	1.09	1.29	-.23*	.09	.04
Difference	.08	1.05	.83	.16	-.12	.08
Lower performance	-.33**	1.48	1.25	-.02	-.03	.09
Higher performance	-.04	.97	1.16	-.23*	.06	.08
Difference	-.30*	1.53	1.08	.21	-.10	.01
Male	-.09	1.00	.95	-.11	.12	.15
Female	-.20	1.22	1.47*	-.22*	-.03	.02
Difference	.11	.82	.65	.11	.15	.13
K-8	-.14	1.07	1.15	-.18*	.06	.11
9-12	-.11	1.40	1.43	-.07	-.13	-.06
Difference	-.04	.76	.80	-.11	.19	.17
Cohort 2	-.17	1.14	1.21	-.19*	.03	.09
Cohort 1	-.03	.98	1.09	-.06	.07	.06
Difference	-.13	1.16	1.11	-.12	-.05	.02

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

Effect sizes for categorical variables are expressed as odds ratios, which describe the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for these variables see Appendix F.5.

NOTES: Valid *N* for Student Aspirations = 1,027, including: SINI ever *N* = 516, SINI never *N* = 511, Lower performance *N* = 458, Higher performance *N* = 704, Male *N* = 498, Female *N* = 529, K-8 *N* = 866, 9-12 *N* = 161, Cohort 2 *N* = 765, Cohort 1 *N* = 262. Valid *N* for Attendance = 1,403, including: SINI ever *N* = 580, SINI never *N* = 823, Lower performance *N* = 444, Higher performance *N* = 959, Male *N* = 685, Female *N* = 718, K-8 *N* = 1,250, 9-12 *N* = 153, Cohort 2 *N* = 1,142, Cohort 1 *N* = 261. Valid *N* for Tardiness = 1,380, including: SINI ever *N* = 577, SINI never *N* = 803, Lower performance *N* = 439, Higher performance *N* = 941, Male *N* = 678, Female *N* = 702, K-8 *N* = 1,229, 9-12 *N* = 151, Cohort 2 *N* = 1,116, Cohort 1 *N* = 264. Valid *N* for Reads for Fun = 1,104, including: SINI ever *N* = 551, SINI never *N* = 553, Lower performance *N* = 347, Higher performance *N* = 757, Male *N* = 539, Female *N* = 565, K-8 *N* = 933, 9-12 *N* = 171, Cohort 2 *N* = 828, Cohort 1 *N* = 276. Valid *N* for Engagement in Extracurricular Activities = 1,037, including: SINI ever *N* = 520, SINI never *N* = 517, Lower performance *N* = 330, Higher performance *N* = 707, Male *N* = 505, Female *N* = 532, K-8 *N* = 874, 9-12 *N* = 163, Cohort 2 *N* = 770, Cohort 1 *N* = 267. Valid *N* for Frequency of Homework = 1,072, including: SINI ever *N* = 539, SINI never *N* = 533, Lower performance *N* = 337, Higher performance *N* = 735, Male *N* = 522, Female *N* = 488, K-8 *N* = 790, 9-12 *N* = 167, Cohort 2 *N* = 801, Cohort 1 *N* = 271. 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 "Reads for Fun" are reported as marginal effects. Data regarding student aspirations, reads for fun, engagement in extracurricular activities, and frequency of homework were drawn from student surveys and therefore limited to students in grades 4-12.

- Four of the five student subgroups that had statistically significant reading impacts may also have experienced a negative Program impact on the likelihood that they read for fun. Compared to control group students, students in the treatment group were less likely to report that they read for fun among the SINI-never (ES = -.23), higher baseline performance (ES = -.23), female (ES = -.22) and K-8 (ES = -.18) subgroups of students. None of these treatment impacts remained statistically significant after adjustments for multiple comparisons, suggesting that they may be false discoveries.

- The cohort 2 (ES = -.19) subgroup of students, which did not have impacts on reading achievement, also may have had a lower likelihood of reading for fun as a result of being offered a scholarship. This subgroup impact may be a false discovery.

- There were a few instances where an insignificant overall finding may have had significant subgroup effects. Females, for whom the OSP improved reading achievement, had higher rates of parent-reported tardiness (odds ratio = 1.47)[60] if they had been offered a scholarship. The Program may also have reduced the future educational aspirations of students who entered the Program with relatively lower academic performance (ES = -.33) but had no significant impact on the educational aspirations of higher baseline performers (ES = -.04), a differential impact that itself is statistically significant. None of these impacts remained statistically significant after adjustments for multiple comparisons.

Instructional Characteristics

As was true for students overall, there was no impact for subgroups of students on their views of teachers' attitudes or their use of tutors in school (table 4-4). There were subgroup impacts on other instructional characteristics, both for students whose reading achievement was affected by the OSP and those whose achievement was unaffected.

- Overall, students offered a scholarship experienced a lower likelihood that their school offered tutoring. That effect was consistent across all subgroups although impacts for three groups were not statistically significant. The impact was significant for four of the five subgroups for whom the OSP improved reading achievement, including students who were higher-performing at baseline (ES = -.49), female (ES = -.52), entering grades K-8 (ES = -.29), or in cohort 1 (ES = -.68). Statistically significant negative impacts on school tutors were also found for three groups that did not experience reading impacts: SINI-ever (ES = -.63), students entering high school (ES = -.87), and cohort 2 (ES = -.31). Adjustments for multiple comparisons indicate that none of these impacts are in danger of being false discoveries.

- Overall, students offered a scholarship experienced a lower probability that their school offered special programs for students with learning problems. That effect was consistent and statistically significant across all five of the subgroups with reading achievement impacts (ES from -.34 to -.89). Statistically significant impacts also were found for three of the five groups without reading impacts (ES from -.28 to -.43). There was no significant impact for SINI-ever students (ES = -.33), and the impact on students entering grades 9-12 at baseline could not be determined due to data limitations. Adjustments for multiple comparisons indicate that the impacts for cohort 1, male, and lower baseline performance students could be false discoveries.

[60] The effect size for this categorical variable is expressed as an odds ratio, which describes the extent to which being in the treatment group increases (if above 1.0) or decreases (if below 1.0) the likelihood of giving a higher-category response. For a complete description of the treatment and control group response patterns for this variable see appendix F.5.

Table 4-4. Year 3 Effect Sizes for Subgroups: Instructional Characteristics (ITT)

Subgroup:	Student/ Teacher Ratio	Teacher Attitude	Ability Grouping	School Provides Tutors	In School Tutor Usage	Programs for Learning Problems	Programs for English Language Learners	Programs for Advanced Learners	Before- or After-School Programs	Enrichment Programs
Overall Impact	**.01**	**-.04**	**.02**	**-.38****	**.04**	**-.36****	**-.61****	**.27***	**-.11**	**.23****
SINI ever	.10	.00	.21	-.63**	.02	-.33	-.55**	.58**	-.21*	.24*
SINI never	-.05	-.08	-.15	-.20	.06	-.38**	-.65**	.09	.00	.23*
Difference	.14	.08	.35*	-.44*	-.04	.05	.09	.49*	-.21	.01
Lower performance	.04	-.02	-.05	-.09	.13	-.43*	-.31	.60*	.03	.23
Higher performance	.00	-.05	.04	-.49**	-.01	-.34**	-.73**	.17	-.18*	.23**
Difference	.04	.03	-.09	.40	.14	.08	.41*	.43	.21	-.00
Male	.00	-.06	.03	-.25	.09	-.28*	-.39**	.57**	-.07	.22*
Female	.03	-.02	.01	-.52**	-.00	-.44**	-.85**	.03	-.18	.25*
Difference	-.02	-.04	.02	.27	.09	.16	.46**	.53*	.11	-.03
K-8	-.06	-.02	-.08	-.29**	.08	-.37**	-.58**	.17	.04	.23**
9-12	.62**	-.22	.87*	-.87**	-.30	N/A	-.84*	1.36**	-.49**	.31
Difference	-.69**	.21	-.95*	.58	.38	N/A	.26	-1.19**	.53**	-.08
Cohort 2	-.04	-.04	-.13	-.31**	.08	-.32**	-.60**	.33**	-.17*	.31**
Cohort 1	.27	-.05	.39*	-.68**	-.10	-.89*	-.63**	.03	.17	-.16
Difference	-.31	.01	-.52*	.37	.17	.57	.03	.30	-.34*	.47*

*Statistically significant at the 95 percent confidence level

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for Student/Teacher Ratio = 1,034, including: SINI ever *N* = 427, SINI never *N* = 607, Lower performance *N* = 313, Higher performance *N* = 721, Male *N* = 501, Female *N* = 533, K8 *N* = 917, 9-12 *N* = 117, Cohort 2 *N* = 833, Cohort 1 *N* = 201. Valid *N* for Teacher Attitude = 1096, including: SINI ever *N* = 551, SINI never *N* = 545, Lower performance *N* = 347, Higher performance *N* = 749, Male *N* = 541, Female *N* = 555, K-8 *N* = 926, 9-12 *N* = 170, Cohort 2 *N* = 821, Cohort 1 *N* = 275. Valid *N* for Ability Grouping = 848, including: SINI ever *N* = 324, SINI never *N* = 524, Lower performance *N* = 244, Higher performance *N* = 604, Male *N* = 418, Female *N* = 430, K-8 *N* = 786, 9-12 *N* = 62, Cohort 2 *N* = 702, Cohort 1 *N* = 146. Valid *N* for School Provides Tutors = 854, including: SINI ever *N* = 328, SINI never *N* = 526 Lower performance *N* = 243, Higher performance *N* = 611, Male *N* = 425, Female *N* = 429, K-8 *N* = 792, 9-12 *N* = 62, Cohort 2 *N* = 706, Cohort 1 *N* = 148. Valid *N* for In-School Tutor Usage = 1,384, including: SINI ever *N* = 578, SINI never *N* = 806, Lower performance *N* = 443, Higher performance *N* = 941, Male *N* = 687, Female *N* = 697, K-8 *N* = 1224, 9-12 *N* = 160, Cohort 2 *N* = 1116, Cohort 1 *N* = 268. Valid *N* for Learning Problems = 819, including: SINI ever *N* = 305, SINI never *N* = 514, Lower performance *N* = 236, Higher performance *N* = 583, Male *N* = 412, Female *N* = 407, K-8 *N* = 757, 9-12 *N* = 62, Cohort 2 *N* = 672, Cohort 1 *N* = 147. Valid *N* for Programs for English Language Learners = 867, including: SINI ever *N* = 330, SINI never *N* = 537, Lower performance *N* = 245, Higher performance *N* = 622 Male *N* = 431, Female *N* = 436, K-8 *N* = 805, 9-12 *N* = 62, Cohort 2 *N* = 719, Cohort 1 *N* = 148. Valid *N* for Programs for Advanced Learners = 866, including: SINI ever *N* = 330, SINI never *N* = 536, Lower performance *N* = 244, Higher performance *N* = 622 Male *N* = 431, Female *N* = 435, K-8 *N* = 804, 9-12 *N* = 62, Cohort 2 *N* = 719, Cohort 1 *N* = 147. Valid *N* for Before- or After-School Programs = 867, including: SINI ever *N* = 330, SINI never *N* = 537, Lower performance *N* = 245, Higher performance *N* = 622, Male *N* = 431, Female *N* = 436, K-8 *N* = 805, 9-12 *N* = 62, Cohort 2 *N* = 719, Cohort 1 *N* = 148. Valid *N* for Enrichment Programs = 867, including: SINI ever *N* = 330, SINI never *N* = 537, Lower performance *N* = 245, Higher performance *N* = 622, Male *N* = 431, Female *N* = 436, K-8 *N* = 805, 9-12 *N* = 62, Cohort 2 *N* = 719, Cohort 1 *N* = 148. 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 student surveys and therefore limited to students in grades 4-12. Regression runs for Before-/After-School Programs omit the grade level 4 dummy variable because it predicted success perfectly. Because of data limitations, analysis of Programs for Learning Problems was not possible for the 9-12 subgroup.

- Overall, students offered a scholarship experienced a lower likelihood that their school offered special programs for English language learners. That effect also was consistent and statistically significant across all five of the subgroups with reading achievement impacts (ES from -58 to -.85). Four of the five subgroups that did not demonstrate reading impacts from the Program also were less likely to attend a school with special programs for English language learners if offered a scholarship (ES from -.39 to -.84), the exception being students with lower baseline performance (ES = -.31). Adjustments for multiple comparisons indicate that none of these impacts are in danger of being false discoveries.

- Overall, students offered a scholarship experienced a higher probability that their school offered programs for advanced learners. That effect was not concentrated among the subgroups with achievement impacts, but among the groups that did not experience impacts on reading test scores. Statistically significant impacts were found for none of the five subgroups with achievement impacts (ES from .03 to .17), but were found for all five of the groups without achievement impacts: SINI ever (ES = .58), lower performing at baseline (ES = .60), male (ES = .57), entering grades 9-12 (ES = 1.36), and in cohort 2 (ES = .33). Adjustments for multiple comparisons indicate that the impact on students with lower baseline performance may be a false discovery.

- Seven subgroups of students—four that also demonstrated reading impacts (SINI never, ES = .23; higher baseline performance, ES = .23; female, ES = .25; and K-8, ES = .23) and three that did not demonstrate reading impacts (SINI ever, ES = .24; male, ES = .22; and cohort 2, ES = .31)—experienced a higher likelihood that their school offered enrichment programs (music, art, foreign language) if they had been offered a scholarship. Adjustments for multiple comparisons indicate that the impacts on SINI ever and male student subgroups may be false discoveries.

- Higher performing students (for whom the OSP produced positive impacts on reading achievement) experienced a lower likelihood that their school offered a before- or after-school program (ES = -.18) if in the treatment group, as did SINI-ever (ES = -.21), grade 9-12 (ES = -.49), and cohort 2 (ES = -.17) students (for whom the Program had no effect on achievement). These subgroup impacts remained statistically significant after adjustments for multiple comparisons.

School Environment

The primary element of the school environment that changed for students as a result of the scholarship offer was the size of their school's student population (table 4-5).

- All five of the student subgroups with reading impacts experienced schools that, on average, were significantly smaller than those attended by the control group (table 4-5) (ES range from -.27 to -.37). Four of these five subgroup impacts on school size remained statistically significant after adjustments for multiple comparisons, with the exception of the cohort 1 impact, which might be a false discovery.

Table 4-5. Year 3 Effect Sizes for Subgroups: School Environment (ITT)

Subgroup:	School Communication Policies	School Size	Percent Non-White	Peer Classroom Behavior
Overall Impact	**-.06**	**-.29****	**-.10**	**.09**
SINI ever	.04	-.20	-.40*	-.06
SINI never	-.16	-.35**	-.05	.21*
Difference	.20	.15**	-.34	-.28*
Lower performance	.05	-.33**	-.05	.06
Higher performance	-.10	-.27**	-.12	.11
Difference	.15	-.06	.07	-.05
Male	.00	-.20*	-.22*	.17
Female	-.14	-.37**	-.03	.03
Difference	.14	.18	-.19	.14
K-8	-.03	-.31**	-.08	.11
9-12	-.28	-.03	-.51	-.02
Difference	.25	-.28*	.43	.13
Cohort 2	-.12	-.29**	-.07	.11
Cohort 1	.14	-.36*	-.28	.01
Difference	-.26	.07	.21	.10

* Statistically significant at the 95 percent confidence level.

** Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for School Communication Policies = 873, including: SINI ever *N* = 330, SINI never *N* = 543, Lower performance *N* = 247, Higher performance *N* = 626, Male *N* = 433, Female *N* = 440, K-8 *N* = 810, 9-12 *N* = 63, Cohort 2 *N* = 723, Cohort 1 *N* = 150. Valid *N* for School Size = 1,186, including: SINI ever *N* = 491, SINI never *N* = 695, Lower performance *N* = 373, Higher performance *N* = 813, Male *N* = 584, Female *N* 602, K-8 *N* = 1,051, 9-12 *N* = 135, Cohort 2 *N* = 957, Cohort 1 *N* = 223. Valid *N* for Percent Non-White = 1,188, including: SINI ever *N* = 491, SINI never *N* = 697, Lower performance *N* = 373, Higher performance *N* = 815, Male *N* = 586, Female *N* = 602, K-8 *N* = 1,053, 9-12 *N* = 135, Cohort 2 *N* = 958, Cohort 1 *N* = 230. Valid *N* for Peer Classroom Behavior = 1,099, including: SINI ever *N* = 551, SINI never *N* = 548, Lower performance *N* = 348, Higher performance *N* = 751, Male *N* = 541, Female *N* = 558, K-8 *N* = 928, 9-12 *N* = 171, Cohort 2 *N* = 824, Cohort 1 *N* = 275. 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 student surveys and therefore limited to students in grades 4-12.

- In addition, three of the five subgroups that did not demonstrate reading impacts also experienced schools that were smaller by a statistically significant margin (lower performing at baseline ES = -.33, male ES = -.20, cohort 2 ES = -.29), though only the impacts for the lower baseline performance and cohort 2 subgroups remained statistically significant after adjustments for multiple comparisons.

- Although the treatment group may have experienced a smaller percentage of non-White classmates than the control group among the SINI-ever (ES = -.40) and male (ES = -.22) subgroups, and SINI-never students may have experienced a positive impact on the level of disciplined and respectful peer classroom behavior as a result of the scholarship offer (ES = .21), statistical adjustments suggest that all three of those impact estimates might be false discoveries.

4.3 Chapter Summary

This chapter presents the results of an experimental analysis of the impacts of the OSP on specific features of students' educational experience and environment, viewed as "intermediate outcomes" of the OSP. The analysis determined that, as result of being offered an Opportunity Scholarship, students attended smaller schools that were more likely to have programs for enrichment and advanced learners but less likely to have programs for students with learning problems and English Language Learners. As a result of the scholarship offer, students were less likely to attend a school with tutors on staff and also less likely to use a tutor outside of school. None of the intermediate outcomes explored here seem to be likely mediators of the impact of the scholarship treatment on year 3 reading outcomes.

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.

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

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

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.

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

Reardon, Sean F., and John T. Yun. *Private School Racial Enrollments and Segregation*. Cambridge, MA: Harvard Civil Rights Project, 2002. Available online at [http://www.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.

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].

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

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

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

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

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

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

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

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

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa. *Evaluation of the DC Opportunity Scholarship Program: Impacts After Two Years*. (NCEE 2000-4023). U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, NCEE 2008-4023. Washington, DC: U.S. Government Printing Office, 2008. Available online at [http://ies.ed.gov/ncee/].

Wolf, Patrick 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.

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

APPENDIX A
RESEARCH METHODOLOGY

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

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

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

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

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

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

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

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

A.2 Study Power

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

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

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

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

- Although this evaluation examines a variety of outcomes, including student test scores in every year post-baseline, in this report we present the power analysis numbers only for the third outcome year. Power estimates for earlier study years are available elsewhere (e.g., Wolf et al. 2007, appendix B).

- Central to analytic power is the sample size of study participants *who actually provide outcome information in a given year*. In order to produce highly precise power estimates 3 years into this longitudinal study, here we use the actual counts of student observations obtained from the year 3 data collection. Sample size is one of three parameters of these power estimates that are fixed based upon actual numbers from this evaluation.

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

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

- The analysis also takes account of the estimated correlation between baseline test scores and outcome test scores, derived from a previous experimental analysis.[2] 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.

- 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

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

programmatic impacts simply because fewer than expected members of the treatment group actually use the programmatic treatment.[3]

- 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;

  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 3, 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 representative of the degree to which sibling test scores are correlated. The MDEs are estimated for math impacts, but would be approximately similar for reading impacts as well.

In the third year of the evaluation, the study has sufficient power to detect an overall test score impact of .12 of a standard deviation or higher (table A-1). The MDEs are .20 of a standard deviation or less for 7 of the 10 subgroups of policy interest—SINI ever, SINI never, higher baseline performance, male, female, K-8, and cohort 2. The study has less power to detect impacts on the lower baseline performance (MDE = .22), grade 9-12 (MDE = .38), and cohort 1 (MDE = .30) subgroups.

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

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

Table A-1. Minimum Detectable Effects in Year 3, Overall and by Subgroup

	Sample Size				MDE
Impact Sample	Treatment	Control	Treatment/ Control Ratio	Total	
All K-12 (Third Year Evaluation)	909	551	1.65	1,460	0.124
Subgroup					
School					
SINI ever	397	213	1.86	610	0.196
SINI never	512	338	1.51	850	0.161
Performance					
Lower	299	166	1.80	465	0.223
Higher	610	385	1.58	995	0.150
Gender					
Male	458	256	1.79	714	0.180
Female	451	285	1.58	736	0.174
Grade					
K-8	855	432	1.98	1,287	0.136
9-12	54	119	0.45	173	0.378
Cohort					
2	724	462	1.57	1,186	0.137
1	185	89	2.08	274	0.297

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

In summary, the power analysis shows that we are able to estimate treatment effects of reasonable magnitudes in year 3. 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 many of the policy-relevant subgroups of participants.

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

Sources of Data

Comparable data were collected for each student in the impact sample regardless of whether the student was in cohort 1 or 2 or was randomly assigned to the treatment or control group. However, the temporal separation of the two study cohorts leads to the relationship between the actual timing of data collection and the impact analysis samples shown below in table A-2. 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-2. 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 Stanford Achievement Test 9th Edition (SAT-9) standardized assessment administered by the District of Columbia Public Schools (DCPS) as part of its spring testing program for cohort 1 and from the SAT-9 standardized assessment administered by the evaluation team in the spring for cohort 2.[5] 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 members of the control group who did not receive a scholarship.[6] 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 69 percent for the third year follow-up assessments.[7]

[5] 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.

[6] 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).

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

- **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.[8] 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 68 percent for the third 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 third year follow-up data collection, the survey response rate among students in grade 4 or higher was 67 percent.

- **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 the final evaluation report 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 57 percent in the third year follow-up data.

Outcome Measures

Congress specified in the Program statute that the rigorous evaluation study possible impacts regarding academic achievement, school safety, and satisfaction. For this third 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 (one measure) and student self-reports of satisfaction (one measure). All outcome data were obtained from impact sample respondents in the spring and include the following:

[8] 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.

- **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.[9] 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.[10] 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 safety and an orderly school climate.** Parents were asked about the perceived seriousness of a number of problems at their child's school commonly associated with danger and rule-breaking. The specific items, all drawn from the surveys used in previous experimental evaluations of scholarship programs, were:

  – Property destruction;
  – Tardiness;
  – Truancy;
  – Fighting;
  – Cheating;
  – Racial conflict;
  – Weapons;
  – Drug distribution;
  – Drug and alcohol use; and
  – Teacher absenteeism.

  Parents were asked to label these conditions as "very serious," "somewhat serious," or "not serious" at their child's school. Responses to these items subsequently were categorized as "yes" (very or somewhat serious) or "no" (not serious). The number of "yes" responses for each parent were then summed to create a parental danger index or count that ranged from 0 to 10. Finally, the index was reverse coded to transform it from a "danger" measure to a "safety" (i.e., lack of danger) measure.[11]

---

[9] 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.

[10] 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.)

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

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

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

  Responses to these items were categorized as "yes" (at least once) or "no" (never) to create a count of the number of reported events that ranged from 0 to 8. The index was reverse coded to transform it from a "danger" measure to a "safety" (i.e., lack of danger) measure. [12]

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

  To generate the primary measure of school satisfaction, parents were asked "What overall grade would you give this child's current school?" A response of "F" was assigned the value "1", a "D" was assigned a "2" and so on up to a value of "5" for an "A." Observations with the value "5" or "4" were then recoded "1" and all other values were recoded "0" for the binary variable "graded school A or B" used in the main analysis. The original, full grade scale was preserved and the impact of the Program on that measure of parent satisfaction is presented in Appendix D, Table D-6.

  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;

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

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

The responses to this set of items were combined into a single parent satisfaction scale using maximum likelihood IRT. IRT is a procedure 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.4 below for a more detailed description of IRT.) The consistency and reliability of scaled measures of traits such as satisfaction can be determined by a rating statistic called Cronbach's Alpha (Spector, 1992). The completed parent satisfaction scale exhibited very high reliability with a Cronbach's Alpha of .93.[13] The impact of the Program on the parent satisfaction with school scale is presented in appendix D, table D-7. Program impacts on individual scale items appear in appendix D, table D-13.

- **Student self-reports of satisfaction.** Students were also asked to grade their school using the same question asked of parents, and two outcomes were created—a grade range and a dichotomous variable—as discussed above for parents. The results of the analysis of the impact of the Program on a student's likelihood of assigning their school a grade of A or B appear in Chapter 3 as the primary measure of student satisfaction with their school. The impact of the OSP on the average grade given across the full grade range appears in appendix D, table D-9.

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

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

  A single composite satisfaction scale was created for students using the same IRT procedures used to create the parent satisfaction scale. (See section A.4 below for a more detailed description of IRT.) The student scale also exhibited a high level of reliability; it had a Cronbach's Alpha of .85. The impact of the Program on the student satisfaction with school scale is presented in appendix D, table D-10. Program impacts on individual scale items appear in appendix D, table D-14.

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

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

assignment via the application form (including a parent survey) and administration of the SAT-9 in reading and math. Such "baseline" covariates are important in the context of an experimental evaluation because they permit researchers to (1) verify the integrity of the random assignment, (2) inform the generation of appropriate nonresponse weights, and (3) include the covariates in regressions to improve the precision of the estimations of treatment impacts and adjust for any baseline differences across the treatment and control groups.[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,
- 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.

---

[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).

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, Swaminathan, and Rogers 1991). IRT has gained increasing attention in the development of standardized academic tests and, most recently, in the development of scales measuring a wide variety of "subjective traits" such as satisfaction with treatment and individual perceptions of health status and overall quality of life.

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

"satisfaction" with the child's school. The results of the IRT analysis can be used to determine the extent to which the items included in the scale (or test) are good measures of the underlying construct, and how well the items "hang together" (show common relationships) to characterize the underlying, and unobserved, construct.

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.

Some students provided some, but not all, outcome subtest scores within the reading and math domains in year 3 because they either missed or skipped entire subtests. This included 82 students in reading and 17 students in math.[17] The total number of individual students who provided incomplete test score data was 96, since three students provided only subtest scores in both reading and math.

When the problem of incomplete test scores first emerged during the initial stages of the evaluation, the research team conducted an analysis to determine how closely subtest reading and math scores correlated with composite scores for the over 1,600 respondents for whom both subtest and composite scores were available. The correlations between subtest and composite scores within particular domains and grades were very strong, ranging from a low of r =.79 to a high of r = .92.[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 all cases where only the subtest scores were available. In year 3, these 96 cases were considered respondents for the purposes of calculating the test score nonresponse weights and were therefore included in the test score impact analysis.

A.6 Imputation for Missing Baseline Data

One difficulty that arose regarding the baseline data was the extent to which data were missing. Although some important baseline covariates (e.g., family income, grade, race, and gender) were available for all students, other baseline covariates contained some missing values. Importantly, nearly 20

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

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

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]

A.7 Sampling and Nonresponse Weights

Sampling weights were used in the impact analyses to account for the fact that the study sample was selected differently in the 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 nonresponse to data collection.

Base Weights

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

[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 were 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.

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-3).

Table A-3. Base Weights by Randomization Strata

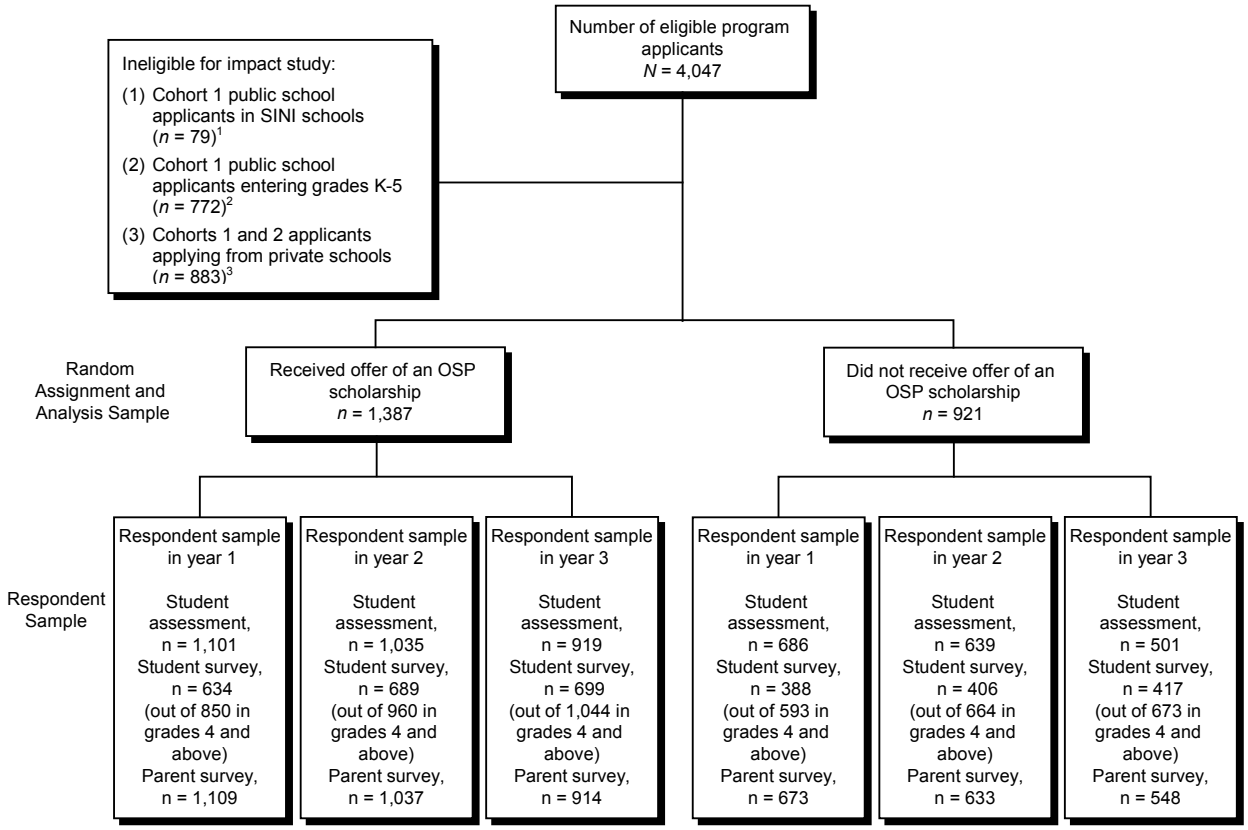
Stratum	Cohort	SINI Status	Grade Band	Treatment Sampling Rate (%)	Base Weight for Control Group	Base Weight for Treatment Group
1	Cohort 1	Non-SINI	6th to 8th	75.89	4.15	1.32
2	Cohort 1	Non-SINI	9th to 12th	28.21	1.39	3.54
3	Cohort 2	SINI	K to 5th	78.34	4.62	1.28
4	Cohort 2	SINI	6th to 8th	75.00	4.00	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 Nonresponse

The members of the treatment and control groups were offered similar inducements to cooperate in outcome data collection. Treatment students were invited to data collection events to renew their scholarships, and their parents were given a small cash payment for their time and transportation costs in responding. Control students were made eligible for follow-up scholarship lotteries, and their parents were provided with a compensation payment for attending follow-up data collection sessions. The initial base weights were adjusted for nonresponse, where a "respondent" was considered a student with reading or mathematics test data in year 3 (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 nonresponse response bias by compensating for different data collection response rates across various demographic groups of students organized within classification "cells." In effect, the nonresponse adjustment factor "spreads the weight" of the nonresponding students over the responding students in that cell, so that they represent not only students who responded (i.e., themselves), but also students who

[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: 3 Years After Application and Random Assignment



[1]The program operator offered a scholarship to all eligible public school applicants in cohort 1 applying from SINI schools.
[2]The program operator awarded scholarships to all eligible public school applicants in cohort 1 entering grades K-5 because there were sufficient slots in private schools to accommodate all the applicants in these grades.
[3]The evaluation design is intended to estimate the impact of giving students the opportunity to attend private school, so applicants to the Program who were already in private schools were excluded from the study.

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 nonresponse (see Howell et al. 2006, pp. 209-216; U.S. Department of Health and Human Services 2005). As a last step, the nonresponse-adjusted base weights were trimmed. Trimming prevents extremely large

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

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 nonresponse described above, initially there was a large differential between the response rates of the two experimental groups, which could have undermined their comparability and therefore biased the impact analysis. For year 3, after four invitations to attend data collection events, the evaluation team had obtained responses from nearly 68 percent of the treatment group but only about 58 percent of the control group (table A-4).

Table A-4. Test Score Response Rates for Third Year Outcomes Before Drawing Subsample

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)
Cohort 1 C	174	72	41.4
Cohort 1 T	290	185	63.8
Cohort 2 C	693	429	61.9
Cohort 2 T	1,066	734	68.9
Cohort 1 total	464	257	55.4
Cohort 2 total	1,759	1,163	66.1
C total	867	501	57.8
T total	1,356	919	67.8
Combined total	2,223	1,420	63.9

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Recently, a technique was developed to help reduce nonresponse bias in longitudinal impact analyses. Nonresponse subsampling is a strategy to reduce the differences between the characteristics of baseline and outcome samples by way of random sampling and nonresponse conversion. After the regular period of outcome data collection is over, a subsample of nonrespondents is drawn and subjected to intensive efforts at nonresponse conversion. If initial nonresponse was significantly higher in one experimental group compared with the other, as was the case in this evaluation, then the subsample can be drawn exclusively from the underresponded group (e.g., controls). Each initial nonrespondent who converts to a respondent by providing outcome data counts as one more respondent for purposes of the "actual" response rate but counts as 1/sampling rate (r) respondents for purposes of the "effective"

[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.

response rate. Through a simple weighting algorithm, the random sampling permits the respondent to also "stand in" for members of the initial nonrespondent group who were not selected for the subsample but who presumably would have converted to respondent status if they had been selected to receive the intensive recruiting efforts and incentives that were the conversion "treatment." In other words, the proportion of subsampled nonrespondents that converts represents themselves as well as the same proportion of nonsampled nonrespondents.

This technique was applied for the spring 2008 data collection, as it had been in 2007 and 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 nonresponders. A random sample of 177 of the 353 control group nonrespondents was drawn (50 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 nonrespondent to respondent status. A total of 51 initial nonrespondents (29 percent) were converted to respondents as a result of this effort, 17 from cohort 1 and 34 from cohort 2 (table A-5)**.** 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 nonrespondents who were not randomly selected for targeted conversion efforts but who would have responded if they had been targeted (see Kling, Ludwig, and Katz 2005; Sanbonmatsu et al. 2006).[26] The weights ensure that each converted member of the subsample represents him or herself as well as another study participant: a nonrespondent like him or her who would have converted had he/she been included in the subsample. As a result of implementing this approach, the combined cohort control group response rate increased to an effective rate of 70 percent for outcome testing in math and reading, and the treatment-control response differential decreased to 2 percentage points. The response-rate differential also decreased to 2 percentage points for parent surveys and 0 percentage points for student surveys (tables A-6 through A-8).

The What Works Clearinghouse (WWC) considers a Randomized Control Trial (RCT) such as this evaluation to meet evidence standards for claims of causality without reservations if study sample attrition is neither severe overall nor significantly different across the treatment and control groups. Even

[25] There were 100 control group nonresponders from cohort 1 and 253 from cohort 2. The random sample of 177 consisted of 50 from cohort 1 and 127 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 nonresponders and subjected them to intense recruitment efforts that resulted in nearly half of them responding, thereby increasing their response rate to 81 percent. The evaluators then assumed that the second-wave respondents were similar to the half of the larger nonrespondent group that they did not pursue aggressively and thus estimated and reported an "effective response rate" of 90 percent, even though actual data were obtained for only 81 percent of the respondents.

Table A-5. Subsample Conversion Response Rates for Third Year Outcomes

	Subsample Members	Actual Response Conversions	Actual Conversion Rate (%)
Cohort 1	50	17	34.0
Cohort 2	127	34	26.8
Total	177	51	28.8

Table A-6. Final Test Score Response Rates for Third Year Outcomes, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	174	89	51.1	106	60.9
Cohort 1 T	290	185	63.8	185	63.8
Cohort 2 C	693	463	66.8	497	71.7
Cohort 2 T	1,066	734	68.9	734	68.9
Cohort 1 total	464	274	59.1	291	62.7
Cohort 2 total	1,759	1,197	68.1	1,231	70.0
C total	867	552	63.7	603	69.5
T total	1,356	919	67.8	919	67.8
Combined total	2,223	1,471	66.2	1,522	68.5

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Table A-7. Parent Survey Response Rates for Third Year Outcomes, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	174	89	51.1	106	60.9
Cohort 1 T	290	189	65.2	189	65.2
Cohort 2 C	693	459	66.2	489	70.5
Cohort 2 T	1,066	725	68.0	725	68.0
Cohort 1 total	464	278	59.9	295	63.6
Cohort 2 total	1,759	1,184	67.3	1,214	69.0
C total	867	548	63.2	595	68.6
T total	1,356	914	67.4	914	67.4
Combined total	2,223	1,456	65.5	1,509	67.9

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

Table A-8. Student Survey Response Rates for Third Year Outcomes, Actual and Effective

	Impact Sample Members	Actual Respondents	Actual Response Rate (%)	Effective Respondents	Effective Response Rate (%)
Cohort 1 C	173	88	50.9	105	60.7
Cohort 1 T	290	191	65.9	191	65.9
Cohort 2 C	500	329	65.8	346	69.3
Cohort 2 T	754	508	67.4	508	67.4
Cohort 1 total	463	279	60.3	296	63.9
Cohort 2 total	1,254	837	66.7	854	68.1
C total	673	417	62.0	452	67.1
T total	1,044	699	67.0	699	67.0
Combined total	1,717	1,116	65.0	1,151	67.0

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample totals or response rate calculations.

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 despite the study attrition or that acceptable methods have been used to re-equate the study samples (What Works Clearinghouse 2006, pp. 6-7). In practice, the WWC considers overall sample responses that are below 70 percent, or rates that differ between the treatment and control group by more than 5 percentiles, as constituting a possible attrition problem. The test score effective response rate differential in year 3 met the WWC standard of less than 5 percentage points for all three major data collection instruments. The overall response rates for year 3 data collection of 69 percent (student tests), 68 percent (parent surveys), and 67 percent (students surveys) were just short of the WWC standard of 70 percent. In this study, the nonresponse 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 3 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 nonrespondents (with X_i equal to 2.0 for cohort 1 (100 divided by 50) and 1.992 for cohort 2 (253 divided by 127) for this set, and equal to 1 otherwise), NR_j is the nonresponse adjustment (the reciprocal of the response rate) for the classification cell to which student *i* belongs, and TR_i is the

trimming adjustment (usually equal to 1, but in some cases equal to 4.5 times median cutoff divided by the untrimmed weight).

Subgroup Sample Sizes and Response Rates

Because this evaluation examines programmatic impacts across a predefined set of participant subgroups, study response rates and subsequent analytic sample sizes are presented for each of those subgroups and for all three primary data collection instruments (student tests, parent surveys, and student surveys). The year 3 subgroup-level effective response rates for student test scores ranged from a low of 59 percent for participants entering the high school grades at baseline to a high of 71 percent for students who attended non-SINI schools at baseline (table A-9). The subgroup of students entering a high school grade at baseline was the smallest subgroup sample size for the analysis, 192 observations compared to 1,330 observations in the K-8 subgroup.

Table A-9. Effective Test Score Response Rates for Third Year Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	966	632	65.4
SINI never	1,257	890	70.8
Lower performance	737	490	66.5
Higher performance	1,486	1,032	69.4
Male	1,104	745	67.5
Female	1,119	777	69.4
K-8	1,896	1,330	70.1
9-12	327	192	58.7
Cohort 2	1,759	1,231	70.1
Cohort 1	464	291	62.7

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

The year 3 subgroup-level effective response rates for parent surveys ranged from a low of 57 percent for participants entering the high school grades at baseline to a high of 70 percent for their counterparts entering grades K-8 at baseline and students from non-SINI schools (table A-10).

The year 3 subgroup-level effective response rates for student surveys ranged from a low of 58 percent for participants in the 9-12 subgroup to a high of 70 percent for the students in the SINI-never subgroup (table A-11).

Table A-10. Effective Parent Survey Response Rates for Third Year Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	966	625	64.7
SINI never	1,257	884	70.3
Lower performance	737	485	65.8
Higher performance	1,486	1,024	68.9
Male	1,104	744	67.4
Female	1,119	765	68.3
K-8	1,896	1,324	69.8
9-12	327	185	56.6
Cohort 2	1,759	1,214	69.0
Cohort 1	464	295	63.6

NOTES: A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

Table A-11. Effective Student Survey Response Rates for Third Year Outcomes, by Subgroup

	Impact Sample Members	Effective Respondents	Effective Response Rate (%)
SINI ever	884	569	64.3
SINI never	833	582	69.9
Lower performance	572	368	64.3
Higher performance	1,145	783	68.4
Male	860	568	66.1
Female	857	582	68.0
K-8	1,389	959	69.0
9-12	328	192	58.4
Cohort 2	1,254	854	68.1
Cohort 1	463	296	63.9

NOTES: Student surveys administered to students in grades 4-12. A total of 85 students initially in the impact sample were no longer grade eligible for the Program by year 3. These "grade outs" were not invited to data collection events and are not counted in the impact sample subgroup totals or response rate calculations.

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 Opportunity 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:

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

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

[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.

[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.

schooling.[29] Thus, the models that produced the regression-adjusted impact estimates for this analysis took the general form:[30]

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

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:

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

where *P* is an index for whether a student is a member of a particular subgroup (the *P* must be part of the X's). 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:

---

[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.

[31] After the initial impacts were obtained in the year 1 impact analysis, a second set of estimates were run to test the sensitivity of the results to the set of covariates included in the model. This sensitivity model used only cohort, grade, special needs, number of children in the household, African American race, baseline reading, baseline math, and days until test as control variables, as these variables tended to be significant predictors of test score outcomes in the first set of models. No important differences regarding test score impacts were found (Wolf et al. 2007, pp. 43, 49-50). As a result and upon the recommendation of our Expert Advisory Panel, the limited covariate model was subsequently dropped from the sensitivity testing.

- 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.

Computation of Standard Errors

In computing standard errors it is necessary to factor in the stratified sample design, clustering of student outcomes within individual families, and nonresponse adjustments. As a consequence, all of the impact analyses were completed using sampling weights in STATA.[33] The effects of family clustering, which is not part of the sample design, but which may have a measurable effect on variance, were taken into account using robust regression calculations (i.e., "sandwich" variance estimates) (see Liang and Zeger 1986; White 1982).[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

[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.

[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 were currently attending. Baseline school clustering reduced the standard errors of the various impact estimates by an average of 2 percent, compared to an average reduction of less than 1 percent due to clustering by family. These results indicate that the student outcome data are almost totally independent of the most likely sources of outcome clustering. This may appear to be 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.

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 3 outcome test scores, 86 percent had used a scholarship for all or part of the 3 years after random assignment. The 14 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 the random assignment, even though scholarship decliners likely experienced no impact from the Program. Fortunately, there is a way to estimate the impact of the OSP on the average participant who actually used a scholarship, or what we refer to as the IOT estimate. This approach does not require information about why 14 percent of the individuals declined to use the scholarship when awarded, or how they differ from other families and children in the sample. But if one can assume that decliners experience zero impact from the scholarship Program, which seems reasonable given that they did not use the scholarship, it is possible to avoid these kinds of assumptions about (or analyses of) selection into and out of the Program.

This is possible by using the original comparison of **all** treatment group members to **all** control group members (i.e., the ITT estimates described above) but re-scaling it to account for the fact that a known fraction of the treatment group members did not actually avail themselves of the treatment and therefore experienced zero impact from the treatment. The average treatment impact that was generated from a mix of treatment users and nonusers is attributed only to the treatment users, by dividing the average treatment impact by the proportion of the treatment group who used their scholarships. For this report, depending on the specific outcome being rescaled, this "Bloom adjustment" (Bloom 1984) will increase the size of the ITT impacts by 11-43 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 70-90 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 and

[35] The Bloom adjustment is generated by dividing the ITT estimate by the usage rate for that outcome. Any number that is divided by .70 will generate a dividend that is 43 percent larger. Any number that is divided by .90 will generate a dividend that is 11 percent larger.

lotteries made it possible for some control group members to attend participating private schools, above and beyond the rate at which low-income students would have done so in the absence of the Program. Statistical techniques that take this "program-enabled crossover" into account are necessary for testing the sensitivity of the evaluation's impact estimates.

In a social experiment, even as some students randomized into the treatment group will decline to use the treatment, some students randomized into the control group will obtain the treatment outside of the experiment. For example, in medical trials, this control group "crossover" to the treatment can occur when the participants in the control group purchase the equivalent of the experimental "treatment" drug over the counter and use it as members of the treatment group would. The fact that crossovers have obtained the treatment does not change their status as members of the control group—just as treatment decliners forever remain treatments—for two reasons: (1) changing control crossovers to treatments would undermine the initial random assignment, and (2) control crossover typically represents what would have happened absent the experimental program and therefore is an authentic part of the counterfactual that the control group produces for comparison. If not for the medical trial, the control crossovers would have obtained the similar drug over the counter anyway. Therefore, under normal conditions, any effect that the crossover to treatment has on members of the control group is factored into the ITT and Bloom-adjusted IOT estimates of impact as legitimate elements of the counterfactual.

In the case of the OSP experiment, control crossover takes place in the form of students in the control group attending private school. Among the members of the control group for whom we knew their school attended in year 3, 15.3 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 (17.0 percent) compared to those without treatment siblings (13.6 percent).[37] 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.

---

[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.

The data suggest that 1.7 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 13.6 percent of the control group students without treatment siblings are attending private schools, whereas 15.3 percent of the control group overall is in private schools. Since the 13.6 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 15.3 percent to arrive at the hypothesized Program-enabled crossover rate of 1.7 percent. To adjust for the fact that this small component of the control group may have actually received the private-schooling treatment by way of the Program, the estimates of the impact of scholarship use in chapter 3 include a "double-Bloom" adjustment. We rescale the pure ITT impacts that are statistically significant by an amount equal to the treatment decliner rate (~14 percent), as described above plus the estimated Program-enabled crossover rate (~1.7 percent) to generate the IOT estimates.

Appendix B
Benjamini-Hochberg Adjustments for Multiple Comparisons

The following series of tables (tables B-1 through B-39) present the original *p*-values from the significance tests conducted in the analysis for all outcome domains in which multiple comparisons were made that produced statistically significant results. The sources of the multiple comparisons were either various subgroups of the impact sample (chapters 3 and 4), or the multiple comparisons made within the conceptual groupings of mediating effects (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 make up 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 math, student reports of school climate and safety, or student satisfaction in year 3, so there could not have been false discoveries in those domains. Significant impacts for the entire sample were observed regarding student outcomes in reading, parental perceptions of school climate and safety, and parental satisfaction with their child's school, but they were not the result of multiple comparisons. In chapter 4, no statistically significant impacts were found for the SINI-ever subgroup across the indicators of home educational supports and student motivation; the lower baseline performance subgroup across the indicators of home educational supports; the male subgroup across student motivation and engagement; the grade 9-12 subgroup across student motivation and school environment; and the cohort 1 subgroup across home education supports and student motivation. Thus, there could not have been false discoveries among those subgroups across those domains and no adjustments for multiple comparisons were applied to those particular subgroup results.

Table B-1. Multiple Comparisons Adjustments, Reading

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.59	.65
SINI never	.01**	.04*
Lower performance	.47	.59
Higher performance	.02*	.05*
Male	.15	.21
Female	.04*	.08
K-8	.01**	.04*
9-12	.98	.98
Cohort 2	.09	.16
Cohort 1	.04*	.08

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-2. Multiple Comparisons Adjustments, Parental Perceptions of Safety and an Orderly School Climate

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.00**	.00**
SINI never	.00**	.00**
Lower performance	.01**	.01**
Higher performance	.00**	.00**
Male	.00**	.00**
Female	.00**	.00**
K-8	.00**	.00**
9-12	.02*	.02*
Cohort 2	.00**	.00**
Cohort 1	.03*	.03*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

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

Subgroup	Original *p*-value	False Discovery Rate *p*-value
SINI ever	.17	.21
SINI never	.00**	.00**
Lower performance	.21	.23
Higher performance	.00**	.00**
Male	.01**	.02*
Female	.01**	.02*
K-8	.00**	.00**
9-12	.88	.88
Cohort 2	.02*	.03*
Cohort 1	.00**	.00**

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-4. Multiple Comparisons Adjustments, Home Educational Supports

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.06	.12
Parent Aspirations	.69	.69
Out-of-school Tutor Usage	.00**	.02*
School Transit Time	.29	.39

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-5. Multiple Comparisons Adjustments, Student Motivation and Engagement

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.06	.17
Attendance	.36	.43
Tardiness	.19	.32
Reads for Fun	.03*	.17
Engagement in Extracurricular Activities	.62	.62
Frequency of Homework (days)	.21	.32

*Statistically significant at the 95 percent confidence level.

Table B-6. Multiple Comparisons Adjustments, Instructional Characteristics

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.83	.83
Teacher Attitude	.54	.67
Ability Grouping	.83	.83
Availability of Tutors	.00**	.00**
In-school Tutor Usage	.50	.67
Programs for Learning Problems	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.01*	.02*
Before-/After-School Programs	.09	.15
Enrichment Programs	.00**	.01**

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-7. Multiple Comparisons Adjustments, School Environment

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.46	.46
School Size	.00**	.00**
Percent Non-White	.13	.23
Peer Classroom Behavior	.17	.23

**Statistically significant at the 99 percent confidence level.

Table B-8. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for SINI-Ever Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.34	.42
Teacher Attitude	.97	.97
Ability Grouping	.12	.17
Availability of Tutors	.00**	.00**
In-school Tutor Usage	.86	.96
Programs for Learning Problems	.06	.10
Programs for English Language Learners	.00**	.01**
Programs for Advanced Learners	.00**	.01**
Before-/After-School Programs	.01*	.03*
Enrichment Programs	.05*	.10

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-9. Multiple Comparisons Adjustments, Impacts on School Environment for SINI-Ever Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.74	.74
School Size	.07	.15
Percent Non-White	.04*	.15
Peer Classroom Behavior	.50	.67

*Statistically significant at the 95 percent confidence level.

Table B-10. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for SINI-Never Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.04*	.07
Parent Aspirations	.88	.88
Out-of-school Tutor Usage	.02*	.07
School Transit Time	.46	.61

*Statistically significant at the 95 percent confidence level.

Table B-11. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for SINI-Never Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.08	.23
Attendance	.58	.63
Tardiness	.14	.27
Reads for Fun	.02*	.14
Engagement in Extracurricular Activities	.36	.54
Frequency of Homework (days)	.63	.63

*Statistically significant at the 95 percent confidence level.

Table B-12. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for SINI-Never Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.58	.65
Teacher Attitude	.43	.59
Ability Grouping	.22	.44
Availability of Tutors	.12	.30
In-school Tutor Usage	.46	.59
Programs for Learning Problems	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.47	.59
Before-/After-School Programs	.96	.96
Enrichment Programs	.01*	.04*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-13. Multiple Comparisons Adjustments, Impacts on School Environment for SINI-Never Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.15	.20
School Size	.00**	.00**
Percent Non-White	.51	.51
Peer Classroom Behavior	.03*	.05

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-14. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for Lower Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.01**	.05
Attendance	.07	.20
Tardiness	.33	.66
Reads for Fun	.91	.91
Engagement in Extracurricular Activities	.81	.91
Frequency of Homework (days)	.46	.69

**Statistically significant at the 99 percent confidence level.

Table B-15. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Lower Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.75	.86
Teacher Attitude	.86	.86
Ability Grouping	.79	.86
Availability of Tutors	.62	.86
In-school Tutor Usage	.20	.39
Programs for Learning Problems	.02*	.09
Programs for English Language Learners	.07	.24
Programs for Advanced Learners	.01*	.09
Before-/After-School Programs	.77	.86
Enrichment Programs	.17	.39

*Statistically significant at the 95 percent confidence level.

Table B-16. Multiple Comparisons Adjustments, Impacts on School Environment for Lower Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.76	.76
School Size	.00**	.00**
Percent Non-White	.61	.76
Peer Classroom Behavior	.63	.76

**Statistically significant at the 99 percent confidence level.

Table B-17. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for Higher Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.01**	.02*
Parent Aspirations	.97	.97
Out-of-school Tutor Usage	.02*	.05*
School Transit Time	.07	.10

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-18. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for Higher Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.69	.82
Attendance	.83	.83
Tardiness	.33	.67
Reads for Fun	.01*	.07
Engagement in Extracurricular Activities	.46	.69
Frequency of Homework (days)	.34	.67

*Statistically significant at the 95 percent confidence level.

Table B-19. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Higher Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.97	.97
Teacher Attitude	.52	.74
Ability Grouping	.69	.86
Availability of Tutors	.00**	.00**
In-school Tutor Usage	.88	.97
Programs for Learning Problems	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.17	.29
Before-/After-School Programs	.02*	.05*
Enrichment Programs	.01**	.02*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-20. Multiple Comparisons Adjustments, Impacts on School Environment for Higher Performance Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.28	.28
School Size	.00**	.00**
Percent Non-White	.14	.26
Peer Classroom Behavior	.20	.26

**Statistically significant at the 99 percent confidence level.

Table B-21. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for Male Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.48	.55
Parent Aspirations	.28	.55
Out-of-school Tutor Usage	.04*	.15
School Transit Time	.51	.51

*Statistically significant at the 95 percent confidence level.

Table B-22. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Male Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.97	.97
Teacher Attitude	.53	.66
Ability Grouping	.80	.89
Availability of Tutors	.07	.15
In-school Tutor Usage	.33	.46
Programs for Learning Problems	.02*	.07
Programs for English Language Learners	.00**	.02*
Programs for Advanced Learners	.00**	.00**
Before-/After-School Programs	.22	.37
Enrichment Programs	.05*	.12

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-23. Multiple Comparisons Adjustments, Impacts on School Environment for Male Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.99	.99
School Size	.04*	.08
Percent Non-White	.03*	.08
Peer Classroom Behavior	.08	.10

*Statistically significant at the 95 percent confidence level.

Table B-24. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for Female Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.04*	.09
Parent Aspirations	.57	.57
Out-of-school Tutor Usage	.04*	.09
School Transit Time	.38	.50

*Statistically significant at the 95 percent confidence level.

Table B-25. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for Female Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.06	.11
Attendance	.20	.30
Tardiness	.03*	.10
Reads for Fun	.03*	.10
Engagement in Extracurricular Activities	.78	.81
Frequency of Homework (days)	.81	.81

*Statistically significant at the 95 percent confidence level.

Table B-26. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Female Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.78	.99
Teacher Attitude	.80	.99
Ability Grouping	.95	.99
Availability of Tutors	.00**	.00**
In-school Tutor Usage	.99	.99
Programs for Learning Problems	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.83	.99
Before-/After-School Programs	.19	.38
Enrichment Programs	.01*	.03*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-27. Multiple Comparisons Adjustments, Impacts on School Environment for Female Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.27	.53
School Size	.00**	.00**
Percent Non-White	.77	.79
Peer Classroom Behavior	.79	.79

**Statistically significant at the 99 percent confidence level.

Table B-28. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for K-8 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.21	.28
Parent Aspirations	.81	.81
Out-of-school Tutor Usage	.01**	.02*
School Transit Time	.13	.25

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-29. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for K-8 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.08	.24
Attendance	.57	.57
Tardiness	.31	.46
Reads for Fun	.03*	.19
Engagement in Extracurricular Activities	.44	.52
Frequency of Homework (days)	.14	.28

*Statistically significant at the 95 percent confidence level.

Table B-30. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for K-8 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.40	.50
Teacher Attitude	.82	.82
Ability Grouping	.37	.50
Availability of Tutors	.01**	.03*
In-school Tutor Usage	.20	.33
Programs for Learning Problems/ELL	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.14	.27
Before-/After-School Programs	.53	.59
Enrichment Programs	.01**	.02*

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-31. Multiple Comparisons Adjustments, Impacts on School Environment for K-8 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.75	.75
School Size	.00**	.00**
Percent Non-White	.24	.32
Peer Classroom Behavior	.14	.28

**Statistically significant at the 99 percent confidence level.

Table B-32. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for 9-12 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.05*	.19
Parent Aspirations	.68	.68
Out-of-school Tutor Usage	.44	.59
School Transit Time	.41	.59

*Statistically significant at the 95 percent confidence level.

Table B-33. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for 9-12 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.00**	.00**
Teacher Attitude	.17	.20
Ability Grouping	.04*	.05
Availability of Tutors	.01**	.01*
In-school Tutor Usage	.20	.20
Programs for Learning Problems	N/A	N/A
Programs for English Language Learners	.01*	.02*
Programs for Advanced Learners	.00*	.01*
Before-/After-School Programs	.00**	.00**
Enrichment Programs	.14	.18

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-34. Multiple Comparisons Adjustments, Impacts on Home Educational Supports for Cohort 2 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parental Involvement	.10	.20
Parent Aspirations	.88	.88
Out-of-school Tutor Usage	.00**	.00**
School Transit Time	.40	.53

**Statistically significant at the 99 percent confidence level.

Table B-35. Multiple Comparisons Adjustments, Impacts on Student Motivation and Engagement for Cohort 2 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student Aspirations	.05	.16
Attendance	.28	.33
Tardiness	.17	.33
Reads for Fun	.03*	.16
Engagement in Extracurricular Activities	.74	.74
Frequency of Homework (days)	.25	.33

*Statistically significant at the 95 percent confidence level.

Table B-36. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Cohort 2 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.62	.62
Teacher Attitude	.61	.62
Ability Grouping	.27	.33
Availability of Tutors	.01**	.01*
In-school Tutor Usage	.25	.33
Programs for Learning Problems/ELL	.00**	.00**
Programs for English Language Learners	.00**	.00**
Programs for Advanced Learners	.00**	.01*
Before-/After-School Programs	.03*	.05*
Enrichment Programs	.00**	.00**

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-37. Multiple Comparisons Adjustments, Impacts on School Environment for Cohort 2 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.18	.24
School Size	.00**	.00**
Percent Non-White	.32	.32
Peer Classroom Behavior	.15	.24

**Statistically significant at the 99 percent confidence level.

Table B-38. Multiple Comparisons Adjustments, Impacts on Instructional Characteristics for Cohort 1 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Student/Teacher Ratio	.07	.13
Teacher Attitude	.72	.80
Ability Grouping	.03*	.11
Availability of Tutors	.00**	.04*
In-school Tutor Usage	.49	.62
Programs for Learning Problems/ELL	.05*	.12
Programs for English Language Learners	.01**	.05*
Programs for Advanced Learners	.91	.91
Before-/After-School Programs	.14	.24
Enrichment Programs	.40	.57

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

Table B-39. Multiple Comparisons Adjustments, Impacts on School Environment for Cohort 1 Subgroup

Intermediate Outcome	Original *p*-value	False Discovery Rate *p*-value
Parent/School Communication	.49	.65
School Size	.02*	.08
Percent Non-White	.14	.27
Peer Classroom Behavior	.93	.93

*Statistically significant at the 95 percent confidence level.

Appendix C
Sensitivity Testing

In any evaluation, decisions are made about how to handle certain data or analysis issues (e.g., nonresponse differentials, sampling weights, etc.). While there are some commonly accepted approaches in research and evaluation methodology, sometimes there are multiple approaches, and any could be acceptable. The evaluation team chose its approach in consultation with a panel of methodology experts before analyzing the data and seeing the results. However, in an effort to be both transparent and complete, each presentation of analyses is followed by a discussion of the sensitivity testing conducted to determine how robust the estimates are to specific changes in the analytic approach. These different specifications include:

- *Trimmed sample*: 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 group nonrespondents. Since the actual response rate of the treatment group was higher (68 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 58 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 nonresponse weighting of subsampled initial nonrespondents.

- *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.

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

Neither sensitivity test produced changes in the overall findings for reading and math impacts (table C-1). For subgroup reading impacts, the findings were not sensitive to the trimmed sample analysis. 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 generated results that differed from the primary analysis in two of five subgroup estimations. Both the female subgroup and cohort 1 subgroup reading impacts are not statistically significant when estimated using this method.

Table C-1. Year 3 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	4.46*	.01	5.21**	.01	4.46*	.03
Full sample: math	.81	.62	1.51	.40	.81	.64
SINI never: reading	6.57**	.01	8.14**	.00	6.57**	.01
Higher performing: reading	5.45*	.02	5.50*	.02	5.45*	.03
Female: reading	5.07*	.04	6.67*	.01	5.07	.06
K-8: reading	5.23**	.01	5.71**	.01	5.23**	.01
Cohort 1: reading	8.70*	.04	11.17*	.03	8.70	.10

*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 = 1,460; math = 1,468. Trimmed sample valid *N* for reading = 1,296; math = 1,303. Separate reading and math sample weights were used.

In sum, the finding from the primary analysis of no significant programmatic impact overall on math achievement but a significant impact overall on reading achievement was consistent across the analysis approaches. The finding from the primary analysis of a significant programmatic impact in reading for five subgroups was consistent across specifications, except in the case of clustering on current school attended for the female and cohort 1 subgroups.

Sensitivity Checks for ITT Impacts on Parent Perceptions of Safety and an Orderly School Climate

The programmatic impacts on parental reports of safety and an orderly school climate discussed in chapter 3 were consistent across analytic approaches with one exception (table C-2). The positive impact of the Program on parent perceptions of safety and an orderly school climate, statistically significant for the cohort 1 subgroup in the primary analysis, loses significance when estimated using the smaller trimmed sample.

Table C-2. Year 3 Parent Perceptions of Safety and an Orderly School Climate: 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 safety and climate: parents	1.01**	.00	.99**	.00	1.01**	.00
SINI ever	1.16**	.00	1.15**	.00	1.16**	.00
SINI never	.90**	.00	.87**	.00	.90**	.00
Lower performance	1.02**	.01	.98**	.01	1.02**	.01
Higher performance	1.01**	.00	1.00**	.00	1.01**	.00
Male	1.06**	.00	.86**	.00	1.06**	.00
Female	.96**	.00	1.12**	.00	.96**	.00
K-8	.93**	.00	.86**	.00	.93**	.00
9-12	1.51*	.02	1.84**	.01	1.51*	.03
Cohort 2	.97**	.00	1.01**	.00	.97**	.00
Cohort 1	1.20*	.03	.92	.13	1.20*	.05

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Original estimates valid $N = 1{,}423$. Trimmed sample valid $N = 1{,}248$. Parent survey weights were used.

Sensitivity Checks for ITT Impacts on Student Reports of Safety and an Orderly School Climate

The primary analysis discussed in chapter 3 found no treatment impact on students' perceptions of a safe school climate. This result is consistent across different analytic approaches (table C-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 safety.

Table C-3. Year 3 Student Reports of Safety and an Orderly School Climate: 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 safety and climate: students	.12	.36	.07	.64	.12	.36

NOTES: Original estimates valid *N* = 1,098. Trimmed sample valid N = 968. Student survey weights were used.

Sensitivity Checks for ITT Impacts on Parent Reports of School Satisfaction

The finding of a positive impact of the Program on parent satisfaction for the full sample and for 7 of 10 subgroups was not sensitive to different analytic approaches with one exception (table C-4). The positive impact of the Program on parents likelihood of grading their child's school A or B, statistically significant for the female student subgroup in the primary analysis, loses significance when estimated using the smaller trimmed sample.

Table C-4. Year 3 Parent Satisfaction ITT Impact Estimates and *P*-Values with Different Specifications

Parent gave school grade of A or B	Original Estimates		Trimmed Sample		Clustering on Current School	
	Impact	*p*-value	Impact	*p*-value	Impact	*p*-value
Full sample	.11**	.00	.08**	.01	.11**	.00
SINI never	.14**	.00	.11**	.01	.14**	.00
Higher performance	.13**	.00	.12**	.00	.13**	.00
Male	.11**	.01	.08*	.04	.11*	.02
Female	.10**	.01	.07	.07	.10*	.01
K-8	.13**	.00	.10**	.00	.13**	.00
Cohort 2	.08*	.02	.07*	.05	.08*	.02
Cohort 1	.20**	.00	.12*	.05	.20**	.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 = 1,410. Trimmed sample valid *N* for school grade = 1,239. Parent survey weights were used. Impact estimates reported for the dichotomous variable "parents who gave school a grade of A or B" are reported as marginal effects.

Sensitivity Checks for ITT Impacts on Student Reports of School Satisfaction

The results of the primary analysis found no programmatic impact on overall student self-reports of satisfaction. That finding is consistent across the different methodological approaches (table C-5). In every specification, there are no differences in the likelihood of a student grading his/her school A or B.

Table C-5. Year 3 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
Student gave school a grade of A or B	-.03	.41	-.04	.26	-.03	.49

NOTES: Original estimates valid *N* for school grade = 1,014. Trimmed sample valid *N* for school grade = 897. Student survey weights were used. Impact estimates are reported as marginal effects. Survey given to students in grades 4-12.

Appendix D
Detailed ITT Tables

Table D-1. Year 3 Test Score ITT Impacts: Reading

	Unadjusted 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	635.69	630.98	4.71	.16	4.46*	.01	.13
	(35.06)	(34.52)	(3.35)		(1.82)		(34.52)
Subgroups							
SINI ever	655.11	648.25	6.87	.11	1.52	.59	.05
	(31.74)	(32.79)	(4.33)		(2.78)		(32.79)
SINI never	621.90	618.72	3.18	.49	6.57**	.01	.19
	(36.04)	(34.60)	(4.64)		(2.40)		(34.60)
Difference	33.22	29.53	3.69	.56	-5.05	.17	-.15
	(3.95)	(4.96)	(6.31)		(3.69)		(34.52)
Lower performance	613.19	612.38	0.81	.16	2.10	.47	.07
	(26.13)	(28.08)	(5.18)		(2.93)		(28.08)
Higher performance	646.57	639.29	7.28	.08	5.45*	.02	.17
	(33.31)	(32.37)	(4.08)		(2.23)		(32.37)
Difference	-33.37	-26.90	-6.47	.33	-3.35	.35	-.10
	(4.21)	(5.16)	(6.62)		(3.61)		(34.52)
Male	633.08	627.48	5.60	.23	3.83	.15	.11
	(34.03)	(34.09)	(4.69)		(2.64)		(34.09)
Female	638.31	634.24	4.07	.39	5.07*	.04	.15
	(35.22)	(33.24)	(4.74)		(2.46)		(33.24)
Difference	-5.23	-6.75	1.53	.82	-1.23	.73	-.04
	(4.22)	(5.07)	(6.65)		(3.58)		(34.52)
K-8	627.41	622.07	5.34	.12	5.23**	.01	.15
	(35.75)	(35.40)	(3.44)		(1.97)		(35.40)
9-12	685.01	682.50	2.51	.64	-.10	.98	-.00
	(30.93)	(29.39)	(5.31)		(3.92)		(29.39)
Difference	-57.61	-60.43	2.83	.65	5.33	.21	.15
	(4.90)	(3.99)	(6.21)		(4.25)		(34.52)
Cohort 2	625.41	622.27	3.14	.40	3.37	.09	.09
	(35.46)	(35.45)	(3.73)		(2.01)		(35.45)
Cohort 1	674.84	664.17	10.67*	.05	8.70*	.04	.31
	(33.09)	(27.72)	(5.36)		(4.16)		(27.72)
Difference	-49.42	-41.90	-7.53	.25	-5.34	.25	-.15
	(3.95)	(5.24)	(6.53)		(4.59)		(34.52)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for reading = 1,460. Reading sample weights were used.

Table D-2. Year 3 Test Score ITT Impacts: Math

	Unadjusted 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	629.50	629.34	.15	.96	.81	.62	.03
	(29.82)	(31.70)	(3.39)		(1.64)		(31.70)
Subgroups							
SINI ever	651.91	646.56	5.35	.23	.17	.95	.01
	(26.07)	(27.73)	(4.46)		(2.51)		(27.73)
SINI never	613.61	617.12	-3.51	.45	1.27	.58	.04
	(31.57)	(33.59)	(4.63)		(2.26)		(33.59)
Difference	38.29	29.44	8.85	.17	-1.10	.75	-.03
	(3.90)	(5.16)	(6.38)		(3.45)		(31.70)
Lower performance	612.98	615.08	-2.10	.71	.35	.90	.01
	(21.41)	(23.96)	(5.72)		(2.79)		(23.96)
Higher performance	637.55	635.65	1.90	.64	.98	.64	.03
	(30.85)	(31.06)	(4.09)		(2.09)		(31.06)
Difference	-24.57	-20.57	-4.00	.57	-.63	.86	-.02
	(4.47)	(5.38)	(7.05)		(3.57)		(31.70)
Male	629.77	629.31	0.46	.93	.04	.99	.00
	(29.29)	(31.14)	(4.97)		(2.44)		(31.14)
Female	629.22	629.38	-0.15	.97	1.54	.50	.05
	(29.46)	(30.73)	(4.62)		(2.25)		(30.73)
Difference	0.55	-0.07	0.61	.93	-1.50	.65	-.05
	(4.31)	(5.17)	(6.78)		(3.34)		(31.70)
K-8	620.76	620.73	0.03	.99	1.01	.57	.03
	(30.87)	(33.37)	(3.56)		(1.79)		(33.37)
9-12	681.56	679.18	2.38	.55	-.41	.92	-.02
	(23.60)	(22.07)	(3.99)		(3.93)		(22.07)
Difference	-60.80	-58.45	-2.35	.66	1.42	.74	.04
	(3.92)	(3.64)	(5.26)		(4.28)		(31.70)
Cohort 2	618.18	619.53	-1.35	.73	-.21	.91	-.01
	(31.54)	(34.21)	(3.84)		(1.86)		(34.21)
Cohort 1	672.63	666.74	5.89	.15	4.74	.19	.23
	(22.67)	(20.77)	(4.09)		(3.59)		(20.77)
Difference	-54.45	-47.21	-7.24	.20	-4.95	.23	-.16
	(3.44)	(4.56)	(5.61)		(4.08)		(31.70)

NOTES: Impacts displayed in terms of scale scores and effect sizes in terms of standard deviations. Valid *N* for math = 1,468. Math sample weights were used.

Table D-3. Year 3 Parental Perceptions of School Safety and Climate: ITT Impacts

Parental Perceptions of Safety and an Orderly School Climate (0-10 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	8.18	7.06	1.12**	.00	1.01**	.00	.29
	(3.03)	(3.50)	(.21)		(.20)		(3.50)
Subgroups							
SINI ever	7.92	6.75	1.17**	.00	1.16**	.00	.32
	(3.17)	(3.67)	(.35)		(.34)		(3.67)
SINI never	8.37	7.29	1.08**	.00	.90**	.00	.27
	(2.92)	(3.36)	(.25)		(.25)		(3.36)
Difference	-.45	-.54	.09	.84	.26	.53	. 07
	(.24)	(.36)	(.43)		(.42)		(3.50)
Lower performance	7.85	6.80	1.05**	.01	1.02**	.01	.28
	(3.18)	(3.62)	(.38)		(.36)		(3.62)
Higher performance	8.34	7.18	1.16**	.00	1.01**	.00	.29
	(2.94)	(3.45)	(.25)		(.25)		(3.45)
Difference	-.49	-.39	-.11	.81	.01	.99	.00
	(.26)	(.38)	(.45)		(.43)		(3.50)
Male	8.19	7.06	1.14**	.00	1.06**	.00	.30
	(2.98)	(3.52)	(.30)		(.30)		(3.52)
Female	8.17	7.07	1.10**	.00	.96**	.00	.28
	(3.08)	(3.49)	(.27)		(.26)		(3.49)
Difference	.03	-.01	.04	.92	.10	.78	.03
	(.22)	(.34)	(.39)		(.38)		(3.50)
K-8	8.31	7.28	1.04**	.00	.93**	.00	.27
	(2.98)	(3.40)	(.22)		(.21)		(3.40)
9-12	7.33	5.78	1.55*	.01	1.51*	.02	.40
	(3.20)	(3.82)	(.62)		(.62)		(3.82)
Difference	.98	1.49	-.51	.43	-.58	.37	-.17
	(.49)	(.44)	(.65)		(.65)		(3.50)
Cohort 2	8.37	7.32	1.04**	.00	.97**	.00	.29
	(2.90)	(3.37)	(.22)		(.21)		(3.37)
Cohort 1	7.47	6.06	1.41*	.01	1.20*	.03	.31
	(3.40)	(3.80)	(.56)		(.56)		(3.80)
Difference	.90	1.27	-.37	.53	-.23	.70	-.07
	(.34)	(.51)	(.60)		(.60)		(3.50)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Effect sizes are in terms of standard deviations. Valid N = 1,423. Parent survey weights were used.

Table D-4. Year 3 Student Reports of School Safety and Climate: ITT Impacts

Student Perceptions of Safety and an Orderly School Climate (0-8 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	6.29	6.06	.23	.10	.12	.36	.06
	(1.68)	(1.90)	(.14)		(.13)		(1.90)
Subgroups							
SINI ever	6.24	5.99	.25	.24	.08	.66	.04
	(1.67)	(2.04)	(.21)		(.19)		(2.04)
SINI never	6.32	6.11	.21	.25	.14	.41	.08
	(1.68)	(1.79)	(.18)		(.17)		(1.79)
Difference	-.08	-.12	.04	.90	-.06	.83	-.03
	(.14)	(.24)	(.28)		(.26)		(1.90)
Lower performance	6.24	5.90	.34	.17	.14	.55	.07
	(1.70)	(2.04)	(.25)		(.24)		(2.04)
Higher performance	6.30	6.13	.18	.30	.10	.50	.06
	(1.67)	(1.83)	(.17)		(.15)		(1.83)
Difference	-.06	-.22	.16	.59	.04	.90	.02
	(.15)	(.26)	(.30)		(.29)		(1.90)
Male	6.11	5.86	.25	.20	.10	.59	.05
	(1.76)	(1.95)	(.19)		(.18)		(1.95)
Female	6.45	6.25	.20	.27	.13	.44	.07
	(1.57)	(1.83)	(.18)		(.17)		(1.83)
Difference	-.34	-.39	.05	.85	-.04	.88	-.02
	(.14)	(.22)	(.26)		(.25)		(1.90)
4-8	6.24	6.01	.22	.14	.12	.39	.06
	(1.68)	(1.91)	(.15)		(.14)		(1.91)
9-12	6.57	6.32	.25	.41	.11	.74	.06
	(1.63)	(1.80)	(.31)		(.31)		(1.80)
Difference	-.33	-.30	-.03	.93	.01	.97	.01
	(.25)	(.22)	(.33)		(.34)		(1.90)
Cohort 2	6.30	6.02	.28	.06	.16	.29	.08
	(1.62)	(1.89)	(.15)		(.15)		(1.89)
Cohort 1	6.24	6.22	.02	.95	-.04	.88	-.02
	(1.88)	(1.92)	(.35)		(.28)		(1.92)
Difference	.06	-.20	.26	.50	.20	.54	.10
	(.19)	(.34)	(.38)		(.32)		(1.90)

NOTES: Effect sizes are in terms of standard deviations. Valid N = 1,098. Student survey weights were used. Survey given to students in grades 4-12.

Table D-5. Year 3 Parental Satisfaction ITT Impacts: Parents Who Gave School a Grade of A or B

Parents Who Gave School a Grade of A or B	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	.77	.63	.13**	.00	.11**	.00	.22
	(.42)	(.48)	(.03)		(.03)		(.48)
Subgroups							
SINI ever	.72	.63	.09	.06	.06	.16	.13
	(.45)	(.48)	(.05)		(.04)		(.48)
SINI never	.80	.64	.17**	.00	.14**	.00	.29
	(.40)	(.48)	(.04)		(.04)		(.48)
Difference	-.09	-.01	-.08	.21	-.09	.17	-.18
	(.04)	(.05)	(.07)		(.06)		(.48)
Lower performance	.68	.57	.10*	.04	.06	.21	.12
	(.47)	(.50)	(.05)		(.05)		(.50)
Higher performance	.81	.66	.15**	.00	.13**	.00	.27
	(.39)	(.47)	(.03)		(.04)		(.47)
Difference	-.14	-.09	-.06	.40	-.07	.26	-.15
	(.04)	(.05)	(.06)		(.07)		(.48)
Male	.75	.60	.14**	.00	.11**	.01	.22
	(.43)	(.49)	(.04)		(.04)		(.49)
Female	.79	.67	.12**	.00	.10**	.01	.22
	(.41)	(.47)	(.04)		(.04)		(.47)
Difference	-.05	-.06	.02	.78	.01	.93	.01
	(.04)	(.04)	(.06)		(.05)		(.48)
K-8	.79	.64	.15**	.00	.13**	.00	.27
	(.41)	(.48)	(.03)		(.03)		(.48)
9-12	.63	.60	.02	.77	-.01	.88	-.03
	(.48)	(.49)	(.08)		(.08)		(.49)
Difference	.18	.03	.13	.12	.14	.11	.28
	(.08)	(.05)	(.08)		(.08)		(.48)
Cohort 2	.78	.68	.11**	.00	.08*	.02	.16
	(.42)	(.47)	(.03)		(.03)		(.47)
Cohort 1	.72	.48	.22**	.00	.20**	.00	.41
	(.45)	(.50)	(.07)		(.06)		(.50)
Difference	.06	.19	-.12	.13	-.13	.06	-.27
	(.05)	(.07)	(.08)		(.07)		(.48)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for school grade = 1,410. 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 3 Parental Satisfaction ITT Impacts: Average Grade Parent Gave School

Average Grade Parent Gave School (5.0 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	4.05	3.79	.26**	.00	.20**	.00	.22
	(.96)	(1.11)	(.07)		(.07)		(1.11)
Subgroups							
SINI ever	3.97	3.76	.20	.07	.13	.19	.12
	(1.01)	(1.09)	(.11)		(.10)		(1.09)
SINI never	4.12	3.81	.30**	.00	.25**	.00	.23
	(.92)	(1.12)	(.09)		(.08)		(1.12)
Difference	-.15	-.05	-.10	.46	-.12	.36	-.11
	(.08)	(.11)	(.14)		(.13)		(1.11)
Lower performance	3.80	3.66	.15	.26	.06	.61	.05
	(1.05)	(1.17)	(.13)		(.12)		(1.17)
Higher performance	4.18	3.85	.32**	.00	.27**	.00	.25
	(.88)	(1.07)	(.08)		(.07)		(1.07)
Difference	-.37	-.20	-.18	.23	-.21	.14	-.19
	(.08)	(.12)	(.15)		(.14)		(1.11)
Male	4.02	3.76	.26**	.01	.19*	.04	.18
	(1.01)	(1.09)	(.10)		(.09)		(1.09)
Female	4.09	3.83	.26**	.01	.21*	.02	.19
	(.91)	(1.12)	(.09)		(.09)		(1.12)
Difference	-.07	-.06	-.01	.96	-.02	.88	-.02
	(.07)	(.11)	(.13)		(.13)		(1.11)
K-8	4.10	3.82	.28**	.00	.23**	.00	.21
	(.95)	(1.10)	(.07)		(.07)		(1.10)
9-12	3.75	3.63	.12	.52	.04	.84	.03
	(.98)	(1.11)	(.19)		(.19)		(1.11)
Difference	.35	.19	.16	.42	.19	.33	.17
	(.16)	(.14)	(.20)		(.20)		(1.11)
Cohort 2	4.10	3.87	0.22**	.00	.16*	.02	.15
	(.93)	(1.06)	(.07)		(.07)		(1.07)
Cohort 1	3.90	3.49	0.41*	.02	.36*	.02	.30
	(1.05)	(1.20)	(.17)		(.16)		(1.20)
Difference	.20	.38	-0.19	.32	-.19	.26	-.17
	(.11)	(.16)	(.19)		(.17)		(1.11)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for school grade = 1,410. Parent survey weights were used.

Table D-7. Year 3 Parental Satisfaction ITT Impacts: School Satisfaction Scale

School Satisfaction Scale (IRT Scored .05-3.0)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	2.14	1.95	.20**	.00	.17**	.00	.24
	(.67)	(.72)	(.05)		(.04)		(.72)
Subgroups							
SINI ever	2.09	1.90	.18*	.02	.15*	.03	.20
	(.70)	(.76)	(.07)		(.07)		(.76)
SINI never	2.19	1.98	.21**	.00	.18**	.00	.26
	(.64)	(.70)	(.06)		(.06)		(.70)
Difference	-.10	-.08	-.03	.78	-.03	.76	-.04
	(.05)	(.08)	(.09)		(.09)		(.72)
Lower performance	2.05	1.91	.15*	.05	.12	.13	.15
	(.69)	(.76)	(.07)		(.08)		(.76)
Higher performance	2.19	1.96	.23**	.00	.19**	.00	.27
	(.65)	(.71)	(.06)		(.05)		(.71)
Difference	-.14	-.06	-.08	.39	-.08	.38	-.11
	(.05)	(.08)	(.09)		(.09)		(.72)
Male	2.12	1.95	.17**	.01	.14*	.03	.19
	(.69)	(.71)	(.06)		(.06)		(.71)
Female	2.17	1.95	.23**	.00	.20**	.00	.27
	(.65)	(.73)	(.06)		(.06)		(.73)
Difference	-.05	.00	-.06	.50	-.07	.43	-.09
	(.05)	(.07)	(.09)		(.08)		(.72)
K-8	2.17	1.96	.22**	.00	.19**	.00	.25
	(.67)	(.74)	(.05)		(.05)		(.74)
9-12	1.96	1.89	.07	.58	.04	.72	.07
	(.66)	(.60)	(.12)		(.11)		(.60)
Difference	.22	.06	.15	.24	.15	.24	.20
	(.11)	(.08)	(.13)		(.12)		(.72)
Cohort 2	2.18	1.97	.21**	.00	.18**	.00	.26
	(.65)	(.71)	(.05)		(.05)		(.71)
Cohort 1	2.01	1.85	.16	.21	.11	.35	.14
	(.74)	(.78)	(.12)		(.12)		(.78)
Difference	.17	.12	.05	.69	.07	.55	.10
	(.08)	(.11)	(.13)		(.12)		(.72)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for parent satisfaction = 1,438. Parent survey weights were used.

Table D-8. Year 3 Student Satisfaction ITT Impacts: Students Who Gave School a Grade of A or B

Students Who Gave School a Grade of A or B	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	.72 (.45)	.73 (.44)	-.01 (.03)	.71	-.03 (.03)	.41	-.06 (.44)
Subgroups							
SINI ever	.64 (.48)	.72 (.45)	-.07 (.04)	.12	-.08 (.04)	.07	-.18 (.45)
SINI never	.78 (.42)	.74 (.44)	.04 (.05)	.41	.02 (.04)	.60	.05 (.44)
Difference	-.13 (.04)	-.02 (.05)	-.11 (.07)	.12	-.11 (.07)	.10	-.25 (.44)
Lower performance	.64 (.48)	.68 (.47)	-.04 (.07)	.50	-.06 (.06)	.29	-.13 (.47)
Higher performance	.75 (.43)	.75 (.43)	-.00 (.04)	1.00	-.01 (.04)	.81	-.02 (.43)
Difference	-.11 (.04)	-.07 (.06)	-.04 (.07)	.60	-.05 (.07)	.46	-.12 (.44)
Male	.68 (.47)	.72 (.45)	-.04 (.04)	.34	-.05 (.04)	.27	-.11 (.45)
Female	.76 (.43)	.74 (.44)	.02 (.05)	.62	-.00 (.04)	.95	-.01 (.44)
Difference	-.08 (.04)	-.02 (.05)	-.07 (.07)	.32	-.05 (.07)	.46	-.11 (.44)
4-8	.75 (.44)	.76 (.43)	-.02 (.04)	.67	-.03 (.03)	.44	-.06 (.43)
9-12	.57 (.50)	.57 (.50)	-.00 (.07)	.99	-.02 (.07)	.76	-.04 (.50)
Difference	.18 (.07)	.19 (.06)	-.01 (.08)	.86	-.01 (.07)	.94	-.01 (.44)
Cohort 2	.75 (.37)	.78 (.38)	-.03 (.04)	.42	-.05 (.04)	.24	-.11 (.38)
Cohort 1	.59 (.43)	.56 (.44)	.03 (.06)	.63	.03 (.05)	.63	.05 (.44)
Difference	.16 (.05)	.22 (.07)	-.06 (.07)	.41	-.07 (.06)	.28	-.16 (.44)

NOTES: Valid *N* for school grade = 1,014. 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 3 Student Satisfaction ITT Impacts: Average Grade Student Gave School

Average Grade Student Gave School (5.0 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	3.98	3.99	-.01	.84	-.03	.68	-.03
	(.97)	(.91)	(.07)		(.07)		(.91)
Subgroups							
SINI ever	3.81	3.93	-.12	.24	-.13	.20	-.13
	(.95)	(.97)	(.10)		(.10)		(.97)
SINI never	4.11	4.04	.07	.49	.05	.60	.06
	(.98)	(.86)	(.10)		(.09)		(.86)
Difference	-.29	-.11	-.18	.19	-.18	.19	-.20
	(.09)	(.11)	(.14)		(.13)		(.91)
Lower performance	3.87	3.98	-.10	.43	-.13	.32	-.14
	(.99)	(.90)	(.13)		(.13)		(.90)
Higher performance	4.02	4.00	.02	.79	.01	.86	.02
	(.96)	(.92)	(.08)		(.08)		(.92)
Difference	-.15	-.03	-.12	.43	-.14	.35	-.15
	(.10)	(.12)	(.15)		(.15)		(.91)
Male	3.93	4.03	-.10	.31	-.08	.37	-.09
	(.96)	(.88)	(.10)		(.09)		(.88)
Female	4.03	3.96	.07	.51	.02	.85	.02
	(.99)	(.94)	(.10)		(.10)		(.94)
Difference	-.10	.06	-.16	.24	-.10	.45	-.11
	(.09)	(.11)	(.14)		(.13)		(.91)
4-8	4.06	4.07	-.01	.87	-.02	.77	-.02
	(.95)	(.88)	(.07)		(.07)		(.88)
9-12	3.56	3.59	-.03	.88	-.07	.71	-.07
	(1.03)	(.98)	(.19)		(.19)		(.98)
Difference	.50	.48	.02	.93	.05	.80	.05
	(.16)	(.13)	(.20)		(.19)		(.91)
Cohort 2	4.07	4.08	-.01	.86	-.03	.74	-.03
	(.95)	(.88)	(.08)		(.08)		(.95)
Cohort 1	3.65	3.68	-.03	.84	-.04	.79	-.04
	(1.00)	(.95)	(.15)		(.14)		(.88)
Difference	.42	.41	.02	.93	.01	.94	.01
	(.11)	(.13)	(.17)		(.16)		(.91)

NOTES: Valid *N* for school grade = 1,014. Student survey weights were used. Survey given to students in grades 4-12.

Table D-10. Year 3 Student Satisfaction ITT Impacts: School Satisfaction Scale

School Satisfaction Scale (IRT Scored .54-2.8)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D./S.E.)	Control (S.D./S.E.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Full sample	2.07	2.02	.05	.16	.01	.74	.02
	(.39)	(.41)	(.03)		(.03)		(.41)
Subgroups							
SINI ever	2.04	1.99	.05	.26	.01	.73	.03
	(.37)	(.43)	(.05)		(.04)		(.43)
SINI never	2.10	2.06	.04	.36	.01	.89	.02
	(.40)	(.39)	(.05)		(.04)		(.39)
Difference	-.06	-.07	.01	.93	.01	.89	.02
	(.04)	(.06)	(.07)		(.06)		(.41)
Lower performance	2.02	1.99	.03	.64	-.02	.66	-.06
	(.35)	(.39)	(.06)		(.05)		(.39)
Higher performance	2.10	2.04	.06	.18	.02	.50	.06
	(.40)	(.41)	(.04)		(.04)		(.41)
Difference	-.08	-.05	-.03	.68	-.05	.45	-.12
	(.04)	(.06)	(.07)		(.06)		(.41)
Male	2.06	2.02	.04	.34	.01	.83	.02
	(.37)	(.37)	(.04)		(.04)		(.37)
Female	2.08	2.03	.05	.28	.01	.80	.03
	(.41)	(.43)	(.05)		(.04)		(.43)
Difference	-.02	-.01	-.01	.82	-.00	.97	-.01
	(.04)	(.05)	(.06)		(.06)		(.41)
4-8	2.09	2.05	.03	.36	.00	1.00	.00
	(.39)	(.41)	(.04)		(.03)		(.41)
9-12	2.01	1.92	.09	.16	.05	.41	.14
	(.35)	(.37)	(.06)		(.06)		(.37)
Difference	.08	.13	-.05	.46	-.05	.48	-.12
	(.05)	(.05)	(.07)		(.07)		(.41)
Cohort 2	2.10	2.07	.02	.67	-.01	.73	-.03
	(.37)	(.38)	(.04)		(.04)		(.38)
Cohort 1	2.01	1.91	.10	.16	.07	.22	.16
	(.43)	(.44)	(.07)		(.06)		(.44)
Difference	.09	.16	-.07	.34	-.08	.24	-.20
	(.04)	(.07)	(.08)		(.07)		(.41)

NOTES: Valid *N* for student satisfaction = 886. Student survey weights were used. Survey given to students in grades 4-12.

Table D-11. Year 3 Parental Perceptions of School Safety and Climate: ITT Impacts on Individual Items

Parental Safety: NOT Current School Problems	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D.)	Control (S.D.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Kids destroying property	.81	.69	.11**	.00	.10**	.00	.22
	(.40)	(.46)	(.03)		(.03)		(.46)
Kids being late for school	.67	.51	.15**	.00	.15**	.00	.30
	(.47)	(.50)	(.03)		(.03)		(.50)
Kids missing classes	.74	.61	.13**	.00	.13**	.00	.27
	(.44)	(.49)	(.03)		(.03)		(.49)
Fighting	.58	.75	.17**	.00	.17**	.00	.33
	(.44)	(.49)	(.03)		(.03)		(.49)
Cheating	.84	.73	.11**	.00	.11**	.00	.24
	(.37)	(.44)	(.03)		(.03)		(.44)
Racial conflict	.88	.83	.05*	.02	.05*	.03	.12
	(.33)	(.38)	(.02)		(.02)		(.38)
Guns or other weapons	.87	.77	.10**	.00	.08**	.00	.20
	(.34)	(.42)	(.02)		(.02)		(.42)
Drug distribution	.87	.78	.09**	.00	.07**	.00	.18
	(.34)	(.42)	(.02)		(.02)		(.42)
Drug and alcohol use	.87	.76	.11**	.00	.09**	.00	.21
	(.33)	(.42)	(.03)		(.02)		(.42)
Teacher absenteeism	.84	.72	.12**	.00	.11**	.00	.25
	(.37)	(.45)	(.03)		(.03)		(.45)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N*s for the individual items range from 1,371 to 1,406.

Table D-12. Year 3 Student Reports of School Safety and Climate: ITT Impacts on Individual Items

Student Safety: Did NOT Happen This Year	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D.)	Control (S.D.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Something stolen from desk, locker, or other place	.59 (.49)	.59 (.49)	.00 (.04)	.97	-.03 (.04)	.32	-.08 (.49)
Taken money or things from me by force or threats	.91 (.28)	.85 (.35)	.06* (.02)	.02	.04* (.02)	.01	.13 (.35)
Offered drugs	.89 (.31)	.88 (.33)	.01 (.02)	.58	.01 (.02)	.60	.03 (.32)
Physically hurt by another student	.82 (.38)	.79 (.41)	.03 (.03)	.26	-.00 (.03)	.93	-.01 (.41)
Threatened with physical harm	.89 (.32)	.84 (.37)	.05 (.03)	.08	.03 (.02)	.16	.08 (.37)
Seen anyone with a real/toy gun or knife at school	.82 (.38)	.76 (.43)	.06* (.03)	.05	.07* (.03)	.03	.16 (.43)
Been bullied at school	.87 (.34)	.83 (.37)	.03 (.03)	.25	.01 (.02)	.34	.03 (.37)
Been called a bad name	.47 (.50)	.48 (.50)	-.01 (.04)	.71	-.04 (.04)	.33	-.07 (.50)

*Statistically significant at the 95 percent confidence level.

NOTES: Valid *N*s for the individual items range from 1,067 to 1,090.

Table D-13. Year 3 Parental Satisfaction ITT Impacts on Individual Items

School Satisfaction Scale: Items (1-4 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D.)	Control (S.D.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Location	3.21	3.05	.16**	.01	.15**	.01	.16
	(.89)	(.96)	(.06)		(.06)		(.96)
Safety	3.23	3.00	.22**	.00	.21**	.00	.24
	(.81)	(.90)	(.05)		(.05)		(.90)
Class sizes	3.17	2.91	.27**	.00	.22**	.00	.24
	(.84)	(.94)	(.06)		(.06)		(.94)
School facilities	3.10	2.92	.18**	.00	.14*	.01	.15
	(.83)	(.92)	(.06)		(.05)		(.92)
Respect between teachers and students	3.17	2.97	.20**	.00	.18**	.00	.20
	(.82)	(.92)	(.06)		(.06)		(.92)
Teachers inform parents of students' progress	3.21	3.04	.17**	.00	.14*	.01	.15
	(.83)	(.91)	(.06)		(.05)		(.91)
Amount students can observe religious traditions	3.23	2.95	.28**	.00	.28**	.00	.25
	(.94)	(1.15)	(.07)		(.07)		(1.15)
Parental support for the school	3.17	2.94	.23**	.00	.19**	.00	.21
	(.79)	(.88)	(.06)		(.05)		(.88)
Discipline	3.16	2.93	.23**	.00	.20**	.00	.21
	(.85)	(.94)	(.06)		(.06)		(.94)
Academic quality	3.21	2.99	.22**	.00	.16**	.00	.17
	(.82)	(.95)	(.06)		(.05)		(.95)
Racial mix of students	3.09	3.00	.09	.11	.06	.30	.06
	(.85)	(.92)	(.06)		(.06)		(.92)
Services for students with special needs	3.76	3.57	.18*	.04	.17*	.05	.13
	(1.24)	(1.29)	(.09)		(.08)		(1.29)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N*s for the individual items range from 1,349 to 1,407.

Table D-14. Year 3 Student Satisfaction ITT Impacts on Individual Items

School Satisfaction Scale: Items (1-4 Scale)	Unadjusted Mean Differences				Regression-Based Impact Estimates		
	Treatment (S.D.)	Control (S.D.)	T-C Difference (S.E.)	*p*-value	Estimated Impact (S.E.)	*p*-value	Effect Size (S.D.)
Students are proud to go to this school	2.94	2.93	.00	.96	-.05	.41	-.06
	(.86)	(.85)	(.06)		(.06)		(.85)
There is a lot of learning at this school	3.40	3.32	.08	.13	.05	.38	.06
	(.70)	(.74)	(.06)		(.05)		(.74)
Rules of behavior are strict	3.25	3.16	.09	.20	.06	.37	.07
	(.87)	(.91)	(.07)		(.07)		(.91)
When students misbehave, they receive the same treatment	2.82	2.75	.07	.42	.03	.73	.03
	(1.00)	(1.06)	(.08)		(.08)		(1.06)
I feel safe	3.16	3.13	.03	.68	-.00	.99	-.00
	(1.00)	(1.00)	(.08)		(.08)		(1.00)
People at my school are supportive	3.18	3.12	.05	.40	-.01	.80	-.02
	(.83)	(.83)	(.06)		(.06)		(.83)
I do not feel isolated at my school	3.19	3.19	.00	.96	-.02	.80	-.02
	(.92)	(.94)	(.07)		(.07)		(.94)
I enjoy going to school	3.26	3.26	-.00	.95	-.03	.60	-.04
	(.81)	(.83)	(.06)		(.06)		(.83)
Students behave well with the teachers	2.83	2.75	.08	.15	.06	.30	.07
	(.83)	(.84)	(.06)		(.06)		(.84)
Students do their homework	2.53	2.36	.17*	.01	.12	.07	.13
	(.89)	(.91)	(.07)		(.07)		(.91)
I rarely feel made fun of by other students	3.11	3.01	.10	.22	.05	.49	.05
	(1.04)	(1.08)	(.08)		(.08)		(1.08)
Other students seldom disrupt class	2.28	2.16	.13	.05	.05	.40	.06
	(.95)	(.91)	(.07)		(.07)		(.90)
Students who misbehave rarely get away with it	2.83	2.72	.11	.17	.04	.59	.04
	(1.00)	(1.04)	(.08)		(.07)		(1.04)
Most of my teachers really listen to what I have to say	3.15	3.14	.01	.88	.00	1.00	.00
	(.86)	(.93)	(.07)		(.07)		(.93)
My teachers are fair	3.07	3.06	.02	.78	.01	.92	.01
	(.84)	(.88)	(.06)		(.06)		(.88)
My teachers expect me to succeed	3.60	3.47	.13*	.02	.12*	.02	.17
	(.63)	(.74)	(.05)		(.05)		(.74)
Teachers punish cheating when they see it	3.33 (.90)	3.09 (1.03)	.25** (.07)	.00	.22** (.07)	.00	.21 (1.03)

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N*s for the individual items range from 960 to 1,087. Higher values indicate a greater degree of agreement with the statement.

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 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 3. 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 tends to place 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 3 and the following outcomes for groups of students and parents (table E-1):

- For the full sample of students, reading achievement for students who attended a private school in year 3 was 7.11 scale score points higher (ES = .22) [2] than that of students who were not in private school in year 3.
- Students who applied from non-SINI schools who were enrolled in private school in year 3 scored 10.25 scale score points higher (ES = .30) in reading than that of non-SINI students who were not in private school in year 3.
- Reading achievement for students who applied with relatively higher academic performance who were enrolled in private school in year 3 was 9.52 scale score points higher (ES = .30) than that of like students who were not in private school in year 3.
- Reading achievement for students who were entering grades K-8 in the application year and were enrolled in private school in year 3 was 8.28 scale score points higher (ES = .25) than that of K-8 students who were not in private school in year 3.
- Though the ITT results in chapter 3 found statistically significant treatment impacts in reading achievement for the female and the cohort 1 subgroups of students, these same effects were not significant through the IV estimation of the outcomes of private schooling.
- Parental perceptions of school climate and safety were higher for those enrolled in private schools in year 3 (ES = .55) than for those with children in public schools.

---

[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 (71.6 percent) and to each member of the control group (12.3 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 distinguishes only between treatments and controls (who were randomly assigned) and cannot distinguish between private school enrollees and nonprivate school enrollees (who were self-selected), the use of the lottery as the instrumental variable in this analysis generates unbiased estimates of the effects of private schooling.

[2] ES stands for Effect Size and is measured as a fraction of a standard deviation of the distribution of control group values (in this case, those who did not attend private school) of the outcome variable.

- Parents of students who attended private schools were more likely (21 percentage points) to give their child's school a grade of A or B (ES = .43) than if the child was in a public school.

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

Outcomes	IV Regression Estimate	*p*-value	Effect Size
Student Achievement			
Full sample: reading	7.11*	.04	.22
SINI never: reading	10.25*	.02	.30
Higher performing: reading	9.52*	.02	.30
Female: reading	6.08	.15	.19
K-8: reading	8.28*	.02	.25
Cohort 1: reading	15.75	.05	.57
School Safety and Climate: Parents	2.04**	.00	.55
School Satisfaction: Parents			
School grade of A or B	.21**	.00	.43

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for reading = 1,365. Reading sample weights were used. Difference displayed in terms of scale scores. Valid *N* for school danger = 1,345. Parent survey weights were used. Valid *N* for school grade = 1,333. 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.

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).[3]

- The overall private schooling effect on reading remains statistically significant under the trimmed sample procedure, but is not significant in the estimation procedure that generates robust standard errors by clustering on current school attended.

- All five subgroups that had significant reading achievement impacts under the ITT analysis (chapter 3) also demonstrated statistically significant private schooling effects using IV estimation on the trimmed sample.

- The same two subgroups of students with significant ITT reading impacts that did not show statistically significant reading achievement effects in the main IV analysis (female and cohort 1 students) also did not demonstrate significant reading effects when clustering on current school attended.

- The finding that parental perceptions of school climate and safety were higher for those who enrolled their child in private school is not sensitive to different analytic methods.

[3] For a description of the sensitivity tests, see appendix C.

- The finding that parental satisfaction is higher for those who enrolled their child in a private school is not sensitive to different analytic methods.

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						
Full sample: reading	7.11*	.04	9.38*	.02	7.11	.06
SINI never: reading	10.25*	.02	13.10**	.01	10.25*	.01
Higher performing: reading	9.52*	.02	9.09*	.03	9.52*	.03
Female: reading	6.08	.15	10.56*	.02	6.08	.17
K-8: reading	8.28*	.02	10.22*	.02	8.28*	.02
Cohort 1: reading	15.75	.05	25.10*	.02	15.75	.10
School Safety and Climate: Parents	2.04**	.00	1.95**	.00	2.04**	.00
School Satisfaction: Parents						
School grade of A or B	.21**	.00	.20**	.00	.21**	.00

*Statistically significant at the 95 percent confidence level.

**Statistically significant at the 99 percent confidence level.

NOTES: Valid *N* for reading = 1,365; trimmed sample valid *N* = 1,220. Reading sample weights were used. Difference displayed in terms of scale scores. Valid *N* for school danger = 1,345; trimmed sample valid *N* = 1,203. Parent survey weights were used. Valid *N* for school grade = 1,333; trimmed sample valid *N* = 1,195. 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.

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 three questions about how often during the school year the parent volunteered in school, attended a school organization meeting, or accompanied students on class trips. Parental involvement was chosen because it has been shown to vary between public and private schools (Bryk et al. 1993; Witte 1993; Bauch and Goldring 1995) and to have a relationship to student achievement (Henderson and Berla 1994; Sui-Chu and Willms 1996; Fan and Chen 2001; Wu and Qi 2006).

The parental involvement variable ranges from .67 to 7.78 with a mean of 2.82 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 1991). The critical value of Cronbach's Alpha is .70, above which a scale is considered to have a satisfactory level of reliability.

2. Parent Aspirations

Parent aspirations is a measure of how many years of education a parent expects his/her child to receive. Taken from the parent survey, the variable is treated as a continuous variable with the following values:

a. Some high school, but will not graduate=11
b. Complete high school=13
c. Attend a 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; Fan and Chen 2001; Wu and Qi 2006). The measure of parent aspirations ranges from 11 to 19. The mean of parent aspirations is 17.28, 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 or not the student receives help on schoolwork from tutoring held outside of the child's school. Out-of-school tutor usage is one of two measures of tutor usage, along with in-school tutor usage. These measures were chosen because tutor usage has been shown to vary across public and private schools (Howell et al 2006) and to be associated with student achievement (Cohen, Kulik, and Kulik 1982; Ritter 2000). As a dichotomous variable, out-of-school tutor usage can take the value of 0 or 1. The mean value of out-of-school tutor usage is .11, and the standard deviation is .31.

4. School Transit Time

School transit time seeks to measure the length of the school commute that a parent provides for his/her child. The variable is taken from the parent survey and is an ordinal variable with values assigned as:

a. Under 10 minutes= 1
b. 11-20 minutes=2
c. 21-30 minutes=3
d. 31-45 minutes=4
e. 46 minutes to an hour=5
f. More than one hour=6

This variable was chosen because it has been shown to be associated with student achievement (Dolton, Marcenaro, and Navarro 2003). Commuting time has a negative effect on student achievement because it is unproductive time that is not being spent on student learning. The school transit time variable ranges from 1 to 6 with a mean of 2.70 and a standard deviation of 1.36. Due to the ordinal design of this variable, details regarding the impact of the offer of a scholarship on school transit time are presented by response category in tables F-1 to F-11 below.

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, Reads for Fun, Engagement in Extracurricular Activities, and Frequency of Homework (measured in days per week).

1. Student Aspirations

Student aspirations is a measure of how many years of education the student expects to receive. Taken from the student survey, the variable is treated as continuous with the following values:

a. Some high school, but will not graduate=11
b. Complete high school=13
c. Attend a 2 year college=14
d. Attend a 4 year college=15
e. Obtain a certificate=15
f. Obtain a bachelor's degree=17
g. Obtain a master's degree or other higher degree=19

Student aspirations is one of two measures of educational aspirations, along with parent aspirations. These factors were chosen as potential mediators because educational aspirations have been shown to vary

across public and private schools (Plank et al. 1993) and to be associated with student achievement (Natriello and McDill 1986; Singh et al. 1995). The student aspirations variable ranges from 11 to 19 years of education. The mean of student aspirations is 16.80, and the standard deviation is 1.95.

2. Attendance

Attendance is a measure of how often the student has missed school. Attendance is an ordinal variable taken from the parent survey that measures how many school days the student missed in the preceding month:

a. None=0
b. 1-2 Day =1
c. 3-4 Days=2
d. 5 or more days=3

Attendance was chosen as a possible mediator because it has been shown to be associated with student achievement (Lamdin 1996). The attendance variable ranges from 0 to 3. Attendance has a mean of .79 and a standard deviation of .83. Due to the ordinal design of this variable, details regarding the impact of the offer of a scholarship on attendance are presented by response category in tables F-12 to F-22.

3. Tardiness

Tardiness is a measure of how often the student has missed school. Taken from the parent survey and measuring how many days the student arrived late in the preceding month, tardiness is an ordinal variable with the following values:

a. None=0
b. 1-2 Days=1
c. 3-4 Days=2
d. 5 or more days=3

Tardiness was chosen as a possible mediator because it has been associated with student achievement (Mulkey, Crain, and Harrington 1992). The tardiness variable ranges from 0 to 3. Tardiness has a mean of .50 and a standard deviation of .78. Due to the ordinal design of this variable, details regarding the impact of the offer of a scholarship on tardiness are presented by response category in tables F-23 to F-33.

4. Reads for Fun

Reads for fun seeks to measure whether the student reads for personal enjoyment. The variable is taken from the student survey and is a dichotomous variable that equals 1 if the student claims to read for fun and 0 if not. The variable was chosen as a possible mediator because it has been shown to be associated with student achievement (Mulkey et al. 1992; Mullis et al. 2003). Reads for fun has a mean of .43 and a standard deviation of .49.

5. Engagement in Extracurricular Activities

Engagement in extracurricular activities seeks to measure the student's involvement in programs that are not a required part of the school's educational program. Taken from the student survey, the variable is a count of the number of activities in which a student reports participating from a list of five items, 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 2.25 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 zero to five, of the number of school days per week that the student said that he or she typically works on homework. Frequency of homework was chosen because it has been shown to vary across public and private schools (Hoffer, Greeley, and Coleman 1985) and to be associated with student achievement (Rutter et al. 1979; Natriello and McDill 1986; Rumberger and Palardy 2005; Wolf and Hoople 2006). The mean of frequency of homework is 3.73, and the standard deviation is 1.53.

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, Ability Grouping, Availability of Tutors, In-school Tutor Usage, Programs for Students with Learning

Problems, Programs for English Language Learners, Programs for Advanced Learners, Before- or After-School Programs, and Enrichment Programs.[2]

1. Student/Teacher Ratio

Student/teacher ratio is the number of students at the child's school divided by the full-time equivalency of classroom teachers at the school. The variable is a continuous measure taken from the National Center for Educational Statistics' Common Core of Data (NCES CCD) and Private School Universe Survey (NCES PSS). Student/teacher ratio was chosen as a possible mediator because it has been shown to vary across public and private schools and to be associated with student achievement (Arum 1996; Nye, Hedges, and Konstantopoulos 2000). Student/teacher ratio ranges from 2.60 to 26.20. The mean of student/teacher ratio is 12.93, and the standard deviation is 4.23.

2. Teacher Attitude

Teacher attitude measures the extent to which students report being treated with consideration by their classroom teachers. Taken from the student survey, the variable is an IRT scale that combines student evaluations of four items involving how well teachers listen to them, are fair, expect students to succeed, and encourage students to do their best. Teacher attitude was chosen because it has been shown to differ across public and private schools (Ballou and Podgursky 1998; Gruber et al. 2002) and to be associated with student achievement (Hanushek 1971; Card and Krueger 1992; Wayne and Youngs 2003; Wolf and Hoople 2006). Teacher attitude ranges from .44 to 10.39 with a mean of 2.61 and a standard deviation of 1.97. The Cronbach's Alpha for teacher attitude is .75.

3. Ability Grouping

Ability grouping is a measure of the ways in which a school differentiates instruction based on student ability. Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school differentiates instruction by either organizing classes with similar content but different difficulty levels or organizing classes with different content. The variable equals 0 if neither of these

[2] The offer of an Opportunity Scholarship has a negative impact on the likelihood that a student participates in the federal government's free or reduced-price lunch program. Although the families in the treatment and control groups were equally income disadvantaged at baseline, 3 years after random assignment only 27 percent of the treatment group but 50 percent of the control group was participating in the federal lunch program based on parent reports. The negative impact of the scholarship offer on participation in the lunch program of 23 percentiles has an effect size of -.13 of a standard deviation and is statistically significant beyond $p < .01$. The federal lunch program was not included as a possible mediator in the analysis because participation in the federal school lunch program per se is not associated with student achievement. The treatment does not affect family income, only the likelihood of participating in a certain program (free/reduced-price lunch) that is sometimes used as an imperfect indicator of low family income.

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 .73 and a standard deviation of .45.

4. Availability of Tutors

Availability of tutors measures whether the school a student attends has tutors available for its students. Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school makes tutors available to its students and 0 if not. Though not entirely comparable to the two measures of tutor *usage* analyzed as possible mediators, this variable was chosen for similar reasons: 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 .58 and a standard deviation of .49.

5. In-school Tutor Usage

In-school tutor usage is a measure of whether a child actually uses a tutor provided by the school. Taken from the parent survey, the measure is a dichotomous variable that equals 1 if the student uses a school-provided tutor and 0 if not. In-school tutor usage is one of two measures of tutor usage, along with out-of-school tutor usage, analyzed as possible mediators. These measures were chosen because tutor usage has been shown to vary across public and private schools (Howell et al 2006) and to be associated with student achievement (Cohen et al. 1982; Ritter 2000). In-school tutor usage has a mean of .24 and a standard deviation of .43.

6. Programs for Students with Learning Problems

Programs for students with learning problems is an indicator of an affirmative response to a question in the principal survey about providing distinctive instructional activities for students with learning problems. This measure of special school programs was chosen for analysis because the availability of such programs has been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the principal gave an affirmative response when asked if his or her school offers such programs and 0 if not. The mean of this variable is .81 and the standard deviation is .39.

7. Programs for English Language Learners

Programs for English language learners is an indicator of an affirmative response to a question in the principal survey about providing special instruction for non-English speakers. This measure of special programs was chosen for analysis because the availability of such programs has been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the principal gave an affirmative response when asked if his or her school offers such programs and 0 if not. The mean of this variable is .42 and the standard deviation is .49.

8. Programs for Advanced Learners

Programs for advanced learners is a measure of whether a principal reports offering any of a list of three items in the principal survey, including Advanced Placement (AP) courses, International Baccalaureate (IB) programs, and special instructional programs for advanced learners or a gifted and talented program. The variable is one of four potential mediators that measure special school programs. These factors were chosen for analysis because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). Taken from the principal survey, the measure is a dichotomous variable that equals 1 if the school reported offering any of the three types of programs and 0 if it reported offering none. The mean of this variable is .47 and the standard deviation is .50.

9. Before-/After-School Programs

Before- or after-school programs was taken from the principal survey and is a dichotomous variable that equals 1 if the school offers a program for students either before or after school and equals 0 if not. The variable is one of four that measure the availability of special school programs. These programmatic variables were chosen for the mediator analysis because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). The mean of before-/after-school programs is .87, indicating that almost every student in the impact sample attended a school with a before- or after-school program and the standard deviation is .33.

10. Enrichment Programs

Enrichment programs is a count of how many programs a school reports offering out of three items: foreign language programs, music programs, and arts programs. The variable is one of four that measures the availability of special school programs. These factors were chosen for analysis as possible mediators because they have been shown to vary across public and private schools (Gruber et al. 2002; Howell et al. 2006) and to be associated with student achievement (Rumberger and Palardy 2005). The enrichment programs variable ranges from 0 to 3 with a mean of 2.60 and a standard deviation of .63.

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: School Communication Policies, School Size, Percent Non-White, and Peer Classroom Behavior.

1. School Communication Policies

School communication policies measures the number of distinct policies a school has regarding required school-parent communications. Taken from the principal survey, the variable is a count of the number of communication policies a school reports having out of four items: informing parents of their 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. School communication policies was chosen for analysis as a possible mediator because it has been shown to vary across public and private schools (Bauch and Goldring 1995; Howell et al. 2006) and to be associated with student achievement (Henderson and Berla 1994; Sui-Chu and Willms 1996). The variable for school communication policies ranges from 1 to 4 with a mean of 3.04 and a standard deviation of .84.

2. School Size

School size is the total reported student enrollment in the attended school and is taken from the National Center for Education Statistics (NCES) Common Core of Data (CCD) (http://nces.ed.gov/ccd) and the NCES Private School Survey (PSS) (http://nces.ed.gov/surveys/pss). The variable was included in the analysis as a possible mediator because it has been shown to vary across public and

private schools (Wasley 2002) and to be associated with student achievement (Sander 1999; Lee and Loeb 2000). School size ranges from 10 to 3,514. The mean of school size is 463.83 and the standard deviation is 527.44.

3. Percent Non-White

Percent non-White is the percentage of enrolled students at the attended school who were identified as American Indian/Alaska Native, Asian Pacific Islander, Black non-Hispanic, and Hispanic. The data for the variable were taken from the NCES' CCD and PSS. The variable was included in the analysis as a possible mediator because it has been shown to vary across public and private schools (Plank et al. 1993; Reardon and Yun 2002; Schneider and Buckley 2002) and to be associated with student achievement (Coleman 1966; Coleman 1990; Hanushek et al. 2002; Nielsen and Wolf 2002). Percent non-White ranges from .12 to 1.00 with a mean of .96 and a standard deviation of .11.

4. Peer Classroom Behavior

Peer classroom behavior seeks to measure the degree to which the other students in the child's class are well behaved. Taken from the student survey, the variable is an IRT scale composed of student evaluations of five statements about their peers: whether students behave well with teachers, students neglect their homework, students tease them, other students often disrupt class, and students get away with bad behavior. Peer classroom behavior was chosen for the analysis as a possible mediator because it has been shown to vary across public and private schools (Lee, Dedrick, and Smith 1991; Harris 1998) and to be associated with student achievement (Card and Krueger 1992). Peer classroom behavior ranges from 2.93 to 12.91 with a mean of 8.31 and a standard deviation of 2.29. The Cronbach's Alpha for peer classroom behavior is .68.[3]

[3] 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.

F.5 Impacts on Intermediate Outcomes for Ordinal Variables by Variable Category

Table F-1. Marginal Effects of Treatment: School Transit Time for Full Sample

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.20	.22	-.02
11-20 minutes	.30	.31	-.01
21-30 minutes	.21	.20	.01
31-45 minutes	.17	.16	.01
46 minutes to an hour	.09	.07	.02
More than one hour	.03	.03	.00

NOTES: Ordered logit beta = .12. Effect is not statistically significant (p-value = .29). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-2. Marginal Effects of Treatment: School Transit Time for SINI-Ever Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.19	.21	-.02
11-20 minutes	.29	.30	-.01
21-30 minutes	.22	.21	.01
31-45 minutes	.19	.17	.02
46 minutes to an hour	.09	.08	.01
More than one hour	.04	.03	.00

NOTES: Ordered logit beta = .15. Effect is not statistically significant (p-value = .44). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-3. Marginal Effects of Treatment: School Transit Time for SINI-Never Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.23	-.02
11-20 minutes	.31	.31	-.01
21-30 minutes	.20	.20	.01
31-45 minutes	.17	.16	.01
46 minutes to an hour	.08	.07	.01
More than one hour	.03	.03	.00

NOTES: Ordered logit beta = .10. Effect is not statistically significant (p-value = .46). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-4. Marginal Effects of Treatment: School Transit Time for Lower-Performing Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.21	.19	.02
11-20 minutes	.31	.29	.01
21-30 minutes	.20	.21	-.01
31-45 minutes	.17	.18	-.02
46 minutes to an hour	.08	.09	-.01
More than one hour	.03	.04	-.00

NOTES: Ordered logit beta = -.14. Effect is not statistically significant (p-value = .48). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-5. Marginal Effects of Treatment: School Transit Time for Higher-Performing Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.20	.24	-.04
11-20 minutes	.29	.32	-.02
21-30 minutes	.21	.20	.01
31-45 minutes	.18	.15	.03
46 minutes to an hour	.08	.07	.02
More than one hour	.03	.03	.01

NOTES: Ordered logit beta = .25. Effect is not statistically significant (p-value = .07). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-6. Marginal Effects of Treatment: School Transit Time for Male Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.20	.22	-.02
11-20 minutes	.30	.31	-.01
21-30 minutes	.21	.20	.01
31-45 minutes	.18	.16	.01
46 minutes to an hour	.08	.08	.01
More than one hour	.04	.03	.00

NOTES: Ordered logit beta = .11. Effect is not statistically significant (p-value = .51). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-7. Marginal Effects of Treatment: School Transit Time for Female Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.21	.23	-.02
11-20 minutes	.30	.31	-.01
21-30 minutes	.21	.20	.01
31-45 minutes	.17	.16	.01
46 minutes to an hour	.08	.07	.01
More than one hour	.03	.03	.00

NOTES: Ordered logit beta = .14. Effect is not statistically significant (p-value = .38). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-8. Marginal Effects of Treatment: School Transit Time for K-8 Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.21	.24	-.03
11-20 minutes	.30	.32	-.02
21-30 minutes	.21	.20	.01
31-45 minutes	.17	.15	.02
46 minutes to an hour	.08	.07	.01
More than one hour	.03	.03	.01

NOTES: Ordered logit beta = .19. Effect is not statistically significant (p-value = .13). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-9. Marginal Effects of Treatment: School Transit Time for 9-12 Subgroup

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.15	.11	.04
11-20 minutes	.27	.24	.02
21-30 minutes	.22	.23	-.01
31-45 minutes	.21	.23	-.03
46 minutes to an hour	.11	.12	-.02
More than one hour	.05	.05	-.01

NOTES: Ordered logit beta = -.26. Effect is not statistically significant (p-value = .41). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-10. Marginal Effects of Treatment: School Transit Time for Cohort 2

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.22	.24	-.02
11-20 minutes	.31	.32	-.01
21-30 minutes	.20	.20	.01
31-45 minutes	.16	.15	.01
46 minutes to an hour	.07	.07	.01
More than one hour	.03	.03	.00

NOTES: Ordered logit beta = .11. Effect is not statistically significant (p-value = .40). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-11. Marginal Effects of Treatment: School Transit Time for Cohort 1

School transit time	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
Under 10 minutes	.13	.16	-.03
11-20 minutes	.27	.28	-.02
21-30 minutes	.23	.22	.01
31-45 minutes	.22	.20	.02
46 minutes to an hour	.11	.09	.01
More than one hour	.05	.04	.00

NOTES: Ordered logit beta = .18. Effect is not statistically significant (p-value = .49). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-12. Marginal Effects of Treatment: Parent Reported Attendance for Full Sample

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.42	.45	-.03
1-2 days	.38	.37	.01
3-4 days	.14	.13	.01
5 or more days	.06	.06	.01

NOTES: Ordered logit beta = .11. Effect is not statistically significant (p-value = .36). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-13. Marginal Effects of Treatment: Parent Reported Attendance for SINI-Ever Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.43	.46	-.03
1-2 days	.38	.37	.01
3-4 days	.13	.12	.01
5 or more days	.06	.05	.01

NOTES: Ordered logit beta = .13. Effect is not statistically significant (p-value = .48). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-14. Marginal Effects of Treatment: Parent Reported Attendance for SINI-Never Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.41	.44	-.02
1-2 days	.38	.38	.01
3-4 days	.14	.13	.01
5 or more days	.06	.06	.00

NOTES: Ordered logit beta = .08. Effect is not statistically significant (p-value = .58). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-15. Marginal Effects of Treatment: Parent Reported Attendance for Lower-Performing Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.35	.45	-.09
1-2 days	.42	.38	.04
3-4 days	.16	.12	.04
5 or more days	.07	.05	.02

NOTES: Ordered logit beta = .39. Effect is not statistically significant (p-value = .07). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-16. Marginal Effects of Treatment: Parent Reported Attendance for Higher-Performing Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.45	.44	.01
1-2 days	.37	.37	-.00
3-4 days	.12	.13	-.00
5 or more days	.06	.06	-.00

NOTES: Ordered logit beta = -.03. Effect is not statistically significant (p-value = .83). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-17. Marginal Effects of Treatment: Parent Reported Attendance for Male Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.44	.44	-.00
1-2 days	.37	.37	.00
3-4 days	.13	.13	.00
5 or more days	.06	.06	.00

NOTES: Ordered logit beta = .00. Effect is not statistically significant (p-value = .99). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-18. Marginal Effects of Treatment: Parent Reported Attendance for Female Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.40	.45	-.05
1-2 days	.39	.37	.02
3-4 days	.14	.13	.02
5 or more days	.07	.06	.01

NOTES: Ordered logit beta = .20. Effect is not statistically significant (p-value = .20). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-19. Marginal Effects of Treatment: Parent Reported Attendance for K-8 Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.43	.45	-.02
1-2 days	.38	.37	.01
3-4 days	.13	.13	.01
5 or more days	.06	.06	.00

NOTES: Ordered logit beta = .07. Effect is not statistically significant (p-value = .57). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-20. Marginal Effects of Treatment: Parent Reported Attendance for 9-12 Subgroup

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.36	.45	-.08
1-2 days	.41	.38	.04
3-4 days	.15	.12	.03
5 or more days	.07	.05	.02

NOTES: Ordered logit beta = .34. Effect is not statistically significant (p-value = .39). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-21. Marginal Effects of Treatment: Parent Reported Attendance for Cohort 2

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.43	.46	-.03
1-2 days	.38	.37	.01
3-4 days	.13	.12	.01
5 or more days	.06	.05	.01

NOTES: Ordered logit beta = .13. Effect is not statistically significant (p-value = .28). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-22. Marginal Effects of Treatment: Parent Reported Attendance for 9-12 Cohort 1

Days absent	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.38	.38	.00
1-2 days	.39	.40	-.00
3-4 days	.15	.15	-.00
5 or more days	.07	.07	-.00

NOTES: Ordered logit beta = -.02. Effect is not statistically significant (p-value = .96). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-23. Marginal Effects of Treatment: Parent Reported Tardiness for Full Sample

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.63	.67	-.04
1-2 days	.25	.23	.02
3-4 days	.07	.06	.01
5 or more days	.05	.04	.01

NOTES: Ordered logit beta = .17. Effect is not statistically significant (p-value = .19). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-24. Marginal Effects of Treatment: Parent Reported Tardiness for SINI-Ever Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.64	.66	-.01
1-2 days	.25	.24	.01
3-4 days	.07	.06	.00
5 or more days	.04	.04	.00

NOTES: Ordered logit beta = .06. Effect is not statistically significant (p-value = .74). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-25. Marginal Effects of Treatment: Parent Reported Tardiness for SINI-Never Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.63	.68	-.06
1-2 days	.26	.23	.03
3-4 days	.07	.06	.01
5 or more days	.04	.04	.01

NOTES: Ordered logit beta = .25. Effect is not statistically significant (p-value = .14). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-26. Marginal Effects of Treatment: Parent Reported Tardiness for Lower-Performing Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.59	.64	-.05
1-2 days	.28	.25	.03
3-4 days	.08	.07	.01
5 or more days	.05	.04	.01

NOTES: Ordered logit beta = .22. Effect is not statistically significant (p-value = .33). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-27. Marginal Effects of Treatment: Parent Reported Tardiness for Higher-Performing Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.65	.68	-.03
1-2 days	.25	.23	.02
3-4 days	.06	.06	.01
5 or more days	.04	.04	.01

NOTES: Ordered logit beta = .15. Effect is not statistically significant (p-value = .33). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-28. Marginal Effects of Treatment: Parent Reported Tardiness for Male Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.65	.64	.01
1-2 days	.24	.25	-.01
3-4 days	.06	.07	-.00
5 or more days	.04	.04	-.00

NOTES: Ordered logit beta = -.05. Effect is not statistically significant (p-value = .77). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-29. Marginal Effects of Treatment: Parent Reported Tardiness for Female Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.61	.70	-.09*
1-2 days	.27	.22	.05*
3-4 days	.07	.05	.02*
5 or more days	.05	.03	.01*

NOTES: Ordered logit beta = .39. Effect is statistically significant (p-value = .03). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-30. Marginal Effects of Treatment: Parent Reported Tardiness for K-8 Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.64	.67	-.03
1-2 days	.25	.23	.02
3-4 days	.07	.06	.01
5 or more days	.04	.04	.00

NOTES: Ordered logit beta = .14. Effect is not statistically significant (p-value = .31). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-31. Marginal Effects of Treatment: Parent Reported Tardiness for 9-12 Subgroup

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.60	.68	-.08
1-2 days	.28	.23	.05
3-4 days	.08	.06	.02
5 or more days	.05	.04	.01

NOTES: Ordered logit beta = .36. Effect is not statistically significant (p-value = .36). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-32. Marginal Effects of Treatment: Parent Reported Tardiness for Cohort 2

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.64	.68	-.04
1-2 days	.25	.23	.03
3-4 days	.07	.06	.01
5 or more days	.04	.04	.01

NOTES: Ordered logit beta = .19. Effect is not statistically significant (p-value = .17). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.

Table F-33. Marginal Effects of Treatment: Parent Reported Tardiness for Cohort 1

Days tardy	Treatment Group Mean	Control Group Mean	Difference (Estimated Impact)
None	.62	.64	-.02
1-2 days	.26	.25	.01
3-4 days	.07	.07	.00
5 or more days	.05	.05	.00

NOTES: Ordered logit beta = .09. Effect is not statistically significant (p-value = .79). The treatment and control group means are mean percentages of respondents in each group giving each category of response. The "Difference (Estimated Impact)" is the marginal effect, that is, impacts of treatment on the probabilities of respondents giving that specific category of response; the estimated impact is the difference between the treatment and control group means.