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

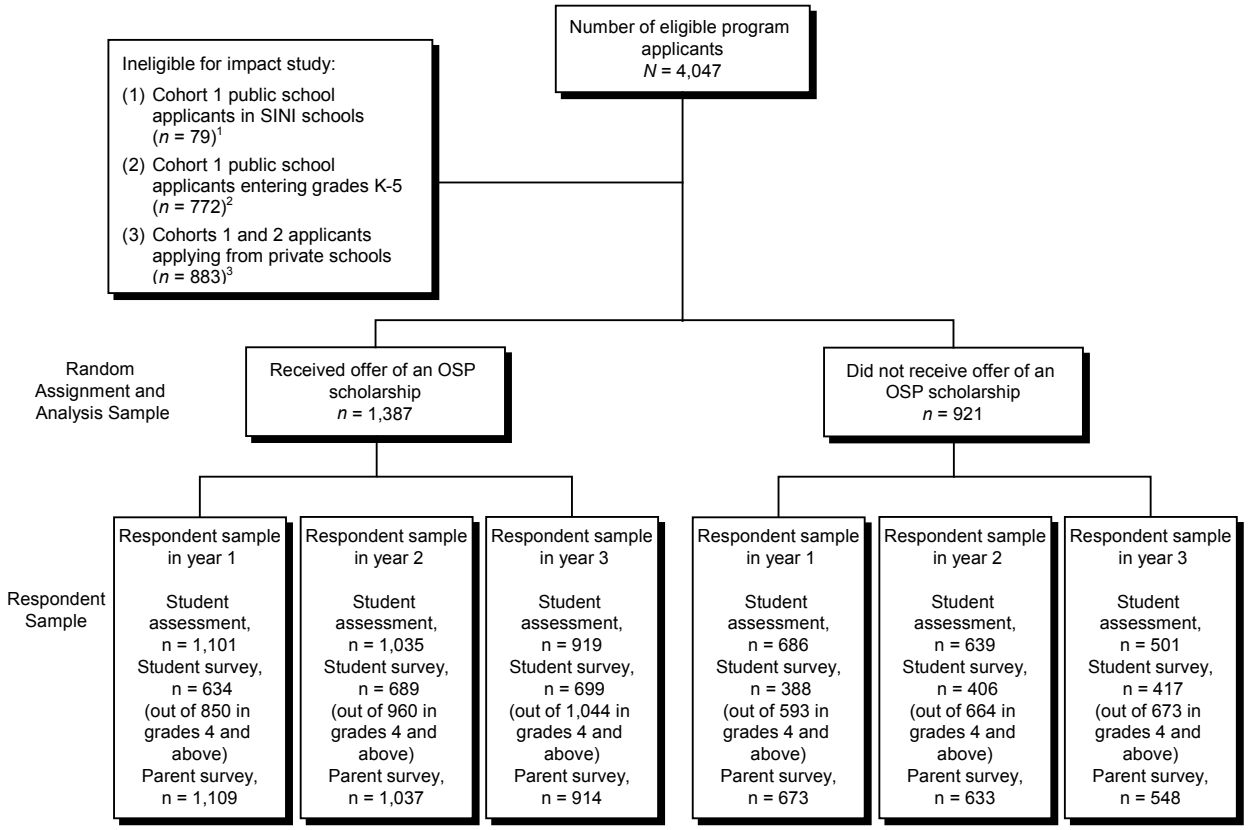
Stratum	Cohort	SINI Status	Grade Band	Treatment Sampling Rate (%)	Base Weight for Control Group	Base Weight for Treatment Group
1	Cohort 1	Non-SINI	6th to 8th	75.89	4.15	1.32
2	Cohort 1	Non-SINI	9th to 12th	28.21	1.39	3.54
3	Cohort 2	SINI	K to 5th	78.34	4.62	1.28
4	Cohort 2	SINI	6th to 8th	75.00	4.00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