Evaluation of the Florida Tax Credit Scholarship Program
Participation, Compliance and Test Scores in 2009-10

David N. Figlio
University of Florida
Northwestern University
and
National Bureau of Economic Research

August 2011

Executive summary

This is the fourth in a series of reports evaluating the Florida Tax Credit Scholarship (FTC) Program, as required by the Florida Statutes, s. 1002.395(9)(j). This report provides information on private school compliance with program rules regarding required testing, describes the attributes of eligible students who participate in the program, and presents data on student test score levels and gains in the program, as well as compared with the eligible population of non-participating students. This report does not include new information about parental satisfaction; the satisfaction survey reported in last year's report was only to be conducted once.

During the 2009-10 academic year, David Figlio, the Project Director, collected test score data from private schools participating in the FTC Program in real time. This is the fourth year for which program participants' test score data were collected, and the third year in which this data collection occurred in real time.

Compliance with program testing requirements, 2009-10:

● Compliance with program testing requirements in 2009-10 remains at very high levels, and private school reporting errors continue to decline. Private schools provided usable test scores for 91.3 percent of program participants in grades 3-10. Another 5.8 percent of participants were ineligible for testing or were not enrolled in the school at the time of testing; this is largely driven by the fact that some students arrived in schools after fall testing (for schools that test in the fall, principally those that administer the Iowa Test of Basic Skills) and some students who began the year in a school left the school prior to the more typical spring testing. The 0.8 percent rate of reported illness/absence is the lowest it has been since the beginning of data collection. Test administration compliance errors by participating schools are at the same level as 2008-09, and well below earlier years, with reporting problems involving only 1.3 percent of participants in 2009-10.

● The vast majority (69.2 percent) of test-takers took the Stanford Achievement Test. Other popular tests were the Iowa Test of Basic Skills (20.6 percent) and the TerraNova (3.7 percent).

● Scholarship students whose test scores were received are modestly more advantaged than are those scholarship students whose scores were not received. It is not known whether the gains of those without score reports would have been higher or lower than those with score reports.

Selection into the FTC Program:

● Program participants tend to come from less advantaged families than other students receiving free or reduced-price lunches.

● Unlike in prior years, program participants are no more likely to come from lower-performing public schools prior to entering the program. However, as in prior years, they

tend to be among the lowest-performing students in their prior school, regardless of the performance level of their public school. The selection from the bottom of the prior test score distribution of a school is becoming stronger over time.

Test scores of program participants, 2009-10:

● The typical student in the program scored at the 45th national percentile in reading and the 46th percentile in mathematics, about the same as in 2008-09. The distribution of test scores is similar whether one considers the entire program population or only those who took the Stanford Achievement Test in the spring of 2010. The Stanford Achievement Test is the most commonly administered test and is the test most directly comparable to the FCAT.

● The mean reading gain for program participants is -1.2 national percentile ranking points in reading and -1.7 national percentile ranking points in mathematics. These mean gains are indistinguishable from zero. In other words, the typical student participating in the program tended to maintain his or her relative position in comparison with others nationwide. It is important to note that these national comparisons pertain to all students nationally, and not just low-income students.

● Because families can choose whether to participate in the program, it is inappropriate to consider the differences in test score gains between FTC Program participants and their public school counterparts to be *caused* by program participation. Credible comparisons of program participants and non-participants must take into account this *selection problem*. This report makes use of the best available statistical tools (given the nature of the data) for determining the causal effect of program participation.

● The best possible statistical estimates (using a tool called regression discontinuity design) of the effects of program participation indicate that participation is associated with small improvements in reading and mathematics, relative to public school students who applied for participation in the program, though these differences are not always statistically significant. The results are consistent with a finding of small but positive differences between program participants and non-participants.

● Recent statistical research has shown that the FTC Program has improved the performance of Florida public schools to a modest degree. Therefore, the correct interpretation of the findings in this report are that students participating in the program have kept pace with the improvements in the public schools associated with the FTC Program.

I. Background

This is the fourth in a series of reports evaluating the Florida Tax Credit Scholarship Program, as required by the Florida Statutes, s. 1002.395(9)(j). This report provides information on private school compliance with program rules regarding required testing, describes the attributes of eligible students who participate in the program, and presents data on student test score levels and gains in the program, as well as compared with the eligible population of non-participating students.

The Florida Department of Education first awarded a contract to the University of Florida as the Independent Research Group and Professor David Figlio as the Project Director in October 2007 to collect program participants' test scores directly from the private schools. Therefore, the first year in which test score data collection could take place in real time was the 2007-08 academic year; data from the 2006-07 academic year, the first year in which testing was required, could only be collected retrospectively from private schools. It was unclear at the time the degree to which the 2006-07 academic year would make an acceptable baseline for evaluation, but it was decided that to accelerate the possibility of providing concrete information regarding testing and compliance amongst participating schools an attempt would be made to retrospectively collect as complete information from 2006-07 test scores as possible. The results of that effort were presented in the program report dated March 2008. Later reports, released in June 2009 and June 2010, presented data from the 2007-08 and 2008-09 academic years, with the 2010 report being the first to present gain scores for program participants where all test scores were collected in real time.

This report presents the results of the real-time test score collection in 2009-10. This report details key information about program participation and test scores, and compares test score gains for program students to comparable students in Florida public schools.

II. Test score collection in 2009-10

Data collection protocol

As required by s. 1002.395(8)(c)(2), participating schools administered to students an approved nationally norm-referenced test as identified by the Florida Department of Education, including the Stanford Achievement Test, Basic Achievement Skills Inventory, Metropolitan Achievement Test, Iowa Test of Basic Skills, Terra Nova, or the Preliminary Scholastic Aptitude Test and ACT/PLAN (for students in high school grades) or made provisions for participating students to take statewide assessments at a public school in accordance with s. 1002.395(7)(e). This testing was first required in the 2006-07 academic year, and the Independent Research Organization attempted to collect retroactively as many of these test scores as possible.

The 2009-10 academic year was the third year in which it was possible to collect participant test score data in real time. Pursuant to s. 1002.395(8)(c)(2), in Fall 2009 and again in Winter 2010 the Independent Research Organization contacted the 1,028 private schools that had participating students in grades three through ten during the 2009-10 school year, as reported on the October roster of program participants. The Florida Department of Education provided the Project Director with a list of all participating

students in 2009-10, as of the October participant roster, and refreshed and cross-checked against the January participant roster; of these, 15,151 were in the relevant grades, according to the state records. Schools were provided lists of the relevant students and were instructed to submit test scores to the Independent Research Organization. Schools were also informed that they must provide explanations for any missing or invalid student test scores.

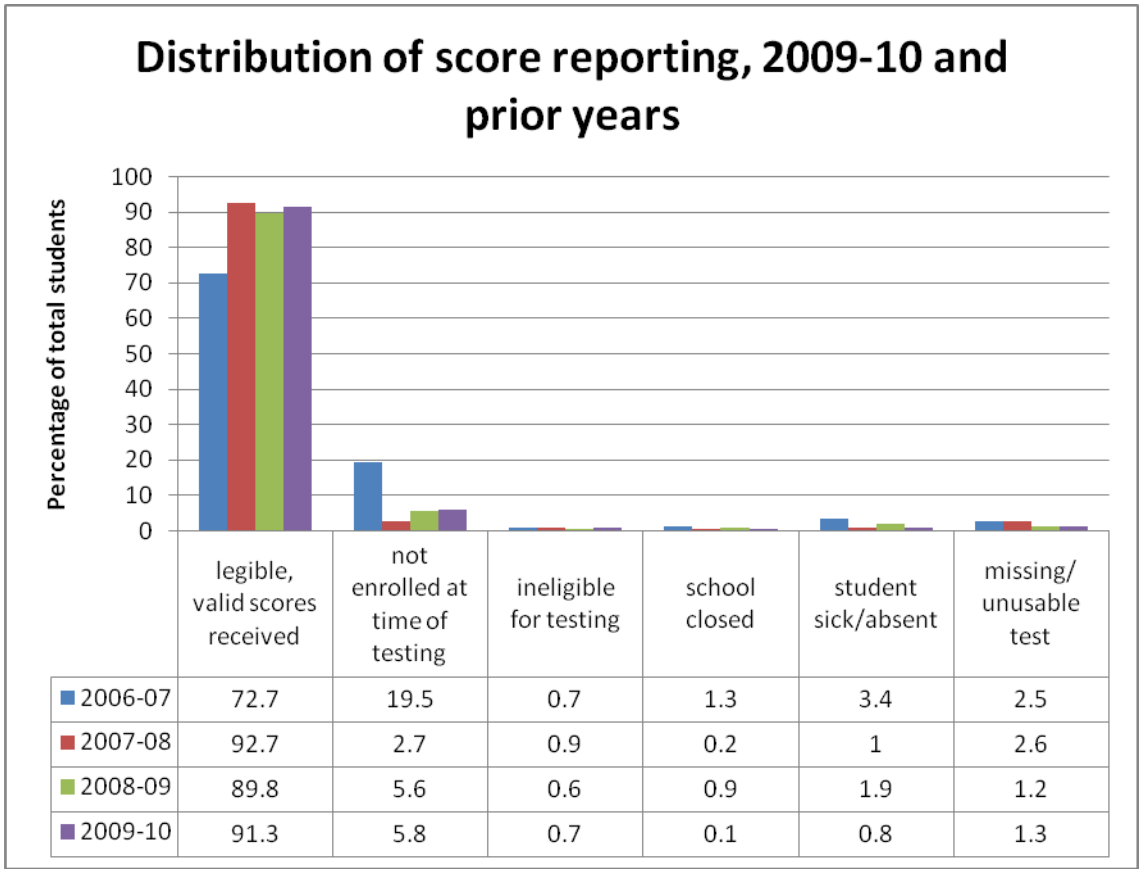
Private school compliance

In over 99 percent of cases, schools submitted photocopies of official score sheets provided to them by the relevant testing company (e.g., Pearson Assessments). In a small number of schools, the schools scored the tests themselves and forwarded to the Project Director detailed information regarding the nature of test administration and scoring. The Independent Research Organization followed up with schools that had provided partial or incomplete data, or that did not provide data regarding students who had attended school in the relevant grades but for whom no valid test score was received. Upon receipt of the test scores, the Project Director and his staff double-entered, audited and reconciled the scores, and once the scores were confirmed, the original score sheets were destroyed and the resulting electronic databases stored in accordance with s. 1002.22(3)(d)(5) of the Florida Statutes. These data were then matched with student FCAT, public schooling, subsidized lunch and disability history, when available, from the Education Data Warehouse, and with information from student scholarship applications provided by the Scholarship Funding Organizations, and then were stripped of individual identifiers such as names, social security numbers or birthdates, for the purposes of analysis.

Of the 1,028 schools with students in the relevant grades in 2009-10, the overwhelming majority provided evidence of test administration according to the specifications of the program. A small fraction of participating schools closed following the 2009-10 school year and did not provide test scores to the Project Director. In a handful of other cases, the schools administered unapproved tests or neglected to administer tests to participating students; in the case of the small number of non-compliant schools, the Project Director reported the schools to the Florida Department of Education for disciplinary action.

Of the 15,151 students in relevant grades participating in the program in 2009-10, the Independent Research Organization received valid, legible test scores for 13,829 students, or 91.3 percent of all expected students;[1] virtually all of these scores were from tests administered by the private schools themselves. This is modestly lower than the 92.7 percent figure for 2007-08, though still in the same vicinity and easily explainable for reasons described below, but higher than the 89.8 percent rate from 2008-09, and it represents maintenance of the dramatic improvement in score reporting rates over the retrospective 2006-07 score reporting, in which the comparable figure was 72.7 percent. The difference between the retrospective score reporting in 2006-07 and the real-time score reporting in 2007-08 and later years underscores the importance of collecting test score data in real time.

[1] We received 9 additional test scores following the January 21, 2011 date in which we merged score records with school records. This report excludes these 9 test scores, because they cannot be merged with the state records for the purposes of analysis.



	legible, valid scores received	not enrolled at time of testing	ineligible for testing	school closed	student sick/absent	missing/ unusable test
2006-07	72.7	19.5	0.7	1.3	3.4	2.5
2007-08	92.7	2.7	0.9	0.2	1	2.6
2008-09	89.8	5.6	0.6	0.9	1.9	1.2
2009-10	91.3	5.8	0.7	0.1	0.8	1.3

The difference between the 2007-08 and 2008-09/2009-10 percentage of program participants with valid test score gains can be explained by an uptick in the percentage of students who either arrived in the private school after the testing took place -- there is a larger fraction of students attending schools that administered the Iowa Test of Basic Skills in the fall in 2008-09 and 2009-10 as opposed to what occurred during 2007-08 -- or left the school before the time in the academic year in which the school administered testing. In 2008-09 and 2009-10, the percentage of students falling into one of these two categories increased to 5.6 percent and 5.8 percent, respectively, as opposed to the comparable figure of 2.7 percent in 2007-08. In addition, 0.6 percent of 2008-09 program participants and 0.7 percent of 2009-10 participants listed on the official roster were deemed ineligible for test score reporting pursuant to s. 1002.395(8)(c)(2) -- slightly

lower than the 0.9 percent in 2007-08. Few schools closed before reporting their scores in 2009-10; only 0.1 percent of scores in 2009-10 are missing in this way, a lower fraction than in 2008-09 and similar to the 0.2 percent in 2007-08. Taken together, the percentage of students in 2009-10 with either legible, valid score reporting or one of these other explanations was 97.9 percent, above the 96.9 percent in 2008-09 and the 96.5 percent in 2007-08.

In the remaining cases, the private school either reported the student was absent (0.8 percent, as compared with 1.9 percent in 2008-09 and 1.0 percent in 2007-08) or had some problem with test reporting (1.3 percent, as compared with 1.2 percent in 2008-09 and 2.6 percent in 2007-08.) This last category includes the school providing test scores that were illegible, not providing scores that could be compared with national norms, testing students using an unapproved test, or failing to test students at all. The percentage of schools falling into these categories continues to fall with each successive round of testing, implying that private school compliance with the testing requirements continues to improve.

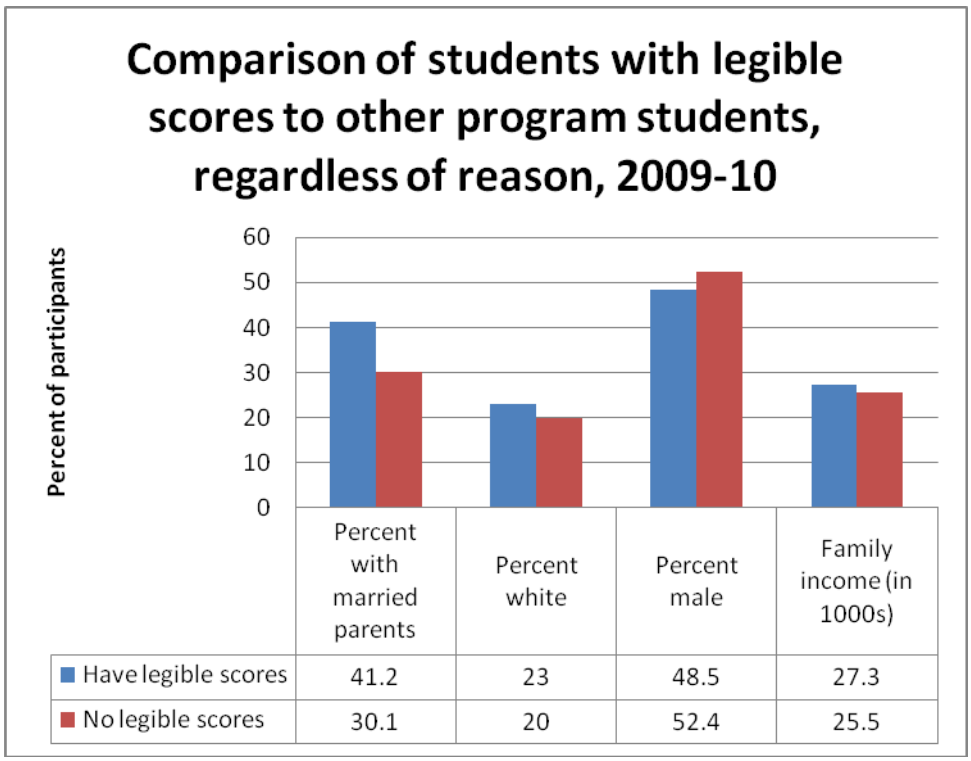
Of the students who have taken tests that were reported to the Independent Research Organization, virtually 100 percent took a test approved by the Florida Department of Education. The vast majority of the students (69.2 percent) took the Stanford Achievement Test, the nationally norm-referenced test administered to all public school students in the relevant grades in Florida through 2007-08, while another 20.6 percent took the Iowa Test of Basic Skills and 3.7 percent took the Terra Nova test. The other students took a number of other tests, most notably the Basic Achievement Skills Inventory, taken by 1.7 percent of students, the PSAT/NMSQT, taken by 1.3 percent of

students, the ACT/PLAN, taken by 1.3 percent, and the Metropolitan Achievement Test, taken by 0.5 percent. 1.7 percent took other approved tests. No students took a test that was not approved by the Florida Department of Education.

Schools have flexibility as to when they administer their exams, and 20 percent of participating students took their exam in the fall months. These scores are less likely to be directly comparable to public school students' tests than are those taken during the time immediately surrounding the public schools' test administration. The tests most typically taken in the fall months are the PSAT/NMSQT and the Iowa Test of Basic Skills. The latter case is driven strongly by Florida Catholic schools' uniform assessment of students in October using the Iowa Test of Basic Skills. It is likely to be inappropriate to directly compare status scores of tests administered in March to tests administered in October, as they likely have very different purposes. This speaks to the importance of measuring student learning gains rather than levels comparisons, and also indicates that it would be useful to conduct a fall-spring concordance study if at all possible.

Similarity of students with received legible tests to the overall scholarship population

In 2009-10, the rate of successful test reporting remained at the high levels of previous years. However, around eight percent of the potentially-tested population of students was not tested (due in large part to students arriving at school after testing or leaving a school before testing, or to students being sick or absent during the testing period), so it is important to gauge whether the students whose test scores were successfully reported are comparable to the overall population of students enrolled in the scholarship program at any time during 2009-10.

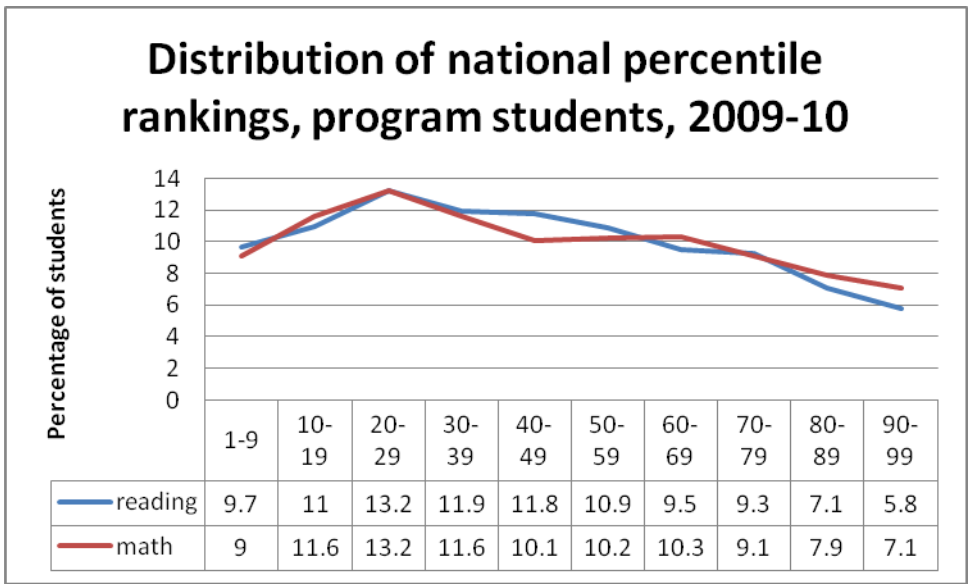


As can be seen from the accompanying figure, there is evidence that students whose test scores were successfully reported are modestly more advantaged than other program participants whose scores were not successfully reported, based on data from the families' scholarship applications. Students whose scores were successfully reported come from families with somewhat higher incomes, with parents considerably more likely to be married, and are more likely to be white, than are students whose scores were not successfully reported, for whatever reason. These differences may have been expected, as highly transient students are likely to be less advantaged, and are more likely to have not been tested because they changed schools. However, even among students who were still in the school at the time of testing, those missing score reports tend to be less advantaged (with family incomes eight percent lower), with unmarried parents (23 percent married versus 41 percent married), nonwhite (19 percent white versus 23 percent white), and male (52 percent male versus 49 percent male.) These differences therefore

underscore the importance both (1) of obtaining as full a collection of test score data as possible, and (2) of measuring student test score gains. It is not obvious that students with missing test scores would have had higher or lower gain scores than those with test scores available. It is also important to note that while public school records do not include data on family income or parental marital status, we observe that those missing public school test scores are also more likely to be nonwhite and eligible for free or reduced price lunches.

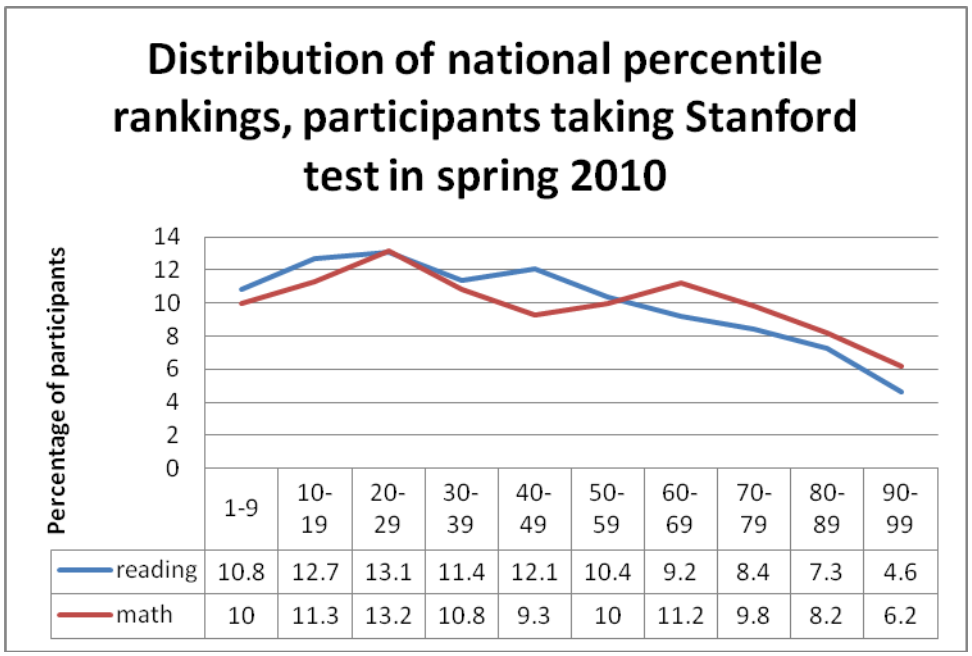
III. Test scores of 2009-10 program participants

Because program participants may take any number of nationally norm-referenced tests and because private schools have some flexibility in the form in which these test scores are reported and the time of year the test is administered, the only way to ensure reasonable comparability across schools and program participants is to report national percentile rankings. National percentile rankings are desirable because they are compared against a nationally-representative group of students; so long as the national norms for one test (such as the Stanford Achievement Test) are comparable to the national norms for another test (such as the Iowa Test of Basic Skills) then there is no inherent bias associated with comparing the national percentile rankings of one student taking a certain test to those of another student taking a different test.



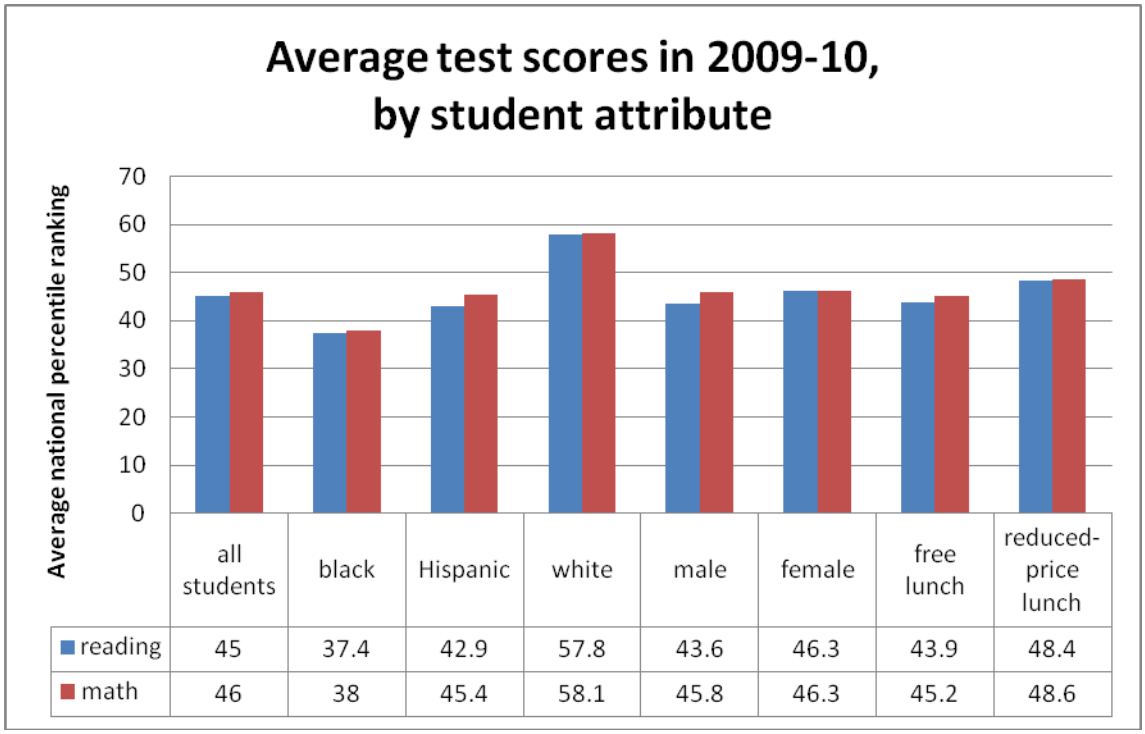
	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
reading	9.7	11	13.2	11.9	11.8	10.9	9.5	9.3	7.1	5.8
math	9	11.6	13.2	11.6	10.1	10.2	10.3	9.1	7.9	7.1

The chart above presents the basic distribution of national percentile rankings among FTC students participating in the program in 2009-10. The typical student in the program scored at the 45th percentile in reading and the 46th percentile in mathematics. This is basically unchanged from 2007-08 or 2008-09, in which the typical student in the program scored at the 44.8th (45.3rd in 2008-09) percentile in reading and the 46.3rd (46.2nd in 2008-09) percentile in mathematics. Were the distributions to be limited to those taking the Stanford Achievement Test in the spring -- the most comparable to the students in the public schools -- the typical student would have scored at the 44th percentile in reading and the 47th percentile in mathematics. Given that the distributions of test scores are sufficiently similar for those taking the Stanford Achievement Test in the spring versus the full set of scholarship recipients, this report will focus on the full set of students for whom data are available, regardless of test administered.



	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
reading	10.8	12.7	13.1	11.4	12.1	10.4	9.2	8.4	7.3	4.6
math	10	11.3	13.2	10.8	9.3	10	11.2	9.8	8.2	6.2

The next chart presents average norm referenced test scores, expressed in terms of national percentile rankings, for various subsets of the FTC Scholarship recipient population, stratified by race, sex, income, and parental marital status. Income is expressed in terms of fraction of the poverty line, to reflect the fact that families of different sizes have different official measures for poverty; those with family incomes below 130 percent of the federal poverty line are eligible for free school meals, while those with incomes between 130 and 185 percent of the poverty line are eligible for reduced-price meals. As can be observed in the table, white participants tend to score better than do minority participants, females tend to perform better than do males, students with married parents tend to score better than do students with unmarried parents, and relatively high-income families tend to score better than do relatively low-income families. These averages closely mirror the figures presented in previous years' reports.

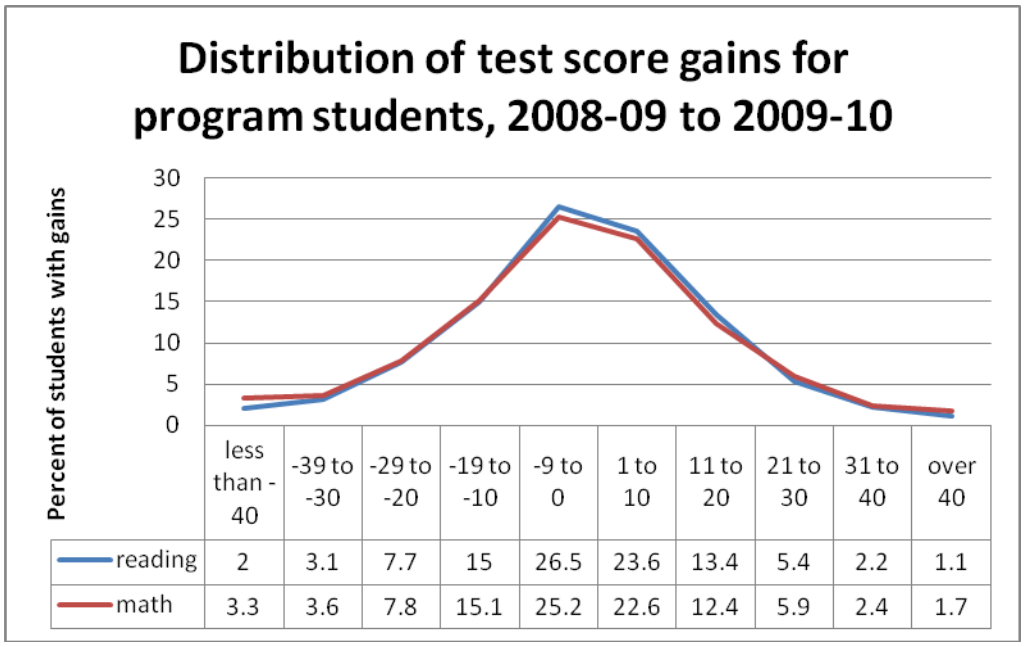


	all students	black	Hispanic	white	male	female	free lunch	reduced-price lunch
reading	45	37.4	42.9	57.8	43.6	46.3	43.9	48.4
math	46	38	45.4	58.1	45.8	46.3	45.2	48.6

Test score gains for FTC Scholarship program participants

The relevant statutes call for comparisons of test score gains for FTC Scholarship Program students to similar students in public schools. Because the test scores in both 2008-09 and 2009-10 are measured in terms of national percentile rankings, gain scores can only be interpreted as changes in national percentile rankings, and are therefore subject to issues regarding ceiling effects (where students whose scores are already in the high percentiles cannot gain much more) and floor effects (where students whose scores are already in the low percentiles cannot lose much more ground.) Ceiling and floor effect concerns are mitigated for students whose initial national percentile ranking falls in the middle portions of the initial test score distributions, which is the case for the vast

majority of students participating in the FTC Scholarship Program (as well as in the public schools.)



	less than -40	-39 to -30	-29 to -20	-19 to -10	-9 to 0	1 to 10	11 to 20	21 to 30	31 to 40	over 40
reading	2	3.1	7.7	15	26.5	23.6	13.4	5.4	2.2	1.1
math	3.3	3.6	7.8	15.1	25.2	22.6	12.4	5.9	2.4	1.7

The chart above presents information on the distribution of program participants' test score gains in reading and mathematics for the set of 6,667 students with legible reading scores and 6,687 students with legible mathematics scores in both 2008-09 and 2009-10. The mean gain for program participants is -1.2 national percentile ranking points in reading and -1.7 national percentile ranking points in mathematics, numbers that are numerically slightly worse but statistically indistinguishable from past years' average gains scores. In other words, the typical student participating in the program tended to maintain his or her relative position in comparison with others nationwide. It is important to note that these national comparisons pertain to all students nationally, and not just low-income students -- the students eligible to participate in the FTC Scholarship Program. It is also important to note that while the typical gain in national percentile rankings compared with the nation as a whole is essentially zero for program participants,

this statistic masks considerable variation in individual students' gains. For instance, 10.8 percent of students participating in the program lost 20 or more percentile points in reading relative to the nation as a whole between 2008-09 and 2009-10, while 8.7 percent of program participants gained 21 or more percentile points in reading over this same time period. Furthermore, these comparisons are very similar when limited to students taking the Stanford Achievement Test during the spring: -1.4 national percentiles in reading and -1.2 national percentiles in mathematics.)

IV. Comparisons with public school test-takers

One important purpose of this evaluation is to compare the relative year-to-year gains in the test score of FTC Scholarship Program students to those of comparable public school students. This report compares the distribution of test score gains between 2008-09 and 2009-10 for the two groups of students. It is very important to note, however, that differences in the gains should not be interpreted as causal, for two principal reasons.

One reason to not interpret differences in test score gains between public school students and FTC Scholarship Program students as causal per se involves the fact that students and families choose whether to participate in the program, and these choices introduce "selection bias" into any comparison of test score gains.[2] In addition, selection into a public school comparison group is not random. All FTC Scholarship Program students are certified to be low-income, but only three percent of public school free- or

[2] A technical description of selection into the FTC Scholarship Program is provided in David Figlio, Cassandra Hart, and Molly Metzger, "Who Uses a Means-Tested Scholarship, and What Do They Choose?" published in the *Economics of Education Review* in 2009. A brief summary of the key points of that paper is provided in this report.

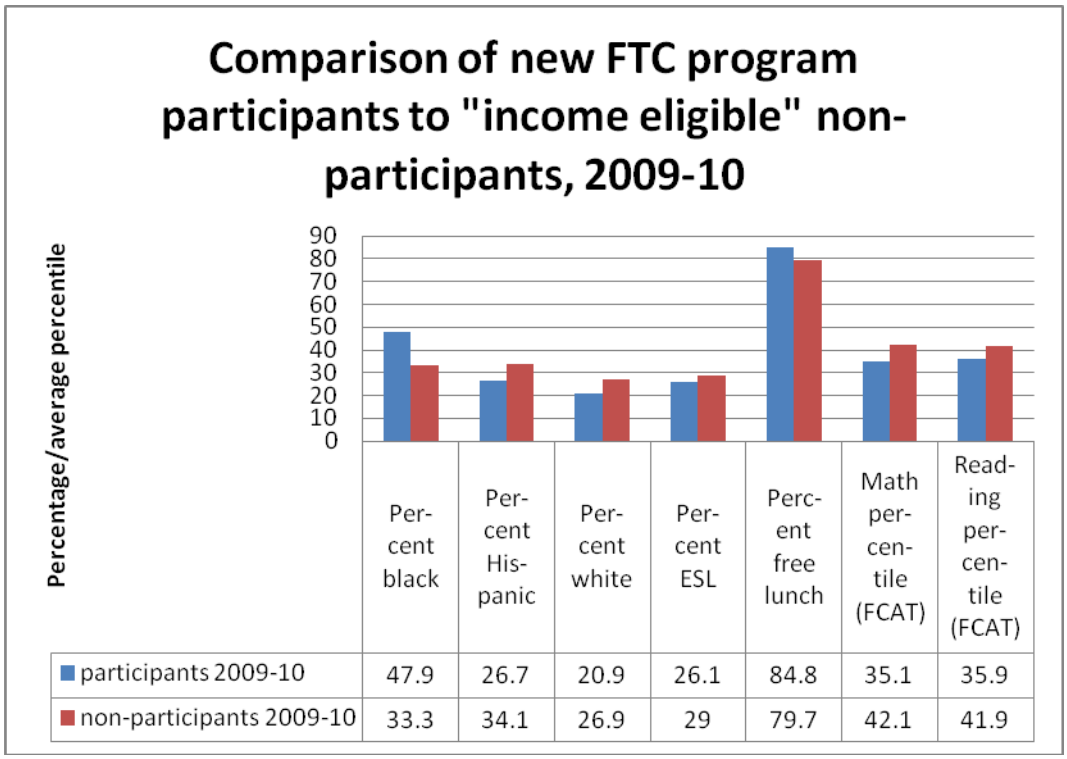
reduced-price lunch students' family incomes are audited, so some fraction of the public school comparison population may actually be of higher income than the program allows. The results of these audits strongly suggest that many public school students receiving free or reduced-price lunches are not from families with comparable incomes to those participating in the FTC Scholarship Program. Therefore, it seems to be clear that school meals recipients in the public schools are not a very effective comparison group for FTC Scholarship Program participants, because their family incomes are likely to be considerably different. While it is impossible to measure just how large these differences are, the results of the audits indicate that they may be substantial.

Taken together, these two factors indicate that direct comparisons of average test score gains in the public sector versus FTC Scholarship Program participants, while informative, should not be interpreted as effects of the program on student test score gains. This report presents these basic comparisons of student test score gains in the public and private sectors, and then presents the results of more sophisticated empirical methods aimed at more compellingly deducing the causal effect of participating in the FTC Scholarship Program.

Summary of key selection findings

Before directly comparing student test score gains between FTC Scholarship Program participants and others in the public sector, who may or may not be ultimately eligible for program participation, it is important to gauge the degree to which these comparisons are likely to be apples-to-apples comparisons. This report therefore begins with a brief summary of some of the key findings of the technical paper mentioned above

that describes selection into the program. Any selection findings could reflect either of the two factors -- differential self-selection amongst eligible students; or systematic ineligibility amongst non-participating students who still receive subsidized school meals -- but these findings are highly informative in either case.



The most natural way to make comparisons is to consider a set of students who all spent the prior year in Florida public schools and who received subsidized school meals, making them plausibly eligible to participate in the program. This report employs the most recent data available at the time of writing -- students who spent the 2008-09 academic year in the Florida public schools, so one can compare the students who entered the FTC Scholarship Program in 2009-10 versus potentially comparable students who did not enter the program in that year but remained free or reduced-price lunch eligible in 2009-10, according to Department of Education records. We exclude students with disabilities who could participate in the McKay Scholarship Program. The chart above

presents some basic facts about FTC Scholarship Program participants relative to other potentially income-eligible students. In order to compare similar populations across bars, we restrict analysis to students who had taken either a reading or math test in public school in 2008-09; prior research suggests that this is very similar to the overall population of potential program participants who spent the prior year in a public school. We also limit the analysis to students who would be in grade 10 or below in 2009-10, so that this reflects the set of students for whom a test score is possible. By these standards, there were 2,408 new students in the FTC Scholarship program from this sample and 641,873 students who remained in the public schools and continued on subsidized school lunches in 2009-10.

One observes that FTC Scholarship Program participants differ from non-participants on all of the characteristics easily observed in the administrative record. Scholarship participants are more likely than non-participants to be black, and less likely to be Hispanic or white, and participants are less likely than are non-participants to speak English as a second language. Scholarship participants are more economically disadvantaged than are non-participants on average. While all children in both the participant and non-participant groups were self-reported to be eligible for subsidized lunch at some point in the 2008-09 school year, participants were more likely to qualify for free lunch as of the last survey taken, while non-participants were more likely to qualify only for reduced-price lunch, indicating that scholarship participants were relatively disadvantaged, even conditional on reported income eligibility. Finally, and perhaps most importantly, scholarship participants have significantly poorer test performance in the year prior to starting the scholarship program than do non-

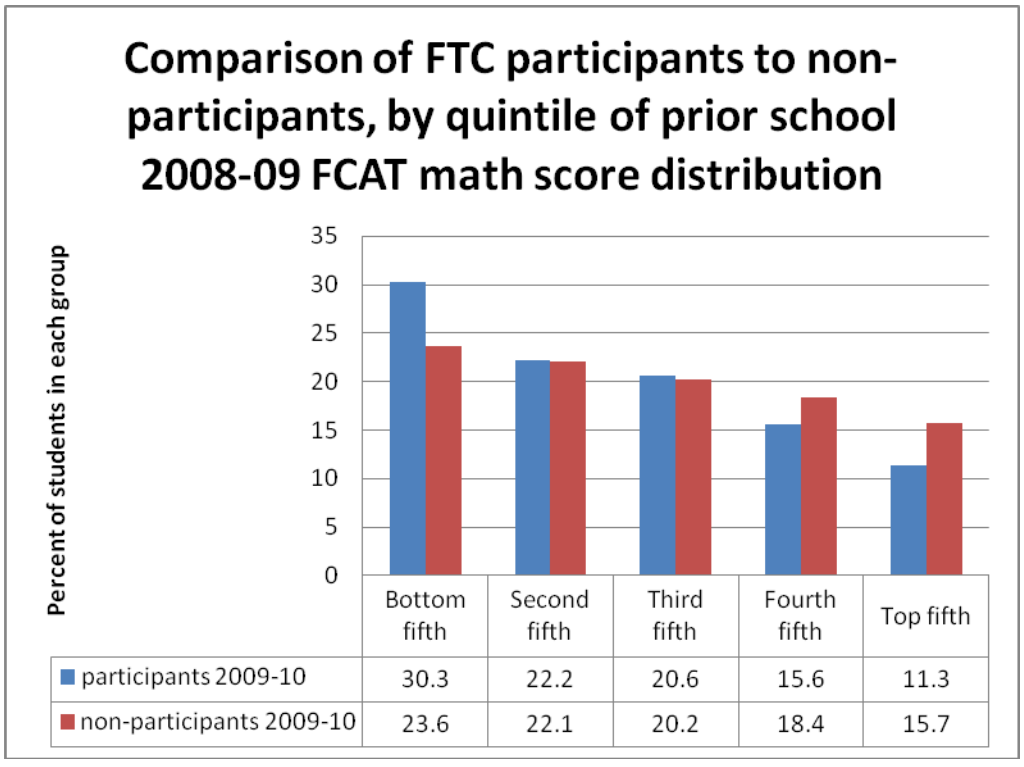
participants. On both the FCAT mathematics and FCAT reading tests, 2009-10 non-participants out-performed 2009-10 scholarship participants in the 2008-09 school year, when both groups were still attending public schools.[3] All of these differences are large in magnitude and are statistically significant, and indicate that scholarship participants tend to be considerably more disadvantaged and lower-performing upon entering the program than their non-participating counterparts. These differences are very similar to those observed in years past and reported in prior program reports.

The mean differences in 2008-09 performance between public school students who would ultimately participate in the FTC Scholarship Program in 2009-10 and those who are plausibly income-eligible but who remained in Florida public schools in 2009-10 are compelling, but there are numerous remaining selection questions. For instance, these results are consistent both with the idea that relatively high-performing students from low-performing schools are the ones selecting into the scholarship program, as well as with the idea that relatively low-performing students, regardless of school, are the ones selecting into the program. It is clear that these two possibilities have very different implications for the interpretation of differential selection into the program.

Unlike previous years, in which FTC Scholarship Program participants came disproportionately from lower-performing schools, the newcomers to the program in 2009-10 came from comparable schools, according to Florida Department of Education school grades in 2009, as did eligible students who did not participate in the program. Amongst the students new to the program in 2009-10, 49.6 percent came from schools

[3] Note that the numbers reported in the test score comparisons are different in this report from those in previous reports. In previous reports, I reported the prior-year norm referenced test national percentile. That is not possible to do in 2008-09, as students in public schools took only the FCAT. Therefore, in this report, I present information based on the state percentile ranking on the FCAT. All comparisons are qualitatively very similar to those presented in prior years' reports.

graded "A" by the Florida Department of Education in 2009, as compared with 50.8 percent of those public school students eligible for free or reduced-priced lunches who did not participate. At the other extreme, 9.0 percent came from schools graded "D" or "F" by the Florida Department of Education in 2009, as compared with 9.8 percent of those public school students eligible for free or reduced-priced lunches, and 21.3 percent came from schools graded "C" or below by the Florida Department of Education in 2009, as compared with 20.0 percent of those public school students eligible for free or reduced-priced lunches.



One selection pattern that remains from prior years, and in fact has gotten stronger, is that regardless of the performance level of the public school that FTC Scholarship Program participants came from, these students tended to be lower-performing before they entered the program. As can be seen in the accompanying figure, 30.3 percent of students who would select into the program were in the bottom fifth of

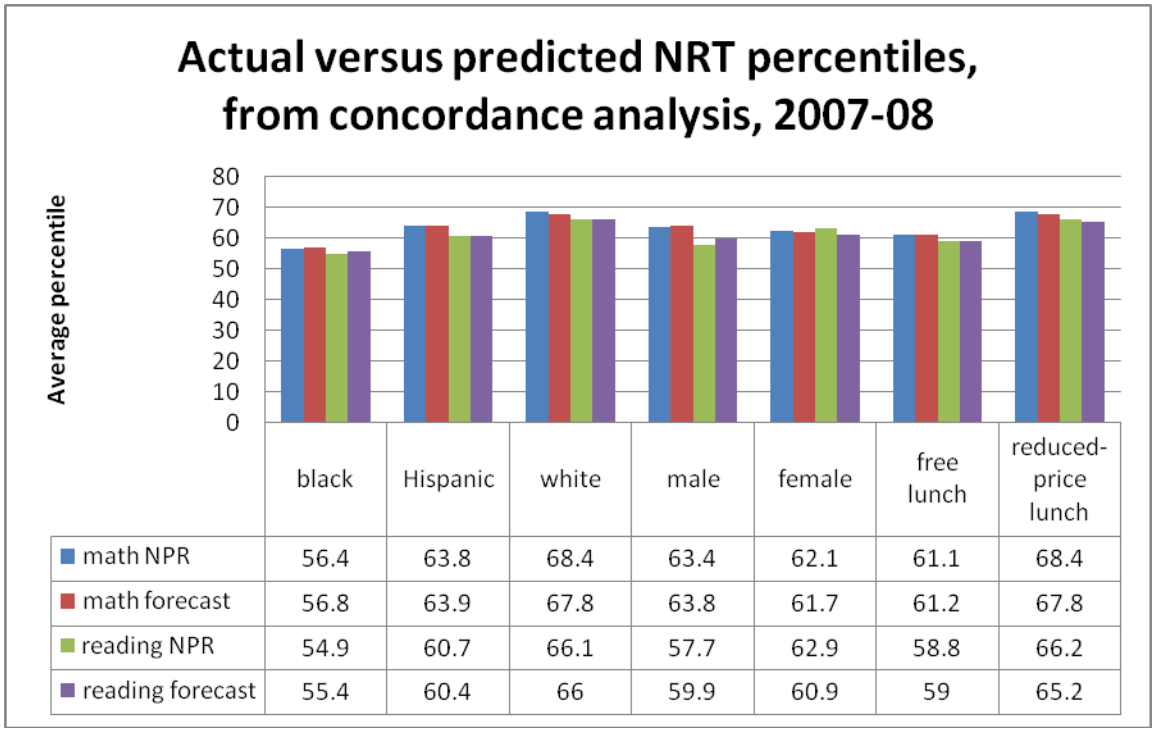
their prior public school's mathematics FCAT test score distribution, while only 23.6 percent of non-participating free- or reduced-price lunch students were in the bottom fifth of the distribution in the prior public school. This gap of 6.7 percentage points is considerably more pronounced than the 4.4 percentage point gap observed in the previous year's report. (Similar differences are present in terms of reading scores.) At the top of the test score distribution, only 11.3 percent of students who would select into the program were in the top fifth of their prior public school's mathematics test score distribution, as compared with 15.7 percent of free- or reduced-price lunch students in the top fifth of the distribution in the prior public school; the 4.3 percentage point gap is larger than the 3.3 point gap observed in last year's report. Clearly, public school students who ultimately became program participants are more likely to be the relatively lower-performing students in their schools.

Computing gains of public school students

The fact that program participants are not a random sample of potential students makes clear that direct comparisons of gains of program participants to non-participants will not yield causal estimates of the effects of the program on participating students. Nonetheless, it is still very worthwhile to benchmark the distribution of measured student learning gains amongst program participants against the distribution of learning gains amongst potentially eligible public school students who elected not to participate in the program.

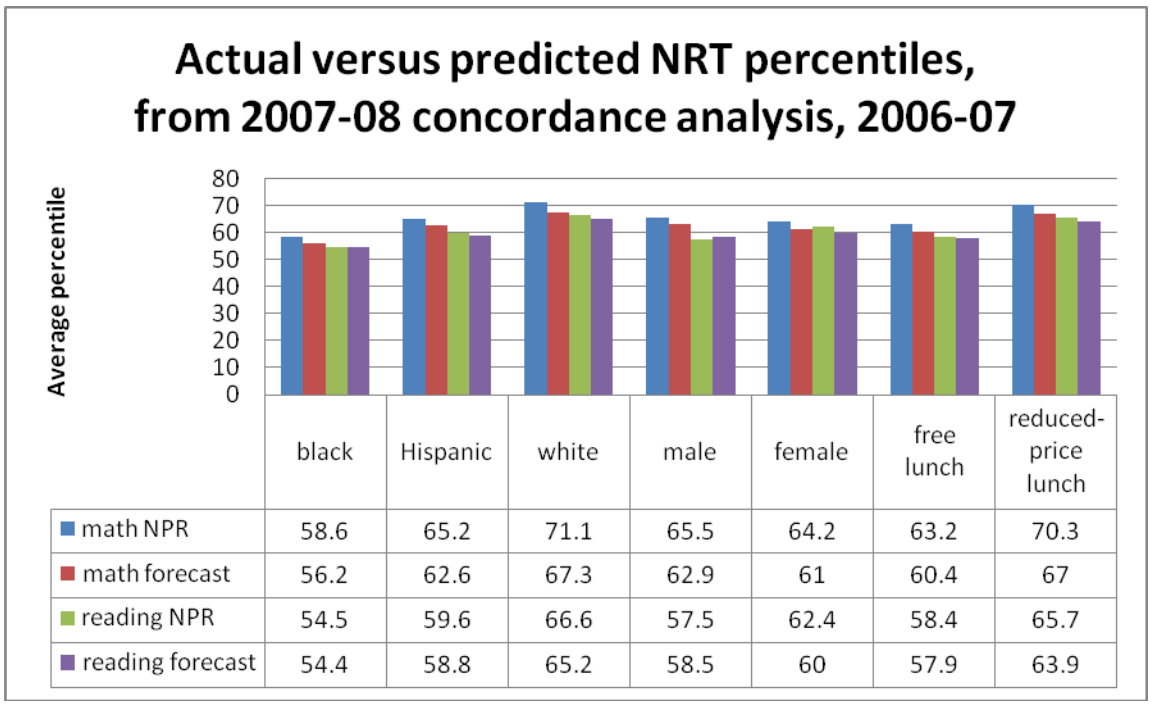
An additional complication is that public school students no longer take a directly comparable nationally norm-referenced test, making comparisons across sectors

somewhat more challenging. Through the 2007-08 academic year, public school students took both the criterion-referenced FCAT as well as the norm-referenced Stanford Achievement Test, but the norm-referenced test administration was ended due to budgetary concerns. That said, it is still possible to make comparisons between program participants and non-participants by performing an analysis of the concordance between FCAT scores and Stanford Achievement Test scores. In principle, a concordance analysis predicts what the norm-referenced national percentile would have been given the level of the FCAT score. This concordance analysis was conducted with the most recent data -- the 2007-08 academic year -- for which the same Florida students took both the FCAT and the norm-referenced test. In practice, for every value of the FCAT developmental scale score in each grade level, I computed the mean NRT national percentile ranking and assigned this mean national percentile ranking as the predicted NRT score to accompany a given FCAT developmental scale score for a given grade level. Because students from different groups might have different concordances between the two tests, the predictions were made using the set of students who were eligible for subsidized school means in both 2007-08 and 2008-09. The results of this concordance analysis are highly robust to other population definitions.



	black	Hispanic	white	male	female	free lunch	reduced-price lunch
math NPR	56.4	63.8	68.4	63.4	62.1	61.1	68.4
math forecast	56.8	63.9	67.8	63.8	61.7	61.2	67.8
reading NPR	54.9	60.7	66.1	57.7	62.9	58.8	66.2
reading forecast	55.4	60.4	66	59.9	60.9	59	65.2

The above figure compares mean actual national percentile rankings from the 2007-08 Stanford Achievement Test to predicted national percentile rankings for the same students, based on the concordance analysis conducted in 2007-08, for several subgroups of students. As can be seen in the figure, the actual and predicted scores line up closely across the subgroups. The only place where the match is not as precise involves reading across the genders: The concordance analysis tends to modestly overpredict male reading scores and modestly underpredict female reading scores. However, in general, the concordance analysis using 2007-08 data tends to predict norm-referenced test scores very well. Indeed, the correlation between actual and predicted math scores in 2007-08 is 0.84 and the correlation between actual and predicted reading scores in 2007-08 is 0.78.

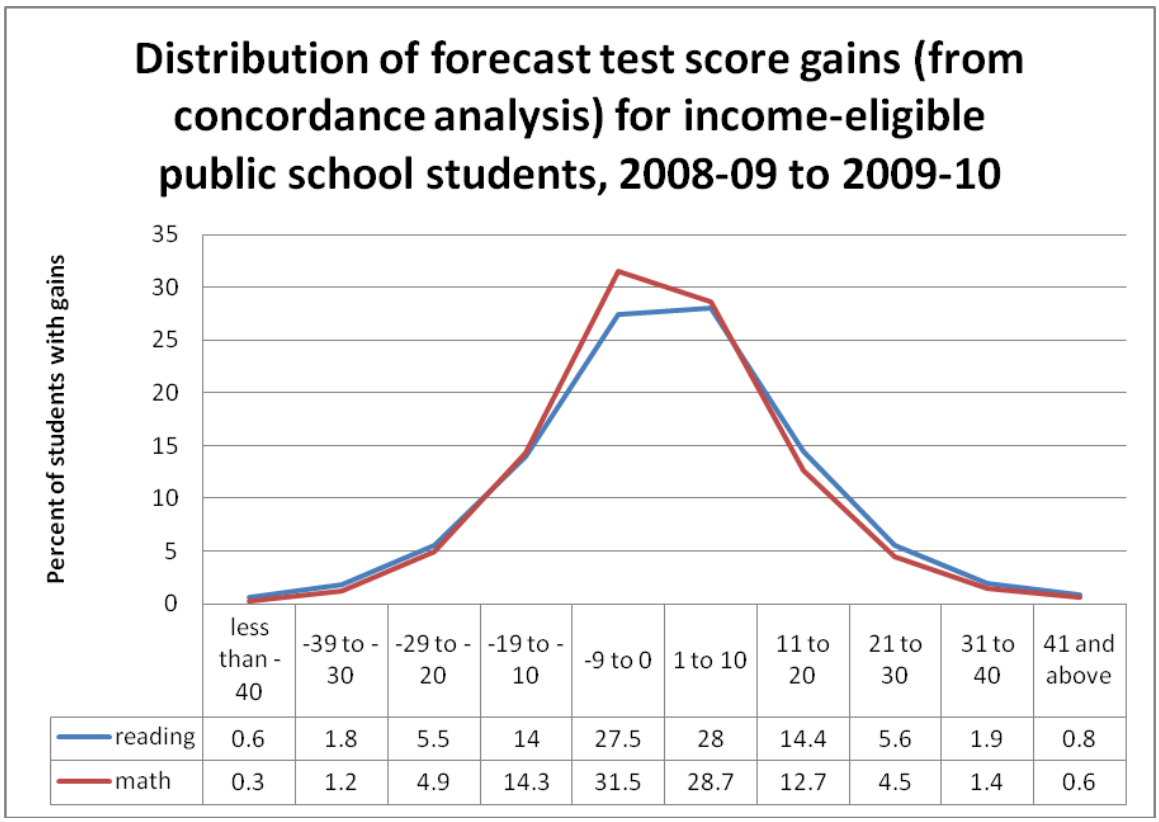


	black	Hispanic	white	male	female	free lunch	reduced-price lunch
math NPR	58.6	65.2	71.1	65.5	64.2	63.2	70.3
math forecast	56.2	62.6	67.3	62.9	61	60.4	67
reading NPR	54.5	59.6	66.6	57.5	62.4	58.4	65.7
reading forecast	54.4	58.8	65.2	58.5	60	57.9	63.9

Of course, the purpose of the concordance analysis is to predict norm-referenced test scores in years when there are no norm-referenced scores. To test the potential validity of the concordance analysis, we back the analysis up a year, and predict 2006-07 norm-referenced test scores using 2006-07 FCAT scores, but with the concordance metrics developed using 2007-08 data. As can be seen in the above figure, the relationship between actual NRT scores and predicted NRT scores based on the concordance analysis remains very high: The correlation between 2006-07 predicted scores and 2006-07 actual scores is 0.82 for math and 0.79 for reading. In practice, it appears as if the concordance analysis modestly underpredicts math scores in 2006-07, so the relationship is not perfect, but the correlations are very strong. One can draw similar conclusions when comparing the realized gain scores on the NRT to the forecast gains on the NRT between 2006-07 and 2007-08: In reading, the mean forecast gain based on the FCAT concordance analysis is 2.0 percentile points while the mean realized NRT gain is a very similar 1.4 percentile points. In mathematics, the difference is greater: The mean

forecast gain is 2.1 percentile points while the mean realized gain is -0.6 percentile points. It is not clear whether this implies that the forecasts for the concordance analysis will overstate or understate the true gains between 2007-08 and 2008-09 -- as both are possible, depending on the interpretation of the differences between 2006-07 and 2007-08 -- but the results do indicate that the concordance analysis is perhaps more successful in the case of reading rather than mathematics. The good news, from the point of view of making valid comparisons between gain scores of private school students who take exclusively norm-referenced exams and those of public school students who take exclusively the FCAT, is that is appears possible to make reasonable comparisons across these two sectors even when the examinations taken are different.

With these provisos in mind, one can now turn to measuring test score gains for the public school students who received subsidized school meals in both 2008-09 and 2009-10. This report employs the concordance metrics described above to compute predicted NRT scores in 2008-09 and 2009-10 based on the student's actual FCAT scores in the two years.



	less than -40	-39 to -30	-29 to -20	-19 to -10	-9 to 0	1 to 10	11 to 20	21 to 30	31 to 40	41 and above
reading	0.6	1.8	5.5	14	27.5	28	14.4	5.6	1.9	0.8
math	0.3	1.2	4.9	14.3	31.5	28.7	12.7	4.5	1.4	0.6

The distribution of test score gains amongst public school students is very similar to the distribution of gains amongst program participants. The mean gain in the public school comparison group is 2.5 percentile points higher than the mean gain amongst program participants in reading and 2.4 percentile points higher in mathematics, but given the selection issues mentioned earlier in this report, these mean gain differences should not be considered to be meaningful. Participating schools have more students in the tails of the distribution -- those with gains or losses of more than 20 percentile points -- than the public school students, but the differences in the extremes may be due in part to the concordance analysis. In summary, both distributions of test score gains are in the same ballpark with some modest evidence that public school gains are mildly larger than private school gains. We turn next to a more causal analysis to gauge the degree to which these differences in test score gains can be attributable to program participation.

V. Causal estimates of the effects of program participation on student test score gains using regression discontinuity models

As mentioned above, families choose to participate in the FTC Scholarship Program for a wide variety of reasons, and selection into the program is definitely not random. Indeed, there is strong evidence that those who participate in the program are substantially more disadvantaged and lower-achieving than are those who are likely income-eligible but do not participate in the program. These patterns have been observed in every cohort studied to date, and if anything are more pronounced in more recent cohorts.

The purest way to gain estimates of the causal effect of program participation on the scores of the participants is to conduct an experiment, in which people apply for scholarships and are randomly selected for participation in the program via lottery. Comparisons between program applicants that win the lottery and those that lose the lottery can then be interpreted as causal estimates of the effects of program participant on student outcomes. Such an experiment has high *internal validity* -- it can be clearly interpreted as causal -- but it may not have high *external validity* -- as the people who apply for a scholarship may not be representative of the overall candidate population. That said, at the least, this type of analysis would provide causal estimates of the effects of program participation for the set of people who wanted to participate in the program -- arguably still an important population.

Of course, participation in the FTC Scholarship Program is not governed by lotteries, and therefore an experimental evaluation of the consequences of participation is not possible. However, it is possible to evaluate the program using *quasi-experimental*

statistical methods that emulate experimental conditions. This section of the report provides the best available attempt to use quasi-experimental methods to estimate the causal consequences of program participation. Specifically, we use a technique called *regression discontinuity design* to measure the effects of program participation.

Regression discontinuity methods are most useful when program participation is based on strict programmatic rules, where two very similar individuals who would be virtually identical *but for* where they stack up along the dimension where selection takes place end up receiving very different treatment. The FTC Scholarship Program is a perfect example of this type of situation: In order to participate, families must have incomes not greater than 185 percent of the poverty line. It is unlikely that an individual with family income of 186 percent of the poverty line is really any different than an individual with a family income of 185 percent of the poverty line, so if it is possible to directly compare these individuals we might be able to get stronger purchase on the causal question at hand.

One important potential problem with this type of analysis in the present setting is that we only observe family income for individuals who *apply* for the scholarship program. But in a world with perfect information, only income-eligible families would apply for scholarships. Therefore, this analytic approach will only work in the present situation if a sufficiently large number of people are confused about their family's potential eligibility for the program. This could happen if many people believe that partial scholarships may be available -- something that is now possible for those renewing the scholarship -- or that the scholarship income rules are not firm. One reason why this confusion could possibly take place is that there are different income cutoffs for

participation depending on whether a student is a new or returning student; since some families can receive full scholarships with incomes of 200 percent of the poverty line (or renewals of up to 230 percent of the poverty line) and others must have an income of 185 percent of the poverty line or below, some families may erroneously believe that they are eligible when they are not.[4] Families may also be confused by the fact that the federal poverty line depends on household size rather than just family income. Any analysis would have to demonstrate that there are a sizeable number of people who applied for the program but could not participate because of ineligibility.

Another important potential problem with this type of analysis is that families may change their behaviors in order to qualify for the policy. In this case, a family with income that would be around 185 percent of the poverty line might choose to work less in order to qualify for the program, because the value of a scholarship to the family would generally be much higher than the lost wages associated with having an income of, say, 180 percent of the poverty line rather than 190 percent of the poverty line. If people are making these types of choices, one would observe the attributes of families just barely eligible to be quite different from families that are just barely ineligible. Therefore, any analysis would also have to gauge the degree to which this is the case.

This regression discontinuity analysis concentrates on students who spent 2007-08 in the public schools and applied for a scholarship for the 2008-09 academic year. All told, there were 4,612 students for whom this was true *and* the student took standardized tests in the public schools in 2007-08. Only 1,740 (37.7 percent) of these applicants

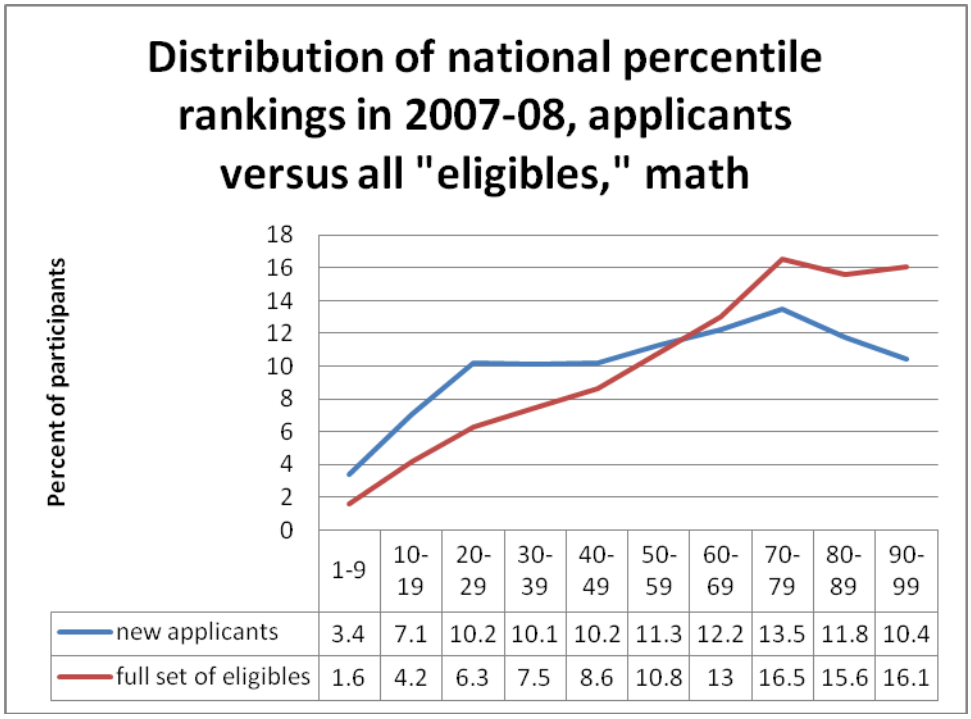
[4] Partial scholarships were not available at the time of the applications that we consider for the purpose of this analysis. However, they are currently available and were being discussed at the time that people were deciding whether to apply for the scholarship.

ultimately participated in the FTC Scholarship Program in 2008-09.[5] Students might not participate in the program for any number of reasons, including an inability to find a good match with a school, financial constraints, and other reasons, but the proposed regression discontinuity model relies on there being a substantial number of applicants whose incomes rendered them ineligible to participate in the program. In the study population there were 341 (7.4 percent of the total) with family income above 185 percent of the poverty line. Of these, 17.9 percent had family incomes between 185 and 190 percent of the poverty line and 41.4 percent had family incomes between 185 and 200 percent of the poverty line. On the other hand. 17.9 percent had incomes over 250 percent of the poverty line, and 5.0 percent had incomes over 300 percent of the poverty line. The fact that there exists a reasonably large number of applicants above the family income cutoff implies that there may be sufficient sample size to conduct the proposed regression discontinuity analysis. Moreover, the threshold of 185 percent of the poverty line is consequential for applicants: Only a trivial number of applicants with family incomes over 185 percent of the poverty line ultimately participate in the program for the first time in 2008-09.[6] (We are excluding returning scholarship students from this analysis.)

On the other hand, in the matter of the external validity of the analysis, there is strong evidence to suggest that the applicants for the FTC Scholarship Program for the 2008-09 school year are not representative of the overall population of potential

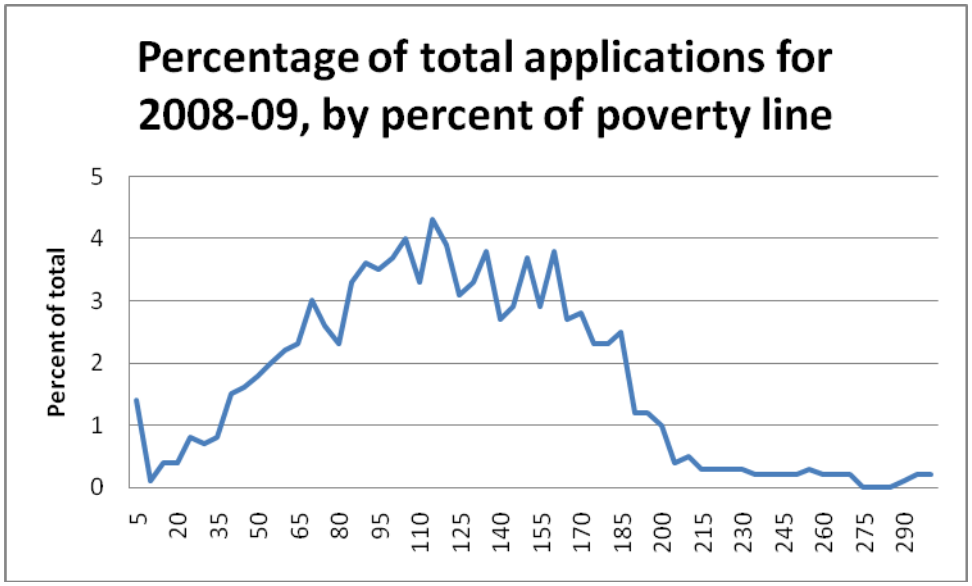
[5] 40.4 percent of income eligible students who applied participated in the program in 2008-09.
[6] Based on the figures that I received from the scholarship funding organizations, 18 students in the sample had recorded family incomes over 185 percent of the poverty line but still participated in the program in 2008-09. It is possible that some people who initially reported incomes that would have been ineligible ultimately appealed the decision to be excluded from the program, and were later included. I do not have information about whether this could explain why a small fraction of people who are recorded with ineligible incomes still participate in the program.

participants. The figure below presents the distributions of the public school mathematics norm-referenced test percentile rankings in 2007-08 of new applicants to the program for 2008-09 and the full set of potential program eligibles. As can be seen, the set of applicants tends to be considerably lower performing than the set of potentially eligible students. (The differences for reading are equally pronounced.) This result is unsurprising, given the previously-reported results regarding differential selection into the program, with program participants being considerably lower-performing in prior years than non-participants. That said, these results suggest that this analysis is probably best thought of as the estimated effects of program participation for the types of students who would apply to participate in the program. This may be exactly the right population to consider, but it is important to note that the results should not be seen as necessarily generalizable to the population of eligible students as a whole. Moreover, the regression discontinuity analysis is best considered an estimate of the students on the margin of eligibility; therefore, while we have solid estimates of the causal effect of program participation for students around 185 percent of the poverty line, it is not known whether these estimates are relevant for students far from the eligibility cutoff.



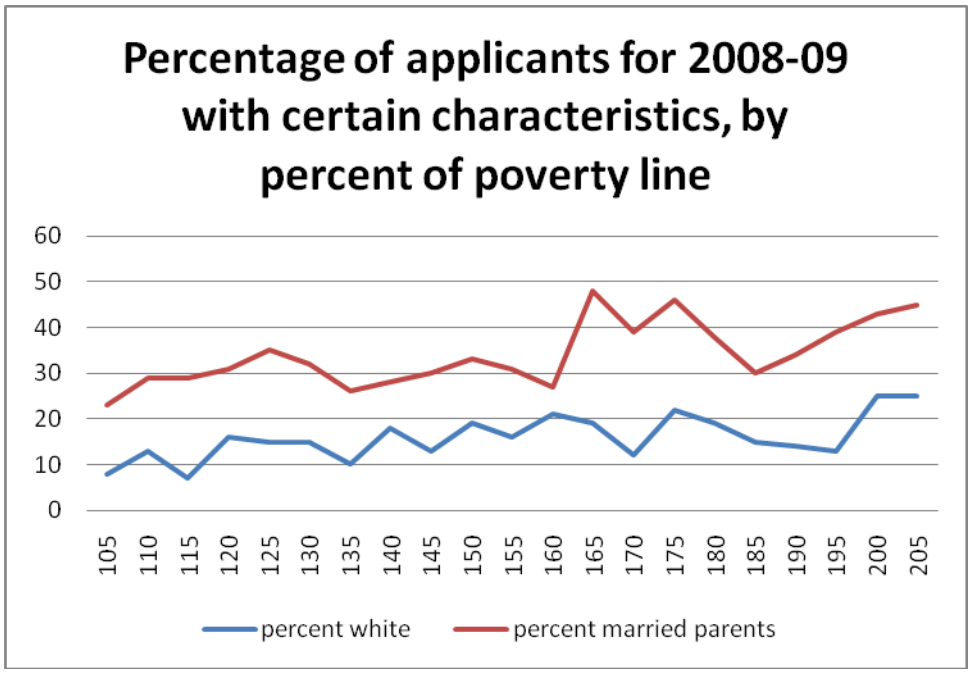
We next turn to the second potential threat to identification in the regression discontinuity model -- the potential for "bunching" of individuals just below the 185 percent of poverty threshold. If one were to observe a large number of applications with incomes just below this threshold, it could raise concerns that individuals are changing their income-earning behaviors in order to qualify for the scholarship. We would, of course, expect a considerable dropoff in applications immediately above the 185 percent of poverty threshold because those with incomes above that level are ineligible to participate in the program, but there would not ideally be a much larger number of applicants immediately below the income threshold versus farther away from the threshold. As can be seen in the graph below, there is no evidence of bunching of applications just below the 185 percent of poverty line. There is the expected sharp dropoff in applications at 185 percent of poverty, but one notes that the decline in applications among the income-ineligible is gradual. This provides some support for a

regression discontinuity model.



A regression discontinuity design could also be challenged if applicants are fundamentally different above versus below the critical value of 185 percent of the poverty line. To gauge the degree to which this is true, we plot two applicant attributes -- race and parental marital status -- against income levels on a graph. Specifically, we investigate whether the percentage of applicants who are white or the percentage of applicants with married parents appears to differ substantially around the critical threshold (also called the "discontinuity."). We limit the analysis to those above 100 percent of the poverty line and those below 210 percent of the poverty line so that we can hone in more clearly on the area around the discontinuity. As can be seen in the following graph, there is no apparent difference along either dimension around the discontinuity, implying that at least along these dimensions there is no fundamental difference between those with incomes just above the threshold and those with incomes just below the threshold. The applicant attributes are not smoothly distributed -- as

would be expected because of sample size -- but there is no evidence that the applicants are different in any substantial way.



Given that it has been established that applicant attributes appear to be similar around the discontinuity, and that there is no bunching of incomes around the discontinuity, it is now possible to measure whether student test score gains are different on either side of the discontinuity. The most basic regression discontinuity model involves estimating a linear regression in which the dependent variable is the student's test score gain and there are two key explanatory variables -- the student's family income as a percentage of poverty and an indicator for whether the student's family is income-eligible for the FTC Scholarship Program (i.e., the income is 185 percent of the poverty line or below.) Other models described below also include student-level control variables and more complicated specifications of the relationship between family income and test score gains. The regression discontinuity model does not distinguish between eligible students who used the scholarship and eligible students who did not use the scholarship;

rather, in order to identify causal effects, the eligibility criterion serves as an *instrumental variable* for the actual participation decision. But as mentioned above, participation conditional on eligibility (so long as an application was made) is over 40 percent, so this is a strong instrumental variable for participation.

In order to be in the regression discontinuity sample, one must observe public school test scores in 2007-08 as well as a test score gain in either the public sector or the private sector between 2008-09 and 2009-10. The sample size for the analysis is 2,229 students in mathematics and 2,222 students in reading.[7] Of the students who contributed gain scores to the analysis, 39.6 percent were enrolled in the program, indicating that gain scores are observed at a slightly higher rate for FTC Scholarship Program participants than for their counterparts who did not participate in the program. Of students with observed gain scores, 162 students (in both reading and math) are ineligible, based on their application data, to participate in the program. While this is a small sample, it is adequate to detect moderate differences in performance between program eligibles and program ineligibles.

Because the primary purpose of this report is not to provide a technical treatment of the causal estimated effects of program participation, we do not provide the technical details of the model estimation, but rather present the results of the regression discontinuity analysis. The table below presents only the key coefficient estimates, standard errors and statistical significance levels of the estimated effects of program participation on reading and mathematics scores.

[7] In the prior year's report, I restricted the sample to only those applicants with income less than 500 percent of poverty. This year, the highest percentage of poverty of an applicant was 442 percent, so this restriction is superfluous.

Model specification	Estimated effect on participation	Estimated effect on math gains	Estimated effect on reading gains
Linear model, no controls except for family income	0.412 (0.032) [p=0.000]	2.400 (1.580) [p=0.129]	2.527 (1.590) [p=0.112]
Linear model, controlling for 2007-08 reading and math NRT scores	0.411 (0.032) [p=0.000]	2.458 (1.588) [p=0.122]	2.510 (1.600) [p=0.117]
Linear model, also controlling for student race, gender, household size, and family marital status	0.411 (0.032) [p=0.000]	2.047 (1.595) [p=0.199]	2.227 (1.609) [p=0.167]
Quadratic model, also controlling for student race, gender, household size, and family marital status	0.324 (0.037) [p=0.000]	3.918 (1.863) [p=0.036]	2.634 (1.881) [p=0.161]
Cubic model, also controlling for student race, gender, household size, and family marital status	0.268 (0.039) [p=0.000]	4.437 (1.948) [p=0.023]	3.625 (1.966) [p=0.065]

In the table above, each cell represents the estimated effect of program participation on student test score gains between 2008-09 and 2009-10 in a regression discontinuity framework. Standard errors are in parentheses beneath coefficient estimates, and statistical significance levels are in square brackets. The first row presents estimated causal effects in a model with no control variables except for family income as a percentage of the poverty line, the variable used to determine program eligibility. As can be seen, being eligible to participate, according to our calculations, conditional on application for 2008-09 is strongly related to participation in 2008-09 -- eligible participants are 41 percentage points more likely to participate in the program than are those who appear to be ineligible. This provides strong first-stage evidence that the regression discontinuity model provides a valid instrument for program participation.

The other two columns suggest statistically insignificant positive effects of participation on mathematics and reading, though the significance levels of the point estimates are close to conventional levels. The estimated effects are of about two national percentile points suggest modest substantive difference between the outcomes of program participants and non-participants. The magnitudes of the results should be treated with caution given that program participants and program non-participants take different examinations.

The second and third rows of the table include additional control variables -- the second row includes 2007-08 percentile rankings of reading and math norm-referenced tests taken in public schools, and the third row also includes controls for student race/ethnicity, gender, household size and parental marital status, all reported on the scholarship application. One observes that only controlling for 2007-08 test scores does nothing to the estimated effects of participation on test score gains. Further controlling for a richer set of covariates also does very little to the magnitudes of the estimated effects of program participation.

It is important to gauge the sensitivity of regression discontinuity results to changes in how the researcher estimates the relationship between the underlying "forcing" variable (income as a percentage of poverty in this case) and the outcome variable. Therefore, the fourth and fifth rows of the table present the same model specification as the third row, but with the relationship between family income as a share of poverty and test score gains modeled either as a quadratic function or a cubic function. The results become larger in magnitude and statistical significance as one adopts a more flexible underlying relationship. The estimated effects of program participation on math

performance are statistically significantly positive at conventional levels in both models, and the estimated effects on reading performance are significantly positive in the case of reading.

The regression discontinuity model, while a substantial step forward from simple comparisons of test score gains between participants and non-participants, still could yield biased estimates. First, as mentioned above, while the concordance analysis used to provide comparable gains for public school students appears to have strong validity, the constructed NRT equivalents to FCAT scores are still just estimates, and the matches between actual and predicted NRT scores were somewhat better for reading than for mathematics. It is uncertain whether errors in the concordance analysis would bias these comparisons upward or downward. Second, if the FTC Scholarship Program is providing competitive pressure for public schools, the public school performance might be different than it would have been absent the FTC Scholarship Program. It is therefore important to interpret these results as the estimated effects of program participation for the types of students who apply to the program, and should not be seen as a more general effect of program participation.

In summary, the regression discontinuity model suggests that there may be positive effects on FTC Scholarship Program participants in terms of reading and mathematics test score gains. These differences, while not large in magnitude, are larger and more statistically significant than in the past year's results, suggesting that successive cohorts of participating students may be gaining ground over time. However, given the fact that the tests are different between public school students and FTC Scholarship Program participants, these results must still be interpreted with caution. Nonetheless,

these results, taken together with the weaker positive evidence from the previous cohort of new participants in the program, suggests that participation in the program likely has small positive effects for students on the margin of participation.

VI. Conclusion

This report presents empirical evidence on the compliance and performance of private schools that participate in the Florida Tax Credit Scholarship Program. The report analyzes data from 2009-10, and compares these data to prior years of test score collection and public school data from the Education Data Warehouse of the Florida Department of Education. There is strong evidence of high degrees of compliance with testing requirements for program participants.

Simple comparisons of the distribution of test score gains between FTC Scholarship Program participants and plausibly-eligible non-participants indicate that the test score gains in both populations are comparable in magnitude, though the raw gains are modestly smaller amongst scholarship participants than for non-participants. These are not causal estimates of differences, and the true effect of program participation may be more positive or more negative than the simple means comparisons. There is strong and compelling evidence that relatively low-performing students from low-income schools tend to be the students to participate in the FTC Scholarship Program, and causal analysis of these differences would need to take this differential selection into account.

With this in mind, this report makes use of regression discontinuity models to estimate the causal impact of program participation. These models rely on data for those who apply for the program, so they may not be representative of the population of

potentially eligible students (and there is evidence to suggest that applicants are indeed different from the overall population of free or reduced-price lunch recipients) and are best thought of as representative of the set of students who applied for the program. Nonetheless, the general pattern of small estimated effects of program participation on test score gains persists. In the regression discontinuity models, the estimated effects of program participation are modestly but consistently positive, and the results must be interpreted with considerable caution. That said, these results, coupled with those from the previous year, indicate that participation in the FTC Scholarship Program may benefit participating students relative to their default public schools.