Education Working Paper

No. 2 August 2003

When Schools Compete: The Effects of Vouchers on Florida Public School Achievement

Jay P. Greene, Ph.D.

Senior Fellow, Manhattan Institute for Policy Research

Marcus A. Winters

Research Associate, Manhattan Institute for Policy Research

EMBARGOED UNTIL 12:01 AM 8-20-03

C C i

CENTER FOR CIVIC INNOVATION

AT THE MANHATTAN INSTITUTE

EXECUTIVE SUMMARY

Florida's A+ Program may be the most controversial education reform program in the country, because it combines two extremely contentious education reforms: vouchers and high-stakes testing. Florida's high-stakes test, the Florida Comprehensive Assessment Test (FCAT), is used to grade schools on a scale from A to F. If a school receives two F grades in any four-year period, it is considered to be chronically failing and its students become eligible to receive vouchers they can use to attend other public or private schools.

The theory behind the A+ Program is that chronically failing public schools will have an incentive to improve if they must compete with other schools for students and the funding they generate. This study identifies five categories of low-performing schools based on the degree of threat each school faces from voucher competition: Voucher Eligible Schools (where students are already receiving vouchers), Voucher Threatened Schools (where one more F will make vouchers available), Formerly Threatened Schools (which used to be Voucher Threatened but no longer are), and two categories of similarly low-performing schools not facing any immediate threat of voucher competition. It then examines test score improvements on the FCAT and on the Stanford-9, a nationally respected standardized test, to see whether low-performing schools facing a greater degree of threat from voucher competition made better improvements than low-performing schools facing a lesser degree of threat from vouchers.

The results demonstrate the following:

- Florida's low-performing schools are improving in direct proportion to the challenge they face from voucher competition. These improvements are real, not the result of test gaming, demographic shifts, or the statistical phenomenon of "regression to the mean."
- Schools already facing competition from vouchers showed the greatest improvements of all five categories of low-performing schools, improving by 9.3 scale score points on the FCAT math test, 10.1 points on the FCAT reading test, and 5.1 percentile points on the Stanford-9 math test relative to Florida public schools that were not in any low-performing category.
- Schools threatened with the prospect of vouchers showed the second greatest improvements, making relative gains of 6.7 scale points on the FCAT math test, 8.2 points on the FCAT reading test, and 3.0 percentile points on the Stanford-9 math test.
- Low-performing schools that have never received any grade other than a D, or that have received at least one D since FCAT grading began, produced small and indistinguishable gains, respectively, relative to Florida public schools that were not low-performing. While these schools were similar to schools facing voucher competition, they failed to make similar gains in the absence of competitive incentives.
- Some researchers theorize that failing schools improve because of the stigma of a failing grade rather than the threat of voucher competition. The results of this study contradict this thesis. Schools that received one F in 1998-99 but none since are no longer exposed to the potential of voucher competition. These schools actually lost ground relative to non-low-performing Florida public schools, supporting the conclusion that once the threat of vouchers goes away, so does the incentive for failing schools to improve.

About the Authors

Jay P. Greene is a Senior Fellow at the Manhattan Institute for Policy Research where he conducts research and writes about education policy. He has conducted evaluations of school choice and accountability programs in Florida, Charlotte, Milwaukee, Cleveland, and San Antonio. He also recently published a report and a number of articles on the role of funding incentives in special education enrollment increases.

His research was cited four times in the Supreme Court's opinions in the landmark *Zelman* v. *Simmons-Harris* case on school vouchers. His articles have appeared in policy journals, such as *The Public Interest*, *City Journal*, and *Education Next*, in academic journals, such as *The Georgetown Public Policy Review*, *Education and Urban Society*, and *The British Journal of Political Science*, as well as in major newspapers, such as the *Wall Street Journal* and the *Washington Post*.

Greene has been a professor of government at the University of Texas at Austin and the University of Houston. He received his B.A. in history from Tufts University in 1988 and his Ph.D. from the Government Department at Harvard University in 1995. He lives with his wife and three children in Weston, Florida.

Marcus A. Winters is a Research Associate at the Manhattan Institute's Education Research Office where he studies and writes on education policy. He received his B.A. in Political Science with departmental honors from Ohio University in 2002.

Acknowledgements

The authors would like to thank the Florida Department of Education for providing the data necessary for our analysis.

About Education Working Papers

A working paper is a common way for academic researchers to make the results of their studies available to others as early as possible. This allows other academics and the public to benefit from having the research available without unnecessary delay. Working papers are often submitted to peer-reviewed academic journals for later publication.

Table of Contents

When Schools Compete: The Effects of Vouchers on Florida Public School Achievement

Introduction

Florida's A+ Program is perhaps the most aggressive and most controversial education reform measure in the country. The state offers vouchers redeemable at private schools to students in public schools that chronically fail the state's accountability test. The theory behind the A+ Program is that the prospect of losing students and the dollars they generate to vouchers will motivate low-performing schools to improve. But critics of the program argue that vouchers will hinder public schools by depriving them of financial resources and the best and brightest of their students.

Florida's A+ Program provides us with a unique opportunity to study the systemic effects vouchers have on public schools. Since Florida schools face vouchers only if they are failing, and they remove that threat only by improving academically, we can measure what effect the voucher threat has on their performance. In addition, since Florida schools already subject to vouchers must compete to attract and retain students, we can measure what effect voucher competition has on their performance.

The purpose of this study is to examine whether the existence or threat of competition causes public schools to improve. The results of this study suggest that schools are improving in response to competition—the amount Florida schools are improving is directly related to the degree of threat they face from vouchers. This study also finds that these improvements are caused by incentives arising from vouchers and not by other aspects of the A+ Program.

A Brief Description of the A+ Program

Florida's A+ Program is a hybrid of two of the most contentious education reform policies, vouchers and high-stakes testing. All Florida public school students enrolled in grades 3-10 take the state's accountability test, the Florida Comprehensive Assessment Test (FCAT). Test results have consequences for students and schools alike. Students must pass the reading portion of the FCAT in order to be promoted to the 4th grade and must pass the 10th grade test in order to graduate. The results of the test are also used to grade schools on a scale from A to F.[1] If a school receives an F twice in any four-year period, it is considered chronically failing and its students become eligible to receive vouchers they can use at other public schools or at private schools. Going into the 2002-03 administration of the FCAT, which is the focus of this study, 129 schools had received at least one F and ten schools have had their students become eligible for vouchers since school grading based on the FCAT began in the 1998-99 school year.[2]

Previous Research

While a great deal of research has shown that vouchers improve the performance of the students using them, there is little evidence on the effect school choice policies have on public schools. This lack of research is unfortunate because of the importance of the question at hand. If the availability of vouchers helps students who receive them but does not benefit, or even harms, the larger number of students who remain in public schools, the desirability of school choice would be greatly decreased.

In a previous analysis of Florida's A+ Program similar to this study, Greene found that F schools made extraordinary gains on the FCAT (see Greene, 2001). However, his findings were limited by the newness of the A+ Program. At the time of his study only two schools had received multiple failing grades and had vouchers offered to their students, so he was unable to separate schools that had already had vouchers offered to their students from schools only under threat of such a sanction. Also, since the

program was less than four years old, every school that had ever received an F from the state was still under the threat of vouchers, making it impossible to examine whether schools that receive an F continue to improve even after the threat of vouchers expires.

Harvard's Caroline Minter Hoxby has also performed some research on the systemic effects of school choice. In one study Hoxby looked at the effects of public school choice by comparing areas with several school districts to those with very few (see Hoxby, 1998). Areas with numerous school districts give parents more educational options because they can more easily move from one school district to another. Such residential school choice is so widely used that it is rarely even recognized at all as a type of school choice. Hoxby found that areas with greater residential school choice consistently have higher test scores at a lower per-pupil cost than areas with very few school districts.

Hoxby has also studied the effects of vouchers in Milwaukee and of charter schools in Arizona and Michigan on nearby public schools (see Hoxby, 2001). She found that public schools forced to compete with these schools of choice made greater test score gains than schools not faced with such competition.

Greene and Forster confirmed Hoxby's Milwaukee finding that competition from vouchers caused the public schools there to improve, and also found that charter schools were associated with improvements in nearby public schools (see Greene and Forster, 2001). Additionally, they looked at voucher competition in a small Texas school district serving a low-income, predominantly-Hispanic population. Since 1998 a private organization has offered vouchers to students in the Edgewood school district. Greene and Forster calculated expected test score gains for every Texas school district based upon their demographic characteristics and resources. They then compared this expected gain to the actual gains made by each district. They found that Edgewood's actual performance relative to its expected performance was better than that of 85% of Texas school districts.

Another study by Greene finds that states that provide their residents with more education freedom have better test scores (see Greene, 2001). He developed an index for the amount of education freedom available in each state. The index takes into account the availability of school choice as provided by charter schools, vouchers, the ease of home schooling, the ease of relocating to a different school district, and the ease of sending a child to another school district without moving. Greene found that states providing more school choice perform significantly better on the NAEP test.

Though the information provided by previous research on the systemic effects of school choice programs is positive, it remains limited. This study is an attempt to broaden our knowledge of this important aspect of school choice and of voucher programs in particular.

Some may question the results of this study because they do not believe the results from high-stakes standardized tests are accurate indicators of student performance. Many critics of such exams claim that they cause teachers to abandon teaching their students real skills in order to prepare them to "game" a particular test, or worse, that they will directly cheat on the exam. If this were the case, then a study finding gains on the FCAT would give us no clear information about the effectiveness of the A+ Program as a policy. Several recent studies have tackled the question of whether we can believe the results of high-stakes tests.

Greene, Winters, and Forster compared school-level results on high-stakes tests with school-level results on other standardized tests given around the same time with no stakes attached to their results, or "low-stakes tests" (see Greene, Winters, and Forster, 2002). Schools have no incentive to manipulate results on low-stakes tests, so if the scores on the high-stakes and low-stakes tests correlate, this would give us confidence in the validity of high-stakes test results. They looked at nine school systems nationwide, including two states, and found high correlations between the two types of tests in their score levels and mixed results when examining the changes in their scores over time.

Of particular interest to this current study, Greene, Winters, and Forster found impressively strong correlations between high-stakes and low-stakes tests in Florida. School-level scores on the FCAT and Florida's low-stakes test, the Stanford-9, correlated at an average of 0.96 (if the results were identical they would correlate at 1.00). They found that the

gain in scores in Florida also correlated at a high level (0.71). These findings in Florida provide strong confidence that whatever gains are made on the FCAT are the result of gains in real learning, not a school's ability to "beat" a particular test.

Arizona State researchers Audrey Amrein and David Berliner also analyzed whether we can trust the results of high-stakes tests (see Amrein and Berliner, 2002). They compared results on high-stakes tests with results on the NAEP, as well as on AP and college entrance exams. They argued that high-stakes tests are not only unreliable indicators of performance, but also harm school performance on other tests and cause more students to drop out of high school. The results of Amrein and Berliner's study caused a large commotion, especially after they were trumpeted on the front page of the *New York Times*.

Further analysis of their data, however, shows their results to be misleading at best. Raymond and Hanushek take Amrein and Berliner to task for several devastating errors in their study, including violating some of the most basic tenets of social science research and statistical analysis (see Raymond and Hanushek, 2003). Among the most damaging criticisms, they show that Amrein and Berliner improperly (and inconsistently) excluded data from their analysis, incorrectly measured when states implemented a policy of high-stakes testing, and failed to do any significance testing of their findings. Raymond and Hanushek show that correctly applying Amrein and Berliner's method to their data produces the opposite results on the outcomes of high-stakes testing from what Amrein and Berliner actually reported.

Method

The purpose of this study is to find whether school choice incentives cause failing public schools to improve. The A+ Program's combination of high-stakes testing and school choice provides us with a unique opportunity to study whether and to what extent this incentive to improve exists. Since the threat of vouchers is primarily a function of each school's performance on the FCAT, we can be confident that if this program provides the school with an incentive to improve that incentive should manifest itself in gains on the FCAT. We also consider alternative explanations for improvements in test scores to identify whether vouchers or other factors were responsible.

To perform the analysis, we collected school-level test scores on the 2001-02 and 2002-03 administrations of the FCAT and the Stanford-9, a nationally respected norm-referenced test administered to all Florida public school students around the same time as the FCAT. We also collected the most recent demographic information available for every school in Florida, as well as their school grades for every year since the FCAT was first given in the 1998-99 school year.[3] All of this information is available from the Florida Department of Education and was obtained either through its website or through data requests.

We identified schools with different degrees of incentive to improve because of vouchers, in order to see whether schools with more incentive to improve made greater gains on the FCAT than schools not facing the same incentives. We used school grades given out between 1998-99 and 2001-02 to identify five specific categories of schools that we wanted to compare to the rest of Florida's public schools. Comparisons of the changes in FCAT scores between the 2001-02 and 2002-03 school years among these different categories of schools allow us to test various hypotheses about the causes of test score improvements. The demographic characteristics of these schools are reported in Table 1 (see Appendix for all Tables).

The Five Categories

Voucher Eligible Schools

These schools have received at least two Fs since FCAT grades were first given in 1998-99 and have been deemed chronically failing by the state. Students at these schools have already been offered vouchers to attend private schools. Thus, Voucher Eligible Schools are currently competing against private schools in the market for students. They are the group with the greatest incentive to improve and also the greatest likelihood of being harmed by vouchers if vouchers are in fact harmful.

There are nine Voucher Eligible Schools in our analysis. The average scores for these schools on the 2001-02 administration of the FCAT, the administration on which their last grade was based, were 240.3 in reading and 252.4 in math on a scale of 100-500. These schools serve populations that are largely poor and minority—88% of their students are

enrolled in the free or reduced price lunch program, 18% are deemed limited English proficient, and only 1% of their students are white.

Hypothesis: Because Voucher Eligible Schools must attract or retain students we hypothesize that they face the greatest incentives to improve and should outperform all other categories of schools.

Voucher Threatened Schools

Voucher Threatened Schools have received exactly one F in the three school years prior to the administration of the 2002-03 FCAT. If these schools received another F in 2002-03, that would have been their second F in four years and their students would have been offered vouchers. These schools were not yet forced to compete against private schools in the market, but they faced the prospect of future competition if they did not make sufficient improvement. They therefore had an incentive to improve so that the prospect of competition did not become a reality.

There are 50 Voucher Threatened Schools in our analysis. These schools are similar to the Voucher Eligible Schools in many ways. Their average 2001-02 FCAT scores were 252.4 in reading and 258.2 in math. Students in these schools are 9% white, 8% limited English proficient, and 69% enrolled in the subsidized lunch program.

Hypothesis: Because Voucher Threatened Schools face the prospect of future competition they have incentives to improve, but those incentives should be less powerful than those facing schools that are already voucher eligible and are already competing for students. We therefore hypothesize that Voucher Threatened Schools should make less test score improvement than Voucher Eligible Schools, but greater improvements than all other categories of schools.

Always D Schools

Always D Schools have never received any grade other than D. Since Always D Schools have consistently performed poorly on the FCAT, they are in greater danger than other public schools of receiving their first F in the next administration. Thus, these schools are not Voucher Threatened but they face the prospect of becoming so. Though they do not have the same immediate incentive that Voucher Threatened Schools have to improve, they still have strong reasons to worry about their performance on the FCAT.

There are 63 Always D Schools in our analysis. On the 2001-02 administration of FCAT they averaged scores of 254.1 in reading and 261.2 in math. In these schools an average of 5% of the students are white, 11% are limited English proficient, and 77% are enrolled in the free or reduced lunch program.

Their relatively low initial test scores and disadvantaged student populations make Always D Schools an attractive group to compare to F schools in our analysis. While we can and do control for observable characteristics of schools in our analyses, comparing the Voucher Eligible and Voucher Threatened Schools against this group of schools helps us rule out unobserved factors that may be responsible for school improvement. As can be seen from the demographic profiles of the school categories in Table 1, Always D Schools are very similar to Voucher Eligible and Voucher Threatened Schools in their observable characteristics. It is reasonable to expect that they are similar in unobserved ways as well.

Hypothesis: Always D Schools, which face no voucher competition, should make academic improvements less than those of the Voucher Eligible and Voucher Threatened Schools. But because Always D Schools are in real danger of receiving their first F grade, they face stronger incentives to improve than most other schools and therefore should make above average test score gains.

Ever D Schools

These schools have received at least one D since grades have been given but have never received an F. Ever D Schools includes all the schools in the Always D category. These schools are not currently forced to compete for students since their students do not have vouchers, nor do they face the imminent prospect of having to compete for students. School grades are a function of the percentage of students in a school meeting a certain test-score threshold, the year-to-year test-score gains students in the school are making, and some other non-test-score factors. Because school grades are not simply a function of a

school's level performance on FCAT, many Ever D Schools have similar or even lower test scores than F schools, but have still managed to avoid receiving a failing grade.

There are 570 Ever D Schools in our analysis. In 2001-02 their average FCAT reading score was 273.6 and their average FCAT math score was 284.1. In Ever D Schools 12% of the students are limited English proficient, 72% are enrolled in the free or reduced price lunch program, and 29% of the students are white.

Hypothesis: Ever D Schools, which face no voucher competition, should make academic improvements less than those of the Voucher Eligible and Voucher Threatened Schools. And because Ever D Schools are less likely to slip into an F grade than Always D Schools, we should also expect Ever D Schools to make less improvement than the Always D group.

Formerly Threatened Schools

Formerly Threatened Schools received an F in the first year of FCAT grading, 1998-99, but have not received another F since. These schools once faced the prospect of vouchers but no longer do because they have survived the four-year time period without receiving another F. An F on the 2002-03 administration of the FCAT would not have been their second F in four years; it would have been their first F in a new four-year period. Analyzing this group allows us to see whether schools continue to improve relative to the rest of the public schools in Florida once the threat of vouchers disappears.

Examining Formerly Threatened Schools is also particularly important because those schools are like the Voucher Eligible and Voucher Threatened Schools in that they have received an F grade. Some researchers have suggested that the improvements made by schools facing vouchers are not actually the result of vouchers, but of the stigma those schools experience by receiving a failing grade (see Ladd, 2001). If the stigma of having received a failing grade were sufficient to prod schools to improve, then Formerly Threatened Schools should be improving like the other schools that have received F grades. If, on the other hand, the stigma of a failing grade is insufficient motivation, Formerly Threatened Schools should under-perform other schools that not only have the stigma of a failing grade but also must face the prospect or actuality of voucher competition. While it is true that this comparison is colored by the fact that Formerly Threatened Schools received their F a longer time ago, it is also the case that some of the Voucher Eligible and Voucher Threatened Schools are a few years removed from their most recent failing grade.

There are 59 Formerly Threatened Schools in or analysis. These schools had average 2001-02 FCAT scores of 264.7 in reading and 279.4 in math. In Formerly Threatened Schools 15% of the students are white, 13% of the students are limited English Proficient, and 83% of the students are enrolled in the free or reduced price lunch program.

Hypothesis: Formerly Threatened Schools no longer face voucher competition if they receive an F on the next administration of the FCAT, therefore they do not have any special incentives to improve their academic performance. We expect that their test score improvement should be comparable to that of most other schools.

School-Level Score Calculation

We compared score gains for each of these groups relative to the rest of Florida public schools between the 2001-02 and 2002-03 administrations of the FCAT and the Stanford-9. We did this by following a cohort of students and calculating school gains on these tests for each grade in grades 3-10 in both math and reading. For example, we subtracted a school's third grade reading score on the 2001-02 FCAT from its fourth grade reading score on the 2002-03 FCAT. Following a cohort allows us to measure the performance of roughly the same students on the test over time. We then averaged the score gains for each cohort in the school on each test and subject. This gave us a single cohort change for each school in Florida.

We then performed a regression analysis comparing the performance of our five subgroups to the performance of the rest of Florida's public schools. In our analysis we controlled for demographic characteristics of the schools, including the percentage of students participating in subsidized lunch programs, the percentage of students who were white, the percentage of students who were limited English proficient, and the operating cost the school spends per student.

Results

The results of our analyses were remarkably consistent with our hypotheses. The results show that voucher competition in Florida is leading to significant academic improvements in public schools. Public schools currently facing voucher competition or the prospect of competition made exceptional gains on both the FCAT and the Stanford-9 tests compared to all other Florida public schools and the other subgroups in our analysis.

Our results on the FCAT are reported as the cohort change in mean scale score on a scale of 100-500 and our results on the Stanford-9 are reported as the cohort change in national percentile rank.

The results of the FCAT math test are listed in Table 2. Voucher Eligible Schools improved by 9.3 scale score points more than the gains made by the rest of Florida's public schools between the 2001-02 and 2002-03 administrations. Voucher Threatened Schools made the next highest relative gain of 6.7 scale score points on the FCAT math test. Each of these results is statistically significant at a very high level ($p < 0.01$), meaning that we can have high confidence that the test score gains made by schools facing the actuality or prospect of voucher competition were larger than the gains made by other public schools. As we hypothesized, actual voucher competition produced the largest test score improvements while the prospect of voucher competition produced somewhat smaller gains.

The results for the Always D and Ever D Schools were also consistent with our hypotheses. Always D Schools, which faced some incentive to improve given the real danger of receiving their first F, made some improvements in excess of those made by other Florida public schools—a relative gain of 2.2 scale score points on the FCAT math test. But the Always D Schools gain was statistically significant only according to a relaxed standard ($p < 0.1$), meaning that we do not have high confidence that the improvements made by Always D Schools were actually different from those made by other schools. In addition, the magnitude of the gain by Always D Schools was substantially smaller than the gains made by Voucher Eligible or Voucher Threatened Schools. The Ever D Schools experienced year-to-year changes in FCAT math scores that were indistinguishable from the year-to-year changes experienced by other public schools in Florida.

The small or non-existent gains achieved by Always D and Ever D Schools compared to Voucher Eligible and Voucher Threatened Schools, despite the similar characteristics of all of these schools, strengthens our confidence that voucher competition is the cause of the improvements. Always D Schools in particular are very similar to Voucher Eligible and Voucher Threatened Schools in terms of their initial test scores, student populations, and resources, as well as other unobserved factors that are not controlled for in our model. Yet the schools that faced voucher competition made much larger test score improvements.

The lack of gains among Formerly Threatened Schools also increases our confidence that voucher competition is the cause of test score improvement. Formerly Threatened Schools are like Voucher Eligible and Voucher Threatened Schools in that they have received at least one failing grade. If the stigma of receiving a failing grade were sufficient motivation for schools to make academic progress we would expect Formerly Threatened Schools to make gains comparable to those realized by Voucher Eligible and Voucher Threatened Schools. Instead of making relative gains, Formerly Threatened Schools actually made a relative loss of 2.2 points on the FCAT math test, though the result is barely statistically insignificant using a relaxed standard ($p = 0.103$). As we hypothesized, schools that had a failing stigma but not facing voucher competition did not make gains like those achieved by schools that had the failing stigma and were also facing voucher competition.

On the Stanford-9 math test the story is much the same as it was on the FCAT. The results on this test are reported in Table 3. Schools currently experiencing voucher competition, Voucher Eligible Schools, achieved gains that were 5.1 percentile points greater than the year-to-year gains achieved by other Florida public schools. Schools that faced the prospect of competition if they failed again improved their Stanford-9 math scores by 3.0 percentile points. Both of these gains are statistically significant at a very high level ($p < 0.01$).

The gains made by schools facing the threat or reality of vouchers are again in stark contrast to the results for our comparison D school groups. Neither Always

D Schools nor Ever D Schools did significantly better than the rest of Florida's public schools. We also find a relative loss for schools that had the stigma of a failing grade but no longer faced voucher competition. Formerly Threatened Schools made a loss of 2.1 percentile points on the Stanford-9 math test. This result is statistically significant at a very high level ($p < 0.01$).

The similarity of our findings on the Stanford-9 and FCAT math tests suggests that the gains being made by schools facing voucher competition are the results of real learning and not simply manipulations of the state's high-stakes testing system. Schools have no incentives to "teach to" or otherwise manipulate the Stanford-9 results. Whatever educational improvements Voucher Eligible and Voucher Threatened Schools are making yielded improved results on a low-stakes test as well as on the state's high-stakes test. If schools facing voucher competition were only appearing to improve by somehow manipulating the Florida's high-stakes testing system, we would not have seen a corresponding improvement on another test that no one had incentives to manipulate.

Our results in reading continue on the same trend as our math results. Table 4 contains the results of the FCAT reading test. Voucher Eligible Schools realized an improvement of 10.1 points on the FCAT reading test beyond the gains made by the rest of Florida's public schools. This gain was closely followed by the 8.2 relative point gain made by the Voucher Threatened Schools. These gains are again statistically significant at a very high level ($p < 0.01$).

Always D Schools made a statistically significant relative gain of 2.5, while again the Ever D Schools failed to achieve any improvements over the gains made by the rest of Florida's public schools. We also find a statistically significant relative loss for Formerly Threatened Schools.

The pattern of results for the FCAT reading tests is consistent with what we hypothesized except for the significant loss among Formerly Threatened Schools. Clearly, removing the threat of voucher competition permits backsliding at these schools.

Finally, our results on the Stanford-9 reading test, reported in Table 5, further confirm that voucher incentives are improving Florida's public schools. Again, the greatest relative gain, 2.3 percentile points, is made by Voucher Eligible Schools, though the gain is slightly statistically insignificant ($p = 0.105$). These gains are followed by a statistically significant relative gain by the Voucher Threatened Schools' of 1.6 percentile points. The very small number of Voucher Eligible Schools (only 9 in our analysis) helps explain why their larger gain fell short of statistical significance.

Neither the Always D schools nor the Ever D Schools make any significant relative gains on the Stanford-9 reading test. Formerly Threatened Schools slip by 1.7 percentile points, a relative decline that is statistically significant. Again, the schools facing either the prospect or reality of vouchers make extraordinary gains compared to the gains made by the rest of Florida's public schools and those made by schools with similar test scores serving similar populations.

The results of our analysis are very supportive of the use of vouchers as a means to improve the performance of public schools. The more in danger a school is of having to compete with vouchers, the greater score gains they make on both the FCAT and Stanford-9. As we hypothesized, schools already facing vouchers make the greatest gains, followed by schools faced with the threat of vouchers, followed by schools in danger of encountering the voucher threat.

Possible Explanations of the Results Other than a Voucher Effect

While there is no question that failing schools in Florida that were subject to actual or prospective voucher competition made exceptional gains, there are likely to be some questions about whether these exceptional gains were actually caused by voucher competition. We considered a number of possible alternative explanations.

Failing Stigma

Others have argued (see Ladd, 2001; Carnoy, 2001; and Harris, 2001) that the exceptional gains made by failing schools in Florida are caused by the stigma of failing, not the competitive incentives created by vouchers. But by examining the performance of schools that have the stigma of failing without being subject to any voucher threat, Formerly Threatened Schools, we help eliminate this alternative explana-

tion. Schools that only received an F in 1998-99 would still suffer under the stigma of their F grade, but once the threat of vouchers is removed they not only fail to continue improving, they actually lose ground.

It is possible that the stigma of the F grade fades over time so that schools that received an F in 1998-99 no longer felt the stigma of the grade in 2002-03. But the Voucher Eligible and Voucher Threatened Schools categories include some schools that have not received an F grade for several years, and yet those categories made gains. It is implausible that the stigma effect only exists for three years and then suddenly disappears. The more believable explanation is that the actuality or prospect of voucher competition provides incentives for schools to improve and this effect suddenly disappears when the four year voucher threat period expires.

Regression to the Mean

Another alternative explanation that has been advanced for the exceptional improvements made by schools facing voucher competition is that their extremely low initial scores are affected by a statistical tendency called "regression to the mean" (see Camilli and Bulkley, 2001; and Kupermintz, 2001). Very high and very low-scoring schools may report future scores that return to being closer to the average for the whole population. This tendency is created by non-random error in the test scores, which can be especially problematic when scores are "bumping" against the top or bottom of the scale for measuring results. If a school has a score of 2 on a scale from 0 to 100, it is hard for students to do worse by chance but easier for them to do better by chance. Low-scoring schools that are near the bottom of the scale are likely to improve, even if only by statistical fluke.

In order to test whether regression to the mean is driving our results, we compared gains made by F schools to schools with similar test scores that had never received an F. As mentioned in the method section, there are schools that have received scores similar to or below those of F schools but have never received an F themselves because grades are not entirely a function of test score levels.

We defined Low Performing Non-F Schools as schools that had never received an F and whose test score levels were lower than a benchmark set at one standard deviation above the average score of F schools. Low Performing Non-F Schools had average test score levels very similar to those of F schools. Table 6 compares the test scores and demographic characteristics of these schools to the F school categories in our analysis.

We performed a regression analysis comparing the change in test scores for our three categories of F schools to the change in test scores for Low Performing Non-F Schools on the FCAT and Stanford-9. If regression to the mean is driving our results, then there should be no difference in the change in test scores between F schools, such as Voucher Eligible and Voucher Threatened Schools, and Low Performing Non-F Schools, which have similarly low scores but do not face the threat of vouchers.

The results of our analyses are reported in Tables 7-10. Voucher Eligible and Voucher Threatened schools made statistically significant gains and Formerly Threatened Schools made statistically significant losses relative to the non-F schools with similar test scores in each test and subject except the Stanford-9 reading test, where our result for Voucher Eligible Schools was not statistically significant. This means that the gains we found for Voucher Eligible and Voucher Threatened Schools in this study have not been caused by regression to the mean. Low Performing Non-F Schools had initial test scores that were similarly subject to regression to the mean, yet the Voucher Eligible and Voucher Threatened Schools still made significantly greater year-to-year gains.

Demographic Changes

Others may claim the results of our analysis are driven by changes in the demographic composition of the Voucher Eligible and Voucher Threatened Schools. They may argue that the worst students are leaving Voucher Eligible Schools for private schools when they are offered vouchers, which would improve the school's average score. This means that instead of vouchers "creaming" the best students, as many voucher critics worry will happen, vouchers would instead be "dredging" the worst students from the public schools. Though many voucher advocates may welcome such a result because it would mean that vouchers are specifically serving the students most in need, we have no reason to believe this "dredging" is actually taking place.

Change in demographics cannot explain the gains made by Voucher Threatened Schools. These schools make strong statistically significant gains compared to the rest of Florida's public schools even though they have lost no students to vouchers. However, Voucher Eligible Schools have lost students to vouchers, so for these schools the possibility is worth investigating.

We performed a test to determine whether the demographic changes in Voucher Eligible Schools are driving our results. In our regular analysis we control for demographic characteristics by the level in a single year; for this test we ran a regression analysis controlling for the change in demographics from the 2001-02 administration of FCAT to the 2002-03 administration. In this study we were only able to perform this test controlling for the change in the percentage of students in the free or reduced lunch program and the percentage of students who were white, because the 2002-03 data in the other demographic categories were not available from the Florida Department of Education at the time of this analysis. Though we were unable to control for the change in spending and in the percentage of students deemed limited English proficient, we continued to control for the level of spending and the level of these demographics. Greene, however, did examine the influence of additional resources on test score improvement in his previous analysis of the A+ Program and found that schools threatened by vouchers made large gains even after he controlled for their change in spending (see Greene, 2001).

The results of our analyses are presented in Tables 11-14. Controlling for changes in demographics slightly raises the p values on both reading tests, FCAT and Stanford-9. Our results in math, however, remain quite robust. This leads us to conclude that changes in demographics are not a major factor in our results.

Conclusion

Having ruled out these other possible explanations, we are left with the conclusion that the gains low-performing schools are making on Florida's statewide assessments are the result of the competitive pressure of school vouchers. Since previous research shows convincingly that we can believe the results of these tests, we can have confidence that the gains these failing schools are making are the result of real improvement in the education they are providing their students. Thus, Florida's A+ Program is achieving its goal of providing a better education to the students its school system has failed in the past.

The question of what effect vouchers have on public schools is an increasingly important one as more states and localities consider adopting school choice. While by no means definitive, this study provides us with valuable evidence that public schools improve when they are forced to compete in the market with private schools or are simply threatened with this competition.

Florida's A+ Program shows that public schools improve when they are given an incentive to do so. We will continue to watch the progress of this important school choice program to see whether it continues to improve the education provided to Florida's previously underserved students.

ENDNOTES

1. The precise formula that has been used to grade schools has changed over the course of the program and now is a function of the percentage of students exceeding certain performance goals as well as year-to-year improvement in student performance, as measured by the FCAT.

2. This study's analyses include only schools for which complete information is available. Thus, the number of schools in each category in our analyses will be somewhat lower than the actual number of schools meeting the criteria for inclusion in that category.

3. We were limited by the availability of demographic data. Both the percent of students who were white and the percent of students in the free or reduced price lunch program were from the 2002-03 school year. However, information on the percent of students who were limited English proficient and school operating cost per pupil were only available up to the 2001-02 school year.

REFERENCES

Amrein, Audrey L, and Berliner, David C., "High-Stakes Testing, Uncertainty, and Student Learning," *Education Policy Analysis Archives*, Volume 10 Number 18.

Camilli, Gregory and Bulkley, Katrina, "Critique of 'An Evaluation of the Florida A-Plus Accountability and School Choice Program". *Education Policy Analysis Archives*, Volume 9, Number 7, 2001.

Carnoy, Martin "School Vouchers: Examining the Evidence" Economic Policy Institute, 2001.

Greene, Jay P, "An Evaluation of the Florida A-Plus Accountability and School Choice Program" Florida State University, The Manhattan Institute, and the Harvard Program on Education Policy and Governance, 2001.

Greene, Jay P, "2001 Education Freedom Index", The Manhattan Institute, 2001.

Greene, Jay P and Forster, Greg, "Rising to the Challenge: The Effect of School Choice on Public Schools in Milwaukee and San Antonio" The Manhattan Institute, 2002.

Greene, Jay P., Winters, Marcus A., and Forster, Greg, "Testing High Stakes Tests: Can We Believe the Results of Accountability Tests?" The Manhattan Institute, 2003.

Harris, Doug. "What Caused the Effects of the Florida A+ Program: Ratings or Vouchers?" in Martin Carnoy, "School Vouchers: Examining the Evidence" Economic Policy Institute, 2001.

Hoxby, Caroline, "Rising Tide" *Education Next*, Winter, 2001.

Hoxby, Caroline "Analyzing School Choice Reforms that Use America's Traditional Forms of Parental Choice" in Paul E. Peterson and Bryan C. Hassel eds., *Learning from School Choice*, Brookings Institution, 1998.

Kupermintz, Haggai, "The Effects of Vouchers on School Improvement: Another Look at the Florida Data". *Education Policy Analysis Archives*, Volume 9, Number 8, 2001.

Ladd, Helen F, "Debating Florida's Voucher Effect" *Education Week*, March 14, 2001.

Raymond, Margaret E., and Hanushek, Eric A. "High Stakes Research" *Education Next*, Summer, 2003.

APPENDIX

Table 1: Demographic Characteristics of Schools

	Average FCAT Math Score, 2001-02	Average FCAT Reading Score, 2001-02	Percent White	Percent in Free or Reduced Lunch Program	Percent Limited English Proficient	Number of Schools
Voucher Eligible Schools	252.4	240.3	1%	88%	18%	9
Voucher Threatened Schools	258.2	252.4	9%	69%	8%	50
Always D Schools	261.2	254.1	5%	77%	11%	63
Ever D Schools	284.1	273.6	29%	72%	12%	570
Formerly Threatened Schools	279.4	264.7	15%	83%	13%	59
All Other Florida Public Schools	306.3	297.6	61%	42%	6%	1825

Includes only schools for which complete information is available

Table 2: FCAT Math Test

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	9.3***
Voucher Threatened Schools	6.7***
Always D Schools	2.2*
Ever D Schools	-0.3
Formerly Threatened Schools	-2.2
Reported in Mean Scale Scores	* = statistically significant at p<0.1
Number of Schools: 2504	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 3: Stanford-9 Math Test

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	5.1***
Voucher Threatened Schools	3.0***
Always D Schools	0.8
Ever D Schools	0.2
Formerly Threatened Schools	-2.1***
Reported in Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 2493	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 4: FCAT Reading Test

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	10.1***
Voucher Threatened Schools	8.2***
Always D Schools	2.5**
Ever D Schools	0.4
Formerly Threatened Schools	-2.5**
Reported in Mean Scale Scores	* = statistically significant at p<0.1
Number of Schools: 2503	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 5: Stanford-9 Reading Test

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	2.3
Voucher Threatened Schools	1.6**
Always D Schools	0.6
Ever D Schools	-0.3
Formerly Threatened Schools	-1.7***
Reported in Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 2495	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 6: Demographic Characteristics of F and Low Performing Non-F Schools

	Average FCAT Math Score, 2001-02	Average FCAT Reading Score, 2001-02	Percent White	Percent in Free or Reduced Lunch Program	Percent Limited English Proficient	Number of Schools
Voucher Eligible Schools	252.4	240.3	1%	88%	18%	9
Voucher Threatened Schools	258.2	252.4	9%	69%	8%	50
Formerly Threatened Schools	279.4	264.7	15%	83%	13%	59
Low Performing Non-F Schools (Reading)	265.0	256.6	24%	74%	12%	466
Low Performing Non-F Schools (Math)	266.5	261.3	29%	70%	9%	552

Includes only schools for which complete information is available

Table 7: FCAT Math Test for Regression to the Mean

	Improvements Relative to Low Performing Non-F Public Schools
Voucher Eligible Schools	7.0*
Voucher Threatened Schools	4.9***
Formerly Threatened Schools	-4.7***
Reported in Mean Scale Scores	* = statistically significant at p<0.1
Number of Schools: 664	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 8: Stanford-9 Math Test for Regression to the Mean

	Improvements Relative to Low Performing Non-F Public Schools
Voucher Eligible Schools	4.2**
Voucher Threatened Schools	2.3***
Formerly Threatened Schools	-3.1***
Reported as Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 657	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 9: FCAT Reading Test for Regression to the Mean

	Improvements Relative to Low Performing Non-F Public Schools
Voucher Eligible Schools	8.3**
Voucher Threatened Schools	5.6***
Formerly Threatened Schools	-4.9***
Reported as Mean Scale Scores	* = statistically significant at p<0.1
Number of Schools: 579	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 10: Stanford-9 Reading Test for Regression to the Mean

	Improvements Relative to Low Performing Non-F Public Schools
Voucher Eligible Schools	2.0
Voucher Threatened Schools	1.7**
Formerly Threatened Schools	-2.0***
Reported as Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 574	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 11: FCAT Math Test Controlling for Change in Demographics

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	15.1***
Voucher Threatened Schools	9.3***
Always D Schools	2.0
Ever D Schools	2.4***
Formerly Threatened Schools	0.5
Reported as Mean Scale Scores	* = statistically significant at p<0.1
Number of Schools: 2505	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 12: Stanford-9 Math Test Controlling for Change in Demographics

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	5.9***
Voucher Threatened Schools	3.5***
Always D Schools	0.7
Ever D Schools	0.5**
Formerly Threatened Schools	-1.8***
Reported as Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 2494	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 13: FCAT Reading Test Controlling for Change in Demographics

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	5.2*
Voucher Threatened Schools	6.1***
Always D Schools	2.3**
Ever D Schools	-1.1**
Formerly Threatened Schools	-3.8***
Reported as Mean Scale Score	* = statistically significant at p<0.1
Number of Schools: 2504	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

Table 14: Stanford-9 Reading Test Controlling for Change in Demographics

	Improvements Relative to Other Florida Public Schools
Voucher Eligible Schools	2.2
Voucher Threatened Schools	1.7***
Always D Schools	0.7
Ever D Schools	-0.3
Formerly Threatened Schools	-1.6***
Reported as Percentile Scores	* = statistically significant at p<0.1
Number of Schools: 2496	** = statistically significant at p<0.05
	*** = statistically significant at p<0.01

The Center for Civic Innovation's (CCI) purpose is to improve the quality of life in cities by shaping public policy and enriching public discourse on urban issues.

CCI sponsors the publication of books like The Entrepreneurial City: A How-To Handbook for Urban Innovators, which contains brief essays from America's leading mayors explaining how they improved their cities' quality of life; Stephen Goldsmith's The Twenty-First Century City, which provides a blueprint for getting America's cities back in shape; and George Kelling's and Catherine Coles' Fixing Broken Windows, which explores the theory widely credited with reducing the rate of crime in New York and other cities. CCI also hosts conferences, publishes studies, and holds luncheon forums where prominent local and national leaders are given opportunities to present their views on critical urban issues. *Cities on a Hill*, CCI's newsletter, highlights the ongoing work of innovative mayors across the country.

The Manhattan Institute is a 501(C)(3) nonprofit organization. Contributions are tax-deductible to the fullest extent of the law. EIN #13-2912529

MANHATTAN INSTITUTE FOR POLICY RESEARCH

52 Vanderbilt Avenue • New York, NY 10017
www.manhattan-institute.org