An Evaluation of the Florida A-Plus Accountability and School Choice Program

February 2001

Jay P. Greene, Ph. D.

Senior Fellow,
The Manhattan Institute for Policy Research

Research Associate,
Program on Education Policy and Governance
Harvard University

Program on Education Policy and Governance, Harvard University

TABLE OF CONTENTS

About the Author

Jay P. Greene is a senior fellow at the Manhattan Institute for Policy Research and a research associate at Harvard University's Program on Education Policy and Governance (PEPG). He has conducted evaluations of school choice programs in Milwaukee, Cleveland, Charlotte, and San Antonio. He has also investigated the effects of school choice on civic values and integration. His publications include the chapters, "Civic Values in Public and Private Schools," and "School Choice in Milwaukee: A Randomized Experiment," in the book, *Learning from School Choice*, published by the Brookings Institution in 1998; "The Effect of Private Education on Political Participation, Social Capital, and Tolerance," in the Fall 1999 issue of *The Georgetown Public Policy Review*; and "The Texas School Miracle Is for Real," in the Summer 2000 issue of *City Journal*. He has been a professor of government at the University of Texas at Austin and the University of Houston. He received his Ph.D. from the Government Department at Harvard University in 1995. Dr. Greene lives with his family in Weston, Florida.

Author's Acknowledgements

This report was prepared under contract with Florida State University as part of a grant from the Florida Department of Education to evaluate the A-Plus Program. Additional support was provided by Harvard University's Program on Education Policy and Governance (PEPG). Professors Richard Feiock and Tom Dye of Florida State University and Professor Paul Peterson, Director, PEPG, Harvard University, served as principal investigators on this project. Rob Fusco and Tom Dye provided valuable research assistance.

EXECUTIVE SUMMARY

By offering vouchers to students at failing schools, the Florida A-Plus choice and accountability system was intended to motivate those schools to improve their academic performance. Under this plan, each public school in Florida is assigned a grade, A through F, based on the proportion of its students passing the Florida Comprehensive Assessment Test (FCAT). Students attending schools that receive two "F" grades in four years are eligible to receive vouchers that enable them to attend private schools or to transfer to another public school.

This report examines whether schools that faced the prospect of having vouchers offered to their students experienced larger improvements in their FCAT scores than other schools.

The results show that schools receiving a failing grade from the state in 1999 and whose students would have been offered tuition vouchers if they failed a second time achieved test score gains more than twice as large as those achieved by other schools. While schools with lower previous FCAT scores across all state-assigned grades improved their test scores, schools with failing grades that faced the prospect of vouchers exhibited especially large gains.

The report also establishes that the FCAT math and reading results are highly correlated with the results from a nationally recognized standardized test, the Stanford 9, which suggests that the FCAT is a reliable measure of student performance.

This report shows that the performance of students on academic tests improves when public schools are faced with the prospect that their students will receive vouchers. These results are particularly relevant because of the similarities between the Florida A-Plus choice and accountability system and the education initiatives proposed by President George W. Bush.

An Evaluation of the Florida A-Plus Accountability and School Choice Program

The Purpose of the Study

The Florida A-Plus Program is a school accountability system with teeth. Schools that receive two failing grades from the state during a four-year period have vouchers offered to their students so that those students can choose to leave for a different public or private school. The theory behind such a system is that schools in danger of failing will improve their academic performance to avoid the political embarrassment and potential loss in revenues from having their students depart with tuition vouchers.

Whether the theory behind the A-Plus Program is supported by evidence is the issue addressed in this evaluation. While it is plausible that the incentives provided by an accountability system with teeth should be an impetus for reform, it is also plausible that the A-Plus system would not produce meaningful academic improvement. Perhaps schools would develop strategies for improving the grade they received from the state without actually improving the academic performance of students. Perhaps schools would not have the resources or policy flexibility to adopt necessary reforms even if they had the incentives to do so. Perhaps the incentives of the accountability system interact with the incentives of school politics to produce unintended outcomes. In short, whether the A-Plus system is successful in improving student achievement is a matter that cannot be resolved without reference to evidence.

The evidence presented in this report suggests that the A-Plus Program has been successful at motivating failing schools to improve their academic performance. In addition, the evidence presented in this report suggests that we should have confidence that the improvement in academic achievement is a real improvement and not merely a manipulation of the state's testing and grading system.

A Brief Description of the A-Plus Program

The Florida A-Plus Program assigns each public school a grade based on the performance of its students on the Florida Comprehensive Assessment Tests (FCAT) in reading, math, and writing. Reading and writing FCATs are administered in 4th, 8th, and 10th grades, while the math FCAT is administered in 5th, 8th, and 10th grades. The scale score results from these tests are divided into five categories. The grade that each school receives is determined by the percentage of students scoring above the thresholds established by these five categories or levels. If a school receives two F grades in a four-year period, its students are offered vouchers that they can use to attend a private school. They are also offered the opportunity to attend a better-performing public school.

The FCAT was first administered in the spring of 1998. Following the second administration of the exam in 1999, only two schools in the state had received two failing grades. Both of those schools, located in Escambia County, had

vouchers offered to their students. Nearly 50 students and their families from those two schools chose to attend one of a handful of nearby private schools, most of which were religiously affiliated. When the FCAT was administered in 2000, no additional schools had their students offered tuition vouchers because none had failed for a second time.

Additional information on the FCAT and A-Plus Program can be found at the Florida Department of Education's FCAT web site at http://www.firn.edu/doe/sas/fcathome.htm or its home page at http://www.firn.edu/doe/.

Other Research on Voucher and Accountability Systems

Many states have testing and accountability systems. Some, such as the New York Regents Exam, date back many years. Others, such as the Michigan Educational Assessment Program, are relatively new. States also vary in the difficulty of the tests they administer, the grades to which tests are administered, whether passage is required for promotion or graduation, and whether sanctions or rewards are attached to student and/or school performance.

Despite the increasing prominence of testing and accountability systems as a tool for education reform, the effectiveness of those systems has been the subject of limited systematic research. Additional research in this area is particularly important given the centrality of accountability systems in many state and federal education reform proposals. The attractiveness of such proposals would be increased if stronger empirical evidence were produced to show that widespread testing and grading of schools provided incentives to schools to improve their performance. Evidence on the effects of using vouchers as a sanction for chronically failing schools would speak to whether accountability systems are likely to be more effective at inspiring improvement if vouchers were part of the program. On the other hand, evidence that widespread accountability testing produced results that were subject to manipulation or failed to inspire improvement would argue against the adoption of such policies. And if the evidence failed to show special gains produced by the prospect of vouchers at failing schools, then a voucher component of the policy would be less desirable.

The greatest amount of research attention has been devoted to evaluations of the accountability system in Texas. The Texas Assessment of Academic Skills (TAAS) has been in existence for a decade and is the most comprehensive of the state testing systems. Students in Texas are tested in 3rd through 8th grades in math and reading. In addition, passage of an exam that is first offered in 10th grade is required for graduation. The state is also phasing-in requirements that students pass exams in order to be promoted to the next grade.

The extensiveness of TAAS, its centrality in education policy in Texas, and the fact that the governor was a candidate for president attracted considerable attention to the program. Linda McNeil and Angela Valenzuela of Rice University and the University of Texas, respectively, issued a report with a series of theoretical and anecdotal criticisms of TAAS, but presented no systematic data on the educational effectiveness of the program.[1] Walter Haney of Boston College has written about the relationship between TAAS and minority dropout rates, but again has not systematically evaluated the effect of TAAS on educational achievement.[2]

The most systematic research on TAAS has appeared in two, somewhat contradictory, reports from the Rand Corporation. The first report, with David Grissmer as its chief author, was released in July of 2000.[3] It analyzed scores from the National Assessment of Educational Progress (NAEP), a test administered by the U.S. Department of Education, to identify state policies that may contribute to higher academic performance. It found that states like Texas and North Carolina, with extensive accountability systems, had among the highest and most improved NAEP scores after controlling for demo-

graphic factors. The report featured a lengthy comparison of student performance in California and Texas to highlight the importance of TAAS in improving academic achievement, as measured by the NAEP.

The second report, with Stephen Klein as its chief author, was released in October of 2000. It cast doubt upon the validity of TAAS scores by suggesting that the results do not correlate with the test results of other standardized tests. Because the other standardized tests are "low stakes tests," without any reward or punishment attached to student or school performance, there are few incentives to manipulate the results or cheat. It is therefore reasonable to assume that the low stakes test results are likely to be a reliable indication of student performance.[4] Schools and students, however, might have incentives and opportunities to manipulate the results of high stakes tests, like the TAAS. Because Klein finds that the results of the TAAS do not correlate very well with the results of the low stakes standardized tests, he and his colleagues suggest that the TAAS scores do not represent the true academic performance of students.

Klein, however, cannot rule out alternative explanations for the weak correlation between TAAS results and the results of low stakes standardized tests. It is possible that the TAAS, which is based on the mandated Texas curriculum, tests different skills than those tested by the national, standardized tests. Both could produce valid results and be weakly correlated to each other if they are testing different things. It is also possible that the pool of standardized tests available to Klein is not representative of Texas as a whole. The standardized test results that were compared to TAAS results were only from 2,000 non-randomly selected 5th grade students from one part of Texas. If this limited group of students were not representative of all Texas students, then it would be inaccurate to draw any conclusions about TAAS as a whole.

In addition to comparing TAAS and standardized test results, Klein and his colleagues also analyzed NAEP results in Texas. Contrary to the findings of Grissmer and his colleagues whose Rand report was only released a few months earlier, Klein concluded that the NAEP performance in Texas was not exceptionally strong. This finding contradicted Grissmer's finding that strong NAEP performance in Texas confirmed the benefits of a high stakes testing system, like TAAS.[5]

A third examination of NAEP scores in Texas published in *City Journal* supports Grissmer's claim and refutes Klein's by finding that NAEP improvements were exceptionally strong in Texas while the TAAS accountability system was in place.[6] The fact that these studies differ while all examining NAEP and TAAS results can be explained by the different time periods examined, the grade levels that are compared, and the presence or absence of controls for student demographics. Without discussing these issues at length, it is sufficient to say that there is some ambiguity regarding any conclusions that can be drawn from a comparison of NAEP and TAAS results. This ambiguity is created in part by the fact that the NAEP is administered infrequently and in only certain grade levels.

In addition to ambiguous research results, our expectations for A-Plus based on the experience of TAAS are further limited by the fact that the two accountability systems differ in one very important respect. The A-Plus Program is unique in that it uses vouchers as the potential sanction for low-performing schools, while the accountability systems in Texas, North Carolina, and elsewhere at most threaten schools with embarrassment or reorganization as the sanction for low performance. The incentives for schools to improve when faced with embarrassment or reorganization may not be the same as the incentives produced by the prospect of vouchers.

We could try to look at recent research on school choice to learn more about whether the prospect of vouchers motivates schools to improve. Unfortunately, while there have been

several high-quality studies on the effects of vouchers on the recipients of those vouchers, there has been relatively little research on whether school choice provides the proper incentives to improve academic achievement in an entire educational system.[7] Recent work by Caroline Minter-Hoxby and by the Manhattan Institute attempt to address whether vouchers would improve academic achievement in the education system as a whole by examining variation in the amount of choice and competition currently available in the United States.[8] Some states and metro areas have more school districts, more charter schools, and other types of choice than others. The findings of both studies suggest that areas with more choice and competition experience better academic outcomes than areas with less choice and competition. While these results support the contention that voucher systems would improve the quality of education for the entire educational system, they are not definitive because they involve argument by analogy. It is possible that competition and choice that currently exist contribute to academic achievement while expanding choice and competition would not have similar benefits. A more direct examination of the effects of expanding choice and competition would address the question more definitively.

The Design of the Current Study

The Florida A-Plus Program offers a unique opportunity to researchers to examine the effects of an accountability system as well as the effects of expanding choice and competition. Because the A-Plus Program involves a system of testing with sanctions for failure, we can examine whether such a program motivates schools to improve. And because the sanction that is applied is the prospect of offering choice to families and competition to public schools, we can examine whether the prospect of choice and competition are effective motivators.

To address these issues we will conduct two types of analyses. First, we will want to determine whether the test that is used to determine school grades in the A-Plus accountability system is a valid test of student performance. Given the concerns raised by the Klein study regarding the validity of the TAAS in Texas, we will examine the validity of the Florida Comprehensive Assessment Test (FCAT) using the same analytical technique used by Klein. That is, we will identify the correlation between FCAT results and the results of low stakes standardized tests administered around the same time in the same grade.[9]

During the spring of 2000, Florida schools administered both the FCAT and a version of the Stanford 9, which is a widely used and respected nationally normed standardized test. Performance on the FCAT determined a school's grade from the state and therefore determined whether students would receive vouchers. Performance on the Stanford 9 (or the FCAT Norm Referenced Test as the state refers to it) carried with it no similar consequences. It is therefore reasonable to assume that schools and students had little reason to manipulate or cheat on the Stanford 9. If the results of the Stanford 9 correlate with the results of the FCAT, then we should have confidence that the FCAT is a valid measure of academic achievement. If the two tests do not correlate, one possible explanation for the low correlation would be that the FCAT results were manipulated so that they were no longer valid measures of student performance. Confirming the validity of the FCAT is important for ruling out the concerns raised by Klein and others before proceeding with other analyses.

Second, we will examine whether the prospect of having to compete to retain students who are given vouchers inspires schools to improve their performance. We would expect that the schools that had already received one F grade from the state and whose students would become eligible for vouchers if they received a second F to make the greatest efforts to improve their academic achievement. That is, if the prospect of choice and competition motivates schools to improve, then the schools that

are in the greatest danger of having their students receive vouchers should experience greater test-score improvement than schools for which that prospect is not so imminent.

To test this proposition we examine the average FCAT scale score improvements for schools broken out by the grade they received the year before. If the A-Plus Program is effective, schools that had previously received an F should experience greater gains on the FCAT than schools that had previously received higher grades.

In short, the design of this study is to verify the validity of the FCAT results and then to determine whether those schools that most imminently face the prospect of having to compete to retain their students who have been offered vouchers experience the greatest gains in their FCAT scores.

Data Examined

The FCAT results examined were from the spring of 1999 and spring of 2000. The Stanford 9 results were from the spring of 2000. The Stanford 9 was not administered statewide in 1999. All test results were obtained from the Florida Department of Education.[10] The FCAT was administered in 4th, 5th, 8th, and 10th grades, but not in all subjects. The Stanford 9 (or FCAT NRT, as it is described on the web site) was administered in 3rd through 10th grades, but the reading results from 10th grade were discarded because the state determined that there was a difficulty with their design. Because both kinds of tests were not available in all subjects in all grades, our analyses are confined to those grades and subjects for which results were available.

The Results of Correlating FCAT and Stanford 9 Results

It appears as if the FCAT results are valid measures of student achievement. Schools with the highest scores on the FCAT also have the highest scores on the Stanford 9 tests that were administered around the same time in the spring of 2000. It is also the case that schools with the lowest FCAT scores also tended to have the lowest Stanford 9 scores. We can know this because the school level results from both tests are highly correlated with each other.

If the correlation were 1.00, the results from the FCAT and Stanford 9 test would be identical. As can be seen in Table 1, the correlation coefficient is 0.86 between the 4th grade FCAT and Stanford 9 reading test results. In 8th grade the correlation between the high stakes FCAT and low stakes standardized reading test is 0.95.[11] This demonstrates an extremely high level of correlation between the tests.

Table 1
Verifying the Validity of the FCAT Results

Correlation between …	Grade Level			
	4	5	8	10
FCAT reading and Stanford 9 reading	0.86	na	0.95	na
FCAT math and Stanford 9 math	na	0.90	0.95	0.91
Number of schools	1514	1514	508	356

All correlations are statistically significant at $p < .01$
na=not available

The math results of the two tests are also highly correlated. In 5th grade the correlation coefficient is 0.90. In 8th grade the FCAT and Stanford 9 school level results are correlated at 0.95. In 10th grade the correlation between the results of the two math tests is 0.91.

It is not possible to verify the validity of the FCAT writing test with this technique because there was no Stanford 9 writing test administered.

In the second Rand Corporation study of TAAS in Texas, Stephen Klein and his colleagues never found a correlation of more than 0.21 between the school level results from TAAS and the school level results of a low stakes standardized

tests. In this analysis we never found a correlation between FCAT and standardized tests below .86. All of these correlations in Florida are statistically significant, meaning that the strong relationship between the results of the two tests is very unlikely to have been produced by chance.

While we cannot check the validity of the FCAT writing results, these analyses strongly support the validity of the FCAT reading and math results. Schools in Florida perform on the high stakes FCAT similarly to how they perform on the low stakes Stanford 9. Since schools would have little incentive to manipulate the results of the low stakes test, the fact that they confirm the high stakes test results is important confirmation that the FCAT measures are credible.

FCAT Improvements by State-Assigned Grade

Now that we have confirmed the validity of the FCAT results, is it the case that schools facing the imminent prospect of competing to retain their students experienced the greatest improvement in FCAT results to avoid that prospect? In fact, the incentives appear to operate as expected. Schools that had received F grades in 1999 and were in danger of having their students offered vouchers if they repeated their failure made the largest gains between their 1999 and 2000 FCAT results.

As can be seen in Table 2, the year-to-year changes in FCAT results for schools do not really differ among schools that received A, B, or C grades from the state. Schools that had received D grades and were close to the failing grade that could precipitate vouchers being offered to their students appear to have achieved somewhat greater improvements than those achieved by the schools with higher state grades. But schools that received F grades in 1999 experienced increases in test scores that were more than twice as large as those experienced by schools with higher state-assigned grades.

On the FCAT reading test, which uses a scale with results between 100 and 500, schools that had received an A grade from the state in 1999 improved by an average of 1.90 points between 1999 and 2000. Schools that had received a B grade improved by 4.85 points. Those that had a C in 1999 increased by 4.60 points. But schools that had a D grade in 1999 improved by 10.02 points. And schools that had F grades in 1999 showed an average gain of 17.59 points. The lower the grade that the school received from the state, the greater the improvement it made the following year. This improvement was es-

Table 2
Comparing Test Score Gains by School Grade

School Grade Given by State in 1999	Change in FCAT Scores from 1999 to 2000					
	Reading		Math		Writing	
A	1.90	(202)	11.02	(202)	.36	(202)
B	4.85	(308)	9.30	(308)	.39	(308)
C	4.60	(1223)	11.81	(1223)	.45	(1223)
D	10.02	(583)	16.06	(583)	.52	(583)
F	17.59	(76)	25.66	(76)	.87	(76)

The change for F schools compared to schools with higher grades is statistically significant at $p < .01$

Math and reading scales are from 100 to 500.
The writing scale is from 0 to 6.
Number of schools is in the parentheses.

pecially large for schools that had received a D or F grade the previous year.[12]

Examination of the FCAT math results shows a similar pattern. Schools that had received an A grade experienced an average 11.02 point gain on a scale that ranged between 100 and 500. Schools that had a B gained by 9.30 points. Schools that had received C grades in 1999 showed 11.81 point gains, on average, between 1999 and 2000. While D schools had improved by 16.06 points from 1999 to 2000 on the FCAT math exam, schools that had received an F grade in 2000 made gains of 25.66 points. Again, the year-to-year gains achieved by schools that had previously received a D or F grade were significantly larger than those experienced by higher grade schools. The improvements realized by schools that had previously received an F grade were especially large.[13]

The FCAT writing exam, which has scores that go from 0 to 6, also shows larger gains for schools that had received an F grade. Schools that had received an A grade in 1999 improved by .36 on the writing test. Schools with a B grade had an average gain of .39. For C schools the improvement from 1999 to 2000 was .45. And for schools that had received a D grade, the improvement was .52 points on the FCAT writing exam. However, schools that had received an F in 1999 demonstrated an average gain of .87 points, about double the improvements for the other schools.[14]

The larger improvements achieved by schools that had received an F and were in danger of having vouchers offered to their students are all statistically significant. That is, the gains observed in the F schools differed from those in the other schools by an amount that is very unlikely to have been produced by chance.

A Hard Test of the Voucher Effect

To what extent were the gains produced by failing schools the product of the prospect of vouchers and to what extent were those improvements the product of the pressures of low performance?[15] One technique for isolating the extent to which gains were motivated by the desire to avoid having students offered vouchers is to compare the improvements achieved by higher-scoring F schools to those realized by lower-scoring D schools. The idea behind this comparison is that high-scoring F schools and low-scoring D schools were probably very much alike in many respects.[16] Both groups of schools had low previous scores and faced pressures simply to avoid repeating a low performance. Schools in both groups were also likely to face similar challenges in trying to improve their scores. It is also likely that a fair number of schools near the failing threshold could easily have received a different grade by chance. That is, random error in the testing may have made the difference between receiving a D or F grade for at least some of these schools. To the extent that chance is the only factor distinguishing those schools just above the failing line and those schools just below the failing line we are approximating a random assignment experiment, like those used in medical research.

While the low-scoring D schools and the high-scoring F schools may be alike in many respects and some may only be distinguishable by chance, schools in each category faced very different futures if they failed to improve. The schools with the F grade faced the prospect of having vouchers offered to students at their school if they failed to improve significantly while D schools did not face a similar pressure. A comparison of the gains achieved by low-scoring D schools and high-scoring F schools should help us isolate the gains that are attributable to the prospect of vouchers unique to those with the failing label. This comparison is a hard test for the effect of vouchers in motivating schools to improve because we are not considering all of the failing schools who faced that pressure and we are comparing against D schools that might have experienced some pressure from the prospect of vouchers to the extent that they anticipated the consequences of their experiencing a decline in future performance.

As can be seen in Table 3, the gains realized by high-scoring F schools were greater than the gains realized by low-scoring D schools.[17] The improvement achieved by higher-scoring F schools on the reading test was 2.65 points greater than that achieved by lower-scoring D schools, although this difference fell short of being statistically significant. On the math test the higher-scoring F schools made gains that were 6.09 points greater than those produced by lower-scoring D schools. The difference between the two groups of schools on the writing test was .16, keeping in mind that the scale for the writing test goes from 0 to 6 instead of from 100 to 500 as is the case for the reading and math exams. The differences between these groups on the math and writing tests were statistically significant at $p < .01$, meaning that we can have high confidence that these differences were not produced by chance.

These gains made by the higher-scoring F schools in excess of what were produced by the lower-scoring D schools are what we can reasonably estimate as the effect of the unique motivation that vouchers posed to those schools with the F designation. Given that the higher-scoring F schools were very much like the lower-scoring D schools, the fact that those schools that faced the prospect of vouchers made larger gains suggests that vouchers provide especially strong incentives to public schools to improve.

The excess gains that we can attribute to the prospect of vouchers can be reported in terms of standard deviations, as is conventional in education research. The improvement on the reading FCAT attributable to the prospect of vouchers was a modest 0.12 standard deviations and fell short of being statistically significant. The voucher effect on math scores was a larger 0.30 standard deviations, which was statistically significant. And the prospect of vouchers improved school performance on the writing test by 0.41 standard deviations, an effect that is also statistically significant.

To put the size of these effects in perspective, education researchers generally consider effect sizes of 0.1 to 0.2 standard deviations to be small, effects of 0.3 to 0.4 standard deviations as moderate, and gains of 0.5 or more standard deviations are thought of as large. For comparison, the effect size of reducing class sizes from an average of 25 students to an average of 17 students according to the *Tennessee Star* study was .21 standard deviations.[18] The motivational benefits of the prospect of vouchers were larger than this class size reduction effect, at least on math and writing scores.

Discussion

The most obvious explanation for these findings is that an accountability system with vouchers as the sanction for repeated failure really motivates schools to improve. That is, the prospect of competition in education reveals competitive effects that are normally observed in the marketplace. Companies typically anticipate competitive threats and attempt to make appropriate responses to retain their cus-

Table 3
Isolating the Effect of the Prospect of Vouchers

	Gains in Reading	Math	Writing
Lower-Scoring D Schools	12.87 (251)	18.15 (272)	0.59 (296)
Higher-Scoring F Schools	15.52 (42)	24.24 (41)	0.75 (35)
Voucher Effect	2.65	6.09	0.16
Voucher Effect Measured In Standard Deviations	0.12	0.30	0.41

Number of schools is in the parentheses.
The math and writing results are significant at $p < .01$

tomers before the competition fully materializes. Similarly, it appears as if Florida schools that foresee the imminent challenge of having to compete for their students take the necessary steps to retain their students and stave off that competition.

While the evidence presented in the report supports the claims of advocates of an accountability system and advocates of choice and competition in education, the results cannot be considered definitive. First, the A-Plus Program is still relatively new and its effects might change, for the better or worse, as the program matures. Second, only two schools in the state have actually had vouchers offered to their students because the schools had received two failing grades. It remains to be seen whether the number of schools where students are eligible for vouchers grows in future years. If the number does not grow, it is possible that the prospect of having vouchers offered to students will not seem so imminent to schools and they will not face the same incentives to improve.

Third, one could offer alternative explanations for the results reported in this study. For example, critics might suggest that the findings reported in this study might be produced by manipulation of FCAT results that may be localized among schools that faced the prospect of receiving a second failing grade. That is, perhaps the high correlation between FCAT and Stanford 9 results does not verify the validity of the FCAT among F schools who may face particularly strong incentives to cheat or manipulate results. If one breaks out the correlations between the FCAT and Stanford 9 results by state-assigned grade and grade level of the test, however, we find that the correlations generally remain high even if we only examine F schools. As can be seen in Table 4, the correlation on the reading score is never lower than 0.77 and never below 0.79 on the

Table 4
Verifying the Validity of the FCAT Results For Each State- Assigned Grade

Correlation between …	Grade Level			
	4	5	8	10
A Schools				
FCAT reading and Stanford 9 reading	0.71	na	0.89	na
FCAT math and Stanford 9 math	na	0.82	0.94	0.98
Number of Schools	121	121	68	8
B Schools				
FCAT reading and Stanford 9 reading	0.48	na	0.91	na
FCAT math and Stanford 9 math	na	0.74	0.94	0.89
Number of Schools	207	207	89	12
C Schools				
FCAT reading and Stanford 9 reading	0.62	na	0.86	na
FCAT math and Stanford 9 math	na	0.79	0.89	0.87
Number of Schools	684	684	254	277
D Schools				
FCAT reading and Stanford 9 reading	0.74	na	0.87	na
FCAT math and Stanford 9 math	na	0.83	0.89	0.90
Number of Schools	436	436	92	55
F Schools				
FCAT reading and Stanford 9 reading	0.77	na	0.99	na
FCAT math and Stanford 9 math	na	0.79	0.98	0.99
Number of Schools	66	66	5	4

All correlations are statistically significant at $p < .01$
na = not available

math scores for F schools. And the correlations for the F schools are comparable to the correlations for schools with higher state-assigned grades. Focusing on correlations between the FCAT and Stanford 9 results only among F schools tends to refute the claim that cheating or manipulation may be localized among failing schools.

As another alternative explanation critics might suggest that F schools experienced larger improvements in FCAT scores because of a phenomenon known as regression to the mean. There may be a statistical tendency of very high and very low-scoring schools to 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 very likely to improve, even if it is only a statistical fluke.

In the case of the FCAT results, however, regression to the mean is not a likely explanation for the exceptional improvement displayed by F schools because the scores for those schools were nowhere near the bottom of the scale for possible results. The average F school reading score was 254.70 in 1999, far above the lowest possible score of 100. The average math score for F schools was 272.51 on the 1999 FCAT, also far above the lowest possible score of 100. And on the FCAT writing exam the average F score received a 2.40 on a scale from 1 to 6, also not likely to cause a bounce against the bottom. Given how far the F schools are from the bottom of the scale, regression to the mean does not appear to be a likely explanation of the gains achieved by F schools.

Another way to test for regression to the mean is to isolate the gains achieved by the schools with the very lowest scores from the previous year. If the improvements made by F schools were concentrated among those F schools with the lowest previous scores, then we might worry that the improvements were more of an indication of regression to the mean (or bouncing against the bottom) than an indication of the desire to avoid having vouchers offered to the students in failing schools. We can test this proposition by constructing a simple regression model that predicts the improvement in FCAT scores for those F schools with previous test scores below average for F schools, for those F schools with previous test scores above average for F schools, and for all schools based on how low their previous scores were. The below average F schools are our proxy for a regression to the mean effect. If their gains are not significantly greater than higher-scoring F schools, then we can reasonably exclude regression to the mean as a likely explanation. All F schools should have experienced a similar motivation to improve to avoid vouchers. But if regression to the mean were operating, then the lowest-scoring F schools should have made significantly greater improvements because they would be more likely to be bouncing against the bottom of the scale.

As can be seen in Table 5, the gains achieved by low-scoring F schools are not greater than the gains achieved by higher-scoring F schools. For analyses of the reading, math, and writing results the higher-scoring F schools experienced gains comparable to those gains experienced by low-scoring F schools. This means that all F schools, whether they were "bouncing" against the bottom of the scale or not, produced similar improvements. According to these models, schools that faced the prospect of vouchers by virtue of having received an F grade made improvements on their reading FCAT that were approximately 4 points higher than would be expected simply from how low their previous score was. The exceptional gain achieved by F schools on the math FCAT was approximately 8 points and the exceptional gain on the writing FCAT was approximately one-quarter of a point on a 6-point scale. All of these results are statistically significant. These results are also

consistent with the voucher effect estimated using the analyses reported in Table 3.

It was a general pattern that schools with lower previous scores made larger improvements. This effect of simply having an accountability system in place to put pressure on lower-performing schools operated across all grades, inspiring low-scoring A, B, C, and D schools to improve. But F schools made gains that were even larger than would have been expected simply given how low their previous scores were. The exceptional incentive that existed for schools that had an F grade was the desire to avoid the prospect of vouchers. We might therefore attribute this improvement realized by F schools beyond what would be expected given their low previous score as their "voucher" gain. Because higher-scoring and lower-scoring F schools experienced comparable exceptional improvements, we can have some confidence that this is a voucher effect and not a regression to the mean effect. And all schools, across all grades, faced some motivation to improve lower scores simply by virtue of having an accountability system in place.

It therefore appears as if two forces were in effect to motivate schools to improve. Schools had some motivation to improve simply to avoid the embarrassment of low FCAT scores. This motivation operated across all state-assigned grades. But schools with F scores had a second and very strong incentive to improve to avoid vouchers.

While one cannot anticipate or rule out all plausible alternative explanations for the findings reported in this study, one should follow the general advice to expect horses when one hears hoof beats, not zebras. The most plausible interpretation of the evidence is that the Florida A-Plus system relies upon a valid system of testing and produces the desired incentives to failing schools to improve their performance.

Table 5
Regression Analyses of the Effect of Prior Scores and Failing Status on FCAT Score Improvements

	Reading		Math		Writing	
Variable	Effect	P-Value	Effect	P-Value	Effect	P-Value
Lower Previous Score	0.19	0.00	0.15	0.00	0.14	0.00
Higher-Scoring F Schools	3.92	0.02	7.93	0.00	0.23	0.00
Lower-Scoring F Schools	2.93	0.11	7.24	0.00	0.39	0.00
Constant	61.67	0.00	59.28	0.00	0.89	0.00
Adjusted R-Square	0.16		0.12		0.12	
Number of Schools	2392		2392		2392	

The dependent variable is the change in FCAT scores from 1999 to 2000. P-values below .05 are generally considered statistically significant.

NOTES

1 "The Harmful Impact of the TAAS System of Testing in Texas: Beneath the Accountability Rhetoric," May 1, 2000. Available at http://www.law.harvard.edu/groups/civilrights/conferences/testing98/drafts/mcneil_valenzuela.html. Accessed most recently on December 20, 2000.

2 "The myth of the Texas miracle in education," *Education Policy Analysis Archives*, 8 (41), August 19, 2000. Available at http://epaa.asu.edu/epaa/v8n41. Accessed most recently on December 20, 2000.

3 "Improving Student Achievement: What NAEP State Test Scores Tell Us," by David W. Grissmer, Ann Flanagan, Jennifer Kawata, and Stephanie Williamson, The Rand Corporation, June 25, 2000. Available at http://www.rand.org/publications/MR/MR924/. Accessed most recently on December 20, 2000.

4 Although low stakes also introduce the danger that students or schools will not devote sufficient effort to demonstrating their true level of performance.

5 For a critique of the Klein and Grissmer reports see Eric Hanushek, "Deconstructing RAND," *Education Matters*, Spring 2001. The article is available on-line at www.edmatters.org.

6 "The Texas School Miracle is for Real," by Jay P. Greene, *City Journal*, Summer 2000. Available at http://www.city-journal.org/html/10_3_the_texas_school.html. Accessed most recently on December 20, 2000.

7 For a summary of recent research see "A Survey of Results from Voucher Experiments: Where We Are and What We Know," by Jay P. Greene, *Civic Report 11*, The Manhattan Institute for Policy Research, July 2000. Available at: http://www.manhattan-institute.org/html/cr_11.htm. Accessed most recently on December 20, 2000.

After that summary was written two important voucher studies were released. One is "Test-Score Effects of School Vouchers in Dayton, Ohio, New York City, and Washington D.C.: Evidence from Randomized Field Trials," by William G. Howell, Patrick J. Wolf, Paul E. Peterson and David E. Campbell, August, 2000. Available at: http://www.ksg.harvard.edu/pepg/. The other is "The Effect of School Choice: An Evaluation of the Charlotte Children's Scholarship Fund," by Jay P. Greene, *Civic Report 12*, The Manhattan Institute for Policy Research, August, 2000. Available at http://www.manhattan-institute.org/html/cr_12a.htm.

8 See "Does Competition Among Public Schools Benefit Students and Taxpayers?" by Caroline Minter-Hoxby, *The American Economic Review*, December 2000; and "The Education Freedom Index" by Jay P. Greene, *Civic Report 14*, The Manhattan Institute for Policy Research, September 2000.

9 This technique addresses what is technically known as the concurrent validity of the FCAT. It does not address whether the letter grades assigned by the state are based on appropriate cutoff points in the test results. That is, this report does not address whether schools given an A in Florida truly deserve an A or whether D schools should really receive an F. To use a metaphor

familiar to most students, this report only examines the validity of the test, not the validity of the curve used to assign grades.

10 The Florida Department of Education also has FCAT scores on its web site at http://www.firn.edu/doe/cgi-bin/doehome/menu.pl. However the web site only has scores for standard curriculum students in 1999 and all students in 2000. This study used scores for standard curriculum students in both years. Earlier analyses on these results from the web site do not produce results that are substantively different from those reported here. This suggests that the inclusion or exclusion of test scores from special needs students has little bearing on the conclusions of this evaluation.

11 The correlation between results of test averages for a school will be higher than correlations between the results of individual student test scores. Nevertheless, these school-level correlations are quite high.

12 The within sample standard deviation for the FCAT reading scores is 21.94, making the gain achieved by the F schools the equivalent of .80 standard deviations.

13 The within sample standard deviation for the FCAT math scores is 20.59, making the gain achieved by the F schools the equivalent of 1.25 standard deviations.

14 The within sample standard deviation for the FCAT writing scores is .39, making the gain achieved by the F schools the equivalent of 2.23 standard deviations.

15 For a case study that documents the extent to which improvements at failing schools can be attributed to the prospect of vouchers, see Carol Innerst, "Competing to Win: How Florida's A-Plan Has Triggered Public School Reform," Urban League of Greater Miami, Inc., The Collins Center for Public Polcy, Floridians for School Choice, The James Madison Institute, and the Center for Education Reform, April, 2000.

16 In fact, the high-scoring F schools had slightly higher average test scores from the previous year than did the low-scoring D schools. This is possible because the state-assigned grade is determined by the percentage of students above certain thresholds on the test score, not by the average test score for the school.

17 High-scoring F schools are those with previous scores that were above average for F schools. Low-scoring D schools are those with previous scores below average for their grade.

18 Finn, J.D., and C.M. Achilles (1999), "Tennessee's Class Size Study: Findings, Implications, and Misconceptions," *Educational Evaluation and Policy Analysis*, 21(2): 97-109.

Devoe L. Moore Center for the Study of Critical Issues in Economics and Government &
Askew School of Public Administration and Policy
Florida State University
Tallahassee, Florida 32306
(850) 644-3525
www.fsu.edu/~policy

52 Vanderbilt Avenue, 2nd Floor
New York, New York 10017
(212) 599-7000
www.manhattan-institute.org

Program on Education Policy and Governance
J.F. Kennedy School of Government
T308 Harvard University
79 Kennedy Street
Cambridge, MA 02138
(617) 495-7976
www.ksg.harvard.edu/pepg