DO ACCOUNTABILITY AND VOUCHER THREATS
IMPROVE LOW-PERFORMING SCHOOLS?

David N. Figlio
Cecilia Elena Rouse

Do Accountability and Voucher Threats Improve Low-Performing Schools?
David N. Figlio and Cecilia Elena Rouse
NBER Working Paper No. 11597
August 2005
JEL No. I20, I21

ABSTRACT

In this paper we study the effects of the threat of school vouchers and school stigma in Florida on the performance of "low-performing" schools using student-level data from a subset of districts. Estimates of the change in school-level high-stakes test scores from the first year of the reform are consistent with the early results used by the state of Florida to claim large-scale improvements associated with the threat of voucher assignment. However, we also find that much of this estimated effect may be due to other factors. While we estimate a small relative improvement in reading scores on the high-stakes test for voucher-threatened/stigmatized schools, we estimate a much smaller relative improvement on a lower-stakes, nationally norm-referenced, test. Further, the relative gains in reading scores are explained largely by changing student characteristics. We find more evidence for a positive differential effect on math test scores on both the low- and highstakes tests, however, the results from the lower-stakes test appear primarily limited to students in the high-stakes grade. Finally, we find some evidence that the relative improvements following the introduction of the A+ Plan by low-performing schools were more due to the stigma of receiving the low grade rather than the threat of vouchers.

David N. Figlio
Department of Economics
University of Florida
Gainesville, FL 32611-7140
and NBER
figlio@ufl.edu

Cecilia E. Rouse
Industrial Relations Section
Firestone Library
Princeton University
Princeton, NJ 08544-1013
and NBER
rouse@princeton.edu

I.   Introduction

Requiring that schools be "accountable" for their efforts to educate students is among the latest popular education reforms. Although many states had already implemented their own accountability systems, the national trend toward increased accountability has escalated following the passage of the federal *No Child Left Behind Act of 2001*. Most of these accountability systems share the feature that they require some form of standardized testing of students and that the results be made public. In addition, most have added sanctions for schools that perform poorly and rewards for schools that perform well. Unlike other states, however, Florida embedded a school voucher program within its school accountability system in 1999. As such, Florida's program both enlists stigma and competition to spur improvement in the lowest performing schools.

Florida's program utilizes stigma by grading schools "A" through "F" – those earning an "F" are deemed "failing." Social stigma has the potential to improve schools if local citizens and educators are so outraged and/or embarrassed their school received an "F" that they make attempts to improve. Ladd and Glennie (2001), in an early commentary on the Florida program, suggest that grading stigma is likely a major motivator in the Florida system. They show that schools in North Carolina demonstrated substantial improvement following the receipt of a failing label, and suspect that similar responses to a negative label may occur elsewhere as well.

Further, students in Florida who attend schools deemed "failing" are eligible for taxpayer-funded vouchers (under the "Opportunity Scholarship" Program) to attend a private, or higher-rated public, school. The Opportunity Scholarship program has two rationales, the first of which is "fairness." When faced with a bad school, middle-class families can move to a new neighborhood, but poor families cannot. The program gives poor children, trapped in low-performing schools,

options.[1] The second rationale is based on economic theory: If the cause of poor performance in American education is the monopoly generated by the assignment of students to their neighborhood schools (e.g., Friedman 1962 and Chubb and Moe 1990), the solution is to infuse more competition into the provision of education. Increased competition should force schools to improve, or go out of business, as consumers (parents) seek to buy the highest quality schooling for the price. Critically, few students need actually use the vouchers to generate a response from the public schools, rather even the threat of vouchers should inspire change.

Although vouchers are probably the most hotly debated education reform in the United States, relatively little is known about their educational benefits and risks. A number of recent studies of publicly and privately-funded vouchers study their impacts on the performance of students who are offered (or take up) the voucher. Greene, Peterson and Du (1998), Rouse (1998) and Witte (1997) examine the Milwaukee voucher program and find mixed evidence of participant test score gains. Peterson, Howell, Wolf and Campbell (2003) describe the results of randomized voucher experiments in Dayton, New York and Washington. After three years, African-American voucher program participants in New York apparently experienced significant test score gains. However, a recent reanalysis of the data from New York City by Krueger and Zhu (2004) suggests there were no gains by African-American students. While based on a rigorous (and ideal) experimental design, these studies do not speak to the broader question of whether a large scale voucher program, or the *threat of competition*, affects public school performance.

In any earlier study of Florida's accountability system, Greene (2001) analyzes school

---

[1] Note, however, that the Florida voucher program is school based. All students in voucher-eligible schools are eligible for a voucher regardless of their family's income level.

aggregate data and finds large effects of voucher threat on school performance. At issue is to

determine whether these observed gains were "real" – that is, reflected true improvements in student

learning – or statistical artifact, and if "real," whether they reflect the threat of vouchers or other

elements of the accountability system such as grading stigma.   There are several alternative

hypotheses that might explain positive trends.

The first is that the gains made by the "F"-rated schools were largely due to "mean

reversion." With mean reverting measurement error, gains the following year by schools that score

unusually low in one year are not likely to be normally distributed around the initial score; rather

the schools are likely to experience larger than average gains the following year. This can arise

because of "floor effects" – there is a minimum score a school can obtain on the test such that there

is "nowhere to go but up," and/or to "ceiling effects" – where schools that scored highly on the test

in one year are disproportionately likely to experience lower than average gains the following year.

Further, in order to receive an "F," a school had to have low mean test scores in only one year.

Therefore, it is possible that many of the "F" schools had transitorily low test scores such that their

scores would have increased in subsequent years even in the absence of the A+ Plan.[2] Kane and

Staiger (2001) highlight that measurement error in school-level test scores can be quite severe,

particularly for small schools, thereby generating just such transient swings in test scores. In order

to address mean reverting measurement error, one needs access to data from several years before

the implementation of the policy in question.

A second hypothesis is that the composition of students changed (Camilli and Bulkley

---

[2] In the job training literature this phenomenon was first documented by Ashenfelter (1975, 1978) and has become known as the "Ashenfelter Dip."

2001). Rumors abound that districts attempted to redraw school attendance area boundaries in order to improve the student characteristics of low-rated schools. And, since before 2002 the school grade was based on the level of student test scores, schools had a large incentive to do so. A third hypothesis is that while the improvements along reported dimensions may have been "real," the schools only focused on "high-stakes" grades (i.e., grades 4, 8, and 10 in reading and writing and grades 5, 8, and 10 in math), and taught to the test such that one cannot infer that overall student "learning" improved as a result of the A+ Plan (Goldhaber and Hannaway 2001).[3,4]

We use student-level data from a subset of school districts in Florida to address the question of whether the threat of vouchers and stigma have an immediate effect on public school performance. We also attempt to determine if any observed gains were "real" or due to other behaviors by schools, and we study whether schools focused on the entire spectrum of children or whether they concentrated on certain subgroups in response to the new accountability system. We do so using data from Florida because voucher threat and stigma are tied to school grades in the school accountability system, making it possible to identify schools that are threatened by vouchers and stigma versus those that are not immediately threatened. One disadvantage of these data is that they do not cover the entire state of Florida. In addition, we have no data on writing test scores.

---

[3] Hannaway and Goldhaber (2001) also find evidence that schools focused on writing that is relatively easy to improve quickly. We cannot address this issue in our data as there are no writing scores before 1999.

[4] There have been other documented unintended consequences of accountability systems that may generate test score increases without educational gains. For example, Jacob and Levitt (2003) find that teachers are more likely to cheat when under pressure to produce high test scores; Jacob (2003), Cullen and Reback (2002) and Figlio and Getzler (2002) find that students are more likely to be classified as learning disabled presumably so that such students' scores will not be included in the school's overall assessment; and Figlio (2003) finds that students are more likely to be suspended during the testing cycle under an accountability system.

And we only have data through 2000 such that our results are properly interpreted as the short-run effect of the introduction of the threat of vouchers and stigma on student test scores. However, data covering the entire state do not contain annual test score data for students prior to 2001 when the state began requiring annual testing of all elementary grades. As a result, prior to 2001 one cannot construct a panel data of students and thereby study the effect of the *introduction* of the accountability system on student performance using annual changes in student-level performance. We are able to conduct such an analysis because, while there was not required statewide testing in every grade, all school districts were previously required to administer tests in the "off grades" for diagnostic purposes. We can identify each student's schools attended, and can follow students from one school to another, and across school districts. Fortunately, our district-level microdata cover over one-third of all "F" schools statewide, so despite the limited population, we can address our key research questions using these data for a large portion of the state.

We find that the unadjusted test scores of students in low-performing schools appear to have increased relative to higher-rated schools after the passage of the A+ Plan on both the high-stakes and low-stakes examinations. However, the gains on the nationally norm-referenced metric (the "low-stakes" test) do not appear to be a function of true large-scale general educational improvements. Rather, the relative gains in reading, especially on the low-stakes test, appear to be explained in large part by student characteristics and, while we find more evidence for a positive effect on relative math test scores, the gains appear primarily limited to students in the high-stakes grades. These results hold for a variety of alternative specifications for the comparison group of schools. Further, we find little evidence of differential effects by the initial achievement level of the student, the student's race or ethnicity, whether the student is an English Language Learner, or

whether the student is eligible for the National School Lunch Program.

Finally, we see similar test score increases following the placement of schools on a "critically low performing schools list" in 1996 or 1997 (which stigmatized schools but brought no other sanctions) suggesting that the threat of vouchers was not the main motivator behind the gains in math achievement.[5] Moreover, we find evidence that schools across the board in Florida boosted the test performance of lower-performing students following 1999; this result provides further support for the notion that grading pressures, rather than voucher threats, positively affected test scores in Florida. We therefore conclude that the successes of the school accountability system in improving student test scores in the lowest performing schools were likely more due to the other attributes of the school accountability system (e.g., the performance bonuses for schools, or the competition for good school grades) than they were due to the threat of school vouchers.

We emphasize that our estimates are only those of the *differential* effects of stigma and voucher threats on low-achieving schools, and are not the only effect of the A+ Plan on low-achieving schools. Our methodology does not permit us to investigate the systemic effects of the introduction of the accountability system (i.e., across all schools), and these systemic effects may well have been quite large. Carnoy and Loeb (2002) and Hanushek and Raymond (2004), for instance, show that states with high-stakes accountability systems experienced relatively large improvements on their scores on the National Assessment of Educational Progress (NAEP), another

---

[5] One might be concerned that educators in Florida anticipated vouchers with the publication of the critically low performing schools list. However, this was highly unlikely primarily because the democratic governor at the time, Lawton Chiles, opposed vouchers.

nationally-normed examination, so there may have been overall gains due to the A+ Plan as well.[6]

The rest of the paper is organized as follows. We review the evolution of Florida's school accountability system, the details of how a school is labeled "failing," and other characteristics of the A+ Plan in the next section. In Section III we develop the empirical framework and discuss the data; we present the results in Section IV, and conclude in Section V.

II.   Background on the A+ Plan for Education

A.   General Background on the A+ Plan

Education reform, and specifically a system of school accountability with a series of rewards and sanctions for high-performing and low-performing schools, was the policy centerpiece of Jeb Bush's 1998 gubernatorial campaign in Florida; the resulting A+ Plan for Education was Governor Bush's first legislative initiative upon entering office in 1999. The A+ Plan called for annual curriculum-based testing of all students in grades three through ten, annual grading of all public and charter schools based on aggregate test performance, rewards for high-performing and improving schools, and sanctions (as well as additional assistance) for low-performing schools. The most famous and publicized provision of the A+ Plan, however, was the institution of private school vouchers, called "Opportunity Scholarships," for students attending (or slated to attend) chronically failing schools – those receiving a grade of "F" in two years out of four, including the most recent year.

School grading began in May 1999, immediately following passage of the A+ Plan into law.

---

[6] Florida participated in NAEP until 1996 and then again beginning in 2002. As such one cannot study the effect of the introduction of the accountability system on NAEP scores.

The state was able to roll out school grades, albeit on an interim set of criteria, so quickly following passage of the A+ Plan for Education because Florida began its experimentation with test-based school accountability several years before Bush's election. As seen in the accountability timeline presented in Table 1, Florida first began rating schools based on their aggregate test performance in 1996 with the introduction of the Critically Low Performing Schools List. This list emerged from a state mandate that all school districts use a nationally norm-referenced test (NRT) such as the Iowa Test of Basic Skills (ITBS) or Stanford-8 Achievement Test (SAT-8), chosen at the discretion of the school district, to annually test students in grades three through ten in reading and mathematics. Based on aggregate student NRT scores, the Critically Low Performing Schools List stratified schools into four performance groups. Schools receiving the lowest rating in this system faced the potential for stigma, but did not face significant sanctions for poor performance. Also in 1996, the state began administering the new curriculum standards-based Florida Comprehensive Assessment Test (FCAT), and in 1998 the state began reporting school-level FCAT scores in reading and mathematics, given in grades four, eight and ten in reading and five, eight and ten in mathematics, but these scores were not used for state-level school accountability purposes.

For the first three years of the new A+ accountability program, testing was limited to the subjects and grades that had been the focus of the earlier accountability system. That is, students in grades 4, 8 and 10 were tested in reading and writing, and students in grades 5, 8, and 10 were tested in math.[7] Further, school grading moved from being based on performance on nationally norm-referenced tests taken at the school district's discretion to being based on performance on the

---

[7] Beginning in the 2001-2002 school year, testing was broadened to all grade levels from three to ten.

FCAT. When the A+ Plan was enacted in 1999, the decision was made to maintain a single statewide nationally norm-referenced test, and in 2000, the Stanford-9 Achievement Test (SAT-9) was instituted as the FCAT Norm-Referenced Test (NRT), and assigned to all students in grades three through ten, although the results of this assessment were not used to assess the schools.

While the student-level correlation between test performance on the NRT and the FCAT curriculum-based assessments (known as the Sunshine State Standards (FCAT-SSS) examinations) is quite high in the case of both reading and mathematics (generally at the level of around 0.8), the two tests assess different sets of skills. For instance, elementary school FCAT reading tests assess the student's ability to place words and phrases in context, to understand the main idea, plot and purpose of a piece of text, to evaluate comparisons and cause/effect relationships, and to use reference and research tools. The related NRT assesses initial understanding of a passage and the ability to interpret reading selections, to synthesize and evaluate critical information presented in selections, and to recognize and apply reading strategies in a variety of settings. The FCAT-SSS exams, therefore, test a narrower set of skills than do the broader nationally norm-referenced examinations (NRT), but test these skills in greater depth than their national counterparts.

B.   Incentives in the A+ Plan

While only schools in danger of receiving a grade of "F" faced the threat of vouchers, all schools faced the possibility of financial rewards and recognition by earning a grade of "A" or improving grades from one year to the next.[8] Moreover, schools likely faced pressure from the community, as school grades were apparently highly capitalized into housing values, especially at

---

[8] The appendix describes, in detail, the criteria for receiving various school grades.

the program's beginning (Figlio and Lucas, 2004). In all cases, schools faced pressure to improve the fraction of students scoring above minimal levels on the FCAT-SSS examinations. Schools therefore had the incentive to focus their attention on students in the lower portions of the achievement distribution, at the potential expense of higher-achieving students. In addition, schools had an incentive to concentrate on the subjects and grades tested on the FCAT-SSS examination. In elementary school, reading and writing were high-stakes subjects in grade four, while mathematics was a high-stakes subject in grade five.

C.   Evidence of Test Score Improvements

Table 3 shows the distribution of school grades for the first four years of the A+ Plan. In 1999, 78 schools received an "F" grade – the students in two of these schools became eligible for vouchers in 1999. These two schools had been on the list of "critically low-performing schools" in 1998, and the state had "grandfathered" them into the A+ Plan as having one year of "F" credit prior to the first imposition of school grades in 1999. While many then predicted that vouchers would become widespread in Florida the following year (as a substantial number of these schools were predicted to receive "F"s the following year), in reality only four schools received "F"s in 2000 and none of these schools had previously received an "F" grade. That is, none of the original "F" schools received an additional "F" the following year. And one can potentially see why in Figures 1a and 1b which show the change in school average test scores from 1999 to 2000 in reading, math and writing by the school's grade in 1999. Using the FCAT-SSS scores as a basis for evaluating schools, these figures suggest that schools across the spectrum improved following the implementation of the A+ Plan. These gains were particularly pronounced in the low-rated schools

threatened with sanctions. In fourth grade reading, "F" schools improved by about 12 scale points whereas the other, higher-rated, schools improved by less than one-half of that gain. Similar trends are seen in math and writing (Figure 1b) (for fifth and fourth graders, respectively.)

These relative gains were hailed by many as evidence that the A+ Plan was effective at improving the performance of "low-performing" schools. Table 4 shows the mean FCAT-SSS scores for 4th grade reading and writing and 5th grade math in the 1998-1999 school year, by the school's grade. The table suggests that the average test scores of students attending "F"-rated schools was significantly lower than those of higher-rated schools. For example, the "F"-rated schools scored, on average, 19 points lower than the D-rated schools in reading, 16 points lower in math, and 0.4 points lower in writing. These differences are statistically and educationally significant. In reading the student-level standard deviation in reading was about 60 points and that in math about 52 points, such that the disparities between the "F"-rated and "D"-rated schools was about 0.3 of a standard deviation.[9] The fact that the schools appear to have improved was viewed as a great achievement.

Governor Jeb Bush, in a press release published on the Governor's web site lauding the A+ Plan argued that "...the students who are benefitting most from our reforms are those children who, in the past, had been most likely to be left behind.... It is clear that the state's unprecedented attention to children in low performing schools is producing remarkable results." This optimism only increased the following year when no schools received a failing grade. Indeed, in a speech

---

[9] Unfortunately at this time we are unable to calculate the student-level standard deviation in writing scores. However, the ratio of the student-level to school-level standard deviations in math and reading is about 2.3-2.5 suggesting that the student-level standard deviation in writing may be about 1.0. Thus the difference between the F-rated and D-rated schools was about about 0.4 of a standard deviation.

given February 25, 2003 to the Hoover Institution's Board of Directors, Governor Bush argued that

the threat of vouchers "...has been the greatest catalyst for improvement. In 1999, there were 78 "F"

schools in Florida. That number dropped to four the next year, and to zero the year after that."[10] The

key is to determine whether these gains reflected true improvements in student learning or statistical

artifact; and if they did represent true learning gains, whether they reflect the threat of vouchers or

other elements of the accountability system such as grading stigma.

III.   Empirical Framework and Data

A.   Empirical Framework

We employ the following empirical framework to assess the impact of the threat of vouchers

and school stigma on Florida's students in low-performing schools. We begin by using school level

data from 1999 and 2000 to provide evidence on the representativeness of the data from our subset

of districts using a difference-in-differences framework. We estimate

$$T_{st} - T_{st-1} = a + bF_s + (e_{st} - e_{st-1}) \, , \tag{1}$$

where T_{st} is school s's average test score in year t, F_s is a dummy variable indicating whether or not

the school received a failing grade of "F" in 1999, and e_{st} is a normally distributed error term. The

key parameter of interest is b – a school's test score response to having received an "F." Thus, we

estimate whether the change in test scores experienced by the F-rated schools is significantly

different from that of higher-graded schools.

Our main analysis uses student-level test scores from a subset of districts. We use these data

to estimate models such as

---

[10] Downloaded from http://www-hoover.stanford.edu/homepage/news/022503.html.

$$T_{ist} - T_{ist-1} = \alpha + YEAR_t \lambda + \beta(F_{is} \times POST_t) + \delta(GRADE)_{ist} + \phi_s + \mu_s t + \varepsilon_{ist} \qquad (2)$$

where T_{ist} is student i's test score in school s in year t, F_{ist} is a dummy variable indicating whether

or not the school attended by student i received a failing grade of "F" in 1999. $YEAR_t$ is a vector

of year effects, $POST_t$ is a dummy variable indicating if the year is after the implementation of the

A+ Plan (in these data it is equal to one if the year is 2000 and zero otherwise), $GRADE_{ist}$ is a vector

of dummy variables indicating the student's grade in school, ϕ_s is a vector of school-fixed effects,

μ_s are school-level time trends predicted from the pre-1999 data, and ε_{ist} is a normally distributed

error term. Again, the key parameter of interest is β – the change in a student's test score resulting

from her school having received an "F."

This particular specification of the education production function allows us to control for the

student's prior academic achievement (by controlling for T_{ist-1}), while constraining the effect of the

prior achievement to have a coefficient of one. If the lagged test score (T_{ist-1}) does not fully capture

the human capital the student brings to the current grade and this human capital is correlated with

the school's grade, then our estimates may over or under state the effect of the A+ Plan. We have

estimated models in which we control for multiple lags of the student's test score with qualitatively

similar results.[11]

Because we have multiple observations for each school, we adjust our standard errors for

clustering at the school level. Bertrand, Duflo and Mullainathan (2004) indicate that these standard

---

[11] The coefficient estimates for math in these models are quite similar, those for reading tend to be a little larger. However, we believe that this is due to some heterogeneity in the effect of the A+ Plan as the effects on reading scores are larger for this subset of students in even the simplest models. Similarly, we have also estimated NRT models with student fixed effects as well as with student-specific time trends. The results are qualitatively similar and available from the authors on request.

errors may still be understated if the errors are positively serially correlated. We find, however, that serial correlation in the error terms is not substantial in our application, suggesting that no further error correction is necessary.

B.   Data

        We utilize student-level data from a set of participating school districts from 1995-2000. These data include scores on FCAT-SSS and norm-referenced examinations (either the SAT-8, ITBS or the California Test of Basic Skills (CTBS) prior to 2000, and FCAT-NRT (SAT-9) in 2000), and basic student demographic attributes, including information on student race, ethnicity, poverty status, limited English proficiency status, and disability status.

        Our primary variable of interest is the student's test score. Because the students in the data took as many as three separate norm-referenced examinations during our sample period, our primary dependent variable is the normal curve equivalent of the test score, rather than the scaled score itself.[12] As such, the mean is 50 and the individual-level standard deviation is 21.06. Correlations between normal curve equivalents in the three examinations are extremely high, exceeding 0.80 for the set of students for whom we observe more than one type of NRT test score, suggesting that we can compare across school districts and across time. We have also conducted extensive tests to ensure that the differences in NRT tests employed are not driving any reported results. In the paper, we report the results of regressions that include dummy variables for each NRT test used, but none of the key results are different (to the third significant digit) were we to exclude these dummy variables. Moreover, in Tables 6a and 6b we report the results of model specifications in which we

_____

[12] We have conducted the analysis using national percentile rankings with similar results.

control for lagged test scores. The change in the coefficient on a lagged test score is never more than one percent when we systematically distinguish between students who took the same test in successive years versus students who took different NRT tests in successive years.

We also have access to FCAT-SSS test scores for students in the relevant grades for the years in which the FCAT-SSS was administered. While this last test score is not as useful – because it is impossible to follow the same students from one year to the next (unless a student fails the grade and is forced to retake it) and because its has only been administered for a few years – it is still important to measure changes in student performance on the higher-stakes examination.

In each school district in our data, we use data on third, fourth, and fifth grade NRT scores.[13] Because we collected the data directly from the districts, they required some "cleaning" to deal with errors in student identification numbers.[14] In our analysis sample we observe data for 286,729 year-to-year student transitions in mathematics test scores and 287,444 year-to-year transitions in reading

---

[13] Some school districts test their students in earlier grades. We do not use these data because the testing instruments differ greatly from those used in later grades and we were concerned about comparability in that context. That said, the results are quite similar if we include these data as well.

[14] We found instances in which clearly different students (e.g., a white male in 3rd grade vs. a black female in 5th grade in the same year) shared the same student identification number. Therefore, we adopted the following rules: We first excluded from our analysis any students who held an identification number shared by two or more students during the same academic year; we retained only those student identification numbers with multiple concurrent records (which could legitimately occur due to a within-year move) where the records shared the same race/ethnicity, sex, grade and birthdate. We then excluded the student identification numbers for "students" who were observed in different years as being of different sexes. For students who apparently changed schools within a school district, we cross-checked their information against two sets of identification numbers. But for students who were found in one school district in one year and a different school district in the next year, we retained only those students for whom both records shared the same sex and birthdate, but we considered changes in reported race or ethnicity for the same student identification number across years to be legitimate, so long as sex and birthdate remained unchanged. With these rules we drop 4,016 observations.

test scores, representing 182,135 students. In 7 percent of cases, the school that the student attends received a grade of "F" in 1999, and in another 38 percent of cases the school received a grade of "D."

In addition, we have also compiled administrative data from the Florida Department of Education including test scores for all schools in the state from 1999 and 2000 (representing the 1998-1999 and 1999-2000 school years). We analyze data on FCAT-SSS scores for grade 4 in reading and grade 5 in math. These data include over 1500 schools. Because these data do not allow us to control for changes in school characteristics we analyze these data primarily to assess the likely representativeness of our sample of data using individual student-level records.

Finally, because elementary schools comprised the vast majority of "F" schools, we limit the analysis to students in elementary grades. Because they are more economically and ethnically homogeneous and smaller than secondary schools, elementary schools were more likely to receive extreme grades (i.e., grades other than a "C") in the accountability system.

To assess the extent to which the districts for which we have student-level data are representative of the state as a whole we compare estimates of equation (1) using data for elementary schools across the state and for the subset of districts for which we have microdata. We also assess the quality of our student-level data by comparing estimates using the publicly-available school-level data on the subset of districts to school-level aggregates we computed using the microdata.[15] Table 5 shows these results. The top panel of Table 5 shows the test score gains in reading for 4th grade students, the bottom panel shows the gains in math for 5th grade students; the dependent

---

[15] We show regression results in order to preserve confidentiality of the districts. The results presented pertain only to our analysis sample; however, the analogous results using all available data are very similar to those presented in the table.

variable is the change in the school's average test score from 1999 to 2000. All of the regressions are weighted by the number of students taking the test in 2000.

The results in the columns (1) and (4) suggest that F-rated schools gained about 8.5 scale score points more than higher-rated schools between 1999 and 2000 in math and reading; and these estimates are statistically significant from zero. The results in columns (2) and (5) for the subset of districts for which we have student-level are quite similar as are the results in columns (3) and (6) using the microdata aggregated to the school level. The fact that the coefficient estimates for the subset of districts for which we have microdata are similar to those for the state suggest that the districts are likely representative of the state.

These results also replicate those reported by Greene (2001) as they suggest disproportionately large gains on the FCAT-SSS test by students in schools that received an "F" grade in 1999; a gain of between 0.14σ and 0.19σ (which is respectably large for education interventions). They therefore suggest that the threat of becoming voucher-eligible combined with social stigma can spur school improvement. However, as also discussed above, these gains may be misleading due to teaching to the test, mean-reverting measurement error, and/or changing student characteristics. We address these potentially confounding explanations using the student-level data.

IV. Results

1.   Estimates Controlling for School Fixed Effects

Tables 6a and 6b show estimates from a version of equation (2) that controls for the lagged

test score on the right-hand side.[16] The specification shown here is similar to that in Table 5, except that we use the microdata and include school fixed effects to control for time-invariant characteristics of the schools. In columns (1) and (2) of Table 6a we use the FCAT-SSS score in reading as the dependent variable.[17] The raw gap between "F" schools and higher-rated schools is presented column (1) – again, we estimate relatively large and positive gains by students in the "F" schools of 0.09σ (where 0.09σ is the coefficient estimate divided by the test standard deviation). However, this improvement may overstate the differential effect of the A+ Plan on "F" schools by confusing changing student characteristics with changes occurring because of the accountability system. This bias could occur because of the normal mobility of students between schools, changes in student and parent school choice when a school becomes voucher threatened or stigmatized, regulatory changes in who is required to take the tests, or deliberate changes in school attendance area boundaries. We attempt to account for changing student characteristics by including student demographics, socio-economic status and limited English proficiency status, as well as the student's lagged test score. In column (2) we use the previous year's NRT reading score (there is no FCAT-SSS in reading for 3rd grade). Note that the effect drops by half in column (2) compared to column (1) although the coefficient is still significant at the 10% level. Thus, we find there was a small

---

[16] Our reason for controlling for the prior test score on the right hand side is evident in Tables 6a and 6b; we examine the effect of receiving a failing grade on both the FCAT-SSS and the NRT and we do not have a lagged test score for the FCAT-SSS. Therefore, we estimate a less constrained version of the specification.

[17] In these tables we restrict the sample to students who take both the FCAT-SSS and norm-referenced exams in the same year for fourth-grade reading and fifth-grade mathematics. There are actually between 600 and 1,200 students (depending on the examination) who took one but not the other exam. However, the results are very similar whether we include or exclude these students. We chose to exclude these students from Tables 6a and 6b so that the results can be directly compared across the columns.

increase in reading achievement on the high-stakes test among 4th graders (the high-stakes grade).

In the subsequent columns, we examine two other hypotheses. In columns (3) - (6) we examine whether there appeared to be a focus on the Sunshine State Standards Test by estimating the change in test scores among 4th graders on the norm-referenced test (NRT) – which assesses a broader range of skills than the FCAT-SSS and was not the basis of the school grading under the A+ Plan. Column (3) replicates column (1) but uses the norm-referenced test as the dependent variable. We find a positive, statistically significant increase in reading scores among 4th graders of about 0.07σ – slightly smaller than the effect on the FCAT-SSS. Further, once we control for the prior year's NRT test score, the coefficient estimate drops to 0.628 (or 0.03σ) and becomes statistically insignificant[18]. Thus we have some modest evidence that teachers may have focused attention on the higher-stakes FCAT-SSS.

We also examine whether emphasis was put on the students in the grade level that was tested (i.e., 4th grade in reading and 5th grade in math). In columns (5) and (6) we estimate that 5th graders (who were not tested in reading) in "F" schools experienced test score increases of about 0.025σ as well, although the gains are not distinguishable from zero, suggesting there was not a big shift in emphasis to the high-stakes grade. In results not presented in the table, we combine fourth and fifth grades in the same lagged-dependent-variable model, and directly test for the difference in estimated effects of an "F" grade on NRT reading scores in fourth versus firth grades. Unsurprisingly, given the similarity in the estimated "F" effect coefficients in columns (4) and (6), the difference between

---

[18] Because districts used different NRT tests before 2000, one might wonder whether the coefficients on lagged NRT tests are comparable when referring to different lagged NRT tests. We have found, both here and in the results presented later in the paper, that the coefficient on the lagged NRT test score never varies by more than one percent, regardless of the NRT test employed.

these coefficients is far from statistical significance (p=0.67).

In Table 6b we examine the effect of the voucher-threat and stigma for "F" schools in the A+ Plan on math scores. The table has a layout similar to that in Table 6a. In columns (1) and (2) we estimate an effect of about 0.23σ on the high-stakes math test scores. In contrast, the coefficient estimate falls to 0.08σ-0.11σ on the norm-referenced exam. We again find an emphasis on the narrower Sunshine State Standards testing curriculum. Further, while we estimate a positive and statistically significant effect on 5th grade math scores, we estimate a negative effect on 4th grade math scores, although the effect is not statistically significantly distinct from zero. In a specification in which we combine fourth and fifth grade students in a lagged-dependent-variable model, we find that the difference in the "F" effect between fourth and fifth grades is statistically significant at the four percent level.

The contrast of results between columns (1) and (2), and (3) and (4), in both Tables 6a and 6b suggests that student characteristics explain part of the apparent gain in reading, but not in math.[19] A contrast of columns (2) and (4) in both tables is consistent with educators in "F" schools differentially "teaching to the test," that is putting added emphasis on improving student outcomes on the (high stakes) Sunshine State Standards tests. That said, these efforts nevertheless had some spillover to the broader NRT as well, at least in mathematics.[20] It is also important to note that

---

[19] More directly, the estimated effect of receiving an "F" grade is substantially larger than the estimate presented in column (1) for reading if we do not include student characteristics and slightly larger than that presented in column (3) for math.

[20] It is not unusual to find larger effects on math than reading (see, e.g., Rouse (1998)). One potential explanation is that math skills are more responsive to school-based teaching whereas reading skills develop over a longer time and therefore require more learning outside of regular school hours.

teaching to the test, if present, could be seen as a positive outcome if the high-stakes test better reflects the state's standards, as is the case with the FCAT-SSS examination. Finally a comparison of columns (4) and (6) suggests some emphasis on the achievement of students in the high-stakes grade in mathematics suggesting that schools may have strategically focused resources in response to the incentives of the accountability system.[21]

2.   Alternative Specifications

The results in Tables 6a and 6b focused on the high-stakes grades or an adjacent grade. In Table 7 we combine these samples into one specification.[22] As in the previous analysis, the specifications in columns (1) and (5) compare the achievement gains of students in "F"-rated schools to the achievement gains of students higher-rated schools, controlling for whether a school was not graded in 1999. However, we have now expanded the sample to cover the entire data period from 1995 through 2000. We estimate a small positive, but statistically insignificant effect for both reading (column (1)) and math (column (6)). In columns (2) and (7) we further incorporate pre-FCAT school trends in NRT scores, generated from data before 1999. This innovation, continued in all subsequent specifications reported in the paper, is done to take into account the possibility that some schools may be on an upward or downward trajectory over time. The inclusion of pre-existing

---

[21]As mentioned above, we adjust our standard errors to account for clustering at the school level. As a consequence, our reported standard errors tend to be approximately three times greater than the unadjusted fixed effects OLS standard errors. However, it is important to note that coefficients that we report to be statistically insignificant also tend to have small magnitudes. Thus, the choice of standard error adjustment does not affect the qualitative interpretation of the most important results of this paper.

[22] As a result these specifications include test scores for 3rd, 4th, and 5th graders, although the 3rd grade test scores are not used as an outcome variable, but as the lag for the 4th grade test score.

trends modestly strengthens the estimated effects of receipt of a grade of "F," particularly in mathematics, though these estimates are still not statistically significant at conventional levels.

One concern about these specifications is that schools with higher school grades are included as the counterfactual. However, schools that received an "A" might be quite different than those that received an "F." And although we control for school fixed effects, there may be time varying changes in the schools that confound the true effect of the A+ Plan on low-performing schools. We address this concern by including dummy variables representing the schools that received grades of "A," "B," and "D" in 1999 and interactions of these dummy variables with an indicator for the implementation of the A+ Plan (the variable called "Post 1999"). These results are presented in columns (3) and (8); the difference between the coefficients for the "F" and "D" schools and the standard errors are presented as well. As shown in column (3) we continue to estimate a small, statistically insignificant effect on reading scores when comparing "F" to "C" (the omitted category) schools or when comparing "F" to "D" schools. For math (in column (8)), while we now find that the change in test scores in "F" schools was not statistically different from that in "C" schools, there were larger gains in "F" relative to "D" schools – the effect was 0.07σ and the estimate is statistically significant at the 10% level.

One concern is that the math gain may be due to mean-reverting measurement error. In this case, one should not interpret the relatively large increase in math test scores of students in the F-schools as due to the sanctions in the A+ Plan, but rather because the schools' test scores were unusually low in 1999. To test this possibility, we created "pseudo" "F" grades for schools in 1998 (a year before schools were actually graded) based on the criteria used to grade schools in 1999. If mean-reverting measurement error explains the test score growth following the (true) 1999 grades,

then one should also observe such an increase for schools for which we created the pseudo "F"

grades in 1998. These results are presented in column (4) for reading and column (9) for math. In

both cases we estimate a negative coefficient that is statistically significant at the 1% level in the

case of reading. The fact that this estimated relationship has the opposite sign of that which we

would have expected if mean-reverting measurement error were the explanation indicates that some

other factor is responsible for the observed test score gains associated with receiving a failing grade.

Indeed, the estimated negative coefficients in the 1998-grading experiment indicate schools that

failed in 1999 may have been expected to have even lower scores in 2000 if not for the

accountability system.

We also asked whether the estimated effects of "F" receipt were due to the school grading

system, or some other feature of the school. After all, school grades in 1999 were very highly

correlated with the socio-economic status of students in the school. To try to distinguish between

these two stories, we identified the seven percent of schools with the highest fraction of low-SES

students, as opposed to the "failing" schools.[23] We find that low-SES schools in general did not

experience the same pattern of results as did the schools identified as failing. This exercise, along

with the analysis pertaining to mean-reverting measurement error, suggests that our estimated effects

of an "F" grade in mathematics (as well as reading) may even be mildly understated, given that the

lowest-SES schools experienced a modest movement in the opposite direction from that observed

with the failing schools.

In Table 8 we further examine the effect of the comparison group of schools on our estimated

---

[23] We chose a seven percent threshold because seven percent of the schools in our sample received grades of "F" in 1999.

effects of the A+ Plan. To identify schools that are quite similar to the voucher-threatened ("F"-rated) schools, we identified schools that failed on one or two subjects on the FCAT, but not all three (as did the "F"-schools). Thus in the first two rows we compare the "F" schools to those that failed on either math or reading; the third row compares them to schools that failed in reading and either math or writing, and the fourth row compares them to schools that failed in math and either reading or writing. As is evident from the relative similarity of the coefficient estimates down the columns, the results for both math and reading are robust to these alternative characterizations of the comparison group.

An alternative way of thinking about the comparison group is to implement a regression-discontinuity design, in which we estimate the differential response between "F" and "D" schools in a model controlling directly for the percentage of students failing to meet proficiency in reading or mathematics in the year of school grading (multiplied by a Post 1999 dummy variable). These results are presented in the last row of Table 8. We observe somewhat stronger positive evidence – albeit still modest in magnitude – of a differential response by "F" schools vis-a-vis "D" schools.

3.   Subgroup Estimates

As discussed above, in the early years of the A+ Plan, schools were identified as low-performing if more than 40 percent of students scored at the lowest level (out of 5) in math or reading, or if more than 50 percent of students scored in the lowest two levels in writing. In addition, schools that improved by more than one grade level or retained an "A" were awarded an additional $90-$100 per student. Finally, students were tested in only grades 4, 8, and 10 in reading and writing, and in grades 5, 8, and 10 in math. Given this structure, schools had an incentive to

focus their efforts on particular groups of students; failing schools may have had an even greater incentive to do so. For example, schools had an incentive to focus on improving student scores in only one subject, at least to avoid the stigma of an "F" and the threat of vouchers. Further, schools had an incentive to improve the test scores of students near the threshold for passing as well as the lowest performing students; they had much less of an incentive to focus on the highest performing students (except that they did not want the scores of these students to suddenly decrease). Finally, schools had an incentive to focus on students in the grades that were tested – the "high-stakes grades." The estimates in Table 6b suggests some evidence of focusing on the high-stakes grade in math. In Table 9 we estimate whether the incentives in the A+ Plan differentially affected students along other characteristics. The coefficient estimates shown reflect the difference between "F"- and "D"-rated schools.

The top panel of Table 9 examines whether there were differential gains by students categorized by "ability." We divided students into quartiles based on their third grade norm-referenced test scores in the subject in question.[24] The coefficient estimates suggest that the top-performing students in the "F" schools showed test score declines in both reading and math relative to those in "D" schools, consistent with increased emphasis on lower-performing students although these differences are not statistically significant. Further, we find that the gains among students in the two lowest quartiles were significantly greater than those in the top quartile in math (at the 10% level). In reading, however, the changes across the test score distribution are statistically

---

[24] In the results reported in the table, we have not fully interacted the right-hand side variables by the characteristic in question, although the results are similar if we do. We report these subgroup results separately to facilitate interpretation. The ability quartiles are based on the sample within each district.

indistinguishable.

In the lower three panels of Table 9 we examine whether there were differential responses by the race or ethnicity of the student[25], by whether the student was limited English proficient, and by whether the student qualified for the National School Lunch Program ("Low-income students"). In these specifications we interact the student's characteristic with the school's grade and whether the A+ program had been enacted. Although we would have expected that the sanctioned schools would put additional resources into minority, low-income, and English language learners – traditionally lower performing students – we do not find evidence that they did so relative to D-rated schools. In fact, the only subgroup where we find reading test score effects that, while modest in magnitude, reach statistical significance at traditional levels is the fluent English speaker group.

Overall, we find that students showed significant improvement on the high-stakes exam (the FCAT-SSS) in both reading and math. However, once we examine the effect of the accountability system on a nationally normed test and control for student characteristics (by including the lagged test score), we find that there was no large-scale relative improvement in average NRT reading scores due to the A+ Plan among students in the lowest performing schools. We do find evidence of a modest relative improvement in average NRT math scores, although this improvement appears largely concentrated in the high stakes grade. And while we find some evidence that sanctioned schools put additional resources into low-performing students in math, more generally we find little

---

[25] In the data from Florida, students are categorized by their parents into mutually exclusive categories of Black, Hispanic, Asian, White, Native American, and Mixed. We included dummy variables for each of these groups in our analysis, but for ease of presentation only present the estimated effects of a grade of "F" for the three largest student subgroups of Black, Hispanic, and White. It might seem surprising that the coefficient estimates in Table 9 do not appear to be the average of effects in Table 7. The reason is that in Florida, the lowest-rated districts have large proportions of African American and Hispanic students.

evidence that these schools treated students differently than low-performing schools that did not receive an "F."

4.   Voucher Threat or Stigma?

Because the receipt of an "F" brings both the threat of vouchers and social stigma to the school (and community) one cannot readily distinguish which of these two forces is responsible for the improvements in math test scores that we observe. While we cannot definitively distinguish these two factors, we can shed some light on the issue by studying changes in academic performance that resulted from the earlier policy of placing schools on a "critically low performing list." This list, produced in 1996, 1997 and 1998, resulted in the "stigma" of being identified as a low-performing school, but it did not result in any sanctions (such as the threat of vouchers). The vast majority of schools identified as critically low-performing were only identified as such in the first year of the list, after which point nearly all schools emerged from the list, with only four schools remaining on the list in 1998.[26]   Thus, if we interpret the response to being on the critically low performing list as the stigma effect, then we can interpret the difference between that effect and the changes after the introduction of the A+ Plan as the voucher threat effect. These results are presented in the top panel of Table 10.

The first row of the table shows the difference between the changes in "F"-rated schools (versus all other schools) after the introduction of the A+ Plan; the second row shows the coefficient on whether the school was ever identified as critically low performing, interacted with a post-

[26] None of the 1998 critically low-performing schools are in the school districts covered in this study.

identification dummy variable. As before, we estimate a positive relative effect of the A+ Plan on reading and math scores in "F" schools, although only the effect on math scores is statistically significant.[27] We also estimate positive effects of being on the critically low performing list on reading and math scores, and again only the effect on math scores is statistically significant. Importantly, we estimate no difference between either the math or reading test score gains made following placement on the critically low performing list and being sanctioned under the A+ Plan, suggesting that the main motivator for low-performing schools in the A+ Plan is the stigma rather than the voucher threat.

While we do not observe any schools that were on the critically low-performing schools list in 1998, one in three "F" rated schools in our districts had previously been on the critically low-performing schools list in an earlier year. We, thus, investigate whether these schools had different experiences following receipt of an "F" grade than did "F" schools that had not previously been on this list. One might expect schools that had once been on the critically low-performing schools list, and therefore labeled as either "failing" or "critically low-performing" in at least two years during a four-year window, might be particularly sensitive to the threat of vouchers, and if there were to be a voucher-threat response among the "F" schools, it would most likely be with these schools. Our results, reported in the second and fourth columns of Table 10, indicate that, in both reading and mathematics, the post-1999 interaction between "F" receipt and previous appearance on the critically low-performing schools list has a negative (or trivial positive) and statistically insignificant

---

[27] We have conducted all of these exercises with the A+ Plan comparison being "F" versus "D" schools, and the results are very similar to those presented in Table 10. We chose the "F" versus all other schools comparison here to provide the closest parallel possible to the critically low-performing schools list comparison (for which we have been unable to find statewide data on "near-miss" schools in 1996).

coefficient estimate. However, we continue to find positive, statistically significant independent effects of being on the critically low-performing list on performance in mathematics. Therefore, we do not find evidence that schools that had once been labeled as critically low-performing differentially responded to being labeled as an "F" school in 1999, relative to those who were labeled as an "F" with no previous appearance on the critically low-performing schools list.

We have also investigated whether the effect of being on the "critically low performing list" is subject to mean-reverting measurement error as well. In results not presented here, we have constructed a "pseudo critically low performing list" dummy variable indicating schools that performed poorly in 1995 – the year before the actual critically low performing list was published. We estimate a small, negative, and statistically insignificant result in reading and a small, positive, and statistically insignificant effect in math. Therefore we conclude that the effect is not driven by mean-reversion (which is similar to the grade "F" effect). Further, one might be concerned that stigma under the critically low performing list was quite different from the stigma under the A+ Plan, particularly since there was much publicity (including national media) surrounding the "F" grades in the A+ Plan. However, if the stigma associated with the A+ Plan was greater than that with the critically low performing list, then our exercise understates the relative effect of the earlier accountability system.

We can further evaluate the voucher threat versus grading stigma arguments by revisiting the subgroup analysis mentioned in the previous section. We note that grading "stigma" can really be thought of as a form of grading pressure, and while only "F"-rated schools are subject to potential voucher threat in the accountability system, all schools in Florida faced accountability pressure beginning in 1999. Because of its five levels of accountability grades, the A+ Plan provided

considerable incentive for all schools, even the highest-achieving ones, to improve the performance of the lower-achieving students. As a result, one would expect that if grading pressure were a primary determinant of school performance following the introduction of the accountability system, *all* schools would put more effort into the bottom of the distribution than into the top of the distribution. Indeed, this is exactly what we find. In results not presented in the table, we observe that schools rated "A" through "C" witnessed differential bottom-quartile and second-quartile mathematics performance gains of about 0.10σ and more modest (but still statistically significant) reading performance gains of about 0.04σ, relative to the top quartile of the distribution. The gains of the third quartile of the initial test score distribution are not statistically distinguishable from those of the top quartile in either reading or math. Therefore, all schools concentrated effort on the lower-performing students, with apparent results in the NRT examinations; "F" schools, with the additional threat of vouchers, did not differentially impact lower-performing students in a large way.

We conclude that grading pressures, rather than voucher threats, were the primary determinants of the large observed gains in mathematics (and smaller gains in reading) among lower-performing children.

V.   Conclusion

This paper presents an attempt to systematically study the effects of the threat of school vouchers and school stigma in Florida on the relative performance of "low-performing" schools. Simple estimates of the change in school test scores from the first year of the reform are consistent with the early results used by the state of Florida to claim large-scale improvements associated with the threat of voucher assignment. However, we also find that much of this estimated effect may be

due to other factors. While we estimate a small improvement in reading scores on the high-stakes test, we estimate a much smaller improvement on the nationally norm-referenced test (NRT). Further, the gains in reading scores are explained largely by changing student characteristics. We find more evidence for a positive effect on math test scores on both the low- and high-stakes tests, however, the NRT (the low-stakes test) results appear primarily limited to students in the high-stakes grade. Finally, we find some evidence that the differential improvements following the introduction of the A+ Plan by "F"-rated schools were more due to the stigma of receiving the low grade rather than the threat of vouchers.

Regardless of whether the differential effects of the A+ Plan on low-performing schools was due to voucher threats or grading stigma, the evidence presented in this paper indicates that the short-run general education effects of receiving a grade of "F" were relatively modest. We do find evidence of a modest gain in mathematics test scores in low-performing schools, even when using norm-referenced tests rather than the high-stakes FCAT-SSS examination, but the primary improvements were concentrated in the high-stakes grade. Today, the A+ Plan makes all grades from three through ten high-stakes grades, so there is reason to believe that schools have since responded by spreading effort among a larger number of grades. We have also found that while the differential gains seen in low-performing schools were modest, there may have been somewhat larger gains, especially in mathematics, across the board for lower-performing *students*. That all schools in Florida, regardless of grade, tended to focus attention on lower-performing students (and were successful in doing so) indicates that it was other aspects of the A+ Plan (such as the public reporting of school grades) rather than the threat of vouchers, that led to the improvements in student test scores in Florida.

This paper illustrates some of the desirable features of the A+ Plan in Florida, relative to many other accountability systems. By focusing attention on the outcomes of lower-performing students, the A+ Plan apparently led to test score gains in mathematics (and much smaller gains in reading) in the year following its introduction. The Florida accountability system, with five levels of grading, ensured that all schools faced performance pressure, which this paper suggests was more effective than voucher threats in improving student performance. At the same time, schools apparently concentrated on low-performing students, rather than other measurable subgroups of students. While in the first year of the program only one grade per subject faced high-stakes testing, the results of this paper indicate that the move in Florida (also seen nationally now with No Child Left Behind) to include more grades in high-stakes testing is a desirable policy move, from the perspective of ensuring test score improvements for a larger set of students.

References

Ashenfelter, Orley. "The Effect of Manpower Training on Earnings: Preliminary Results," in *Proceedings of the Twenty-Seventh Annual Winter Meeting of the Industrial Relations Research Association* (1975), pp. 252-260.

Ashenfelter, Orley. "Estimating the Effect of Training Programs on Earnings" *The Review of Economics and Statistics*, 60 (February, 1978), pp. 47-57.

Bagley, Carl, Woods, Philip, and Ron Glatter. "Barriers to School Responsiveness in the Education Quasi-Market." *School Organisation*. Vol. 16, no. 1 (1996), pp. 45-58.

Bertrand, Marianne, Duflo, Esther, and Sendhil Mullainathan. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics*, forthcoming.

Camilli, Gregory and Katrina Bulkley. ""Critique of 'An Evaluation of the Florida A-Plus Accountability and School Choice Program." *Education Policy Analysis Archives*, vol. 9 no. 7 (March, 2001).

Carnoy, Martin and Susanna Loeb. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." *Education Evaluation and Policy Analysis*, vol. 24, no. 4 (Winter 2002), pp. 305-331.

Chubb, John E., and Terry M. Moe. *Politics, Markets, and America's Schools*. The Brookings Institution (1990).

Clark, Melissa A. "Education Reform, Redistribution, and Student Achievement: Evidence from the Kentucky Education Reform Act." Princeton University mimeo, November 2002.

Cullen, Julie Berry and Randall Reback. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." University of Michigan working paper, 2003.

Figlio, David. "Testing, Crime and Punishment." University of Florida working paper, 2003.

Figlio, David and Maurice Lucas. "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review*, vol. 94, no. 3 (June 2004), pp. 591-604.

Figlio, David and Lawrence Getzler. "Accountability, Ability and Disability: Gaming the System?" National Bureau of Economic Research working paper number 9307, (November 2002).

Frey, Donald E. "Can Privatizing Education Really Improve Achievement? An Essay Review." *Economics of Education Review*. Vol. 11, no. 4 (1992), pp. 427-438.

Friedman, Milton. *Capitalism and Freedom*. University of Chicago Press (1982).

Garner, W.T., and Jane Hannaway. "Private Schools: The Client Connection." In *Family Choice in Schooling*. M. Manley-Casimir, ed. D.C. Heath (1992).

Goldhaber, Dan and Jane Hannaway. "Accountability with a Kicker: Observations on the Florida A+ Accountability Plan." Urban Institute mimeo, November 2001.

Greene, Jay P. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." Manhattan Institute for Policy Research publication, February 2001.

Greene, Jay P., Paul Peterson and Jiangtao Du. "School Choics in Milwaukee: A Randomized Experiment." In *Learning from School Choice*, P. Peterson and B. Hassel, eds. Brookings Institution (1998).

Griliches, Zvi. "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica* 45 (January 1977), pp. 1-22.

Haney, Walt. "Lake Woebeguaranteed: Misuse of Test Scores in Massachusetts, Part I." *Education Policy Analysis Archives*, vol. 10, no. 24 (May, 2002).

Haney, Walt. "The Myth of the Texas Miracle in Education." *Education Policy Analysis Archives*, vol. 8, no. 41 (August, 2000).

Hanushek, Eric A. and Margaret E. Raymond. "Improving Educational Quality: How Best to Evaluate Our Schools." Hoover Institution mimeo, June 2002.

Jacob, Brian A. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." National Bureau of Economic Research working paper, 2002.

Jacob, Brian A. And Steven D. Levitt. "Rotten Apples: An Investigation of the Prevalance and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118, no. 3 (August 2003), pp. 843-877.

Kane, Thomas J. and Douglas Staiger. "Volatility in School Test Scores: Implications of Test-based Accountability Systems." National Bureau of Economic Research Working Paper, Number 8156, April 2001.

Krueger, Alan B. and Pei Zhu. "Another Look at the New York City Voucher Experiment," *American Behavioral Scientist*, 2003 forthcoming.

Ladd, Helen C. and Elizabeth J. Glennie. "A Replication of Jay Greene's Voucher Effect Study Using North Carolina Data." In Martin Carnoy, *Do School vouchers Improve Student Performance?* Washington, DC: Economic Policy Institute, 2001.

Peterson, Paul, William Howell, Patrick Wolf and David Campbell. "School Vouchers: Results from Randomized Experiments." In *The Economics of School Choice*, C. Hoxby, ed. University of Chicago Press (2003).

Rouse, Cecilia. "Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics*, 113 no. 2 (May 1998), pp. 553-602.

Schneider, Mark. "The Role of Information in School Choice." SUNY-Stony Brook working paper (1999).

Witte, John. "Achievement Effects of the Milwaukee Voucher Program." University of Wisconsin mimeo, 1997.

Appendix
School Grading Criteria in Detail

Criteria for an "F" Grade

Schools throughout the distribution are subject to accountability pressure, and schools that earn grades of "A" or improve grades from one year to the next receive bonuses of about $90-100 per student. However, the most publicized innovation of the A+ Plan for Education was the treatment of the lowest-rated schools in the system. As intimated above, the criteria for earning a grade of "F" have changed over time since the initial implementation of the grading system. Table 2 presents the critical features in the grading system for the first three years. The initial grading system based school grades on the aggregate performance of all standard curriculum students taking the test, as well as hospital/homebound students, gifted students, language or speech impaired students, and limited English proficient students who have been in an English-as-a-Second-Language (ESOL) program for more than two years, regardless of these students' tenures in the school. Following outcry from districts that this system was unfair to schools with highly mobile populations, in 2000 this student basis for grading was changed to only include students who were in the same school in both October and February of the same academic year, thereby ensuring that only relatively stable students were included in the test pool for state accountability purposes.

In each year from 1999 through 2001, a school would earn a grade of "F" if fewer than 60 percent of students scored at level 2 (out of 5) or above in reading, fewer than 60 percent of students scored at level 2 (out of 5) or above in mathematics, and fewer than 50 percent of students scored at level 3 (out of 6) or above on the Florida Writes! writing evaluation, known from 2001 onward as the FCAT Writing examination.[28] A school could avoid the "F" label by meeting any one of these three standards in 1999; the same was true in 2000 and 2001 provided that at least 90 percent of the test-eligible students took the examination (or that the school could provide the state with a "reasonable explanation" for why fewer than 90 percent of students took the test.)

Criteria for Higher Grades

Schools that met at least one threshold mentioned above would qualify for a grade of "D," while those that met the criteria for reading, mathematics and writing received a grade of "C." To earn a higher grade, more rigorous standards must have been met. First of all, the test performance

---

[28] The state dramatically changed grading criteria in 2002. In 2002, schools received a "grade point average" equaling the sum of six factors: the percentage of students attaining level 3 or above in reading, mathematics and writing, the percentage of students making "learning gains" (defined as either remaining in level 3, 4 or 5, increasing at least one level, or experiencing scale score growth from year to year of at least "one year of learning") from 2001 to 2002 in reading and mathematics, and the percentage of students in the school's bottom quartile who showed any reading gains in scale score points from 2001 to 2002. Schools scoring below 280 on this aggregated scale earned a grade of "F" in 2002, as did schools with fewer than 90 percent taking the test without "reasonable explanation."

thresholds were higher: In order to attain a grade of "A" or "B," at least 50 percent of test-takers must have achieved at level three or above (out of five) on the FCAT-SSS reading test, at least 50 percent of test-takers must have achieved at level three or above (out of five) on the FCAT-SSS mathematics test, and at least 67 percent of test-takers must have achieved at level three or above (out of six) on the Florida Writes! examination. In addition, in order to achieve a grade of "A" or "B," the minimum ("C"-level) criteria must have been met by test-takers in each of six subgroups of students: Economically Disadvantaged, Black, White, Hispanic, Asian, and Native American. Finally, a minimum of 90 percent of standard curriculum students (including language impaired, speech impaired, gifted, hospital homebound, and limited English proficient students who have been in an ESOL program for more than two years) must have taken the examinations in order for a school to earn a grade higher than "C."

To attain a grade of "A" rather than "B" required four additional criteria: (1) the percentage of students absent more than 20 days and the percent suspended must have been below the state average (this second standard was lifted in 2000); (2) at least 95 percent of standard curriculum students must have been tested; (3) reading scores must have "substantially improved" (i.e., the fraction scoring level three or above on the FCAT reading test must have improved by two or more percentage points from the previous year, unless the fraction attaining this level is 75 percent or above already); and (4) math and writing scores must not have "substantially declined" (i.e., the fraction scoring level three or above on the FCAT mathematics test or Florida Writes! must not have fallen by five or more percentage points from the previous year.)

Table 1
Timeline of School Accountability in Florida

Year	Event
1996	First School Accountability Reports
	Establishment of List of Critically Low Performing Schools, four school rating groups (based on performance on Florida Writes! and norm-referenced reading and math tests of district's choosing)
1998	First year of recorded FCAT reading and math exams based on Sunshine State Standards, grades 4 (reading), 5 (math), 8 (both), and 10 (both)
1999	A+ Plan Enacted, five school rating groups and private school vouchers assigned in event of chronic low performance (based on performance on Florida Writes! and FCAT Sunshine State Standards reading and math tests)
2000	School grading criteria amended to exclude mobile students
	FCAT Norm-referenced examination introduced, grades 3-10
2001	FCAT Sunshine State Standards examination introduced, grades 3-10
2002	School grading criteria amended to include both levels and "value added" components in reading and mathematics

Table 2
Elementary School Criteria for Earning a Grade of "F", by Year

Year	Criteria
1999	Based on all standard curriculum students, as well as language impaired, speech impaired, gifted, hospital/homebound, and LEP students in ESOL program for more than two years.
	F earned if fewer than 60% attain level 2+ in reading, fewer than 60% attain level 2+ in math, and fewer than 50% attain level 3+ in writing.
2000	Based on all standard curriculum students, as well as language impaired, speech impaired, gifted, hospital/homebound, and LEP students in ESOL program for more than two years. Only students enrolled in same school in both October and February are counted.
	F earned if fewer than 60% attain level 2+ in reading, fewer than 60% attain level 2+ in math, and fewer than 50% attain level 3+ in writing.
	F also earned if one or two of these criteria are met, but fewer than 90% of eligible students take test without "reasonable explanation."
2001	Same requirements/conditions as in 2000.

Table 3
The Distribution of School Grades, by Year

School Grade	School Year		
	1998-1999	1999-2000	2000-2001
A	187	557	561
B	311	270	414
C	1189	1120	1079
D	594	382	300
F	78	4	0
Total	2359	2333	2354

Source: Authors' calculations from state data.

Table 4
Mean FCAT-SSS Test Scale Scores, by Grade Level/Subject and
School Grade in 1998-1999

| School Grade | Grade Level/Test Subject | | |
	4th Grade Reading	5th Grade Math	4th Grade Writing
A	318.7	330.8	3.34
	[13.3]	[10.7]	[0.24]
B	309.4	325.9	3.22
	[10.5]	[10.0]	[0.23]
C	292.4	305.9	2.98
	[12.2]	[10.6]	[0.24]
D	263.7	281.7	2.75
	[16.1]	[15.7]	[0.26]
F	244.7	265.3	2.33
	[16.5]	[15.0]	[0.17]
Missing	238.6	262.2	2.30
	[46.7]	[43.4]	[0.84]
All	287.8	303.1	2.96
	[24.1]	[22.0]	[0.33]

Notes: Standard deviations in brackets. All means weighted by the number of students taking the test.

Table 5
School-level Fixed-Effects Estimates of School's Receipt of "F" Grade on
Change in FCAT-SSS Reading and Math Test Scale Scores
Between 1999 and 2000

	Entire State	State Data for Districts with Microdata	Data from District Microdata
	4th Grade Reading		
	(1)	(2)	(3)
School Received an "F" Grade in 1999	8.469 (1.540)	9.339 (2.337)	9.075 (2.511)
R^2	0.126	0.074	0.044
Number of Observations	1562	409	411
	5th Grade Math		
	(4)	(5)	(6)
School Received an "F" Grade in 1999	8.824 (1.563)	9.459 (2.295)	9.725 (2.422)
R^2	0.101	0.072	0.039
Number of Observations	1559	409	411

Notes: Dependent variable: Change in test scores between 1999 and 2000. Standard errors in parentheses. All regressions are weighted by the number of students taking the test and all include a constant, a dummy variable indicating if the school was missing a grade in 1998-1999 and district fixed-effects.

Table 6a
Estimated Effects of School Grade on Fourth Grade FCAT-SSS Reading Test Scores:
School Fixed Effects Models for Students Present in Microdata Sample
Coefficients on Post 1999 Variables

	FCAT-SSS Normal Curve Equivalent		NRT Normal Curve Equivalent			
	Grade Level of Student		Grade Level of Student			
	4th Grade	4th Grade	4th Grade	4th Grade	5th Grade	5th Grade
	(1)	(2)	(3)	(4)	(5)	(6)
School Grade F × Post 1999	1.838	0.913	1.449	0.628	0.223	0.496
	(0.761)	(0.553)	(0.617)	(0.450)	(0.614)	(0.500)
Lagged Reading NRT Score		0.803		0.713		0.757
		(0.006)		(0.004)		(0.004)
R^2	0.309	0.691	0.295	0.689	0.304	0.711

Notes: Standard errors adjusted for school-level clustering are in parentheses. The fourth grade sample is the set of students with both FCAT-SSS and norm-referenced tests in the same year, from 1998 to 2000. The fifth grade sample is the set of students with norm-referenced tests between 1998 and 2000. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form used in the norm-referenced exams. (Note that we do not include a school-specific time trend because we only have one year of data before 1999.) Number of observations: 94,606 with FCAT-SSS and norm-referenced tests in fourth grade as well as lagged norm-referenced tests, 95,380 in fifth grade with lagged test scores. In Tables 6a and 6b only we restrict the sample in columns (1) - (4) to those with both the FCAT-SSS and the NRT. For the FCAT-SSS we first converted the student's scale score to a percentile ranking based on the 1998 distribution of test scores statewide; we then converted these scores to normal curve equivalents.

Table 6b
Estimated Effects of School Grade on Fifth Grade FCAT-SSS Math Test Scores:
School Fixed Effects Models for Students Present in Microdata Sample
Coefficients on Post 1999 Variables

	FCAT-SSS Normal Curve Equivalent		NRT Normal Curve Equivalent			
	Grade Level of Student		Grade Level of Student			
	5th Grade	5th Grade	5th Grade	5th Grade	4th Grade	4th Grade
	(1)	(2)	(3)	(4)	(5)	(6)
School Grade F × Post 1999	4.551	5.137	1.610	2.222	-1.019	-0.666
	(1.026)	(0.932)	(0.854)	(0.720)	(1.021)	(0.988)
Lagged Math NRT Score		0.706		0.736		0.659
		(0.004)		(0.004)		(0.003)
R^2	0.341	0.709	0.271	0.665	0.260	0.638

Notes: Standard errors adjusted for school-level clustering are in parentheses. The fifth grade sample is the set of students with both FCAT-SSS and norm-referenced tests in the same year, from 1998 to 2000. The fourth grade sample is the set of students with norm-referenced tests between 1998 and 2000. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form used in the norm-referenced exams. (Note that we do not include a school-specific time trend because we only have one year of data before 1999.) Number of observations: 94,606 with FCAT-SSS and norm-referenced tests in fifth grade as well as lagged norm-referenced tests, 96,727 in fourth grade with lagged test scores. In Tables 6a and 6b only we restrict the sample in columns (1) - (4) to those with both the FCAT-SSS and the NRT. For the FCAT-SSS we first converted the student's scale score to a percentile ranking based on the 1998 distribution of test scores statewide; we then converted these scores to normal curve equivalents.

Table 7
Estimated Effects of School Grade on NRT Test Scores: Student First Difference Models

	Reading Test					Math Test				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
School Grade F × Post 1999	0.235 (0.378)	0.395 (0.387)	0.312 (0.427)			0.722 (0.852)	1.226 (0.896)	0.870 (0.957)		
School Grade D × Post 1999			-0.132 (0.253)					-0.543 (0.482)		
School Grade B × Post 1999			0.336 (0.400)					1.039 (0.761)		
School Grade A × Post 1999			-0.394 (0.522)					-1.697 (0.783)		
Grade F × Post - Grade D × Post			0.443 (0.394)					1.412 (0.904)		
Pseudo School Grade F × Post 1998				-0.941 (0.283)					-0.630 (0.517)	
Low SES School × Post 1999					-0.139 (0.356)					-1.856 (0.790)
Includes School-specific Time Trend?	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes
R^2	0.049	0.049	0.049	0.037	0.049	0.027	0.027	0.027	0.022	0.028

Notes: Standard errors adjusted for school-level clustering are in parentheses. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form

46

used. All specifications except (1) and (6) control for school-specific trends in NRT test scores generated using data from 1995 through 1998. There are 287,444 observations for reading and 286,729 observations for math. The pseudo-grade analyses include only data through 1998-99. Low SES Schools are the 7% of schools with the largest fraction of free lunch-eligible students.

Table 8
Student First Difference Models of Estimated Effects of School Grade on NRT Test Scores:
Different Comparison Groups

Comparison Group	Reading Test	Math Test
Schools that failed in reading in 1999	0.388 (0.394)	1.343 (0.900)
Schools that failed in math in 1999	0.536 (0.386)	1.510 (0.902)
Schools that failed reading plus one other test in 1999	0.431 (0.396)	1.500 (0.905)
Schools that failed math plus one other test in 1999	0.392 (0.392)	1.453 (0.914)
"F" vs. "D" in model controlling for percent of students failing reading (or math) x Post 1999	0.568 (0.410)	1.992 (0.952)

Notes: Standard errors adjusted for school-level clustering are in parentheses. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form used. There are 287,444 observations for reading and 286,729 observations for math. All specifications control for school-specific trends in NRT test scores generated using data from 1995 through 1998.

Table 9
Effect of Receiving an "F" on Reading and Math NRT Score:
Student First Difference Models Comparing "F" Schools to "D" Schools for
Different Student Subgroups

	Reading Test	Math Test
Top quartile in initial test scores	-0.324 (1.451)	-0.322 (1.697)
Second quartile	0.543 (0.643)	1.081 (1.312)
Third quartile	0.480 (0.437)	1.119 (0.903)
Bottom quartile	0.617 (0.490)	0.834 (0.746)
p-value of difference (second vs. top)	0.40	0.17
p-value of difference (third vs. top)	0.34	0.06
p-value of difference (bottom vs. top)	0.17	0.09
White students	-0.021 (0.806)	1.813 (1.426)
Black students	0.711 (0.470)	1.543 (0.948)
Hispanic students	0.129 (0.619)	1.114 (1.617)
p-value of difference (Black vs. White)	0.39	0.82
p-value of difference (Hispanic vs. White)	0.86	0.73
Limited English proficient students	-0.030 (0.692)	1.043 (1.647)
Other students	0.721 (0.439)	1.483 (0.905)
p-value of difference	0.24	0.78
Low-income students	0.537 (0.383)	1.418 (0.898)

|		
Higher-income students	-0.826	2.683
	(1.360)	(1.743)
p-value of difference	0.29	0.40

Notes: Each panel represents a separate regression. Standard errors adjusted for school-level clustering are in parentheses. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form used. There are observations 287,444 for reading and 286,729 observations for math. All specifications control for school-specific trends in NRT test scores generated using data from 1995 through 1998.

Table 10
Effect of Receiving an "F" on Reading and Math NRT Score:
Student First Difference Models Comparing "F" Schools with
Critically Low Performing Schools

	Reading Test		Math Test	
	(1)	(2)	(3)	(4)
"F" grade × Post-1999	0.382	0.567	1.141	1.203
	(0.385)	(0.502)	(0.871)	(1.091)
Critically low performing school × Post-identification	0.314	0.252	2.141	1.786
	(0.259)	(0.281)	(0.610)	(0.633)
"F" grade × Critically low performing school × Post-1999		-0.574		0.081
		(0.729)		(1.721)

Notes: Standard errors adjusted for school-level clustering are in parentheses. All regressions include school and year fixed effects and controls for race, ethnicity, standard curriculum, limited English proficiency, sex and socio-economic status, as well as for the test form used. There are 287,444 observations for reading and 286,729 observations for math. All specifications control for school-specific trends in NRT test scores generated using data from 1995 through 1998.

Figure 1a. Mean Change in Fourth Grade Reading and Fifth Grade Math FCAT-SSS Test Scores from 1999 to 2000, by School Grade in 1999
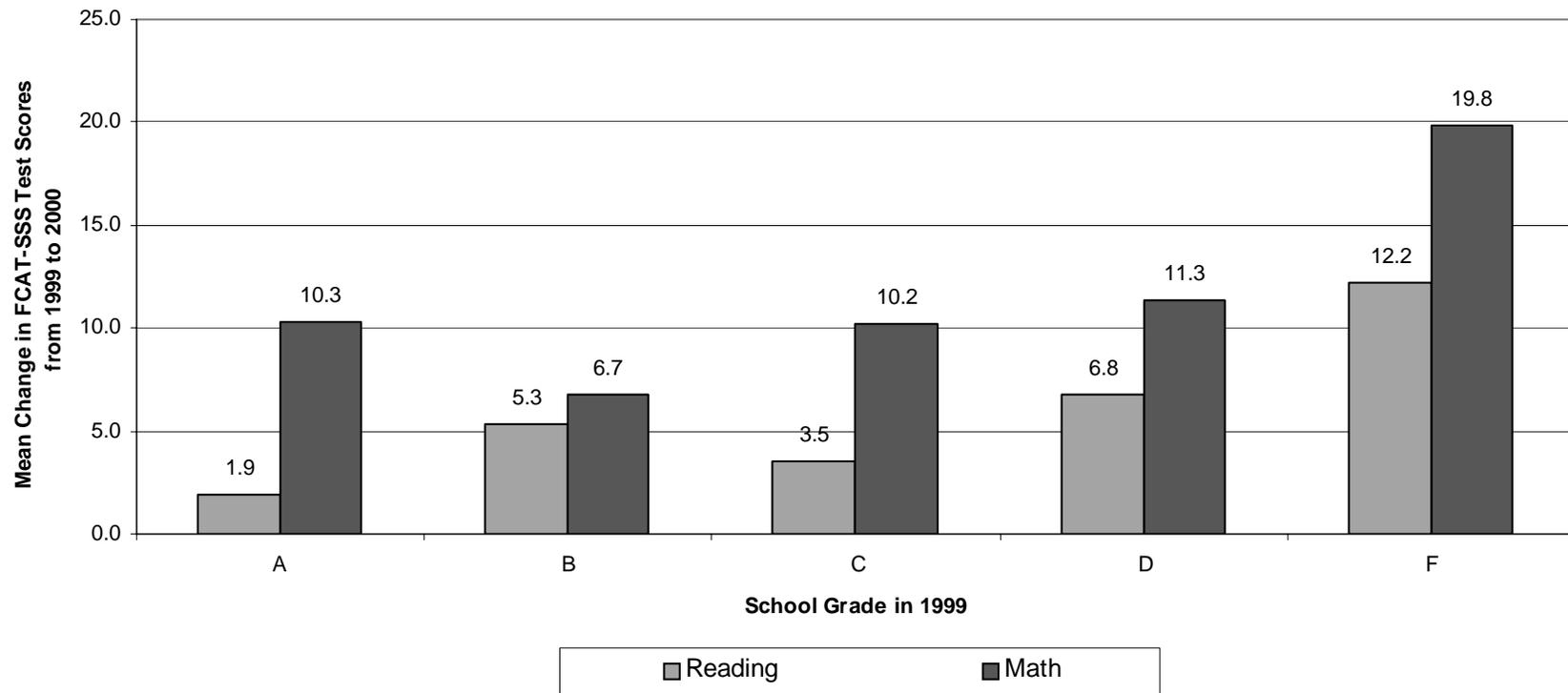
Figure 1b. Mean Change in Fourth Grade FCAT-SSS Writing Test Scores from 1999 to 2000, by School Grade in 1999