Do accountability and voucher threats improve low-performing schools?

David N. Figlio\textsuperscript{a}, Cecilia Elena Rouse\textsuperscript{b,*}

\textsuperscript{a} University of Florida and NBER, USA
\textsuperscript{b} Princeton University and NBER, USA

Received 26 August 2004; received in revised form 19 August 2005; accepted 21 August 2005
Available online 23 September 2005

Abstract

We study the effects of the threat of vouchers and stigma in Florida on the performance of “low-performing” schools. Estimates of the change in raw test scores from the first year of the reform are consistent with the early results which claimed large improvements associated with the threat of vouchers. However, we also find that much of this estimated effect may be due to other factors. The relative gains in reading are largely explained by changing student characteristics and the gains in math—though larger—appear limited to the high-stakes grade. We also find some evidence that these improvements were due more to the stigma of receiving the low grade rather than the threat of vouchers.

Keywords: School accountability; School voucher threat; Florida A+ Plan; No Child Left Behind

1. 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 \textit{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. Under \textit{No Child Left Behind}, students attending “failing” schools are allowed to transfer to a more successful school in the

* Corresponding author. Tel.: +1 609 258 4042; fax: +1 609 258 2907.
E-mail address: rouse@princeton.edu (C.E. Rouse).

0047-2727/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2005.08.005
same district. Florida’s accountability system—implemented in 1999—preceded this move to enlist both stigma and competition to spur improvement in the lowest performing schools. Florida’s system is unique in that it incorporated a school voucher program within its accountability system. Florida’s voucher program, called the Florida Opportunity Scholarship Program, is for students who attend schools that are persistently labeled as failures.\footnote{Currently Florida has two other voucher programs as well: a tax credit for firms that privately fund school vouchers and the McKay Scholarship for students with exceptionalities. Neither of these programs existed at the introduction of the A+ Plan.}

An accountability-tied voucher 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.\footnote{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.} 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, 1982; 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.

But it is not just the threat of vouchers that could inspire change under Florida’s system. By grading schools from “A” through “F” and explicitly labeling schools as “failing,” Florida’s accountability system also uses social stigma as a motivating factor. 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. For example, Figlio and Lucas (2004) find that school grades were apparently highly capitalized into housing values, especially at the program’s beginning, suggesting that schools likely faced pressure from the community. Further, 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.

Since the introduction of Florida’s accountability system, Florida schools across the distribution have improved dramatically on the standards-based Florida Comprehensive Assessment Test (FCAT), indicating that along at least one dimension there has been systemwide improvements in test scores.\footnote{Unfortunately, it is impossible to currently compare Florida’s progress against a national benchmark because Florida did not participate in all rounds of the National Assessment of Educational Progress (NAEP). Carnoy and Loeb (2002) and Hanushek and Raymond (2002), however, show that states with high-stakes accountability systems experienced relatively large improvements on their scores on the NAEP.} This paper does not seek to estimate the magnitude of the systemic effects of the accountability system; rather, this paper attempts to systematically estimate the collective effects of voucher threats and stigma on schools graded “F” in Florida, and to measure the likely impacts of the voucher threat component of the policy.

In an early study of Florida’s accountability system, Greene (2001) analyzed school aggregate data and found 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 reflected the threat of vouchers or
other elements of the accountability system such as grading stigma. There are several alternative hypotheses that might explain these 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 1 year are not likely to be normally distributed around the initial score; rather the “failing” schools are likely to experience larger than average gains the following year. Further, in order to receive an “F,” a school had to have low mean test scores in only 1 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. 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, 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).4,5

It is impossible with state data to study the introduction of the accountability and voucher system, because the state began annual testing of all students in 2001. However, we have acquired student-level demographic and reading and mathematics norm-referenced test score (e.g., CTBS, ITBS, or SAT-8) data from a subset of school districts in Florida from 1995 to 2000 to address the question of whether the threat of vouchers and stigma have an immediate effect on public school performance.6 While our data do not cover all districts, we observe nearly half of all “D”- and “F”-rated schools in the state, and we can follow individual students across schools and districts beginning in 1995, well before the introduction of the grading system. We can therefore estimate the effects of “F” receipt in a model that overcomes the concerns mentioned above. We also attempt to determine if any observed gains were “real” or due to other behaviors by schools, and we study whether vouchers or stigma played the larger role in boosting the scores of low-performing schools.

4 Goldhaber and Hannaway (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.

5 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 (2005), Cullen and Reback (2003) 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; Figlio and Winicki (2005) demonstrate that schools strategically alter the nutritional content of meals on testing days in an apparent attempt to boost test scores; and Figlio (in press) finds that students are more likely to be suspended during the testing cycle under an accountability system. These potential consequences are considerably more likely when accountability systems assess schools based on status test scores, as was the case in Florida during the time period covered in this analysis, rather than based on improvements in test scores, as is the current practice in Florida.

6 Our data end in 2000 so we cannot investigate the longer-term effects of voucher threats on public school performance.
We find that the large unadjusted test score gains in low-performing schools tend to be concentrated in the high-stakes grades for the subject, and in part on the high-stakes examination. In addition, we suspect that the relative gains achieved by “F” schools were driven mainly by the grading stigma component of the accountability system, rather than the threat of vouchers.\footnote{One might be concerned that educators in Florida anticipated vouchers with the publication of the critically low performing schools list—the previous accountability system that enlisted grading stigma but not vouchers. However, this was highly unlikely primarily because the Democratic governor at the time, Lawton Chiles, opposed vouchers.} This makes sense in the context of Florida, since school districts, not the individual schools, bear the financial risk of vouchers, and districts—particularly those where low-performing schools are located—tend to be very large. Grading stigma, on the other hand, affects schools directly so one should expect that they would be highly responsive to grading pressures. In states with smaller school districts, voucher threats may be larger.\footnote{This paper does not address the question of whether recipients of school vouchers benefit as a result of receiving them, but a number of recent studies of publicly- and privately-funded vouchers estimate their impacts on the performance of students who are offered (or take up) the voucher. Greene et al. (1998), Rouse (1998) and Witte (1997) examine the Milwaukee voucher program and find mixed evidence of participant test score gains. Peterson et al. (2003) describe the results of randomized voucher experiments in Dayton, New York and Washington. After 3 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.} That said, we do estimate gains in mathematics on the order of $0.06 \sigma - 0.1 \sigma$ which, though not large by education standards, are noteworthy because the additional expenditures associated with the accountability system are quite low.

2. Background on the A+ Plan for education

2.1. General background on the A+ Plan

The A+ Plan for Education, implemented by Governor Jeb Bush in 1999, requires annual curriculum-based testing of all students in grades 3 through 10, 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 financial and technical assistance) for low-performing schools. The most famous and publicized provision of the A+ Plan, however, are the private school vouchers, called “Opportunity Scholarships,” for students attending (or slated to attend) chronically failing schools—those receiving a grade of “F” in 2 years out of four, including the most recent year.

School grading began in May 1999, immediately following passage of the A+ Plan into law. Schools received grades based on the fraction of students meeting performance thresholds in reading, writing and mathematics in tested grades. Schools where fewer than 60% met the threshold in reading, fewer than 60% met the threshold in mathematics, and fewer than 50% met the threshold in writing received grades of “F.” Schools where one or two of these standards were met received grades of “D.” Schools where all three standards were met received grades of “C,” and those meeting higher standards received grades of “A” or “B.”\footnote{A more detailed description of school grading in Florida is available in Figlio and Rouse (2005).} In the first year, school grades were based on all standard curriculum students, as well as language impaired, speech impaired, gifted, hospital/homebound, and limited-English proficient students in ESOL program.
for more than 2 years. Beginning in 2000, the grading system changed such that only students enrolled in the same school in both October and February were counted.

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. Under Education Commissioner Frank Brogan, who later became Bush’s Lieutenant Governor, 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. 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. At roughly the same time, the state began administering the new curriculum standards-based Florida Comprehensive Assessment Test (FCAT), given in grades 4, 8, and 10 in reading and 5, 8, and 10 in mathematics, but these scores were not used for state-level school accountability purposes.

For the first 3 years of the new A+ accountability program, high-stakes testing was limited to the subjects and grades that had been the focus of the FCAT—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. 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 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 3 through 10, 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 somewhat 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

---

Table 1
The distribution of school grades, by year

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>187</td>
<td>557</td>
<td>561</td>
</tr>
<tr>
<td>B</td>
<td>311</td>
<td>270</td>
<td>414</td>
</tr>
<tr>
<td>C</td>
<td>1189</td>
<td>1120</td>
<td>1079</td>
</tr>
<tr>
<td>D</td>
<td>594</td>
<td>382</td>
<td>300</td>
</tr>
<tr>
<td>F</td>
<td>78</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>2359</td>
<td>2333</td>
<td>2354</td>
</tr>
</tbody>
</table>

Source: Authors’ calculations from state data.

---

10 Beginning in the 2001–2002 school year, testing was broadened to all grade levels from 3 to 10.
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.

2.2. Evidence of test score improvements

Table 1 shows the distribution of school grades for the first 4 years of the A+ Plan. In 1999, 78 schools received an “F” grade—the students in two of these schools became eligible for

![Graph a](image1)

![Graph b](image2)

Fig. 1. (a) Mean change in fourth grade reading and fifth grade math FCAT-SSS test scores from 1999 to 2000, by school grade in 1999. (b) Mean change in fourth grade FCAT-SSS writing test scores from 1999 to 2000, by school grade 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 1 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 Fig. 1a and b 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 (Fig. 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 2 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”- and “D”-

<table>
<thead>
<tr>
<th>School grade</th>
<th>4th grade reading</th>
<th>5th grade math</th>
<th>4th grade writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>318.7</td>
<td>330.8</td>
<td>3.34</td>
</tr>
<tr>
<td></td>
<td>[13.3]</td>
<td>[10.7]</td>
<td>[0.24]</td>
</tr>
<tr>
<td>B</td>
<td>309.4</td>
<td>325.9</td>
<td>3.22</td>
</tr>
<tr>
<td></td>
<td>[10.5]</td>
<td>[10.0]</td>
<td>[0.23]</td>
</tr>
<tr>
<td>C</td>
<td>292.4</td>
<td>305.9</td>
<td>2.98</td>
</tr>
<tr>
<td></td>
<td>[12.2]</td>
<td>[10.6]</td>
<td>[0.24]</td>
</tr>
<tr>
<td>D</td>
<td>263.7</td>
<td>281.7</td>
<td>2.75</td>
</tr>
<tr>
<td></td>
<td>[16.1]</td>
<td>[15.7]</td>
<td>[0.26]</td>
</tr>
<tr>
<td>F</td>
<td>244.7</td>
<td>265.3</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>[16.5]</td>
<td>[15.0]</td>
<td>[0.17]</td>
</tr>
<tr>
<td>Missing</td>
<td>238.6</td>
<td>262.2</td>
<td>2.30</td>
</tr>
<tr>
<td></td>
<td>[46.7]</td>
<td>[43.4]</td>
<td>[0.84]</td>
</tr>
<tr>
<td>All</td>
<td>287.8</td>
<td>303.1</td>
<td>2.96</td>
</tr>
<tr>
<td></td>
<td>[24.1]</td>
<td>[22.0]</td>
<td>[0.33]</td>
</tr>
</tbody>
</table>

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

11 Statewide, 3% of elementary schools received grades of “F” in 1999, while 21% received grades of “D.” Low-rated schools are over-represented in our microdata sample, with 8% receiving grades of “F” and 39% receiving grades of “D.” This is desirable because our analysis explicitly concentrates on the lowest-rated schools.
rated schools were about 0.3 of a standard deviation.\footnote{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”- and “D”-rated schools in writing was about 0.4 of a standard deviation.} 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 given February 25, 2003 to the Hoover Institution’s Board of Directors, Governor Bush claimed 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.”\footnote{Downloaded from http://www-hoover.stanford.edu/homepage/news/022503.html.} 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 reflected the threat of vouchers or other elements of the accountability system such as grading stigma.

3. Empirical framework and data

3.1. 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 to 2000 to provide evidence on the representativeness of the data from our subset of districts using a difference-in-differences framework. We estimate

\[ T_{st} - T_{st-1} = a + bF_s + (e_{st} - e_{st-1}), \]  

where \( T_s \) 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. Since 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.

Our main analysis uses student-level test scores from a subset of districts. We use these data to estimate models such as

\[ T_{ist} - T_{ist-1} = \alpha + \text{YEAR}_t \lambda + \beta (F_{is} \times \text{POST}_t) + \delta (\text{GRADE}_{ist}) + \phi_s + \mu_i t + \epsilon_{ist} \]  

where \( T_{ist} \) is student \( i \)’s test score in school \( s \) in year \( t \), \( F_{is} \) is a dummy variable indicating whether or not the school attended by student \( i \) received a failing grade of “F” in 1999. \( \text{YEAR}_t \) is a vector of year effects, \( \text{POST}_t \) 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), \( \text{GRADE}_{ist} \) is a
vector of dummy variables indicating the student’s grade in school, \( \phi_s \) is a vector of school fixed effects, \( \mu_s \) are school-level time trends predicted from the pre-1999 data, and \( e_{ist} \) is a normally distributed error term. Again, the key parameter of interest is \( \beta \)—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 understate 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.\(^{14}\)

Because we have multiple observations for each school, we adjust our standard errors for clustering at the school level. Bertrand et al. (2004) indicate that these standard 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.

3.2. Data

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, calculated using the entire sample, rather than the scaled score itself.\(^{15}\) 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. Moreover, unlike the relationship between SSS and NRT scores which weakened somewhat with the introduction of the A+ Plan, the correlation between types of NRT tests has remained constant over 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 a subsequent table (Table 4), we report the results of model specifications in which we control for lagged test scores. The change in the coefficient on a lagged test score is never more than 1% when we systematically distinguish between students who took the same test in successive years vs. 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 1 year to the next (unless a student fails the grade

---

\(^{14}\) 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.

\(^{15}\) We have conducted the analysis using national percentile rankings with similar results.
and is forced to retake it) and because it 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.16 Because we collected the data directly from the districts, they required some “cleaning” to deal with errors in student identification numbers.17 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 test scores, representing 182,135 students.

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 student characteristics we analyze these data primarily to assess the likely representativeness of our sample of data using individual student-level records.

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 Eq. (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. Table 3 shows these results. The top panel 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 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 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 data 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 suggests that low-rated schools in these districts are likely representative of those in the state as a whole, despite the fact that lower-rated schools are more heavily represented in the districts for which we have data.

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 0.14σ18 and 0.19σ (which are respectably large for education interventions).

16 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.

17 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 different districts in consecutive years, 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 4016 observations.

18 Note that 0.14σ, for example, is the coefficient estimate divided by the test standard deviation.
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.

4. Results

4.1. Estimated effects on the high-stakes test

We begin by demonstrating the importance of observing multiple rounds of test scores for each individual student. Table 4 shows estimates of a version of Eq. (2).\textsuperscript{19} The specification shown here is similar to that in Table 3, except that we use the microdata and include school fixed effects to control for time-invariant characteristics of the schools. Each column contains two rows—one that controls for lagged scores and one that does not.\textsuperscript{20} In the first row of columns (1) and (4)—that do not control for the prior test score—one observes large and statistically significant estimated effects of receiving a grade of “F” on both reading (0.09σ) and mathematics (0.23σ) scores on the high-stakes FCAT-SSS exam.

\textsuperscript{19} In our main specifications we estimate student first difference models, but estimate lagged test score models here for ease of comparison. The results in question are virtually identical whether we estimate first difference models or lagged dependent variable models.

\textsuperscript{20} In this table we restrict the sample to students who took 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 1200 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 Table 4 so that the results can be directly compared across the columns.

<table>
<thead>
<tr>
<th></th>
<th>Entire state</th>
<th>State data for districts with microdata</th>
<th>Data from district microdata</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4th grade reading</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School received an “F” grade in 1999</td>
<td>8.469</td>
<td>9.339</td>
<td>9.075</td>
</tr>
<tr>
<td></td>
<td>(1.540)</td>
<td>(2.337)</td>
<td>(2.511)</td>
</tr>
<tr>
<td>R\textsuperscript{2}</td>
<td>0.126</td>
<td>0.074</td>
<td>0.044</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1562</td>
<td>409</td>
<td>411</td>
</tr>
<tr>
<td><strong>5th grade math</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School received an “F” grade in 1999</td>
<td>8.824</td>
<td>9.459</td>
<td>9.725</td>
</tr>
<tr>
<td></td>
<td>(1.563)</td>
<td>(2.295)</td>
<td>(2.422)</td>
</tr>
<tr>
<td>R\textsuperscript{2}</td>
<td>0.101</td>
<td>0.072</td>
<td>0.039</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1559</td>
<td>409</td>
<td>411</td>
</tr>
</tbody>
</table>

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.
However, this improvement may overstate the differential effect of the A+ Plan on “F” schools by not fully accounting for changing student characteristics. Although we included student demographics, socio-economic status and limited English proficiency status in the first row of each column, we did not include a measure of student “ability” or previous knowledge. As a result, the estimates may be biased 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. For example, schools rated “F” in 1999 tended to have more mobile populations than did schools rated “D,” with average mobility rates 10% higher than “D” schools and more than 20% higher than in higher-rated schools, suggesting that fully controlling for changing student characteristics could be very important. Thus, in the second row of each column we use the previous year’s NRT score (as the FCAT-SSS is not administered in consecutive grades) to control for student “ability” or previous knowledge. Note that the effect drops by about half in the case of reading, not at all in the case of mathematics, although the coefficients are still significant at the 10% level. Thus, we find there was a small increase in reading achievement on the high-stakes test in the high-stakes grade, and a larger, educationally significant, increase in mathematics achievement on the high-stakes test in the high-stakes grade.\(^{21}\)

\(^{21}\) We have also investigated whether there were differential effects of an “F” grade for students who stayed in a school for two consecutive years vs. those who were new to the school in the current year. We found that the estimated effects on reading and mathematics are statistically indistinguishable between these two groups, once we controlled for prior test scores.
4.2. Estimated effects on the low-stakes test

We next consider whether comparable gains are observed with regard to the low-stakes norm-referenced test. In columns (2) and (5) we repeat the same analysis as before, but consider the low-stakes test in the high-stakes grade (4th grade for reading, and 5th grade for mathematics). In these models, we find results that remain statistically significant, but are smaller than those found with regard to the high-stakes test. In the case of mathematics, the estimated effect sizes are about half of those found for the high-stakes test, and in the case of reading, the estimated effects are only slightly smaller. However, in the case of reading, once we control for the prior year’s NRT test score, the coefficient estimate is no longer statistically significant.\(^{22,23}\) Although the coefficient estimate (controlling for the prior year’s test score) is not statistically distinct from the previous estimate (that does not control for the prior year’s test score), the changes in the magnitudes of the coefficients are 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. There exists additional evidence that some teaching to the test existed: In the year following the introduction of the A+ Plan, the correlation between NRT scores and SSS scores fell somewhat in the districts studied in the analysis, from just over 0.8 in 1998 to closer to 0.7 in 2000, indicating that more attention was paid following the A+ Plan on material emphasized on the high-stakes SSS test. That said, these efforts nevertheless had some spillover to the broader NRT as well, at least in mathematics.\(^{24}\) It is also important to note that 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.

Given that there exists evidence that educators in “F” schools may have taught to the high-stakes test, it is important to gauge whether there existed strategic cross-grade resource allocation as well. We therefore estimate the effects of receipt of an “F” grade on low-stakes grades. In columns (3) and (6) we report that 5th graders experienced no differential increase in reading scores in “F” schools, and 4th graders experienced no differential increase in mathematics scores (and indeed, the point estimate is actually negative, though insignificant). Therefore, the evidence suggests that schools focused additional energy on the high-stakes test in the high-stakes grade, but not at the expense of the other grades.

4.3. Alternative specifications

In Table 5 we present the results of several important alternative specifications of the model. In these models, we expand the time horizon to include the entire data period from 1995 to 2000, so that we may take into account school-specific trends (generated from data from 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 have been on an upward or downward

---

\(^{22}\) 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 1%, regardless of the NRT test employed.

\(^{23}\) In a prior version of this paper (Figlio and Rouse, 2005), we also showed that there were modest positive estimated effects on reading test scores for the population of students who were not limited English proficient.

\(^{24}\) 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.
trajectory over time. In addition, we combine all elementary school grades into the same model (which has the result of combining the positive effects of a grade of \( b \) receipt in the high-stakes grades with the zero effects in the low-stakes grades) and estimate student first difference models. We observe (first row, Table 5) that there is no apparent effect of a grade of \( b \) on overall reading performance - the effect size is only 0.02 \( r \) which is quite small and statistically insignificant. In contrast, we estimate an educationally modest (though statistically insignificant at conventional levels) positive relationship between an \( b \) grade and overall mathematics performance.\(^{25}\)

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 \( b \) schools as due to the sanctions in the A+ Plan, but rather to the schools’ obtaining unusually low test scores in 1999. To test this possibility, we created “pseudo” \( b \) 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 “\( b \)” grades in 1998. These results are presented in the second row of Table 5. 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

\[^{25}\] In a previous version of this paper (Figlio and Rouse, 2005), we estimate separate effects of receiving an “\( b \)” grade for students with different initial achievement levels. We find that the reading and mathematics gains for students are concentrated in the bottom three quartiles of the initial performance distribution, and that students in the top quartile did not apparently gain in either reading or mathematics following their school’s receipt of an “\( b \)” grade. Such results imply that schools may have concentrated their efforts in improving the performance of students closer to, or below, the performance threshold for passing.
schools that failed in 1999 may have been expected to have even lower scores in 2000 if not for the accountability system.

In the next five rows of Table 5 we further examine the effect of the comparison group of schools on our estimated effects of the A+ Plan. To identify schools that are quite similar to the voucher-threatened (“F”-rated) schools, we first compared them to “D”-rated schools (third row) and then identified schools that failed on one or two subjects on the FCAT, but not all three (as did the “F”-schools). Thus in the fourth and fifth rows we compare the “F” schools to those that failed on either math or reading; the sixth row compares them to schools that failed in reading and either math or writing, and the seventh 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 5. We observe somewhat stronger positive evidence—albeit still modest in magnitude—of a differential response by “F” schools vis-a-vis “D” schools.

4.4. 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.

We observe estimated effects of the two accountability systems that are statistically indistinguishable from one another. In a model that includes both the “F” grade and previous critically low-performing identification, we estimate that in mathematics, the receipt of an “F” is associated with a differential gain of 1.141 (with a standard error of 0.871) as compared with 2.141 (0.610) for being called critically low-performing. In the case of reading, we estimate that the receipt of an “F” is associated with a differential gain of 0.382 (0.385) as compared with 0.314 (0.259) for being called critically low-performing. In other results reported in Figlio and Rouse (2005), we also 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. These results suggest that grading pressures, rather than voucher threats, were the

26 None of the 1998 critically low-performing schools are in the school districts covered in this study.
primary determinants of the large observed gains in mathematics (and smaller gains in reading) among lower-performing children.\textsuperscript{27}

5. Conclusion

This paper attempts 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 are explained largely by students’ prior test scores. We find more evidence of a positive effect on math test scores, however the gains 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 were due to voucher threats or grading stigma, the evidence presented in this paper indicates noteworthy short-run general education effects of receiving a grade of “F” in mathematics. We estimate gains of approximately 0.2\(\sigma\) on the high-stakes test in the high-stakes grade, and gains that are approximately 0.1\(\sigma\) once other grades are included. Further, the effects were about 0.06\(\sigma\) on the lower-stakes test (for all grades). While the effects are not as large as those estimated for other interventions, such as class size reduction which had an effect of 0.2\(\sigma\) in the Tennessee STAR experiment (Krueger, 1999), it is not nearly as expensive either. As such, the A+ Plan may be a cost-effective way to improve student achievement in some subjects. Further, the result that the primary test score improvements were concentrated in the high-stakes grades suggests that the move in Florida (also seen nationally now with No Child Left Behind) to include more grades in high-stakes testing is desirable, from the perspective of ensuring test score improvements for a larger set of students.

Acknowledgement

We thank Orley Ashenfelter, Duncan Chaplin, Dan Goldhaber, Jane Hannaway, Jeff Kling, Sunny Ladd, Mel Lucas, and Ken Mease for helpful conversations, and seminar participants at the National Bureau of Economic Research, New York University, Northwestern University, Ohio State University, Princeton University, Syracuse University, the University of British

\textsuperscript{27} 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.
Columbia, the University of California at Davis, the University of Florida, the University of Oregon, and the University of Toronto, as well as two anonymous referees, for insightful comments. We also thank Nina Badgaiyan, Jennifer Graff, Radha Iyengar, and Alice Savage for expert research assistance; the Annie E. Casey, Smith Richardson, and Spencer Foundations, the Atlantic Philanthropic Services Corporation, and the National Institutes for Child Health and Human Development for financial support; and the participating school districts for historical data. Any errors in fact or interpretation are ours.

References