BGPE Discussion Paper

No. 135

What triggers school improvement? Evidence from a court induced change in Florida's A+ accountability plan

Benedikt Siegler

April 2013

ISSN 1863-5733

Editor: Prof. Regina T. Riphahn, Ph.D.
Friedrich-Alexander-University Erlangen-Nuremberg

B G P E Discussion Paper Series

What triggers school improvement? Evidence from a court induced change in Florida's A+ accountability plan*

Benedikt Siegler

April 4, 2013

Abstract

In 2006, the Florida Supreme Court ended the state funded voucher program which allowed students of repeatedly failing public schools to transfer to a private school. This gives us the unique opportunity to evaluate the incentive character of a private school competition threat in school accountability systems. Using administrative student-level data from Florida and a difference-in-discontinuities approach, I analyze whether this reduction in sanction threats led failing schools to lower their effort in raising educational performance. Results indicate that the termination of the voucher program did not attenuate the overall incentive effect of the sanction regime. This leads to conclude that the public school choice option, which remained unaffected by the court's decision, is a sufficient deterrent. This finding has important policy implications.

JEL Classification: H75, I20, I28
Keywords: School accountability; Sanction threats; School choice; Florida public schools; Difference-in-Discontinuities

Benedikt Siegler
Ifo Institute
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1240
siegler@ifo.de

*I thank Paul E. Peterson for suggesting this research project. I also thank Antonio M. Wendland, Matt Chingos, Marty West and Sam Barrows for helpful comments and valuable data assistance. I am particularly grateful to Hanley Chiang for sharing his Stata program on calculating the optimal bandwidth. Financial support from the Bavarian Graduate Program in Economics (BGPE) is gratefully acknowledged. This paper was written while the author was visiting the Program on Education Policy and Governance (PEPG) at Harvard University from 9/2012 - 3/2013. I also thank the PEPG staff for their support. All remaining errors are my own.

1 Introduction

As school accountability systems are becoming more and more widespread throughout the world, it is important to know which kind of incentives are needed to stimulate a school's academic improvement. This paper focuses on the incentives needed to promote improvement of schools at the very low end of the achievement distribution. In most systems, schools that fail to perform above a minimum proficiency level face sanctions such as school competition and ultimately financial cutbacks. In the United States, school accountability was implemented nationwide with the federal No Child Left Behind (NCLB) act of 2001. But even prior to NCLB, many states have had their own accountability programs. Florida's A+ plan is widely regarded as one of the most elaborate and comprehensive school accountability systems in place. A long literature of previous research on A+ has usually found large positive effects on school improvement (e.g. Chiang, 2009; West and Peterson, 2006; Rouse et al., 2007)[1]. But it remained unclear, which specific aspect of the incentive regime was responsible for these results. This paper tries to fill this gap.

The incentive regime of A+ consists of two stages: When a school fails to meet the proficiency requirements for the first time, it is given a letter grade "F". Apart from the stigma effect, this does not have immediate further consequences. However, when the school fails a second time within the following three years, students of that school are given the opportunity to transfer to a better performing public school, either in the same or an adjacent school district. The potential outflow of students ultimately results in lower public funding. Until 2006, students were also offered the opportunity to obtain a state funded voucher to attend a private school of their choice, which right from the start led to a heated debate about its lawfulness. Proponents of the voucher option regarded it the ultimate and most central aspect of the incentive regime. They

[1]Chiang (2009) examines the threat effect using the 2002 school grades. His results show a strong positive effect from being put under sanction threats on both math and reading results in 2003.

argued that the threat of losing students to the private school sector would unfold an incomparable incentive for failing schools to raise performance. Critics attacked the voucher option, because it used taxpayer money to fund private schools. In 2006, the voucher option was eventually struck down by the Florida Supreme Court. However, all other components of the incentive regime remained in effect. This circumstance gives us the unique opportunity to evaluate the importance of the private school voucher threat in the incentive regime of A+.

For this purpose, I use a difference-in-discontinuities design to compare the effect of sanction threats prior and post the 2006 voucher option termination. This approach enables me to use regression discontinuity (RD) analysis to estimate the effect of sanction threats on subsequent student achievement in the academic years 2002-03 (with active voucher option) and 2007-08 (without voucher option), and to compare the magnitude of these effects in a single regression framework. Results suggest that the termination of the voucher option did by no means reduce the positive effect of sanction threats on schools' behavior to improve performance. I conduct several robustness and sensitivity checks, such as dropping a potential outlier from the sample or using low-stakes testing scores as an alternative outcome variable. Results appear robust to these modifications.

This finding suggests that the core driving factor behind the large positive effects produced by the incentive regime of the A+ plan might in fact be the extent to which the public school choice provision in Florida is implemented. Students of failing schools are allowed to transfer to a better performing public school, even beyond their own school district. Under NCLB, for instance, this option is limited to the same school district. The finding that school improvement is closely related to the degree of school competition is also found in a related study by Figlio and Hart (2010) who investigate Florida's Tax Credit Scholarship Program. However, I cannot rule out the possibility that at least part of the effect might be caused by the stigma of receiving the lowest

performance grade as opposed to the threat of competition. Several studies have tried to gauge the impact of stigma by analyzing accountability systems that lack the threat of competition. But while some studies report positive effects of stigma (Ladd and Glennie, 2001; Figlio and Rouse, 2006), others find the opposite (Chakrabarti, 2013; 2008).

The remainder of this paper is organized as follows. Section 2 describes the Florida A+ accountability system and the changes it underwent between 2002 and 2008 in more detail. Section 3 introduces the data and empirical framework used in this study. Results are presented in section 4, while section 5 describes the robustness checks. Section 6 concludes.

2 Florida's A+ accountability system

At the heart of Florida's A+ accountability system lays the statewide annual testing of students in grade 3 through 10 in various subjects. These test scores are the basis for calculating a schools' performance grade of A, B, C, D, or F (highest to lowest). Every school is assessed in several performance categories which measure student achievement and student learning gains. Initially, accountability testing comprised three subjects: mathematics, reading, and writing. In 2007, a fourth subject was added: science. The grading rule is fairly simple: For every performance category the percentage of students that meet a pre-defined proficiency level is calculated from the Florida Comprehensive Assessment Test (FCAT) scores. The sum of these percentages constitutes a school's grade points. These are translated into a letter grade depending on a distinct cutoff value on the grade points scale.

When a school receives the first F in a 4-year-period, this does not have immediate effects other than the stigma of failing minimum achievement requirements. Only the second F in a 4-year-period, triggers ultimate sanctions in the sense that students

are given the opportunity to transfer to a higher scoring public school ("school choice option"). Until 2006, students could also obtain a state funded voucher to attend a private school of their choice ("opportunity scholarship program"). In January 2006, the Florida Supreme Court struck down the voucher option by ruling it unconstitutional. However, the public school choice option remained unaffected.

Florida's A+ accountability system underwent several changes and revisions since its first introduction in 1999. Apart from the termination of the voucher option, these changes applied to adjustments of the grading rule with the purpose of rising accountability standards. The system started off with three categories measuring the percentage of students proficient in mathematics, reading, and writing. In 2002, three more categories were added, measuring the learning gains in mathematics and reading as well as the learning gains of the lowest 25% in reading. In 2005, the range of students who are included in accountability calculations was extended to all students. Prior to this year, students with limited English proficiency and students with certain disabilities were excluded from accountability calculations. In 2007, again more categories were added to school grades calculations: performance in FCAT science, learning gains of students scoring in the lowest 25 percent in mathematics, and performance of FCAT retakes in high school grades 11 and 12.

These adjustments always led to an increase in the number of F-schools in that particular year. In 2002, the number of F-schools totaled 64 (2.5%) and went down to 35 (1.3%) in 2003. In 2005, 78 (2.8%) schools received an F, but only 21 (0.7%) did so in 2006. In 2007, 83 (2.9%) schools were rated as failing, while this number dropped again to 45 (1.6%) in 2008. Table 1 gives an overview of the distribution of Florida public schools graded D or F in a particular year.

3 Data and Empirical Framework

This section first presents the data used in this study. It then describes the empirical strategy for identifying the impact of terminating the voucher option on school's response of an F grade receipt.

3.1 Data

I use administrative, student-level data from the state of Florida. This dataset is provided by the Florida Department of Education Data Warehouse and provides information on all Florida public school students in grades 3 to 10 for school years 2001-02 to 2008-09. The dataset contains a student's annual FCAT scores in math and reading, demographic characteristics such as race, gender, limited English proficiency status, special education information, and free or reduced lunch eligibility. I also know which school a student attended in a particular year. This enables me to merge school specific information. I obtain information on each school's performance grade from the Florida Department of Education School Accountability Report website. In addition, I add school specific information such as the number of incidents at the school or information regarding teacher quality as well as on operating costs which I obtain from the Florida School Indicators Reports.

3.2 Identification Strategy

To estimate the effect of removing the voucher option on the size of the F effect, I use a difference-in-discontinuities design as described by Grembi et al. (2012).[2] Starting point for this analysis is the estimation of receiving an F in year 0 on student's test performance in year 1. School grades in Florida are determined by fixed threshold

[2]The difference-in-discontinuities design combines difference-in-differences with RD estimation. In this study the RD estimates are compared across time periods. Dickert-Conlin and Elder (2010) use a difference-in-discontinuities approach to compare RD estimates across spacial units.

values on a continuous grade points scale. Depending on whether a school scores just above or below the respective threshold for receiving an F, it either is put under the threat of sanctions or just forgoes it. In this context, it is intuitive to use a sharp regression discontinuities (RD) approach to estimate the effect of receiving an F on student's performance in the next year (Chiang, 2009). The equation to be estimated is:

$$Y_{is1} = \alpha_0 + \alpha_1 F_{s0} + \alpha_2 GP_{s0} + \alpha_3 (F_{s0} \times GP_{s0}) + X_{is1}\gamma + Z_{s0}\zeta + \eta_{is1} + \epsilon_{is1} \quad (1)$$

Y denotes the students test scores in year 1. F is an indicator for receiving an F grade in year 0. GP denotes the forcing variable, grade points. It is calculated as the difference of a school's actual grade points and the cutoff value.[3] This ensures that the coefficient α_1 can be directly interpreted as the effect of receiving an F on student performance. X is a vector of student-level covariates and Z is a vector of school-level covariates. η denotes class grade fixed effects and ϵ is an error term.[4] However, one has to keep in mind, that the RD design, although it uses all observations within a certain bandwidth around the threshold to fit local linear regressions on either side of the cutoff, is only able to estimate the F-effect for schools marginally close to the cutoff. A generalization to schools further away from the cutoff is not possible without further assumptions. In addition, it is important to assure that assignment into F and D schools around the cutoff was random. Otherwise, the estimated effect could be biased by confounding (endogenous) factors, such as mean reversion. This is the case when a school experiences a "bad" year due to whatever reason and jumps back to its normal performance in the following year. To circumvent the potential problem of mean reversion in my analysis, I only use those years in which accountability standards

[3] In 2002, this cutoff value was at 280 grade points. In 2007, due to the introduction of additional performance categories, the cutoff value was at 395 grade points.

[4] The inclusion of control variables is not needed in RD-regressions to consistently estimate a treatment effect. However, it increases the precision of the estimation.

were increased by adjustments of the grading rule. Under the stricter grading schemes many schools that previously passed the accountability criteria were now designated as failing. This ensures enough exogenous variation in my data to identify a treatment effect. The downside of this approach is that it limits the scope of available years to only three: 2002-03, 2005-06, and 2007-08. And because the voucher option was struck down during the 2005-06 school year, it seems advisable to also leave aside this year. This leaves us with one year during which the voucher option was active (2002-03) and one year during which the voucher option was no longer available (2007-08).[5] I first run a set of four regressions separately for each year and subject to investigate the size of the F effect in both years. The regressions differ by the control variables used as well as the bandwidth around the cutoff.

In the final step, I evaluate the *difference* in the F effects of both years in a single regression framework. This is done by estimating the following equation:

$$\begin{aligned} Y_{ist} = \alpha_0 + \alpha_1 F_{s,t-1} + \alpha_2 GP_{s,t-1} + \alpha_3 (F_{s,t-1} \times GP_{s,t-1}) + X_{ist}\gamma + \\ Z_{s,t-1}\zeta + T_{t=2008}[\beta_0 + \beta_1 F_{s,t-1} + \beta_2 GP_{s,t-1} + \beta_3 (F_{s,t-1} \times GP_{s,t-1}) + \\ X_{ist}\delta + Z_{s,t-1}\xi] + \eta_{ist} + \epsilon_{ist} \quad (2) \end{aligned}$$

Again, Y stands for the standardized students FCAT score in math (and reading). F is the indicator for receiving an F grade. GP indicates the grade points of school s minus the respective F/D cutoff value. X and Z are vectors of student and school level control variables, respectively. η denotes class grade fixed effects. The indicator T is 1 for observations from 2008 and 0 for observations from 2003. The coefficient of interest now is β_1 which is an estimate for the difference in the F effect between the two years. Adjustments to the grading rule do not induce a confounding effect, since

[5] Using only these two years is a conservative approach, which allows us to minimize the potential bias from mean reversion tendencies. Pooling over the pre-2006 and post-2006 years, however, yields similar results and does not change the interpretation.

it affects all schools equally and thus is already accounted for when calculating the discontinuities. In this sense, the only "adjustment" that affects only schools graded F in 2007 is the termination of the voucher option in 2006.

4 Results

4.1 Graphical Analysis

Following Imbens and Lemieux (2008), I first inspect a graphical illustration of the RD-setup. Figures 1 to 3 show scatter plots of the dependent variable (math and reading, respectively) against school grade points, separately for every year and school type. For clarity reasons, I combine schools within 5-point bins of school grade points, so that every dot in the figures represents the average of the dependent variable within a 5-point bin of school grade points. This means that one dot might in fact represent the average of more than just one school. For ease of presentation, I also normalize the grade points variable by subtracting the cutoff value of the respective year, so that the F/D-threshold is at zero grade points.

When looking at elementary schools (Figure 1) one can easily observe a jump in the dependent variable at the F/D threshold in almost every year. Schools that lay just on the left side of the cutoff exhibit a higher average test score the next year than schools just to the right of the cutoff. This is in line with the hypothesis that the receipt of an F grade triggers actions at the school to improve its performance. The fact that there is obviously also a positive treatment effect in years where the grading rule had not been adjusted is interesting, but could be caused at least in part by mean reversion.

When looking at middle and high schools, however, a different picture emerges. For middle schools there is no indication of a treatment effect whatsoever (Figure 2). One has to keep in mind, however, that the number of middle schools which received an

F in those years was extremely low, making it almost impossible to draw meaningful conclusions. The effect for high schools (shown in Figure 3) is not very distinct. One could be tempted to claim a treatment effect in the first four years. But again, with the exception of 2002-03 and 2007-08 sample sizes are also extremely low for high schools.

The graphical analysis has shown that F graded schools seem to respond differently, depending on whether they are elementary, middle, or high schools. In order to avoid attenuating effects from different school types, I therefore focus on elementary schools in my further analysis. As stated earlier, I also limit the analysis to school years 2002-03 and 2007-08 where adjustments to the grading rule ensure enough exogenous variation around the cutoff. Furthermore, as can be observed from Figure 4 there is no indication of schools' manipulation of the grading rule, as the distribution of schools around the cutoff is fairly smooth.

4.2 Descriptive Analysis

At the end of the school year 2001-02, 38 elementary schools received an F and 121 elementary schools received a D. At the end of the 2006-07 school year, 30 elementary schools were graded F and 51 were graded D. In order to evaluate the effect from sanction threats, I drop those schools from my analysis which already received an F in the previous three years. I also drop those schools that were no longer operating in the following school year.[6] This leaves me with 29 F-schools and 102 D-schools in the 2002-03 schools year, and 24 F-schools and 42 D-schools in the 2007-08 school year.[7]

[6]Form the 2002-03 sample, I drop 7 F and 17 D schools due to prior year F receipt, and 2 F and 2 D schools due to school closure. From the 2007-08 sample, I drop 6 F and 8 D schools due to prior year F receipt, and 1 D school due to school closure.

[7]I include charter schools in my analysis. Charter schools are privately run public schools. Besides being operated by private entities, the same accountability requirements apply to charter schools. Dropping charter schools from my analysis does not change the results.

From this sample of schools, I drop students who are new to their school from my analysis, as they might bias the achievement calculations depending on what school they attended before.[8] In addition, I drop not accountable students. In particular, these are limited English proficiency students in ESOL programs for less than two years, and students with certain disabilities.[9] However, this applies only to the 2002-03 school year, as grading rule adjustments in 2005 led to the inclusion of all students in school grade calculation since then.

Table 2 contains summary statistics for elementary schools graded F and D in school years 2002-03 and 2007-08 respectively. It is noteworthy that the student body in both F and D schools is composed mainly of African American students, with F-schools holding even more African Americans on average than D-schools. The average share of African American students in F-schools was roughly 80% in 2002-03 and nearly 70% in 2007-08 school year. Another striking characteristic of these schools is the high rate of students which are eligible for free or reduced price lunch. In both sample years, about 90% of the students were eligible for free or reduced price lunch. These facts are important when considering the need of these schools to improve their academic performance. Since African American students and poor students are often thought of being disadvantaged in comparison to other social groups, it makes it even more important to improve the academic performance of the schools they attend. F-schools also have more incidents of student violence on average compared to D-schools.

Schools also appear to be fairly similar on average across both sample years. The share of Hispanic students is larger in the 2007-08 school year, but this is equally the case for F and D schools. Also, in the later year, more schools are located in or near a large city. But again, the increase is similar for both treatment and control group (F

[8]Including new students and controlling for new student status does not alter the results.

[9]These disabilities are: autism, deaf or hard of hearing, emotionally handicapped, language impaired, orthopedically impaired, other health impaired, specific learning disabled, traumatic brain injury, visually impaired, intellectual disability.

and D schools, respectively), thus giving support to the parallel trends assumption. Also, almost 50% of the F and D schools from the 2007-08 sample already received a D or F grade in 2002-03.

For the RD-analysis to make sense, it is important that schools close to the cutoff value do not deviate with regard to their characteristics. In Table 3, I therefore compare the predicted values of several school characteristics at the F/D threshold for both sample years. Although these values still deviate in terms of their absolute numbers between F and D schools, the differences are not statistically significant. The only exception is the gender composition. Whereas a D-schools at the cutoff is composed to 50% of males and 50% of females in both years, the F-school counterpart has 4.7% more females in 2002-03 and 4.5% less females in 2007-08. However, these deviations are sufficiently small as to not call into question the validity of the RD-approach.

4.3 Regression Analysis

I now turn to the main results of my analysis. Table 4 shows regression results of the effect of receiving an F in 2002 on the academic performance of this school's students in the 2003 FCAT math and reading, respectively. Columns 1 and 3 do not include any controls other than class grade fixed effects. Columns 2 and 4 in addition to class grades fixed effects include several student- and school-level controls.[10] The inclusion of control variables increases the precision of the estimation. The selection of the optimal bandwidth (h*) to fit the local linear regressions to both sides of the threshold is dealt with quite arbitrarily in different studies. In this paper, I follow Chiang (2009) and use the cross validation criterion method to find h*. The optimal bandwidth used in the regressions displayed in columns 3 and 4 in Table 4 for both math and reading

[10]The control variables used in the regressions are shown in Table 2 and include a cubic polynomial of previous year's math (reading) scores.

is 29 grade points to the left side of the cutoff and 35 grade points on the right of the cutoff. However, since this procedure reduces the number of schools, it gives more weight to the schools close to the cutoff and increases the regression's sensibility to outliers. The F receipt triggered schools to improve student performance by 9.8 to 12.4 percentage points of a standard deviation in math and 8.8 to 10.6 percentage points of a standard deviation in reading. Table 5 shows results of analogous regressions for the 2007-08 school year. Again, effects are positive and at least as large as in the 2002-03 school year. In most of the cases, especially when considering the optimal bandwidth (34 grade points to the left, 59 grade points to the right of the cutoff) the effects for both math and reading are much larger; 19.4 percentage points of a standard deviation in math, and 12.6 percentage points of a standard deviation in reading. This result seems puzzling, since the termination of the voucher option should have worked in the opposite direction.

In order to make a clear statement on the effect of the voucher option termination on student test scores, however, both sample years need to be evaluated together in a single regression framework. Table 6 shows results from the difference-in-discontinuities regression. Neither of the calculated differences between the school years 2002-03 and 2007-08 is statistically significant. This means that, although the voucher option is no longer a potential threat for failing schools in the later year, this does not seem to have altered the school's behavior. In other words, receiving an F grade still induces enough incentives to increase academic performance at failing schools even without the threat of private school vouchers.

5 Robustness Checks

The analysis described in the previous section has shown that the termination of the voucher option did not induce a significant change in a school's behavior upon re-

ceiving an F grade. However, the point estimates suggest an even larger F effect in the post-voucher year 2007-08. Although the difference in the F effects of both years is not statistically significant, this raises the question of potentially confounding effects. When looking at the 2008 scatter plot of test scores on school grade points (Figure 1), one can distinguish a school to the right hand side of the threshold which underperforms the other schools by about 0.4 percentage points of a standard deviation. When this school is dropped from the sample, the estimated F effects resemble those obtained in 2002-03 (Table 7). Finally, applying the difference-in-discontinuities regression approach on the sample without the outlier yields a point estimate of zero (Table 8).

Another robustness check is to evaluate a different outcome variable which also measures student achievement. For this purpose, I use the test scores from a low-stakes exam (SAT-9/10) which is administered together with the FCAT, but is not used for accountability purposes. Table 9 shows the results from the respective difference-in-discontinuities regressions. All estimates are positive (and statistically insignificant) pointing in the opposite direction of what the hypothesis of reduced sanction threats would suggest.

Finally, I also run difference-in-discontinuities regressions on different subgroups of the student population to uncover masking effects of heterogeneous subgroups (Table 10). Since African American and poor students are by far the biggest subgroups in the samples, a negative sign on the respective coefficients would suggest, that schools in fact lowered their performance. However, this is clearly not the case. For both subgroups, point estimates of the difference-in-discontinuities regressions are positive.

6 Discussion and Conclusion

This study analyzes the incentive character of private school vouchers in addition to public school choice as a sanction mean to trigger academic improvement at low-performing public schools. The threat of competition can serve as a powerful incentive for schools to improve academic performance. This study provides evidence that it does not matter whether this threat of competition comes from the private or the public school sector. Florida's A+ accountability plan originally allowed students of repeatedly failing public schools to either attend a better performing public school or to obtain a state funded voucher to transfer to a private school. In 2006, the Florida Supreme Court ruled the private school voucher option unconstitutional. Since then, students of repeatedly failing public schools are left with the option of transferring to a better performing public school. I use a difference-in-discontinuities regression design to evaluate the impact of the court's decision on school's behavior to improve academic performance. Results reveal no change in average school's behavior suggesting that there is no separate incentive effect from the threat of private school over public school competition.

This result seems somewhat puzzling. Since the termination of private school vouchers restricted the choice options for students at failing public schools, this should have also lowered the potential threat of losing students. Data on school choice participation rates in Florida, however, show that only a relatively small fraction (less than 6%) of eligible students actually made use of private school vouchers. On the other hand, more than 10% of eligible usually used the public school choice option. This difference might explain why the termination of the private school vouchers had no effect on school's behavior. However, the threat of vouchers could have had an impact on school's performance when the accountability program was first introduced. But once schools learned that only few students were actually making use of the voucher

option, this incentive diminished. It also remains unclear to what extend the improvement is caused by a stigma effect of receiving the lowest performance grade. Previous research trying to answer this question appears to be quite mixed.

Nevertheless, the finding of this study is important for policy makers around the world trying to set up effective school accountability systems. Knowing that the efficiency of school accountability does not hinge on the competition threat from the private education sector, can make the introduction of such systems publicly more acceptable, because public money can be kept in the public sector. However, apart from setting the right incentives, policy makers should be aware of the fact that accountability pressure might lead to gaming behavior by the schools (e.g. Figlio, 2006; Jacob, 2005; Jacob and Levitt, 2003).

Finally, it remains unclear why schools actually try to avoid competition. Is it because of the potential loss of students (and the funding attached to them) or is it because of the (additional) stigma of having to allow students to transfer to a different school? Since students are assigned to schools depending on the school's catchment area, it might in fact constitute a big disgrace for school officials (principals, teachers, etc.) to lose this quasi-monopolistic privilege. Future research should therefore try to explore the channels through which the school choice threat impacts schools in more detail.

References

Chakrabarti, R. (2008). Impact of voucher design on public school performance: Evidence from Florida and Milwaukee voucher programs. *Federal Reserve Bank of New York Staff Paper Number 315.*

Chakrabarti, R. (2013). Vouchers, public school response, and the role of incentives: Evidence from Florida. *Economic Inquiry*, 51 (1):500–526.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, 93:1045–1057.

Dickert-Conlin, S. and Elder, T. (2010). Suburban legend: School cutoff dates and the timing of births. *Economics of Education Review*, 29:826–841.

Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics*, 90:837–851.

Figlio, D. and Hart, C. (2010). Competitive effects of means-tested school vouchers. *NBER Working Paper No. 16056.*

Figlio, D. and Rouse, C. E. (2006). Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics*, 90:239–255.

Grembi, V., Nannicini, T., and Troiano, U. (2012). Policy responses to fiscal restraints: A Difference-in-Discontinuities Design. *CESifo Working Paper No. 3999.*

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142:615–635.

Jacob, B. (2005). Accountability, incentives, and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89:761–796.

Jacob, B. and Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics*, 118:843–877.

Ladd, H. F. and Glennie, E. J. (2001). A replication of Jay Greene's voucher effect study using North Carolina data. In Carnoy, M., editor, *School vouchers: Examining the evidence*, pages 49–52. Economic Policy Institute, Washington, D.C.

Rouse, C. E., Hannaway, J., Goldhaber, D., and Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *NBER Working Paper No. 13681.*

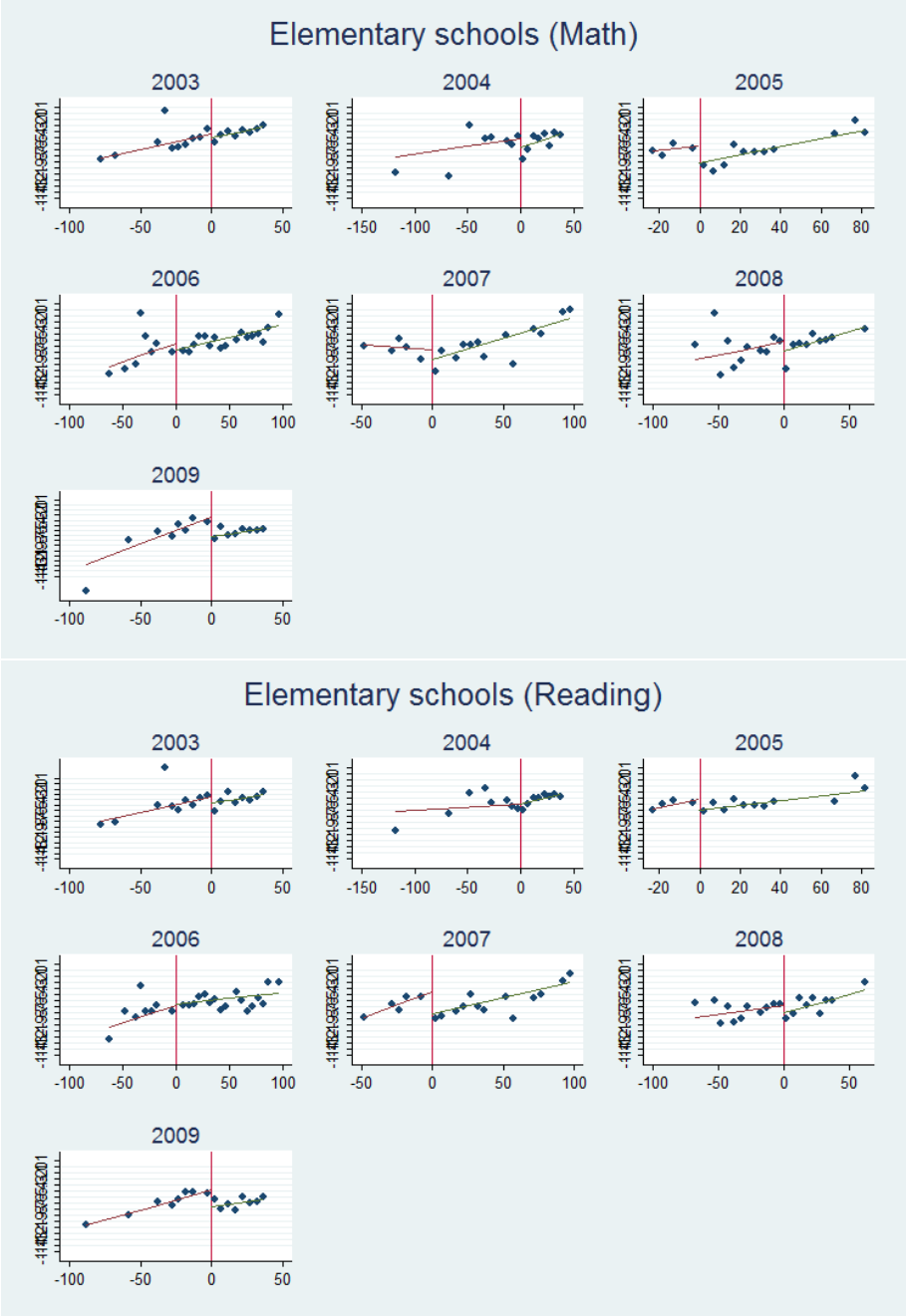
West, M. R. and Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal*, 116:C46–C62.

Table 1
Distribution of Florida public schools graded D or F

		D-schools	F-schools	Total number of D and F schools	Total number of Florida public schools	% graded F
2002*	Elementary	121	38	159	1581	2.4
	Middle	18	6	24	476	1.3
	High	40	19	59	340	5.6
	Combination	7	1	8	118	0.8
2003	Elementary	52	16	68	1592	1.0
	Middle	18	1	19	487	0.2
	High	52	12	64	356	3.4
	Combination	15	6	21	164	3.7
2004	Elementary	62	10	72	1614	0.6
	Middle	25	17	42	502	3.4
	High	83	15	98	364	4.1
	Combination	14	7	21	219	3.2
2005*	Elementary	79	18	97	1651	1.1
	Middle	32	8	40	530	1.5
	High	95	21	116	391	5.4
	Combination	24	31	55	202	15.3
2006	Elementary	36	7	43	1639	0.4
	Middle	6	1	7	531	0.2
	High	67	10	77	400	2.5
	Combination	13	3	16	284	1.1
2007*	Elementary	51	30	81	1691	1.8
	Middle	44	12	56	553	2.2
	High	102	30	132	411	7.3
	Combination	19	11	30	249	4.4
2008	Elementary	54	21	75	1726	1.2
	Middle	20	3	23	558	0.5
	High	70	16	86	394	4.1
	Combination	11	5	16	217	2.3

Notes: * indicates a year in which the grading rule was adjusted. Data from Florida Department of Education school accountability reports.

Figure 1
RD scatter plot of elementary schools

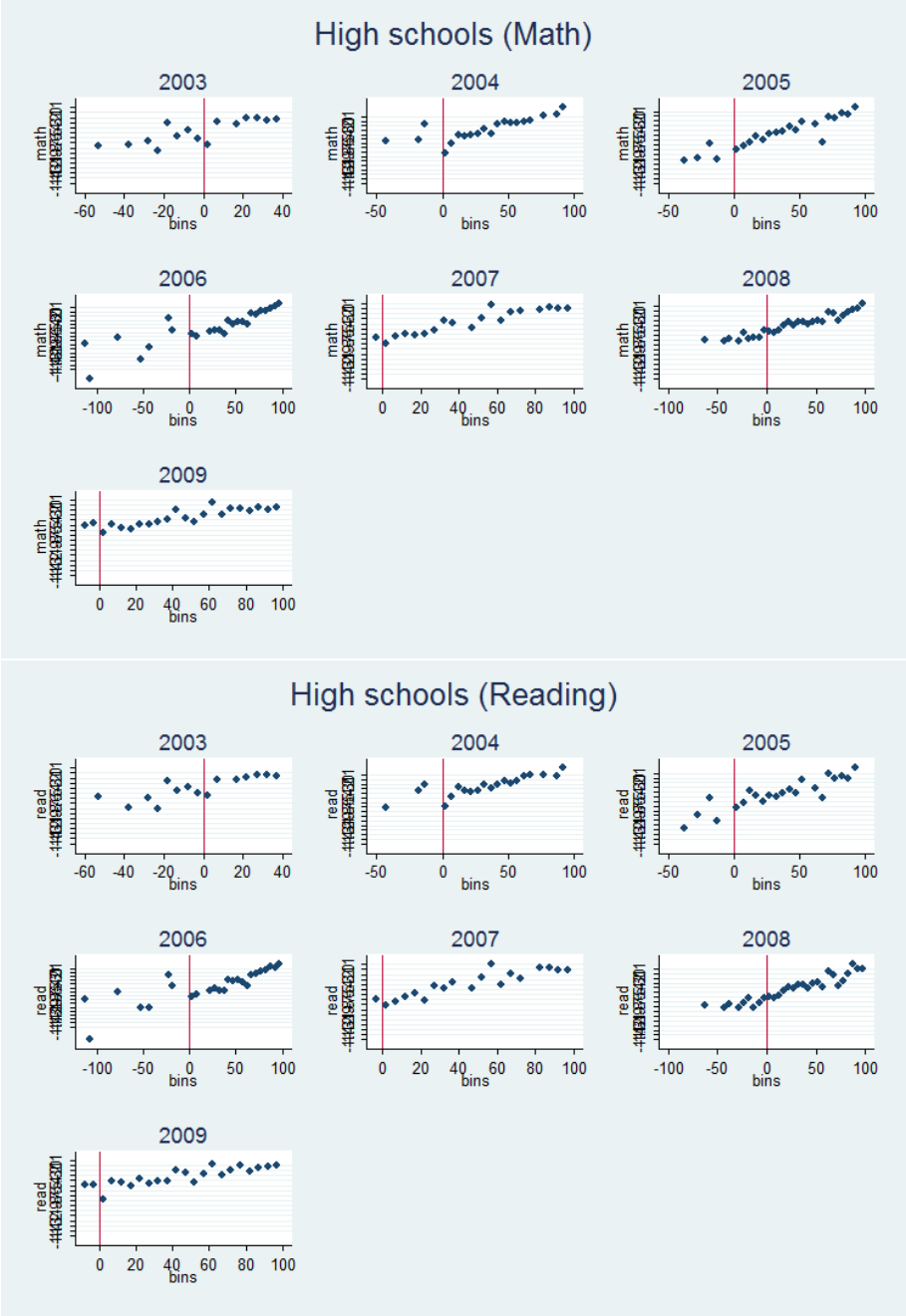


Notes: The figures show scatter plots of average standardizes FCAT test scores (y-axis) (by subject and year) within 5-point bins of previous year's school grade points (x-axis). The vertical line marks the threshold between F and D schools.

Figure 2
RD scatter plot of middle schools

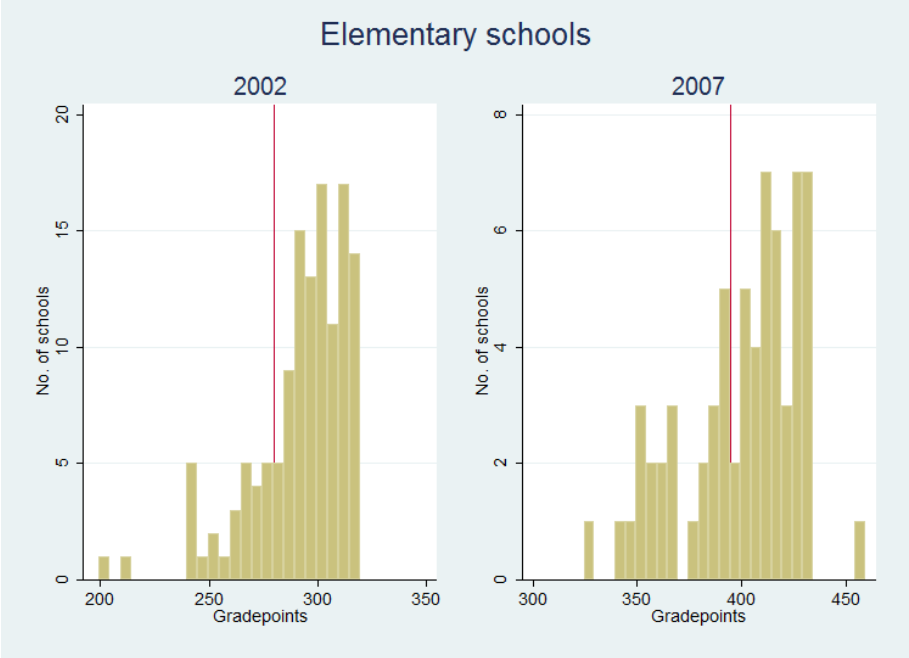
Notes: The figures show scatter plots of average standardizes FCAT test scores (y-axis) (by subject and year) within 5-point bins of previous year's school grade points (x-axis). The vertical line marks the threshold between F and D schools.

Figure 3
RD scatter plot of high schools



Notes: The figures show scatter plots of average standardizes FCAT test scores (y-axis) (by subject and year) within 5-point bins of previous year's school grade points (x-axis). The vertical line marks the threshold between F and D schools.

Figure 4
Distribution of F and D elementary schools around the cutoff



Notes: The vertical line indicates the cutoff value between F and D schools in the particular year.

Table 2
Summary Statistics

Variable	Sample means 2002-03		Sample means 2007-08	
	F-Schools	D-Schools	F-Schools	D-Schools
Black (Dummy)	0.793	0.556	0.694	0.6
Hispanic (Dummy)	0.122	0.19	0.24	0.279
Female (Dummy)	0.528	0.516	0.471	0.499
Free/reduced price lunch (Dummy)	0.894	0.811	0.913	0.892
Fraction in school-by-grade:				
Black	0.77	0.542	0.695	0.607
Hispanics	0.144	0.216	0.236	0.272
Female	0.485	0.489	0.466	0.498
Free/reduced price lunch	0.903	0.832	0.917	0.896
FCAT math score in previous year	-0.572	-0.347	-0.691	-0.48
FCAT reading score in previous year	-0.442	-0.248	-0.607	-0.442
Mean in school-by-grade:				
Previous year FCAT math score	-0.749	-0.521	-0.693	-0.51
Previous year FCAT reading score	-0.651	-0.459	-0.621	-0.456
Grade 4 (Dummy)	0.518	0.473	0.49	0.492
Grade 5 (Dummy)	0.441	0.509	0.51	0.494
Grade 6 (Dummy)	0.041	0.018	0	0.013
Large city (Dummy)	0.467	0.524	0.679	0.655
Pupil-teacher-ratio in previous year	15.29	16.872	14.245	14.893
ln(operating costs per student) in previous year	8.746	8.655	9.126	9.035
Avg. teacher experience in previous year	10.667	10.602	11.492	10.962
Avg. number of incidents in previous year	48.087	39.734	40.038	30.017
Observations	2428	11813	1883	4177
Schools	29	101	22	41

Notes: The table presents mean sample statistics for D and F schools for school years 2002-03 and 2007-08, respectively. FCAT math and reading scores are standardized with mean 0 and sd 1.

Table 3
Predicted values at F/D threshold

Variable	2002-03			2007-08		
	F-Schools	D-Schools	Diff. (D-F)	F-Schools	D-Schools	Diff. (D-F)
Black (Dummy)	0.866	0.732	-0.134	0.625	0.79	0.165
Hispanic (Dummy)	0.048	0.114	0.066	0.337	0.236	-0.101
Female (Dummy)	0.547	0.5	-0.047**	0.453	0.498	0.045*
Free/reduced price lunch (Dummy)	0.884	0.853	-0.031	0.92	0.938	0.017
Fraction in school-by-grade:						
Black	0.831	0.704	-0.127	0.628	0.772	0.143
Hispanics	0.079	0.147	0.068	0.33	0.227	-0.103
Female	0.504	0.486	-0.018	0.46	0.488	0.028
Free/reduced price lunch	0.897	0.862	-0.035	0.924	0.93	0.006
FCAT math score in previous year	-0.446	-0.486	-0.04	-0.547	-0.633	-0.086
FCAT reading score in previous year	-0.308	-0.364	-0.056	-0.521	-0.563	-0.042
Mean in school-by-grade:						
Previous year FCAT math score	-0.649	-0.644	0.005	-0.57	-0.622	-0.052
Previous year FCAT reading score	-0.524	-0.565	-0.041	-0.519	-0.555	-0.036
Grade 4 (Dummy)	0.509	0.495	-0.015	0.483	0.496	0.014
Grade 5 (Dummy)	0.438	0.471	0.034	0.517	0.503	-0.014
Grade 6 (Dummy)	0.053	0.034	-0.019	0	0.001	0.001
Large city (Dummy)	0.517	0.39	-0.127	0.745	0.653	-0.092
Pupil-teacher-ratio in previous year	15.162	16.3	1.138	14.518	14.404	-0.114
ln(operating costs per student) in previous year	8.686	8.638	-0.048	9.105	9.163	0.058
Avg. teacher experience in previous year	10.512	10.113	-0.399	11.954	10.351	-1.602
Avg. number of incidents in previous year	50.854	46.676	-4.178	36.117	39.111	2.994

Notes: The table presents predicted values of covariates at the F/D threshold for D and F schools for school years 2002-03 and 2007-08, respectively. FCAT math and reading scores are standardized with mean 0 and sd 1. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 4
Effect of 2002 F grade on 2003 FCAT scores (Elementary schools)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.0808	.0982*	.1178	.1243**
	(.0783)	(.0541)	(.0850)	(.0517)
Observations	24,683	14,241	21,300	12,228
Reading	.1258*	.1055**	.1170	.0878
	(.0728)	(.0482)	(.0783)	(.0647)
Observations	24,695	14,303	21,311	12,286
D-Schools	102	101	90	89
F-Schools	29	29	21	21

Notes: The table contains results from RD-regressions of FCAT math (reading) scores in 2003 on school grade points in 2002, an indicator for F-school and an interaction term of the two. Each cell represents the result from a separate regression. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The following control variables are used: Dummies for Blacks, Hispanics, Gender, Free or reduced price lunch, fraction in school-by-grade of these variables, previous year FCAT math (reading) score, quadratic and cubic terms of previous year FCAT math (reading) score, mean in school-by-grade of previous math (reading) score, dummies for grade 5 and 6, an indicator term for whether the school is located in a large city or at the fringe of a large city, pupil-teacher-ratio, logarithmic term of operating costs per pupil, teachers average years of experience, number of total incidents at the school. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 5
Effect of 2007 F grade on 2008 FCAT scores (Elementary schools)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.1542*	.1048	.2070**	.1935**
	(.0868)	(.0744)	(.0972)	(.0795)
Observations	12,303	6,060	10,988	5,588
Reading	.1370*	.0868	.1365	.1261**
	(.0739)	(.0525)	(.0832)	(.0596)
Observations	12,276	6,058	10,964	5,587
D-Schools	42	41	41	40
F-Schools	24	22	16	16

Notes: The table contains results from RD-regressions of FCAT math (reading) scores in 2008 on school grade points in 2007, an indicator for F-school and an interaction term of the two. Each cell represents the result from a separate regression. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The following control variables are used: Dummies for Blacks, Hispanics, Gender, Free or reduced price lunch, fraction in school-by-grade of these variables, previous year FCAT math (reading) score, quadratic and cubic terms of previous year FCAT math (reading) score, mean in school-by-grade of previous math (reading) score, dummies for grade 5 and 6, an indicator term for whether the school is located in a large city or at the fringe of a large city, pupil-teacher-ratio, logarithmic term of operating costs per pupil, teachers average years of experience, number of total incidents at the school. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 6
Effect of voucher option termination on student FCAT scores (Elementary schools)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.0734	.0066	.0893	.0692
	(.1165)	(.0916)	(.1286)	(.0944)
Observations	36986	20301	32288	17816
Reading	.0112	-.0186	.0195	.0383
	(.1034)	(.071)	(.1138)	(.0876)
Observations	36971	20361	32275	17873
D-Schools	144	142	131	129
F-Schools	53	51	37	37

Notes: The table shows difference-in-discontinuities estimates of the effect of the voucher option termination on student FCAT scores in math and reading. Test scores are from testing years 2003, when the voucher option was active, and 2008, when voucher option was no longer available. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The optimal bandwidth used to calculate the discontinuity in 2003 is 29 school grade points to the left and 35 school grade points the the right of the cutoff for both math and reading; and 34/59 for the discontinuity in 2008. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 7
Robustness 1: Effect of 2007 F grade on 2008 FCAT scores
(Elementary schools, w/out outlier)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.0876	.0926	.1266*	.1556*
	(.0655)	(.0758)	(.0707)	(.0833)
Observations	12,091	5,958	10,776	5,486
Reading	.1137	.0803	.1039	.1107*
	(.0756)	(.0545)	(.0831)	(.0617)
Observations	12,064	5,957	10,752	5,486
D-Schools	41	40	40	39
F-Schools	24	22	16	16

Notes: The table contains estimation results from RD-regressions of FCAT math (reading) scores in 2008 on school grade points in 2007. The sample excludes one D-school very close to the cutoff, which can be regarded an outlier, since its performance in 2008 is way below the average of comparable schools. Each cell represents the result from a separate regression. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The following control variables are used: Dummies for Blacks, Hispanics, Gender, Free or reduced price lunch, fraction in school-by-grade of these variables, previous year FCAT math (reading) score, quadratic and cubic terms of previous year FCAT math (reading) score, mean in school-by-grade of previous math (reading) score, dummies for grade 5 and 6, an indicator term for whether the school is located in a large city or at the fringe of a large city, pupil-teacher-ratio, logarithmic term of operating costs per pupil, teachers average years of experience, number of total incidents at the school. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 8
Robustness 2: Effect of voucher option termination on student FCAT scores (Elementary schools, w/out outlier)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.0068	-.0056	.0088	.0314
	(.1018)	(.0927)	(.1102)	(.0975)
Observations	36774	20199	32076	17714
Reading	-.0121	-.0251	-.0131	.0228
	(.1046)	(.0725)	(.1138)	(.089)
Observations	36759	20260	32063	17772
D-Schools	143	141	130	128
F-Schools	53	51	37	37

Notes: The table shows difference-in-discontinuities estimates of the effect of the voucher option termination on student FCAT scores in math and reading. The sample excludes one D-school very close to the cutoff, which can be regarded an outlier, since its performance in 2008 is way below the average of comparable schools. Test scores are from testing years 2003, when the voucher option was active, and 2008, when voucher option was no longer available. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The optimal bandwidth used to calculate the discontinuity in 2003 is 29 school grade points to the left and 35 school grade points the the right of the cutoff for both math and reading; and 34/59 for the discontinuity in 2008. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 9
Robustness 3: Effect of voucher option termination on low-stakes test scores (Elementary schools, w/out outlier)

	h=max		h=opt	
	No Controls	With Controls	No Controls	With Controls
Math	.0909	.0798	.1289	.1473
	(.1188)	(.0853)	(.1353)	(.0928)
Observations	20021	20021	17553	17553
Reading	.0989	.0282	.1463	.0698
	(.1275)	(.064)	(.1537)	(.0738)
Observations	20034	20034	17569	17569
D-Schools	141	141	128	128
F-Schools	51	51	37	37

Notes: The table shows difference-in-discontinuities estimates of the effect of the voucher option termination on student low-stakes test scores (SAT-9/10) in math and reading (sample w/out outlier). Test scores are from testing years 2003, when the voucher option was active, and 2008, when voucher option was no longer available. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The optimal bandwidth used to calculate the discontinuity in 2003 is 29 school grade points to the left and 35 school grade points the the right of the cutoff for both math and reading; and 34/59 for the discontinuity in 2008. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.

Table 10
Robustness 4: Effect of voucher option termination by subgroups of the student population (Elementary schools, w/out outlier)

	h=opt; with controls						
	Blacks	Hispanics	Whites	Females	Males	Poor	Rich
Math	.0321	-.1169	.1663	.1229	-.0561	.0571	-.1172
	(.1146)	(.1076)	(.1672)	(.1056)	(.1122)	(.1019)	(.1209)
Observations	10914	3627	2657	9027	8687	15022	2692
Reading	.0387	.0843	.3098	.0707	-.0276	.054	-.0662
	(.1006)	(.1045)	(.1953)	(.0976)	(.1033)	(.0897)	(.1919)
Observations	10961	3630	2667	9059	8713	15065	2707
D-Schools	118	94	94	118	118	117	116
F-Schools	35	26	20	35	35	35	33

Notes: The table shows difference-in-discontinuities estimates of the effect of the voucher option termination on student FCAT scores in math and reading by different subgroups of the student population (sample w/out outlier). The subgroup "Poor" denotes students eligible for free or reduced price lunch. "Rich" refers to students not eligible for subsidized meals. Test scores are from testing years 2003, when the voucher option was active, and 2008, when voucher option was no longer available. h denotes the bandwidth for RD estimation (max = all observations, opt = optimal bandwidth calculated using the cross validation criterion method). The optimal bandwidth used to calculate the discontinuity in 2003 is 29 school grade points to the left and 35 school grade points to the right of the cutoff for both math and reading; and 34/59 for the discontinuity in 2008. Robust standard errors (in parentheses) clustered at the school level. Significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.