THE EFFICACY OF CHOICE THREATS WITHIN SCHOOL ACCOUNTABILITY SYSTEMS: RESULTS FROM LEGISLATIVELY INDUCED EXPERIMENTS

PEPG 05-01

Martin R. West
Research Fellow, Program on Education Policy and Governance
Harvard University
mrwest@fas.harvard.edu

Paul E. Peterson
Henry Lee Shattuck Professor of Government &
Director of the Program on Education Policy and Governance
Harvard University
ppeterso@latte.harvard.edu

A paper presented before the Annual Conference of the Royal Economic Society,
University of Nottingham, March 23, 2005

Abstract

Stigma and school voucher threats under a revised 2002 Florida accountability law have positive impacts on student performance. Stigma and public school choice threats under the U.S. federal accountability law, No Child Left Behind, do not have similar effects in Florida. Significant impacts of stigma, when combined with the voucher threat, are observed on the test score performance of African Americans, those eligible for free lunch, and those with the lowest initial test scores. No significant impacts of the voucher threat on the performances of whites and Hispanics are detected. Estimations rely upon individual-level data and are based upon regression analyses that exploit artificial distinctions created by cliffs within the accountability regimes.

THE EFFICACY OF CHOICE THREATS WITHIN SCHOOL ACCOUNTABILITY SYSTEMS: RESULTS FROM LEGISLATIVELY INDUCED EXPERIMENTS

Martin R. West and Paul E. Peterson

With the growing recognition of the importance of human capital for economic growth, many nations are exploring ways of enhancing the quality of their educational systems.[1] Two of the most widely discussed reforms—school accountability and parental choice—are now being implemented in many parts of the United States (Howell and Peterson, 2002; Peterson and West, 2003). Of particular interest is the effort to combine the two reforms by giving parents a choice of another school when an accountability system indicates that the public school their child attends is inadequate. Such accountability systems typically give schools one year to improve before the parental choice "threat" is implemented. Two prominent examples of the use of parental choice as a threat to stimulate school improvement—the federal program created by the public school choice provisions of No Child Left Behind and the Opportunity Scholarships program created by Florida's A+ Accountability Plan—are currently operating within that state. In this paper, we estimate the impact on student performance of the choice threats as well as of other features of these two accountability systems.

1. Federal and State Accountability Systems in Florida

No Child Left Behind (NCLB), a U.S. federal law enacted in 2002, currently requires states to test all students in grades 3 through 8 in reading and math, with an additional test to be administered in high school. The average performance of all students—and of various student subgroups above a minimum size—on these tests must be reported publicly for all schools within each state. Schools that do not show that their students are making Adequate Yearly Progress (AYP) toward a state-determined level proficiency for two years in succession are said to be "in need of improvement" and

[1] We wish to thank Commissioner John Winn, Christy Hovanetz, Jeff Sellers, and other officials at the Florida Department of Education for supplying the information for the analysis reported in this paper, William G. Howell for his help in designing the analysis, Matthew Chingos for expert research assistance, and the John M. Olin Foundation and the National Research and Development Center on School Choice, Competition, and Achievement for their financial support. The authors alone are responsible for the findings and interpretations reported.

students are given a choice of another public school within the local school district that is not so designated.[2] In Florida, this provision applies only to schools that receive funding through Title I, the federal government's compensatory education program.

NCLB's accountability provisions, first implemented during the 2003 school year, are an outgrowth of previously established accountability systems in several states, each of which had their own distinctive features. (School years are identified by the year in which the student takes the examination, which occurs in the spring of that year.) Of the pioneering state accountability systems (Texas, Massachusetts, North Carolina and others), few have attracted greater interest than Florida's A+ Plan, both for well-known political reasons, and because the original 1999 Florida law served as an important model for NCLB.

As revised by the legislature and fully implemented in 2002, Florida's A+ Plan resembles NCLB in that the average test performance of students in grades three through ten must be reported annually for each school. In addition, students at twice-failed schools are given the opportunity to attend another school.

Despite the similarity of the two laws, certain features of the A+ Plan, as revised, are considerably more rigorous than NCLB. For one thing, students at schools that fail two out of any four years are given the opportunity to receive a voucher to attend any school – public or private – within the school district or elsewhere. The A+ Plan also distinguishes among five levels of school performance, from 'A' to 'F,' a more detailed set of categories than the simple dichotomy between making AYP or not that NCLB draws.

The assessment of the quality of Florida's schools under the A+ Plan and NCLB differed noticeably. Under Florida's grading system in 2002, 39 percent of the state's elementary schools received an 'A', 23 percent were given a 'B,' 28 percent a 'C', 8 percent a 'D,' and 2 percent an 'F.' But when the NCLB accountability system took effect in 2003, nearly 75 percent of all elementary schools were said to be "in need of improvement," often because one or more subgroups within the school was identified as not making adequate progress toward proficiency.

[2] If schools remain in need of improvement for an additional year, families become eligible for supplemental educational services after school, either from the school district or from private or non-profit providers. After four years, the school may be reconstituted.

2. Prior Research

Research on school choice and school accountability within the United States is a rapidly growing cottage industry. Numerous studies have estimated the impact of attending private schools or charter schools on the performance of individual students, and the impact on public-school performance of regimes that allow a wider range of choice (see, for example, Howell and Peterson, 2002; Hoxby, 2004a; Ladd, 2002; Neal, 2002). Similarly, the impact of accountability systems on student achievement and educational productivity is a matter of ongoing investigation (Carnoy and Loeb, 2003; Hanushek and Raymond, 2004.) However, only a few studies, beginning with Greene's pioneering research (Greene, 2001), have attempted to estimate the impact on public school performance of choice threats embedded within accountability systems (Chakrabarti, 2004; Figlio and Rouse, 2004; Greene and Winters, 2003). To our knowledge, no studies have systematically evaluated the effects on school performance of being identified as "in need of improvement" under NCLB.

Choice-threat research has focused mainly on the State of Florida, because that is the place where the most notable policy innovation has taken place. Prior studies have generally found positive impacts on the average performance of schools assigned an 'F,' which placed the school under the threat of the voucher program. However, most of these early studies were limited by the fact that the scholars had access only to school-level data, not the test scores and demographic characteristics of individual students (Chakrabarti, 2004; Greene, 2001; Greene and Winters, 2003). As a result, it is unknown from their results whether gains constituted actual improvements in the performance of individual students or were due to changes in the composition of those taking the test. Such changes could occur as the result of migration between schools or the exclusion of low-performing students from participation in these high-stakes tests. The one study with access to individual-level data (Figlio and Rouse, 2004) was limited to a subset of districts within the state and only examined the "shock" of the less comprehensive accountability system established in 1999, several years before the implementation of the more sophisticated system introduced in 2002 that is the focus of our investigation. Building on these studies, this paper uses individual-level data for all elementary school

students in the state to estimate the effects of several features of school accountability in Florida.

3. Estimating the Impact of the Accountability Shock

When estimating the impact of a policy intervention, the ideal comparison is that of a Randomized Field Trial (RFT), which estimates effects after assigning subjects randomly to treatment and control groups. Unfortunately, within the education policy world, conditions seldom permit the conduct of an RFT. However, policy researchers have in recent years employed a research design that approximates the RFT ideal by comparing subjects that fall on either side of an artificial borderline created for administrative convenience. Whether or not a subject is treated may be due as much to measurement error as to actual differences between the subjects. If so, then the subjects placed in the control group adjacent to the borderline are comparable to those subject to the treatment. For placement of the subjects on either side of the border to be a random act, the policy innovation should be an external shock that the subjects neither anticipated nor helped to shape. Otherwise, subjects could adjust their behavior in such a way as to have anticipated the policy innovation.

Florida A+ Plan. The revised Florida A+ Plan acted as an external shock beyond the ken or control of teachers and administrators at the state's elementary schools. The forces shaping the legislation that revised the A+ Plan in the Spring of 2001, as well as the regulations promulgated under that legislation in December 2001, were mainly political, with the legislature, the governor, and other political leaders – not local schools – determining program guidelines. Admittedly, Florida had had an accountability system in place since 1999, but initially that grading system had been pegged to levels of student achievement and was based upon test scores for just one elementary school year. Under the altered formula introduced in 2002, grades instead assigned as much weight to student gains in performance as to the levels students achieved. This new plan, with its complicated formula, was not put into place until just months before students began taking the tests upon which the new accountability system would be placed. Few, if any, school administrators located in schools on the borderline between grade levels could

have guessed how the new law might impact them. (See Appendix for a fuller discussion of the differences between the pre-2002 and post-2002 accountability systems.)

The overall test score performance of Florida's elementary school children improved subsequent to the implementation of the revised A+ Plan (Table 1), a topic we return to in the conclusion to this paper. But so altered was the new set of regulations that no less than 58 percent of Florida's elementary schools received a grade different from the one they had received the previous year (See Table 2). Thirty-five elementary schools received an 'F' in 2002, despite the fact that not a single school in the state had received that grade the preceding year. At the same time, the share of schools recognized for outstanding performance also increased, with the percentage of elementary schools receiving an 'A' jumping from 24 percent to 38 percent.

No Child Left Behind. Just as the modifications to the A+ Plan in 2002 acted as an external shock that could not be anticipated by local school officials, so NCLB, both in its legislative form and in the key administrative regulations promulgated under the law, was quite beyond the influence of the street-level bureaucrats manning the Florida schools. The passage of the law in Washington, D.C. in January 2002 occurred as the result of a broad set of national political forces and bipartisan compromises that had little to do with circumstances in Florida, which already had put into place its own accountability system. Nor was it easy for school administrators to anticipate how the federal law's central concept of AYP would impact them. Since NCLB regulations were not issued until December 2002 (Peterson and West, 2003), Floridians again did not know the rules by which they would be evaluated until a few months before they were informed in the summer of 2003 whether or not particular schools had made AYP. (See Appendix for details on the way in which AYP was defined in 2003.)

4. Incentives to Improve

Florida A+ Plan. In one key respect, Florida's A+ Plan treated all schools similarly. All schools were awarded $100 per pupil, if they improved their standing by one letter grade. 'A' schools also received this amount simply for retaining their standing. These funds could be spent on teacher bonuses or other non-recurring expenses related to student achievement.

But incentives to improve nonetheless varied, depending on the grade the school received. Florida schools that received an 'F' had the strongest incentives to improve. They bore both the stigma of being among the 2 percent of all schools in Florida given a failing grade as well as the threat that a repeated 'F' would give students at the school the opportunity to use a voucher to go elsewhere. In addition, 'F' schools were assigned a community assessment team made up of parents, business representatives, educators, and community activists who were to write an intervention plan for the school. Schools that received a 'D' were also stigmatized as being among the 10 percent worst performing schools in the state and, like the 'F' schools, were assigned an assessment team.

Schools receiving the other grades were not subject to any sanctions. In addition, 'A' schools received the honor of that designation, which, when first awarded under the old grading system, appeared to enhance local property values (Figlio and Lucas, 2004). High test scores may, at least under some circumstances, facilitate the reelection of school board members (Berry and Howell, 2005). However, 40 percent of schools earned an 'A' in 2002, which may have limited its market and political value. In sum, the Florida system seems to have been designed to give roughly equal rewards to all schools that scored in the higher three categories.

To the extent that the incentives created by the accountability regime work as intended to improve student performance, we expect the impacts to be larger for those perceived to be more difficult to teach. In the absence of an accountability system, educators may be inclined to attend more closely to their more engaged students. With the introduction of effective rewards and sanctions, educators can be expected to focus more resources—time, energy, expectations, and so forth—on students from disadvantaged groups.

No Child Left Behind. As for NCLB, we expect its short-term impact to be minimal, simply because neither the stigma nor the choice threat was particularly consequential. In 2003, no less than 75 percent of the elementary schools in Florida were designated as needing improvement. If most everyone is sanctioned, the embarrassment is less than if only a few are. What's more, the public-school choice sanction turned out to have little bite. School districts did not lose students because all choice was contained within the district. And parental choices were limited to the

relatively few schools within the school district found not to be in need of improvement. In practice, few students exercised the choice. Nationwide, it was less than 1 percent of those eligible (Education Week, 2005; Peterson 2005).

5. Data and Methodology

To estimate the effects of receiving various grades under the A+ and NCLB accountability systems, we obtained from the Department of Education of the State of Florida information concerning test score performance on the reading and math components of the statewide exam (the FCAT), demographic characteristics, and school characteristics for all students tested in grades three through five in Florida elementary schools for the school years ending in 2002, 2003, and 2004. For purposes of analysis, we converted FCAT scale scores to standardized scores with a mean of 0 and a standard deviation of 1. Estimated impacts of an accountability provision can therefore be interpreted as effect sizes, the impact on student performance, as calculated in standard deviations. To increase precision, all results are based upon combined reading and math test scores obtained simply by averaging each student's standardized scores in the two subjects. We obtain similar results when outcomes are estimated separately for each subject.

We also obtained students' test scores on the Stanford Achievement Test, 9th edition (SAT-9), a national norm-referenced test that is administered at the same time as the FCAT but is not used for accountability purposes. Individual performances on the FCAT and the Stanford 9 are highly correlated with one another across grades and subjects (see Table A1). Still, this information allows us to test whether any observed improvements on FCAT performance generalize to other areas of knowledge, or whether perhaps an increased focus on the state's exam system actually serves to lower performance on the more general exam. SAT-9 scores are reported in national percentile rankings. To facilitate comparisons with the results of our FCAT analysis, we also converted these rankings to standardized scores with a mean of 0 and a standard deviation of 1.

Florida A+ Plan. The unique impact of the various features of the A+ Plan is best estimated using results from tests taken by students in the spring of 2003 – well

before NCLB took effect, and when schools were responding to the shock of the new grading system introduced in the summer of 2002.

'F'-Grade/Voucher-Threat Impacts. To isolate the impact of receiving an 'F' and being placed under the threat of vouchers, we set aside the seven 'F' schools whose students were already eligible to receive vouchers as a result of the workings of the old accountability system.[3] We also excluded the four 'F' schools that would have received an 'F' in 2002, had the previous level-based grading system remained in place. This left us with 24 schools that fell unexpectedly under the voucher threat simply as a consequence of the introduction of the new accountability system.[4]

To ensure that our results were robust to alternative classification systems, we undertook three sets of comparisons. Each was intended to identify treated and control schools that closely resembled one another yet were large enough to allow for the precise estimation of treatment effects. In Comparison I, we compared the 24 shocked 'F' schools with all 'D' schools for which the average 2002 test scores of 4th and 5th grade students tested in the school in 2003 did not exceed those of the highest performing treated schools. In Comparison II, we compared the same 'F' schools with only those 'D' schools where the average prior test scores of students tested in 2003 did not exceed the average among treated schools by more than one weighted school-level standard deviation (among treated schools). Comparison III looks only at those shocked 'F' and 'D' schools whose scores on the A+ point system fell within 10 points of the cutoff between the two grades.

Comparison I is the most inclusive in that it compares all shocked 'F' schools with a fairly broad group of 'D' schools. But as is shown in Table A2, the baseline characteristics of the schools in the treated and control groups differ significantly along a number of important dimensions. As Table A2 also shows, differences narrow considerably for Comparisons II and III.

[3] A preliminary analysis of these schools, where students became eligible for vouchers as a result of their 2002 school grades, indicates that voucher eligibility (and the accompanying interventions in school operations) had an additional positive impact on achievement in 2003. The impact on the performance of these very low-performing schools was similar in magnitude to the impact of voucher threat; results are available from authors' upon request.

[4] Our results do not depend on this analytic decision. Analyses that include all newly threatened 'F' schools, regardless of the origins of the threat, yield similar estimates of impact of receiving an 'F' (see Table 5).

Most importantly, no significant differences between treatment and control groups in baseline test scores (those attained in 2001) were observed in Comparisons II and III. This similarity minimizes the problem of "regression to the mean" effect that often bedevils observational studies. The average baseline test scores of the treated and control groups in Comparison I do differ significantly, however. To adjust for such differences, we control in all estimations for multiple measures of students' academic performance the previous year rather than calculating a single gain score.[5]

We initially use four models to estimate the impact on student performance in 2003 of the being newly identified as an 'F' school under the A+ Plan. Model I provides a baseline estimate that controls only for the student's own test score performance the previous year and his or her demographic characteristics. Model II controls for these characteristics plus the aggregated characteristics of the fourth and fifth grade students tested in the school. Model III controls for all the characteristics in Model II plus two measures of the financial and educational resources available to the school. It also reports results for just those students tested in the same school for two consecutive years. Some administrators feel that schools should be held accountable for the learning gains of only those students within their sphere of responsibility for at least this length of time, not for new students who may be more difficult to integrate into the rhythm of the school and whose progress may in part reflect the school they attended the previous year. Model IV controls for the same variables as in Model III, but includes all students tested at the school, regardless of whether they had attended that school the preceding year.

Model I is estimated using the following equation:

$$(1) \qquad T_{ist} = \beta_0 + \beta_1 I^{treat}_{st} + \beta_2 T_{ist-1} + X_{ist}\delta + u_{st} + \epsilon_{ist},$$

Where T is the test score of student i, s indexes school, and t indexes year; I^{treat} is a treatment indicator (i.e. it takes on a value of 1 if the school is operating under the threat of vouchers, receives a 'D' grade, etc); T is a vector of control variables for prior achievement; and X_{ist} is a vector of student-level demographic control variables.

Model II relies upon a modified version of equation (1), where Z_{st} is a vector of school-level aggregate demographic and achievement characteristics:

[5] Specifically, we include a cubic in previous FCAT test performance in math and reading to allow for non-linearity in the relationship between prior and subsequent achievement as well as previous national percentile ranking in SAT-9 math and reading.

$$(2) \qquad T_{ist} = \beta_0 + \beta_1 I^{F\,treat}_{st} + \beta_2 T_{ist-1} + X_{ist}\delta + Z_{st}\gamma + u_{st} + \epsilon_{ist},$$

In Models III and IV, which differ only in the sample of students included, the school-level control vector of control variables Z_{st} is expanded to include per pupil operating costs and pupil-teacher ratio.[6] In all models, standard errors are adjusted for clustering at the school level.

In the interpretation of results, we place the greatest emphasis on those from Model IV. It is the most inclusive both in terms of students for whom estimates are obtained and in the number of controls introduced into the analysis.

Other Grade Effects. To estimate the impact of receiving a grade other than 'F' we compared schools that received a new, lower grade to a comparable set of schools that accumulated enough additional points on the state's grading scale to receive the next higher grade. For these analyses, we first employed the Model IV estimation together with the more inclusive Comparison I approach, as described above. For new-'D' schools, where significant impacts were observed, we also conducted Comparisons II and III as robustness checks. (See Table A3 for descriptive statistics of schools in treated and control groups for the new-'D' analysis.)

Data constraints preclude us from distinguishing those schools who received the new, lower grade simply as a result of the new accountability regime from those who also would have received that grade under the prior system. It is likely that the new grading system was at least partially responsible for most grade changes. When comparing these grade effects to the 'F'-grade/voucher-threat effect, we provide a separate estimate of the effect of receiving a new 'F' regardless of whether a school would have received such a grade under the old accountability system.

No Child Left Behind. The AYP provisions of NCLB shocked Florida schools in two distinct ways. Schools receiving Title I funds that did not make AYP were immediately placed under a public-school choice threat. Unless they made AYP the following year, students at those schools would have the opportunity to attend another public school within the district that had made AYP. The remaining schools in Florida who failed to make AYP still received the stigma of being identified as not performing at the expected level; however, students at these schools not receiving Title I funds would

[6] See the notes to Table 3 for the full list of individual and school-level control variables in each model.

not become eligible for public school choice. Effects for both types of schools are estimated with all four models, using the Comparison I approach described above, which succeeds in producing groups with similar prior test scores. (See Table A3 for descriptive statistics of schools in treated and control groups.)

6. Results

Florida A+ Plan. As the result of the introduction of the new accountability system, a number of Florida schools newly received 'F' grades, identifying them publicly as low-performing schools and subjecting them to the threat of vouchers for continued poor performance. At these schools, students performed at a higher level in the subsequent year than did students at similar schools not so threatened. The size of the impact was about 4 percent of a standard deviation (see Table 3). Consistent across all four model specifications, this result is observed even when controlling for the social composition of the school and available school resources, as measured by the pupil-teacher ratio and operating costs per pupil. Results are also consistent across the three comparisons presented in Table 4 (cols. 1, 2, and 3). Indeed, the two tighter comparisons yield slightly larger estimates—5 percent of a standard deviation—than those obtained from Comparison I, suggesting that mean reversion is not an important problem for the results reported in Table 3.

Impacts of the voucher threat on such disadvantaged groups as African Americans, those eligible for free lunch, and those with low initial test scores were about 6 percent of a standard deviation. All were statistically significant. However, no impacts could be detected for whites, Hispanics, students not eligible for free lunch, or students with higher initial reading test scores (see Tables A5-A6).

Receiving a 'D' has its own impact on student performance, as can be seen in Table 4 (col. 4). Students at schools that received a 'D' also performed roughly 4 percent of a standard deviation better than students at similar schools that received a 'C.' Since only 8 percent of the schools received a 'D', the designation appears to have created a stigma that generated a disproportionately positive school response. The results remain much the same for Comparisons II and III (cols. 5 and 6). Significant impacts of

receiving a 'D' were detected for African Americans, whites, those with low initial test scores, and both those eligible and not eligible for free lunch (see Tables A5-A6).

The results in Tables 4 and 5 imply that 'D' schools, the control group with which the 'F' schools are compared, were themselves affected by Florida's accountability system, perhaps by the stigma of having received such a low grade. The effects on student performance of receiving an 'F,' with its accompanying voucher threat, are over and above the impact of that stigma.

Whether an annual increment in student performance of 4 to 5 percent of a standard deviation is large or small depends on the extent to which such improvement persists over time or is merely a one-year response. However, if new 'F' schools continued to outperform expectations for the three year period they remained immediately threat of vouchers, the accumulated gains would quickly become educationally significant. Given the fact that the costs of test-based accountability systems are a fraction of those of many other prominent reform strategies (Hoxby 2004b), the return on investment is likely to be large.

The improvements associated with receiving an 'F' or 'D' grade had no clear spill-over effect, either positive or negative, on students' performance of the nationally normed SAT-9 (see Table 4). On the other hand, there is no evidence that concentrated attention on the FCAT examination came at the expense of more generalized learning in these subjects; each of the point estimates of the effect of receiving an 'F' or a 'D' on SAT-9 performance is positively signed.

Receiving one of the higher three grades seems, by itself, to have had little differential impact on subsequent student performance (see Table 5, cols. 3-5). Schools that received a 'C' did no better than similar schools that received a 'B.' The same was true for schools that received 'B's and 'A's. This finding should not be interpreted as evidence that the Florida A+ Plan was having no impact on the performance of higher performing schools, however. Rather, it shows only that the impact is consistent across schools receiving 'A's, 'B's and 'C's. Given that the incentives to improve were essentially the same across these categories, a consistent response is not surprising.

No Child Left Behind. There is no indication that designating schools as not having made AYP had any differential impact on student performance in subsequent

years, either in Title I schools subject to the public-school choice threat, or in non-Title I schools (see Table 6).[7]

7. Discussion

Grading systems that target and clearly sanction a relatively small percentage of the school population appear to have a more pronounced differential impact on school performance than those that are less targeted. The Florida A+ Plan, by giving 'D's and 'F's to the lowest 10 percent of all schools, then combining the stigma of the low grade with the threat of vouchers for the lowest 2 percent of all schools, stimulated higher levels of student performance at these schools relative to similarly situated schools not so sanctioned. Notably, the improvements made by 'F' schools came on top of the gains registered by 'D' schools, suggesting that the voucher threat may have an additional impact over and above that of stigma alone. Lacking information on schools that received an 'F' grade but were not threatened by vouchers, however, we cannot test this explanation definitively.

Other elements of the Florida grading system did not have distinctive impacts. Subsequent student performance seems to have been unaffected by whether a school received a grade of 'A', 'B', or 'C.' Arguably, this was the intended purpose of the accountability system, because the incentives provided these schools were essentially the same.

NCLB, on the other hand, is explicitly committed to the principle of raising the level of performance at under-performing schools. Its very title—No Child Left Behind—reveals its strong commitment to closing the gap between students at higher and lower performing schools. An accountability system that identifies problems with many schools, while giving few sanctions or incentives to improve, appears unlikely to be of much consequence. All in all, the Florida A+ Plan seems better tailored to the particulars of that state than NCLB has been thus far.

[7] Since A+ remained in effect after the introduction of NCLB, we cannot exclude the possibility that NCLB effects are contaminated by the simultaneous application of the two accountability systems. However, the cliffs created by NCLB are entirely different from those created by the state accountability system.

Indeed, nothing in the findings reported herein casts doubt on the overall effectiveness of the revised Florida A+ Plan. On the contrary, overall test score performance in Florida has risen since its introduction on both the FCAT and the norm-referenced SAT-9. These improvements do not appear to be simply a function of observable changes in the demographic composition of the student population, as statistically significant improvements are evident in simple models controlling for demographic characteristics (see Table A7).

Since other educationally relevant changes were occurring in Florida at the same time, one cannot attribute the overall gains made in 2003 and 2004 to A+ with any certainty. The improvement may simply be an extension of pre-existing trends due to underlying social or environmental changes. Increments in state funding, mandated class-size reductions, ending social promotion in third grade, or any number of other factors could also have had positive effects on test-score performance. Still, there is nothing in the data that contradicts claims made by Florida officials that the revised A+ Plan had an overall beneficial effect.

References

Berry, Christopher M. and William G. Howell (2005). "Accountability in Public Education," in William G. Howell, ed., *Besieged: School Boards and the Future of Education Politics*. Washington, DC: Brookings.

Carnoy Martin and Susannah Loeb, (2003). "Does External Accountability Improve Student Outcomes? A Cross-State Analysis," *Education Evaluation and Policy Analysis*, 24(4): 305-331.

Chakrabarti, Rajashri. (2005). "Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs." Harvard University.

Education Week Research Center (2005). "NCLB: State Report on Progress." *Education Week*, 24(27): 21.

Figlio, David N. and Cecilia E. Rouse. (2004). "Co Accountability and Voucher Threats Improve Low-Performing Schools?" University of Florida, Princeton University and NBER.

Figlio, David N. and Maurice N. Lucas (2004). "What's in a Grade? School Report Cards and the Housing Market." *American Economic Review*, 94(3): 591-604.

Greene, Jay P. (2000). "The Looming Shadow: Florida Gets its F Schools to Shape Up." *Education Next*, 1(4): 76-82.

Greene, Jay P. and Marcus A. Winters (2003). "Competition Passes the Test," *Education Next*, 4(3): 66-71.

Hanushek, Eric A and Margaret E. Raymond (2004). Does School Accountability Lead to Improved Student Performance? NBER Working Paper 10591. Cambridge, MA: National Bureau of Economic Research.

Howell, William G. and Paul E. Peterson, eds. (2002). *The Education Gap: Vouchers and Urban Schools*. Washington DC: Brookings.

Hoxby, Caroline M. (2004a). The Cost of Accountability. NBER Working Paper 8855. Cambridge, MA: National Bureau of Economic Research.

Hoxby, Caroline M. (2004b). *The Economics of School Choice*. Cambridge, MA: National Bureau of Economic Research.

Ladd, Helen F., (2003). "School Vouchers: A Critical View," *Journal of Economic Perspectives*, 16(4): 3-24.

Neal, Derek M. (2003). "How Would Vouchers Change the Market for Education?" *Journal of Economic Perspectives*, 16(4): 25-44.

Peterson, Paul E. (2005). "A Conflict of Interest: District Regulation of School Choice and Supplemental Services." In John E. Chubb. Ed. *Within Our Reach: How America Can Educate Every Child*. Lanham, MD: Rowman and Littlefield.

Peterson, Paul E. and Martin R. West, eds. (2003). *No Child Left Behind? The Politics and Practice of School Accountability*. Washington, DC: Brookings.

Table 1: Average FCAT Scale Scores of 3rd-5th grade students in Florida, 2002-2004

	FCAT Math			SAT-9 Math			FCAT Reading			SAT-9 Reading		
	2002	2003	2004	2002	2003	2004	2002	2003	2004	2002	2003	2004
3rd Grade	302.1	307.7	310.7	58.9	62.2	63.6	293.2	298.5	303.2	55.4	58.4	59.0
	(66.7)	(67.3)	(65.8)	(28.4)	(27.6)	(27.0)	(65.9)	(62.7)	(64.2)	(28.5)	(27.7)	(27.2)
	[187093]	[187521]	[198800]	[185288]	[185511]	[196887]	[186934]	[187376]	[198688]	[185072]	[185360]	[196803]
4th Grade	293.9	297.9	312.5	59.3	61.0	65.9	299.7	305.3	318.3	55.3	56.1	60.6
	(63.3)	(63.4)	(58.4)	(27.9)	(27.3)	(24.8)	(63.2)	(60.3)	(51.0)	(27.7)	(27.2)	(25.1)
	[189796]	[192166]	[166250]	[188997]	[189859]	[164531]	[190173]	[192207]	[166352]	[188488]	[189727]	[164318]
5th Grade	318.3	320.0	322.3	58.5	59.4	59.9	284.7	290.5	294.6	51.8	53.8	54.4
	(57.9)	(59.2)	(57.4)	(28.7)	(28.3)	(27.8)	(62.7)	(60.6)	(62.3)	(27.7)	(27.5)	(27.2)
	[191148]	[191555]	[185200]	[190008]	[189429]	[182938]	[191545]	[191742]	[185039]	[190766]	[189597]	[183070]

Notes: Standard deviations in parentheses; number of students tested in brackets.

Table 2: Distribution of 2002 School Grades of Elementary Schools by 2001 School Grade

2002 Grade	2001 Grade						Total
	A	B	C	D	F	No grade assigned	
A	235	199	136	5	0	12	587
B	90	69	158	15	0	7	339
C	32	36	250	85	0	7	410
D	0	0	45	63	0	5	113
F	0	0	5	28	0	2	35
Total	357	304	594	196	0	33	1484

Table 3: 'F'-Grade/Voucher-Threat Effect on FCAT Achievement, 2003

	1	2	3	4
	Model I	Model II	Model III	Model IV
Student-level Controls	Yes	Yes	Yes	Yes
School-level Controls	None	No resources	All	All
Sample	All students	All students	Same school both years	All students
F/Threat Effect	0.042*	0.050**	0.060**	0.042*
	(0.024)	(0.025)	(0.026)	(0.024)
N	21531	21531	17114	21531
[Schools]	[125]	[125]	[125]	[125]

Notes: Significant at *10%; **5%; ***1%. Dependent variable is average of FCAT scores in math and reading standardized by grade, year, and subject; standard errors adjusted for clustering by school are in parentheses. Student-level controls in all models control for whether the student is White, African American, Hispanic, or of another ethnicity, gender, special education status, English Language Learner status, eligibility for the federal free/reduced-price lunch program, first year in a new school, grade repeater status, a cubic in previous FCAT test scores in math and reading, and Stanford-9 national percentile rank scores in math and reading. Model II includes school-level aggregate measures of each of these variables for the 4th and 5th grade students tested in the school in 2003; Models III and IV also include per pupil operating costs and pupil-teacher ratio. Model III includes only those students who attended the same school the previous year. Treated schools for each model are schools who received an 'F' in 2003 but would have received a higher grade under the 2001 grading scheme. Control schools defined as in Comparison I; see text for details.

Table 4: 'F'/Threat and New-'D' Effects on Achievement: Robustness Checks and SAT-9 Results

	1	2	3		4	5	6
	Comparison I	Comparison II	Comparison III		Comparison I	Comparison II	Comparison III
Test		FCAT				FCAT	
'F'/Threat	0.042*	0.054**	0.053**	New 'D'	0.038***	0.041***	0.056**
Effect	(0.024)	(0.026)	(0.025)	Effect	(0.014)	(0.014)	(0.023)
N	21531	13378	4452	N	66750	40202	6373
[Schools]	[125]	[80]	[28]	[Schools]	[346]	[212]	[35]
		SAT-9				SAT-9	
'F'/Threat	0.031	0.020	0.060	New 'D'	0.020	0.018	0.002
Effect	(0.025)	(0.027)	(0.037)	Effect	(0.014)	(0.014)	(0.026)
N	21258	13199	4373	N	65945	39706	6300
[Schools]	[125]	[80]	[28]	[Schools]	[346]	[212]	[35]

Notes: Significant at *10%; **5%; ***1%. Standard errors adjusted for clustering by school are in parentheses. For FCAT dependent variable and control variables, see notes to Table 2 (Model IV). Dependent variable for SAT-9 analysis is the average of the student's national percentile ranking in math and reading. Treated schools for 'F'/Threat effect defined as in Table 3; treated schools for New-'D' effect are the 30 highest performing schools that received a 'D' in 2002 after receiving a higher grade in 2001. See text for details of alternate control groups.

Table 5: 'F'-Grade/Voucher-Threat and New School Grade Effects on FCAT Achievement, 2003

	1	2	3	4	5
	New-'F'/Threat vs. D	New-'D' vs. 'C'	New-'C' vs. 'B'	New-'B' vs. 'A'	New-'A' vs. 'B'
	FCAT Combined Math/Reading				
Grade/Sanction Effect	0.049*	0.038***	-0.006	0.000	0.007
	(0.027)	(0.014)	(0.015)	(0.010)	(0.019)
N	22081	66750	42098	111389	26905
[Schools]	[129]	[346]	[196]	[466]	[132]

Notes: Significant at *10%; **5%; ***1%. Standard errors adjusted for clustering by school are in parentheses. For dependent variable and control variables, see notes to Table 2 (Model IV). Treated schools for each model are the 30 highest (lowest for New-'A' analysis) performing schools who received the grade listed in 2002 after receiving a higher grade in 2001. Control schools defined as in Comparison I; see text for details.

Table 6: Effects of NCLB Sanctions on FCAT Achievement, 2004, by Title I Eligibility

	1	2	3	4
	Model I	Model II	Model III	Model IV
Student-level Controls	Yes	Yes	Yes	Yes
School-level Controls	None	No resources	All	All
Sample	All students	All students	Same school both years	All students
AYP Effect (Title 1 Ineligible Schools)	0.002	-0.001	0.005	0.002
	(0.015)	(0.017)	(0.019)	(0.019)
N	27400	27400	24182	26779
[Schools]	[118]	[118]	[113]	[113]
AYP + Public School Choice Effect (Title I Eligible Schools)	0.008	-0.003	-0.003	-0.003
	(0.019)	(0.019)	(0.020)	(0.019)
N	11565	11565	9824	11513
[Schools]	[62]	[62]	[61]	[61]

Notes: Standard errors adjusted for clustering by school are in parentheses. For dependent variable and control variables, see notes to Table 2 (Model IV). Treated schools for each model are the 30 highest performing schools that did not make adequate yearly progress in 2003. Control schools defined as in Comparison I; see text for details. School resource variables are measured in 2003.

Table A1: Correlation between FCAT Scale Scores and Stanford-9 National Percentile Rankings, by grade, subject, and year

	Math			Reading		
	2002	2003	2004	2002	2003	2004
3rd Grade	0.84	0.84	0.84	0.84	0.84	0.83
4th Grade	0.82	0.81	0.79	0.82	0.82	0.79
5th Grade	0.84	0.83	0.83	0.84	0.84	0.83

Table A2: Descriptive Statistics of Treatment and Control Groups for 'F'/Threat Analysis

	1	2	3	4	5	6	7	8	9
	Comparison I			Comparison II			Comparison III		
	Treated [N=24]	Control [N=101]	Treated-Control [p-value]	Treated [N=24]	Control [N=56]	Treated-Control [p-value]	Treated [N=9]	Control [N=19]	Treated-Control [p-value]
2002 FCAT Test Scores, Standardized	-0.67	-0.55	-0.12	-0.67	-0.68	0.01	-0.60	-0.67	0.07
	(0.17)	(0.18)	[0.00]	(0.17)	(0.14)	[0.78]	(0.12)	(0.22)	[0.15]
% African American	79.2	59.5	19.7	79.2	67.8	11.4	79.1	71.4	7.7
	(19.0)	(27.4)	[0.01]	(19.0)	(25.2)	[0.05]	(19.7)	(23.0)	[0.40]
% Hispanic	11.4	22.5	-11.1	11.4	24.1	-12.7	14.2	21.4	-7.2
	(12.8)	(22.3)	[-0.02]	(12.8)	(23.3)	[0.01]	(15.7)	(20.4)	[0.38]
% White	7.5	15.0	-7.5	7.5	6.1	1.4	5.2	4.8	0.4
	(9.7)	(18.9)	[0.06]	(9.7)	(9.3)	[0.54]	(7.6)	(7.0)	[0.90]
% Free Lunch	90.0	87.2	2.8	90.0	93.5	-3.5	89.7	89.3	0.4
	(10.5)	(12.3)	[0.31]	(10.5)	(5.1)	[0.05]	(5.9)	(5.8)	[0.87]
% Special Ed	20.9	16.8	4.1	20.9	17.0	3.9	20.4	17.6	2.8
	(7.3)	(6.6)	[0.01]	(7.3)	(6.5)	[0.02]	(6.3)	(7.0)	[0.34]
% LEP	9.4	15.1	-5.7	9.4	18.4	-9.0	12.1	17.5	-5.4
	(10.5)	(13.2)	[0.05]	(10.5)	(14.5)	[0.01]	(14.2)	(11.7)	[0.33]
% New School	22.5	18.2	4.3	22.5	18.6	3.9	19.1	18.3	0.8
	(6.2)	(6.3)	[0.00]	(6.2)	(7.2)	[0.02]	(6.7)	(6.0)	[0.77]
% Repeater	6.1	3.6	2.5	6.1	3.6	2.5	5.0	5.0	0.0
	(5.5)	(3.6)	[0.01]	(5.5)	(3.9)	[0.02]	(2.9)	(4.7)	[1.00]

Notes: Averages of school characteristics weighted by the number of 4th and 5th grade students tested in 2003. Weighted standard deviations in parentheses. P-value of t-test of difference in prior test scores between treated and control schools in brackets.

Table A3: Descriptive Statistics of Treatment and Control Groups for New-'D' Analysis

	1	2	3	4	5	6	7	8	9
	Comparison I			Comparison II			Comparison III		
	Treated [N=30]	Control [N=317]	Treated-Control [p-value]	Treated [N=30]	Control [N=182]	Treated-Control [p-value]	Treated [N=14]	Control [N=21]	Treated-Control [p-value]
2002 FCAT Test Scores, Standardized	-0.29	-0.20	-0.09	-0.29	-0.30	0.01	-0.28	-0.29	0.01
	(0.12)	(0.12)	[0.00]	(0.12)	(0.07)	[0.52]	(0.08)	(0.07)	[0.90]
% African American	40.4	31.7	8.7	40.4	35.6	4.8	40.2	44.4	-4.2
	(27.6)	(24.7)	[0.07]	(27.6)	(27.0)	[0.37]	(32.3)	(32.9)	[0.72]
% Hispanic	12.3	20.6	-8.3	12.3	25.0	-12.7	11.1	12.0	-0.9
	(12.2)	(21.8)	[0.04]	(12.2)	(24.6)	[0.01]	(10.3)	(15.8)	[0.85]
% White	43.8	43.4	0.4	43.8	35.3	8.5	45.5	40.4	5.1
	(31.6)	(25.3)	[0.94]	(31.6)	(23.8)	[0.09]	(33.1)	(27.2)	[0.62]
% Free Lunch	73.5	67.6	5.9	73.5	74.7	-1.2	73.5	71.7	1.8
	(15.0)	(18.9)	[0.10]	(15.0)	(14.3)	[0.67]	(16.0)	(20.3)	[0.78]
% Special Ed	15.4	16.9	-1.5	15.4	17.5	-2.1	17.3	17.5	-0.2
	(6.1)	(6.7)	[0.24]	(6.1)	(6.9)	[0.12]	(3.6)	(6.0)	[0.91]
% LEP	8.6	11.4	-2.8	8.6	15.0	-6.4	8.9	7.9	1.0
	(9.4)	(12.7)	[0.24]	(9.4)	(14.6)	[0.02]	(10.7)	(10.4)	[0.78]
% New School	20.8	16.6	4.2	20.8	17.6	3.2	16.3	15.4	0.9
	(11.1)	(6.5)	[0.00]	(11.1)	(7.0)	[0.04]	(5.7)	(6.0)	[0.66]
% Repeater	3.1	2.7	0.4	3.1	2.9	0.2	3.0	3.2	-0.2
	(2.5)	(2.6)	[0.42]	(2.5)	(2.8)	[0.71]	(2.5)	(3.2)	[0.85]

See notes to table A2.

Table A4: Descriptive Statistics of Treatment and Control Groups for NCLB Analysis

	1	2	3	4	5	6
	Title I Ineligible (Comparison I)			Title I Eligible (Comparison I)		
	Treated [N=30]	Control [N=88]	Treated-Control [p-value]	Treated [N=30]	Control [N=32]	Treated-Control [p-value]
2002 FCAT Test Scores, Standardized	0.62	0.63	-0.01	0.33	0.36	-0.03
	(0.07)	(0.07)	[0.51]	(0.05)	(0.06)	[0.04]
% African American	7.1	7.9	-0.8	14.1	9.7	4.4
	(7.1)	(7.5)	[0.61]	(9.0)	(11.6)	[0.10]
% Hispanic	14.3	10.8	3.5	9.7	6.9	2.8
	(16.1)	(11.7)	[0.20]	(16.1)	(7.6)	[0.38]
% White	72.9	74.9	-2.0	70.7	70.8	-0.1
	(18.1)	(16.0)	[0.57]	(18.4)	(15.5)	[0.98]
% Free Lunch	14.2	15.3	-1.1	44.0	43.5	0.5
	(6.4)	(8.7)	[0.53]	(9.8)	(8.0)	[0.83]
% Special Ed	12.1	12.1	0.0	14.8	16.5	-1.7
	(3.5)	(4.2)	[1.00]	(4.6)	(3.6)	[0.11]
% LEP	3.9	2.6	1.3	2.5	1.2	1.3
	(5.6)	(4.9)	[0.23]	(4.5)	(1.6)	[0.13]
% New School	7.7	9.2	-1.5	14.6	11.8	2.8
	(3.0)	(8.3)	[0.34]	(9.7)	(3.7)	[0.13]
% Repeater	1.0	0.7	0.3	2.9	2.4	0.5
	(1.2)	(0.8)	[0.12]	(2.4)	(2.6)	[0.44]

See notes to table A2.

Table A5: 'F'-Grade/Voucher-Threat and New-'D' Grade Effects, 2003, by Ethnicity and SES

Sample	1 African Americans	2 Hispanics	3 Whites	4 Free Lunch Eligible	5 Ineligible

Table A7: Student Achievement in Florida, 2002-2004

	1 FCAT Math	2 FCAT Reading	3 SAT-9 Math	4 SAT-9 Reading
2002 (Omitted)	--	--	--	--
2003	3.57	5.49	1.89	1.90
	(0.10)	(0.10)	(0.05)	(0.04)
2004	9.29	11.18	3.71	3.30
	(0.10)	(0.10)	(0.05)	(0.04)
Male	1.45	-3.87	1.59	-2.71
	(0.59)	(0.08)	(0.04)	(0.04)
African American	-33.40	-29.75	-15.33	-14.83
	(0.11)	(0.11)	(0.05)	(0.05)
Hispanic	-8.26	-12.09	4.22	-5.68
	(0.13)	(0.12)	(0.06)	(0.05)
Asian	17.44	6.39	6.75	3.34
	(0.31)	(0.30)	(0.14)	(0.14)
Other non-white	-5.20	-2.48	-2.08	-1.60
	(0.26)	(0.25)	(0.11)	(0.11)
Free/Reduced Price	-24.53	-24.92	-10.67	-11.96
lunch	(0.09)	(0.09)	(0.04)	(0.04)
Eng. Lang. Learner	-27.49	-33.60	-11.67	-12.99
	(0.15)	(0.15)	(0.07)	(0.07)
Special Education	-52.54	-55.52	-22.10	-21.08
	(0.12)	(0.11)	(0.05)	(0.05)
N	1689696	1689644	1673037	1672794

Notes: Dependent variable is FCAT scale score or SAT-9 national percentile rank; standard errors are in parentheses. All coefficients are statistically significant at the 1% level.

Appendix

The Florida A+ Plan, as revised and fully implemented in 2002, acted as a shock on Florida's elementary schools, both in general and, more particularly, on those given a grade or evaluation they would not have received under the prior accountability system. In 2003, when NCLB's accountability system, which found some schools not making Adequate Yearly Progress (AYP), also acted as an external shock. In this Appendix, we provide greater details on the magnitude and timing of these accountability shocks.

Florida A+ Plan.

The modified grading system first used to assign school grades under the Florida A+ Accountability Plan in the summer of 2002 was difficult for schools to anticipate. The new system was not approved by the governor until December 2001, just a few months before students were given the tests that would become the basis of the grades schools received the following summer.

Before the A+ Plan was revised in 2002, no one student was tested in the same subject on the Florida Comprehensive Accountability Test (FCAT) two years in a row, making it impossible to ascertain how much students at any given school had learned during the school year the test was given. In the absence of this information, schools in Florida received a grade, A through F, simply on the basis of the achievement levels attained by students in grades 4 (in reading) or 5 (in math), 8 and 10. For a school to receive an 'A' or a 'B' 50 percent or more of the students at that school had to score in both reading and math at a performance Level 3 on the FCAT, the level at which a student was deemed proficient. (Performance Levels ranged from 1 to 5.) In writing, two-thirds of the students had to perform at this level. 'C's' were awarded to those schools where 60 percent of the students attained Level 2 in reading and math and 50 percent of the students achieved that level in writing. 'D's' were given to schools that missed the requirement in one or two of the subjects. An 'F' was assigned to those who did not reach the minimum in any subject. Other criteria were also considered, including the percentage of students that were tested. To get an 'A', 95 percent of eligible students had to be tested. However, the primary criteria for the determination of grades had to do with the percentage of students scoring above a certain threshold on the three components

of the FCAT. Since levels of achievement are affected not only by school quality but also by family background characteristics, the grades schools received under this old system were highly correlated with the demographic characteristics of the students.

In Spring 2001, new legislation required that A+ take advantage of the fact that students were now being tested in math, reading and writing in all grades, 3-10, to include annual learning gains as a component of Florida's grading system. The revised grading system was approved by the governor in December 2001, just a few months before tests were to be given upon which schools would receive their new grades. The new grading system gives as much as a 50 percent weight to learning gains on a 600 point scale used to calculate a school's grade. A school can attain a maximum of 200 points on this scale, depending upon the percentage of students making learning gains in reading and math. A gain is defined as improving by one performance level, making more than a full year's learning growth, or by maintaining the same performance level, if it is Level 3 or higher. A school can earn another maximum of 100 points, based on the percentage of its lowest performing students (the bottom 25 percent of the school's test takers in reading) making learning gains (as defined above) in reading. A school can receive a maximum of 300 points based upon the percentage of its students achieving Level 3 or higher in reading and math and, in writing, the average of the percentage reaching Level 3.0 or higher and the percentage attaining Level 3.5. To receive an 'A', the school must achieve 410 points; to receive a 'B,' it must receive 380 points; a 'C', 320 points; 'D', 280 points; otherwise an 'F.' 'A' schools must also show that at least half of their lowest performing students have made a year's worth of learning gains, and they must test 95 percent of their students. Otherwise, schools, to receive a grade must test 90 percent of their students and have at least thirty students who have been tested in two consecutive years in both reading and math.

No Child Left Behind

Under No Child Left Behind (NCLB), schools, in order to make AYP, ordinarily must show that the percentage of students achieving a state-determined level of proficiency has risen by an increment large enough that, if the rate is sustained, all students can be expected to be proficient by 2014. The school must also show that the

percentage of students within various subgroups (defined by ethnicity, food-stamp eligibility, English language learning status, and in need of special education) is also increasing at the required rate. In Florida, proficiency is defined as scoring at Level 3 on the FCAT, a standard that is somewhat higher than the one established by the typical state. Ninety-five percent of all students must be tested. Certain exemptions from the rule are allowed for schools with low-performing students, provided the school is showing substantial progress toward achieving proficiency. Schools that do not make AYP for two years in succession are said to be in need of improvement, and students are then given the opportunity to attend another public school within the school district, provided that that school is not also in need of improvement.

In some states, including Florida, a school is designated as in need of improvement only if it is a Title I school, that is, a school receiving Title I services. (Florida requires only that districts serve all schools where 75 percent or more of the students receive free or reduced price lunch. Districts have discretion over which other schools will be served; most serve all those schools where the percentage of students receiving free or reduced price lunch is above the district average.) The rationale for limiting the application of the "in need of improvement" label to Title I schools is based upon the fact that NCLB is simply an amended reauthorization of Title I of the Elementary and Secondary Education Act (ESEA) of 1965, which created a compensatory education program for schools that served disadvantaged students. AYP is determined and reported for all schools, however, regardless of their Title I status.