CALDER

National
Center for Analysis of Longitudinal Data in Education Research

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

Urban Institute

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Feeling the Florida Heat?

How Low-Performing Schools Respond to Voucher and Accountability Pressure

Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio

WORKING PAPER 13 • NOVEMBER 2007

Comments Welcome

FEELING THE FLORIDA HEAT?
HOW LOW-PERFORMING SCHOOLS RESPOND TO VOUCHER AND ACCOUNTABILITY PRESSURE

Cecilia Elena Rouse
Princeton University and NBER

Jane Hannaway
The Urban Institute

Dan Goldhaber
University of Washington

David Figlio
University of Florida and NBER

November 28, 2007

This research could not have been undertaken without the help of many people. We thank Edward Freeland, Craig Deshenski, Kenneth Mease, Rob Santos, and Fritz Scheuren for their exceptional help with the survey and sample development. We appreciate the assistance of the Florida Department of Education in providing us both with administrative student data as well as with sampling frames for our survey analysis. Jay Pfeiffer, Jeff Sellers and others at the Florida Department of Education provided very helpful advice regarding Florida education policy and the administrative data used in the analysis. We also thank Emily Buchsbaum, Cynthia Casazza, Sarah Cohodes, Joseph Gasper, Scott Mildrum, Radha Iyendar, Ty Wilde, Grace Wong, and Nathan Wozny for expert research assistance and Jesse Rothstein, Analia Schlosser, Diane Whitmore Schanzenbach and seminar participants at Harvard University, McMaster University, the University of Chicago, the University of Florida, and the fall 2007 research conference of the National Center for Analysis of Longitudinal Data in Education Research for extremely useful conversations and suggestions. Finally, we are indebted to the Annie E. Casey, Atlantic Philanthropies, Smith Richardson and Spencer Foundations, the U.S. Department of Education, the National Institutes of Health and the National Center for Analysis of Longitudinal Data in Education Research (CALDER is supported by IES Grant R305A060018 to the Urban Institute) for financial support, but the views expressed in this paper do not necessarily represent those organizations supporting this research, the Florida Department of Education, or our host institutions. All errors in fact and interpretation are ours.

ABSTRACT

While numerous recent authors have studied the effects of school accountability systems on student test performance and school "gaming" of accountability incentives, there has been little attention paid to substantive changes in instructional policies and practices resulting from school accountability. The lack of research is primarily due to the unavailability of appropriate data to carry out such an analysis. This paper brings to bear new evidence from a remarkable five-year survey conducted of a census of public schools in Florida, coupled with detailed administrative data on student performance. We show that schools facing accountability pressure changed their instructional practices in meaningful ways. In addition, we present medium-run evidence of the effects of school accountability on student test scores, and find that a significant portion of these test score gains can likely be attributed to the changes in school policies and practices that we uncover in our surveys.

I. Introduction

The current national focus on performance-based accountability in K-12 education began to develop in the early 1990s in response to dissatisfaction with the performance of U. S. schools despite substantial increases in funding (Hanushek, 1994). A conference organized by the National Research Council identified the lack of performance incentives in education as the main culprit (Hanushek and Jorgenson, 1996). Two views emerged about that time on how to introduce performance incentives into education. The first was captured by the work of Chubb and Moe (1990) who argued that problems of academic performance result from the regulation of schools by public bureaucrats who respond to the interests of organized groups and not to interests of students and parents. The solution they proposed was school autonomy, governed by market mechanisms.

The second view, developed by Smith and O'Day (1991), argued that the problem with U. S. education was that the system was organized in ways that put it at odds with itself. Performance standards, curriculum and student assessment were not purposefully integrated and federal and state policies often worked at cross purposes. The solution they argued was alignment. In the early stages, much of the emphasis of the standards movement, as it was called, was on establishing performance standards and aligning the curriculum with those standards. With the exception of a few states and some large school districts, developing new assessments and holding schools accountable for tested student outcomes proceeded slowly (Hannaway and Kimball, 2001). The passage of *No Child Left Behind* (NCLB) spurred the development of test-based accountability systems as the linchpin of standards-based reform.

Market-based systems attempt to provide more accountability by increasing the educational choices available to parents. Most states and school districts now provide increased

options through open enrollment plans or through the provision of independently-operated charter schools. In a handful of places, the market-based system has gone a step further allowing parents, through a voucher system, to send their children to private school at public expense. Test-based incentives include attempts to increase school accountability through the regular testing of students, making the results – aggregated to the school level – public, and rewarding schools with high or increasing aggregate test scores, and imposing sanctions on poor performing schools. Given that accountability ratings are capitalized into housing values (Figlio and Lucas, 2004), constituents have both educational and financial incentives to pressure poor-performing schools to improve. Both the market solution and test-based solution should provide schools with an incentive to operate more efficiently – that is, to improve student outcomes without significant increases in resources.

Previous authors have investigated the short-run effects of both market-based and test-based accountability on student outcomes, reaching mixed conclusions.[1] Relatively few have investigated the actual behavioral responses of schools to these systems. Some of those that do find that schools "game the system" to improve test performance without necessarily increasing effort or true productivity. For example, there is some evidence that schools respond to accountability pressure by differentially reclassifying low-achieving students as learning disabled so that their scores will not count against the school in accountability systems (see, e.g., Cullen and Reback (2007), Figlio and Getzler (2007), Jacob (2004))[2], and Figlio (2006) indicates

[1] Recent nationwide studies by Carnoy and Loeb (2002) and Hanushek and Raymond (2005) find significant improvement in student outcomes as a result of standards-based accountability, whereas the results from some specific state systems have been less positive (see, e.g, Koretz and Barron (1998), Clark (2003) and Haney (2000, 2002)). More recent work by Figlio and Rouse (2006), West and Peterson (2006), and Chakrabarti (2006) find positive short-run effects of accountability on Florida student outcomes, at least in mathematics. The evidence on market-based reforms is also mixed (see, e.g., Howell and Peterson (2002), Krueger and Pei (2004), and Rouse (1998) for evidence from voucher programs).

[2] Chakrabarti (2006), however, does not find that schools respond in this way.

that schools differentially suspend students at different points in the testing cycle in an apparent attempt to alter the composition of the testing pool.[3] Jacob and Levitt (2003) find that teachers are more likely to cheat when faced with more accountability pressure.

Surprisingly, there has been little systematic effort to determine the substantive ways in which schools alter their methods of delivering education in response to school accountability and school choice pressures (see Hannaway and Hamilton, 2007, for a review). This is an important oversight, as some school policy variables have been found to improve student test performance. For example, there is convincing evidence that reducing class size can improve achievement (see, e.g., Angrist and Lavy, 1999; Krueger, 1999; and Krueger and Whitmore, 2001) as can summer school and grade retention (Jacob and Lefgren, 2004).

Does one observe that schools begin to adopt these or other measures in an attempt to improve student outcomes? The answer to this question is of key import for several reasons. First, if one observes gains in student outcomes without changes in how schools operate, one might be suspicious that the gains are not due to improving efficiency of the schools but perhaps to other factors, such as the potentially artificial gains (e.g. changing the tested student population) described above. Second, most students are likely to remain in public schools regardless of the arrangements designed to provide K-12 education. Thus, the majority of the nation's 50 million students will be affected mainly though the institutional response of public schools to these, or other, pressures.

The primary reason for the lack of research on this topic is the lack of data. While information on student test scores and a few crude indicators of student body characteristics (e.g., race/ethnicity, eligibility for the National School Lunch Program, and disability rates) are

[3] Schools may also try to boost student achievement on testing days. Figlio and Winicki (2005), for instance, suggest that Virginia schools facing accountability pressures altered their school meals on testing days to increase the likelihood that students will do well on the exams.

prevalent, there exist no large-scale data sets that systematically describe school instructional policies and practices. In this paper we describe research that addresses the gap in research on school responses to the pressures introduced by school accountability and school choice systems. We do so using an extraordinary data set we have generated. During the 1999-00, 2001-02, and 2003-04 school years, we administered surveys to every principal of every regular public school in the state of Florida. Our surveys, for which we consistently exceeded a 70 percent response rate, addressed a range of questions on the delivery of instruction in public schools in Florida.

At the same time, beginning in summer 1999, Florida's accountability system – the A+ Plan for Education – enlisted stigma (the grading of schools on an "A" to "F" scale), oversight (by the state of Florida), and competition to spur school improvement. The competitive pressure of the system was derived from the Opportunity Scholarship Program: under this system, students attending schools that received an "F" grade in multiple years became eligible for a voucher that allowed them to attend a private or higher-rated public school.[4] While in effect for nearly 7 years, the Opportunity Scholarship Program was declared unconstitutional by Florida's Supreme Court in January 2006. Other components of the A+ Plan, however, remain in effect.

We analyze the impact of the accountability system on Florida's students and schools using a three-part analysis. First, we estimate the effect of the accountability system and the threat of becoming voucher eligible on student test score performance, both in the short-run and in the longer term. Second, we study the effects of the reform on school policies and practices. Finally, we attempt to determine if the policies appear to affect student achievement or explain the change in student performance. We find that student achievement significantly increased in elementary schools that received an "F" grade by between 6 to 14 percent of a standard deviation

[4] Currently Florida has two other voucher programs as well: an income tax credit for corporations to fund vouchers for low-income students and the McKay Scholarship for students with exceptionalities. In our analysis we control for the factors on which eligibility is based for these programs.

in math and between 6 to 10 percent of a standard deviation in reading in the first year. Three years later the impacts persist.

Importantly, we also detect specific school policy changes implemented by the schools that explain part of these increases. Specifically, when faced with increased accountability pressure, schools appear to focus on low-performing students, lengthen the amount of time devoted to instruction, adopt different ways of organizing the day and learning environment of the students and teachers, increase resources available to teachers, and decrease principal control. These, combined with other policies, explain more than 15 percent of the test scores gains of students in reading and over 38 percent of the test scores gains of students in math, depending on the model specification. As such we find evidence that schools respond to accountability pressure in educationally meaningful ways.

II. The Florida School Accountability Program

Florida's 1999 A+ Plan for Education introduced a system of school accountability with a series of rewards and sanctions for high-performing and low-performing schools. The A+ Plan called for annual curriculum-based testing of all students in grades three through ten, and annual grading of all public and charter schools based on aggregate test performance. As noted above, the Florida accountability system assigns letter grades ("A," "B," etc.) to each school based on students' achievement (measured in several ways). High-performing and improving schools receive rewards while low-performing schools receive additional assistance as well as sanctions.

The assistance provided to low-performing schools primarily consists of three components. First, each school district with a school receiving a "D" or "F" is evaluated by a "community assessment team" comprised of local and state leaders, including parents and

business representatives. In theory this team makes recommendations to the state and the local school board on how to improve the local schools. In addition, the Florida Department of Education (FDOE) makes available technical assistance – with needs assessments and the implementation of school improvement plans – to *any* school in the state, with priority given to "D" and "F" schools (as well as those in rural or sparsely populated areas).[5]

However, the most visible form of assistance available for low-performing schools is through Florida's *Assistance Plus* program. While this program offers no direct funding, through the State Board of Education policies are in place that mandate that districts allocate certain resources and targeted funding for these schools. In addition, "F" schools are given priority for the *Just Read, Florida!* program that provides reading coaches trained in scientifically-based reading research to the lowest performing schools (the program was funded at $11 million in 2001-02). The *Assistance Plus* program also consists of a smattering of other forms of assistance such as "Assistance Plus Teams" (teams of experts to help structure improvement efforts); state-coordinated regular conference calls and reports designed to provide continuous progress reporting and feedback to schools and districts; alignment of curriculum, assessment and course materials; analysis of student achievement data to provide technical assistance for schools and instructors; and recommendations regarding professional development for teachers.[6]

While "F" schools receive priority through Florida's *Assistance Plus* program, "D" schools were also targeted. In fact, in the 2007-2008 academic year (the only year for which we were able to obtain this information), 79 "F" schools and 220 "D" schools will receive such assistance and these numbers are quite close to the numbers of potentially-eligible "F" and "D"

[5]The 2002 Florida Statute, Title XLVIII ("K-20 Education Code"), Chapter 1008.345 (Assessment and Accountability; Implementation of state system of school improvement and education accountability). (Downloaded from http://www.leg.state.fl.us/statutes on November 16, 2007).

[6]These statements are based on documents received from the FDOE regarding the program in 2001-2002 (they can also be accessed at http://schoolgrades.fldoe.org/pdf/0102/assist_plus.pdf).

schools. As such, we suspect that while "F" schools receive priority, "D" schools heavily benefit from this state aid as well.[7]

On the sanction side, the most controversial and widely publicized provision of the A+ Plan was the institution of vouchers, called "Opportunity Scholarships," for students attending (or slated to attend) chronically failing schools – those receiving a grade of "F" in two years out of four, including the most recent year. These "Opportunity Scholarships" allowed students to attend a different (higher rated) public school, or an eligible private school. And while poor-performing schools received additional assistance, they also were subject to additional scrutiny and oversight. All "D" and "F"-graded schools are subject to site visits and required to send regular progress reports to the state.[8]

School grading began in May 1999, immediately following passage into law of the A+ Plan. Between 1999 and summer 2001, schools were assessed primarily on the basis of aggregate test score *levels* (and also some additional non-test factors, such as attendance and suspension rates, for the higher grade levels) and only in the grades with existing statewide curriculum-based assessments[9], rather than on the *progress* schools made toward higher levels of student achievement.

Starting in summer 2002, however, school grades began to incorporate test score data from all grades from three through ten and to evaluate schools not just on the level of student test performance but also on the year-to-year progress of individual students. However, while at the beginning of the 2001-02 school year several things were known about the school grades that

[7]We note, as well, that in documents referenced above the FDOE highlights assistance given to "D" as well as "F" schools, reinforcing our belief that "D" schools heavily benefit from such assistance.
[8]Details on oversight and reporting requirements can be found online at http://www.bsi.fsu.edu/PerformanceUpdates/performanceupdates.aspx.
[9]Students were tested in grade 4 in reading and writing, in grade 5 in mathematics, in grade 8 in reading, writing and math, and in 10 in reading, writing and math.

were to be assigned in summer 2002 (school grades were to be based on test scores from all students in all tested grades; the standards for proficiency in reading and mathematics were to be raised; and school grades would incorporate some notion of student learning gains into the formula) the specifics of the formula that would put these components together to form the school grades was unknown. Indeed, the new school grading formula establishing a system of "grade points" with thresholds for letter grades was not announced until March 5, 2002 at the same time that the students were being tested on the Florida Comprehensive Assessment Test (FCAT), leaving schools with virtually no information with which to anticipate their exact grade. A key component of our analytic strategy is to take advantage of the fact that during the 2001-02 school year schools could not necessarily anticipate their school grade in summer 2002 because the specific changes in the grading formula were not decided until the last minute.

Table 1 shows the distribution of schools across the five performance grades for the first six rounds of school grading, for all graded schools in Florida.[10] As is apparent from the variation across years in the number of schools that fall into each performance category, there are considerable grade changes that have taken place since the accountability system was adopted. Most notable is the fact that while 70 schools received an "F" grade in the first year (1998-99) only 4 did so the subsequent year and none did in summer 2001. At the same time, an increasing number of schools were receiving "A"s and "B"s. This is partly due to the fact that schools had learned their way around the system: a school had to fail in all three subjects to earn an "F" grade so as long as students did well enough in at least one subject the school would escape the worst stigma. Goldhaber and Hannaway (2004) and Chakrabarti (2006) find evidence that

[10] In the analysis that follows, we limit the sample to *elementary* schools, which comprised over 50 percent of "F"-graded schools. We do so primarily because our survey asked many questions that focused on elementary schools. The bottom panel of Table 1 presents the change in grade distribution over time for the elementary schools in Florida.

students in failing schools made the biggest gains in writing, which is viewed as one of the easier subjects in which to show improvement quickly. When the rules of the game changed, so did the number of schools caught by surprise. For example, the number of schools earning an "F"-grade increased to 60 in summer 2002, and the number of "A" schools grew as well, largely because the "grade points" system of school grading allowed schools that miss performance goals in one area to compensate with higher performance in another.[11]

A way to judge the extent to which schools might have been caught "off guard" by the new system is to compare the grades that schools actually received in summer 2002 with the grade that they would have been predicted to receive in 2002 based on the "old" grading system (that in place in 2001). In Table 2 we make this comparison for all graded elementary schools using the full set of student administrative test scores provided us by the FDOE.[12] It is possible to make these calculations because the summer 2001 grading system's formula is known. We estimate that about 52 percent of elementary schools experienced no change in their school grade based on the change in the accountability system. The other 48 percent either received a higher or lower grade than they might have "expected." For example, 2 of the 7 schools that might have expected to receive an "F" under the old regime received a "D" under the new one (none of these schools received a higher grade). Importantly, 15 percent of schools that might have expected to receive a "D" under the old system received an "F" under the new one.

We focus our attention on the roughly 2 percent of schools that received an "F" (and therefore became voucher threatened, and faced new stigma and state oversight). Among these

[11]Note that as schools have adapted to the new grading system, the number of "F" schools has also decreased.
[12]The number of observations in Table 2 does not exactly match that in Table 1 because the administrative data on students provided by the FDOE used to simulate each school's grade in 2002 does not include some students in charter schools and "alternative" schools.

35 elementary schools, 6 became voucher eligible.[13,14] Table 3 shows that in 2002 these "F" (elementary) schools were slightly smaller than the higher-rated schools, had a higher proportion of Black students and a higher proportion of students eligible for a free- or reduced-price lunch. However, they also spent slightly more per regular-education student and per special-education student compared to schools that received an "A," "B," or "C" grade. The "F" schools look much more similar, demographically and financially, to those schools that received a "D" grade.[15]

III. Why Might Incentives Improve School Efficiency?

Economic theory would suggest that school superintendents, principals and teachers can produce education more efficiently by using a different mix of inputs; by selecting a different mix of policies. However, why school administrators would have an incentive to make such improvements is less clear. If large numbers of students leave a school, the principals and teachers may be left with the less-motivated (harder-to-teach) students or the school may become so small that it closes and they lose their jobs. Unfortunately, however, this threat is not necessarily so powerful. For example, if dissatisfied students leave for another school, it is possible that the remaining students would be happier, perhaps making the management of the school easier. Further, if take-up of choice options is low, then the risk of job loss due to lack of demand for services is likely to be low as well.[16]

---

[13]Among all public schools, 60 received an "F" in summer 2002 and 8 of them became voucher eligible.

[14]Note that in Table 1 there are 68 elementary schools that received a grade of "N" in 2002. These were new schools in that year. As such, they were not given a formal grade although the state did calculate their accountability points. We therefore impute what their grade would have been. This imputation results in 9 additional "F"-graded schools and 10 additional "D"-graded schools. The results that follow are generally similar if we add a dummy variable controlling for these schools, instead.

[15]This comparison provides additional support for our focus on comparing "F"-graded schools to those receiving a "D" grade which we do, below.

[16]It is difficult to measure the precise take-up rate of the Opportunity Scholarship program because the set of

That said, we believe there are two factors that may sufficiently motivate school administrators to work hard to improve student performance in Florida. First, principals in "F" schools are subject to intensive scrutiny and supervision under the A+ Plan.[17] For example, as discussed earlier, under the *Assistance Plus* Program principals are required to file regular improvement plans and status reports with the state. The long and detailed reports on the "corrective actions" that the school and district are taking to improve the school suggest that, at a minimum, being labeled as a "low-performing school" brings with it many more bureaucratic "headaches" and increased oversight. Principals and teachers are better able to hold off such managerial scrutiny if they increase student achievement. And they may feel that their jobs are more secure if they can show improvement. That said, while the threat of job loss should be a compelling reason for principals to attempt to institute policies that would improve student achievement, Chiang (2007) does not find evidence that principals are, in fact, more likely to lose their jobs following the school's receipt of an "F" grade.

Even if school administrators do not fear the increased oversight or for their jobs, they may experience a loss of utility in at least two other ways that would induce them to make an effort to increase student achievement. First, principals and teachers have chosen a profession where the ideal behavior is to foster the development of children. Presumably conforming to this behavior is part of who they are; it is part of their identity. Evidence suggesting that a principal or teacher is not living up to the ideal would contribute to a loss in what Akerlof and Kranton

---

students who are eligible differs from year to year and depends in some cases on prior take-up. According to personal correspondence with FDOE personnel, Opportunity Scholarship take-up rates among newly-eligible potential participants ranged from five to ten percent, depending on the year, over the course of the program's operations.

[17] An increase in managerial control is likely to occur in many other settings as a result of poor performance as well. Simon (1957) argued that workers in any organization attempt to restrict the "zone of acceptance" that surrounds their position. The zone establishes the boundaries of managerial control that they expect over their behavior as part of their employment contract. Poor performance, however, can increase the boundaries of the zone inviting greater management control and reducing discretion for the employee.

(2005, 2007) refer to as "identity utility." In addition, the public nature of the school grades could lead to a loss of status in the wider community. Indeed, based on interviews and focus groups with principals and teachers, Goldhaber and Hannaway (2004) suggest that principals and teachers considered a school grade of "F" as a social stigma.

IV. Empirical Strategy and Data

A. Empirical Framework

Our empirical framework aims to estimate the effect of having received an "F" grade on student achievement (as measured by test scores) as well as on school policies. We concentrate on the bottom of the school grading distribution because our school surveys were designed primarily to measure the types of school responses that struggling schools might pursue. In our analysis, we drop schools that did not receive any accountability points in summer 2002 (364 schools), one school whose grade was set at "NA," and 7 schools that were not part of our 2002 sampling frame; this leaves a maximum sample size of 1,659 schools.[18]

1. Effects on Student Test Scores

To assess the effect of an "F" grade on Florida's students and schools, we use student-level test scores from the entire state of Florida, to estimate cross-sectional models such as the following:

$$T_{ist} = \alpha + \beta F_{st-1} + \delta(GRADE)_{ist} + \gamma \mathbf{X}_{is} + \gamma' \mathbf{Z}_{st} + \theta f(T_{ist-1}) + \varphi g(POINTS_{st-1}) + \tau_t + \varepsilon_{ist} \qquad (1)$$

[18] Schools that do not have reported accountability points are exempt from the accountability system either because they have too few students for the state to report test scores, even in aggregate form, or because they are a first-year school. The one school with the grade of "NA" had 467 accountability points (which would have earned that school an "A") however the grade, which was originally set in 2002, has been withdrawn by the State.

where F_{st-1} indicates whether a school received an "F" grade in summer 2002, $GRADE_{ist}$ is a vector of dummy variables indicating the student's grade in school, X_{is} is a vector of student characteristics, Z_{st} is a vector of school-level variables (i.e., the school's 2002 "simulated grade" and the school's lagged grades), T_{ist-1} is the student's prior year's test score[19] and ($f(T_{ist-1})$ represents a cubic of this lag), $g(POINTS_{st-1})$ is a cubic in the number of grade points the school received in summer 2002, τ_t is a vector of year effects, and ε_{ist} is a normally distributed error term. The key parameter of interest is β – the change in a student's test score resulting from her school having received an "F."[20] The central assumption of this regression discontinuity approach is that, controlling for $g(POINTS_{st})$, there is little difference in school performance so β identifies the causal impact of a school receiving an "F" grade.[21]

In addition, we compare the performance of a cohort of students affected by the accountability shock to others attending the same school in the previous cohort, estimating models such as the following:

$$T_{isct} = \alpha + \lambda(COHORT_c) + \beta F_{st-1} + \mu(F_{st-1} x COHORT_c) + \gamma X_{is} + \theta f(T_{ist-1}) + \varphi g(POINTS_{st-1}) + \rho(g(POINTS_{st-1}) x COHORT_c) + \phi_S + \varepsilon_{ist} \quad (2)$$

where all notation is as before, but now since we are comparing across multiple cohorts of students, we include a variable (COHORT) reflecting the post-shock cohort, as well as a vector of school fixed effects ϕ_s. A comparison between the first and second cohorts of students in our

[19] Because we control for lagged test scores, we also control for lags of all student background variables. In practice, the results are unaffected by our decision to include or exclude the lags of these background variables, but we include them for completeness. Further, in some model specifications, we control instead for two- or three-year lags of student test scores.

[20] We have also estimated models in which we constrain the effect of the lagged test score to equal 1 (by using the change in test scores as the dependent variable). The results are qualitatively similar, with comparable levels of statistical significance and point estimates that differ from those presented at the level of the second significant digit. These results are available from the authors on request.

[21] Later in the paper, we describe a variety of specification checks and falsification exercises that we conduct to assess the robustness of our results. We also note that because we have multiple observations for each school, we adjust our standard errors for clustering at the school level.

model provides another way to estimate a plausibly causal impact of receipt of an "F" grade on student test score improvements since the first cohort of students would have graduated to middle school before their elementary school would have responded to the school grading system. Thus, the coefficient, μ, on the interaction between the "F"-grade indicator and the post-shock cohort variable is the key variable of interest in this specification.[22]

2. Effects on School Policies

To study the effect of receipt of an "F" grade on school policies, we estimate school-level regressions that are similar to equation (3). In our most saturated model we estimate,

$$P_{st} = a + bF_{st-2} + c(g(POINTS_{st-2})) + dX_{st-2} + (e_{st} - e_{st-2}) \quad (3)$$

where P_{st} indicates whether school *s* implemented policy P in year *t* (the 2003-04 academic year); F_{st-2} is a dummy variable indicating whether or not the school received a failing grade of "F" in summer 2002; g($POINTS_{it-2}$) is a cubic in the number of grade points the school earned in summer 2002; X_{st-2} is a vector of school-level variables, including dummy variables indicating the school's 2002 "simulated grade," the 2002 value of the policy, the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the percentage of the school's student body that is classified as disabled, eligible for free- or reduced-price lunch, limited-English proficient, and gifted; and the "stability rate" of the school[23]; e_{st} and e_{st-2} are normally distributed error terms.

[22] We control for the interaction between summer 2002 grade points and the post-shock cohort indicator to continue the regression-discontinuity methodology in the cross-cohort comparison analysis. Further, in this model, we do not separately estimate the coefficients β and φ, as they are subsumed within the school fixed effect ϕ_s ; we include these variables in the equation for conceptual completeness only.

[23] The stability rate of a school is the fraction of students who were present in the fall census of students who were

The key parameter of interest is b – a school's policy response to having received an "F." Thus, we estimate whether the policy responses of the "F"-graded schools are significantly different from those of higher-graded schools. In some specifications we also compare the "F" schools to the "D" schools. Note that this comparison allows us to net out those policy choices dictated by the state-mandated assistance as "D" schools also heavily benefited from such interventions.

Our survey contained many questions aimed at understanding whether or not a school had adopted a variety of policies and it would be cumbersome and difficult to digest the pattern of potential coefficient estimates that might derive from so many regressions. Thus, we attempt to summarize the results by grouping questions into "domains" that are designed to capture related policies. Grouping policies in this fashion also reduces the severity of what we refer to as the "budget constraint problem." We assume that school superintendents and principals consider which policies to enact subject to a budget constraint. As such, one would not expect that schools adopt *all* possible policies; rather they pick and choose from some feasible set. For example, a principal may attempt to increase instructional time by sponsoring summer school, an extended school year, or after-school tutoring, but not all three. Grouping similarly-intended policies may reduce this problem by allowing us to identify if the administrators adopt a type of policy rather than a specific one.[24]

To see how we analyze the effect of an "F" grade on a "domain," note that we can first rewrite equation (3) to obtain an impact of the accountability system for each policy, where k refers to the kth policy:

---

still at the same school in the spring census.

[24] We have also tried capturing a school's adoption of policies within each domain as the sum of the number of "sub-policies" they adopt. These results are qualitatively similar but less precise than those presented here.

$$P_k = a_k + b_k F_{t-2} + c_k(g(POINTS_{t-2})) + d_k X_{kt-2} + (e_{kt} - e_{kt-2}) = W\Theta_k + \nu_k. \qquad (4)$$

We then aggregate the estimates using a seemingly-unrelated regression approach (Kling and Liebman, 2004). This approach is similar to simply averaging the estimated effect of receiving an "F" grade on school policies if there are no missing values and no covariates.

More specifically, we first estimate equation (4) (or variants) and obtain an item-by-item estimate of *b* (i.e., b_k). We then standardize the estimates of b_k by the standard deviation of the outcome using the responses from the 2002 survey for all (elementary) schools (σ_k). Our estimate of the impact of the school accountability system on school policies is then the average of the standardized *b*'s within each domain,

$$b_{AVG} = \frac{1}{K}\sum_{k=1}^{K} b_k \Big/ \sigma_k \,. \qquad (5)$$

To obtain standard errors for b_{AVG}, we need to account for the covariance between the estimates of b_k within each domain. To do so, we estimate the following seemingly-unrelated regression system:

$$P = (I_K \otimes W)\Theta + \nu \qquad P = (P_1', \ldots, P_k')' \qquad (6)$$

where I_K is a *K* by *K* identity matrix and *W* is defined as in equation (4). We calculate the standard error of the resulting summary measure as the weighted average of the standard errors of the individual estimates.

We present estimates of the summary measure and some of the original underlying regressions.[25] In addition, we also present results using a simple average of the individual items

[25] We emphasize that we do not necessarily believe that the outcomes grouped together within a domain reflect one underlying latent construct. Rather, these are similar policies that schools might adopt.

within each domain where the original data have been normalized by the mean and standard deviation of the variable in 2002.

B. Data

1. Administrative Data

We rely on two sources of administrative data for our analysis. The first are data on individual students throughout the state (including all standardized test scores) from 1998-99 through 2004-05. Students took the criterion-referenced FCAT in grades four, eight and ten (five, eight and ten for mathematics) beginning in 1998-99, and in all grades from three through ten beginning in 2001-02. We refer to this FCAT as the "high stakes test" as the results form the basis of the school's accountability grade. Beginning in 1999-00, all Florida students in grades three through ten also took the Stanford-10 nationally norm-referenced test; we refer to this test as the "low stakes test." The data have been longitudinally linked across years allowing us to follow students over time, as long as they do not leave the public school system or the state of Florida. (Those who leave and return do show up in the dataset and thus are still tracked over time.) In addition to test score results, these data also contain some individual-level characteristics such as the student's race, sex, eligibility for the National School Lunch Program, and English-Language-Learner and disability status.

In each grade and year, anywhere between 177,000 and 207,000 students take the FCAT, depending on cohort size. Few students are lost in the longitudinal analysis of these data: for instance, nearly 95 percent of students who took the fourth-grade FCAT are observed taking the FCAT the next year, and over 87 percent of students who took the fourth-grade FCAT are observed taking the FCAT three years later. All told, our two-cohort analysis sample sizes range

from 346,958 (low-stakes reading) to 349,084 (high-stakes reading) depending on the test. In addition to student test score data, we have obtained data from the FDOE on expenditures per student (by program) and some characteristics of students at the school level for the 2001-02 school year.

2. School Surveys

We also analyze data collected in two rounds of surveys of principals of all "regular" public schools in Florida conducted in the 2001-02 and 2003-04 school years.[26] We achieved response rates that exceeded 70 percent in each year. For the analysis, we focus on elementary schools that received a grade in 2002 and were part of our sampling frame. Of the 1,659 eligible schools in the 2004 sampling frame, we achieved an overall response rate of 75%. The response rate was 76% for schools that received an "A," "B," or "C," in 2002 and 69% and 66% for "D" and "F" schools respectively. Notably, however, the response rate is not statistically different between the "F" and "D" schools. Further, as we show in Appendix Table 1, the characteristics of schools responding to the survey in 2004, by school grade, are quite similar to those of non-respondents.

In the school surveys, we asked principals to identify a variety of policies and resource-use areas which we have divided into several domains: policies to improve low-performing students, lengthening instructional time, reduced class size for subject, narrowing of the curriculum, scheduling systems, policies to improve low-performing teachers, teacher resources, teacher incentives, teacher autonomy, district control, principal control, and school climate.

[26] We excluded "alternative schools" such as adult schools, vocational/area voc-tech centers, schools administered by the Department of Juvenile Justice, and "other types" of schools. Note that we included charter schools serving "regular" students as well. We also conducted the survey in the 1999-00 school year; the results are similar when we control for those results as well. We do not present them here because there is one question in the later surveys that was not asked in the earliest one. The survey instruments are available on request.

We were careful to conduct the surveys in the early spring before schools knew which grade their school would receive for that academic year. Thus, when we conducted the survey in the spring of 2002, the administrators did not yet know their school grade (as school grades were announced on June 12, 2002, after the end of the school year and after our survey data collection had ceased.) Thus, we treat 2001-02 as the "base year" and observe school policy as of the academic year of 2003-04.[27]

V. Results

A. Effects on Student Test Scores

We begin by investigating whether students in "F"-graded schools fared better after the grading policy change in summer 2002 than did their otherwise-equal counterparts in other schools. Throughout this section we present data on four different test outcomes: the high-stakes FCAT "Sunshine State Standards" criterion-referenced tests in reading and mathematics and the low-stakes Stanford-10 national norm-referenced test in reading and mathematics. We report both sets of findings because prior researchers have found that schools have a tendency to teach specialized skills that might improve test performance while possibly not generalizing to a broader knowledge base (e.g., Linn (2005) and Koretz (2003)). The regressions control for student race and ethnicity, lagged FCAT and norm-referenced reading and mathematics tests (in cubics) along with indicators for missing lagged test scores, disability status and history, free/reduced lunch eligibility, sex and English-Language-Learner status, as well as a cubic in the

[27] Note that we do not observe school policies in the first year after the change in grading procedure; that is, in the 2002-03 school year. We believe this likely downward biases our results as schools that received a grade of "F" in summer 2003 may have also changed their educational practices by the 2003-04 school year. As shown in Table 1, 31 schools received an "F" in summer 2003 (18 elementary schools received an "F"). Seven of these schools had received an "F" in summer 2002 but no elementary schools received an "F" in both years.

accountability grade points and a set of lagged school grade variables. Test scores are standardized to have zero mean and unitary variance for ease of presentation.[28]

1. Cross-Sectional Regression-Discontinuity Evidence

The first row of Table 4 presents the estimated effects of receipt of an "F" grade on student test scores in cross-section for the students in fifth grade in 2002-03, the year following the grading system shock in summer 2002. We limit our analysis to the fifth-grade cohort because we will ultimately compare successive cohorts of students – one that was affected by the change in grading standards and the other that was not.[29] Constraining the study to fifth graders is the only way we can ensure that our comparison cohort of students was not affected by the school grading.[30]

We observe that in all four examinations, students fared better in "F"-graded schools than did those in other schools, all else equal. Depending on the test score, we find that students attending "F"-graded schools scored anywhere from 0.07 of a standard deviation to 0.14 of a standard deviation higher than students who attended higher-graded schools in the year following the grade determination, and all are statistically significant at conventional levels. These results suggest that schools awarded "F" grades in summer 2002 improved faster than did their counterparts with higher school grades, all else equal. The fact that test score gains are higher

[28] We eliminate seven schools with test data from our analysis because they are outside our sampling frame, but the results including all schools differ only at the third significant digit.

[29] We have conducted all cross-sectional analyses with a fuller set of students (i.e., both fourth and fifth graders in 2002-03) and the results are very similar. All four test results are of nearly identical magnitude and statistical significance to those presented herein, although the high-stakes math results are modestly below those presented herein (its point estimate is 0.104 with a standard error of 0.032.)

[30] Ninety-nine percent of fourth grade students in the spring of 2001 before the school grading changed (i.e., the potential comparison group) were still in elementary school in the spring of 2002. While we do not believe the schools fully anticipated their grades, this has the potential to affect our analysis. In contrast, 98 percent of fifth graders in the spring of 2001 were in middle school by the spring of 2002 and therefore were not (directly) affected by the grades given to the elementary schools. We have also estimated all models excluding students who repeated fifth grade and find that the results do not substantively change.

for high-stakes tests than for low-stakes tests may lend credence to the notion that schools "taught to the test," but significant gains are seen even for the low-stakes tests.[31,32] While not reported in the table, we have also conducted extensive specification checks, including controlling for third grade test scores rather than, or in addition to, fourth grade test scores. The results are invariant to these decisions, and are available on request.[33]

A handful of "F"-graded schools had previously received a grade of "F," allowing their students to be eligible for vouchers through the Opportunity Scholarship Program as well as imposing much more stringent reporting requirements. These schools arguably face even greater pressure than do newly-"F"-graded schools, both from increased competition and greater scrutiny and oversight from the state. We separated the "F" school effect into the estimated effects for first-time "F" schools and the additional effect of having a second "F" grade; these results are reported in the second and third rows of Table 4. We find that the estimated effect of the "F" for repeat "F" schools is unambiguously larger, and statistically distinct from the first-time "F" effect in the two mathematics specifications (though not in the reading specifications).

---

[31] Furthermore, while studies indicate that teaching to the test is prevalent (e.g., Haney (2000), Jacob (2002)), we note that it is not necessarily "gaming," and may be a desirable outcome, rather than one to be criticized, depending on the alignment of the test in question to the set of standards for which society and policy-makers desire students to have proficiency.

[32] We have also investigated the distributional consequences of the change in school accountability systems. We find that students who are below-average or above-average tend to experience the largest gains in "F" schools following the policy change. These distributional consequences are consistent with Neal and Whitmore (2007) and Krieg (forthcoming) who suggest that proficiency-count-based systems (such as Florida's original system) lead to concentration on the students in the middle of the distribution. A full treatment of distributional consequences is beyond the scope of the present paper.

[33] A picture is worth 1,000 words and in Appendix Figures 1a and 1b we present the mean test score residual on the high-stakes FCAT examination in reading and mathematics for fifth-graders in 2002-03 (Figure 1a) as well as the fourth grade test scores for the same students the year before (1b). It is apparent that schools just to the left of the threshold determining whether a school received an "F" grade or not experienced greater test score gains than did schools just to the right of the threshold in both reading and mathematics in 1a. These results suggest that the estimated differential improvement of "F" schools over "D" schools is not due to regression to the mean but rather to differential responses to the accountability system. Importantly, there is no discontinuity in fourth grade in Figure 1b – before the school grading occurred – for these students; rather, the trend line is smooth around the grade threshold. These results are robust to controlling for cubics or quartics in the grade points rather than quintics. We further explore the mean reversion later in this section.

That said, the estimated effects of "F" receipt for the first-time "F" schools remain large and are statistically significant at conventional levels. These results provide suggestive evidence that schools facing still greater competitive and/or accountability pressure raise student test scores by more than those facing less pressure. That said, these distinctions are based on a very small number of schools, so for the remainder of the paper we do not distinguish between first-time "F" schools and repeat "F" schools in our analysis.

2. Regression-Discontinuity Evidence Using Panel Data

We can further identify the causal effect of receipt of an "F" grade on student test score improvements by comparing the fifth-grade cohort of 2001-02 to the fifth-grade cohort of 2002-03. The students in the first (2001-02) cohort would have graduated to middle school before their elementary school would have responded to the school grading system, so they make a natural comparison group. In the fourth row of Table 4 we include both an affected and an unaffected cohort of students and can therefore control for elementary school fixed effects. The key point estimate of interest then becomes the interaction of a "grade 'F' in summer 2002" variable with a "2002-03 fifth-grade cohort" variable. Compared with the previous cohort of students, students in the affected cohort performed considerably better in fifth grade on all four tests.

We next follow students into middle school, in what would be their seventh grade of school,[34] and further control for elementary school-middle school combination fixed effects. We observe that students from the affected cohort continue to outperform their immediate unaffected

[34] Students who are held back or who skip a grade are counted along with the other members of their cohort; results are no different when we limit our analysis to students who progress normally from grade to grade. We eliminate sixth grade from the table due to space considerations, but the results are statistically and qualitatively similar to seventh grade results.

predecessors who attended the same combination of middle and elementary schools but who did not experience the elementary school accountability shock in summer 2002. Put differently, there is considerable evidence that the accountability shock of summer 2002 had lasting effects on the performance of students who were attending "F"-graded schools at the time.[35]

One might be concerned that despite our reliance on a regression-discontinuity framework, it is unwarranted to compare students attending "F"-graded schools to those attending much higher-rated schools. We therefore estimate the same basic model in two different ways: comparing "F" schools to those that earned a "D" grade in summer 2002 only; and restricting the sample to the set of schools that would have earned a grade of "D" had the old grading system been employed in summer 2002. This specification has the advantage of not only narrowing the focus to the area around the discontinuity in the accountability points, but since "D"-graded schools are also eligible for Florida's *Assistance Plus* Program's benefits, it helps to net out the impact of these additional educational inputs. The results of these two specifications are presented in the sixth and seventh rows of Table 4. We observe that the estimated effects of an "F" grade remain statistically significant when these sampling restrictions are employed, and in seven of eight cases the estimated effects of an "F" grade are larger than when the comparison group is the full set of schools.

3. Mean Reversion and Selection

Even with the cross-cohort regression-discontinuity research design, concerns regarding mean reversion and selection may remain. The eighth and ninth rows of Table 4 include two separate falsification exercises aimed at ruling out measurement error as a candidate explanation

[35] These estimated effects persist even when we also control for the changing accountability pressure faced by the middle schools that the students ultimately attended. For instance, the point estimate on the high-stakes reading exam is 0.083 with a standard error of 0.031.

for our findings. First, we compare the two successive cohorts' performance in *fourth* grade, conditioning on their third grade performance. That is, we compare the "affected" cohort's test scores in 2001-02 to the "unaffected" cohort's test scores in 2000-01. In this comparison, both sets of students would have taken their fourth grade FCAT exams *before* the grade assignment occurred in summer 2002. In the presence of mean-reverting measurement error, students in the affected cohort in schools identified as "F" in 2002 would have experienced an unusually large (random) decline in test scores in 2001-02 that was followed by an unusually large (mean-reverting) increase in 2002-03. Thus, we would therefore expect to see a negative coefficient on the variable identifying "F" schools in 2002 (for the affected cohort). Because only the later cohort took the fourth grade math FCAT exam, we can only conduct this comparison for the high-stakes reading test and the two low-stakes tests. That said, we find no relationship between test performance in fourth grade and what would later become an accountability grading shock. The t-statistics of the falsification test's "F grade effect" are invariably below one, and the sign of the insignificant relationship differs depending on the test employed.

In our second falsification exercise, we calibrate each student's low-stakes norm-referenced test scores to their high-stakes FCAT Sunshine State Standards test scores, and calculate what the school grade points would have been in 2002 had the *norm-referenced tests* been used, rather than the FCAT, to measure student proficiency; this generates a pseudo-"F" grade. If mean-reversion in test scores is what generated the large observed increases in test scores following receipt of an "F," then we would expect to see large increases after identifying schools that received the pseudo-"F" as well. However, as shown in Table 4, this exercise produces "falsification effects" which are trivial in size and statistical significance.

The previous analysis could also be subject to selection bias if students who are observed attending an "F"-graded school after their school received the new school grade differ systematically from the set of students who attended that school the prior year. We investigated the degree to which selection could be a potential problem by comparing the likelihood of remaining in the same school from 2001-02 to 2002-03 for two groups of students – those who were attending schools that would receive an "F" grade and those who were attending schools that would receive a "D" grade. While there was no difference between "F" schools and "D" schools in the fraction of students tested, we did find that there is a statistically significant difference of about three percentage points in the likelihood of leaving a school associated with receipt of an "F" grade. These results are not presented in the table but are available on request.

The question then becomes whether the students who left "F" versus "D" schools (or other better-graded schools) differed in any meaningful way. We therefore estimate the effect of receiving an "F" grade on the differential lagged test scores of stayers versus leavers.[36] As shown in the last row of Table 4, we find that there is virtually no difference, either in magnitude or in a statistical sense, in the lagged test scores of the stayers versus leavers in either "F" schools or "D" schools. Therefore, we are confident that our results are not being driven by differential selection into (or out of) "F"-graded schools.[37]

Thus, we have consistent evidence that when faced with increased stigma, oversight, and the threat of vouchers, student outcomes can improve. However, we know little about what

[36] We implement this comparison by estimating the regression-discontinuity model as before, where the dependent variable is the lagged student test score and all accountability variables are interacted with a "stayer" indicator. Our key variable of interest for this exercise, therefore, is the interaction between an "F" grade indicator and a "stayer" indicator.

[37] Another potential source of selection bias is that students' transitions to certain middle schools might be influenced by their elementary school grade. We find that transition probabilities to any given middle school are exactly the same for both "D" and "F" schools between 2001-02 and 2002-03 as they were (in the same school) between 2000-01 and 2001-02. We therefore contend that middle school selection is not influenced by the student's elementary school grade in 2002, at least between "D" and "F" schools.

generates these improvements. Using an educational production function framework we might conceptualize different ways in which schools increase educational achievement: by increasing the number of inputs into the process (such as would occur if there were more revenues), by changing to a more productive mix of resources, by getting a higher return from the resources employed, or by "gaming" the system. We now turn to an analysis of school responses.

B. What are Schools Doing? Effects on School Policies

1. What Schools Were Doing in 2002

We begin by presenting the policy/production environment of schools in 2002, the base year in our analysis. In Table 5 we report the means of the individual policies organized by their respective "domains" and by the school's grade in 2002. Among the variables that schools identified as strategies they employ to "improve low-performing students"[38] most of the schools required grade retention,[39] in-school supplemental instruction, and tutoring. Slightly less commonly schools required that students attend summer school and a few even required that such students attend Saturday classes, although this policy was more prevalent among the "D"- and "F"-graded schools. Most schools also sponsored (although did not necessarily require) summer school and after-school tutoring. In addition, "F"-graded schools offered "other" school services such as extended day, mentors, or remedial instruction. Interestingly, the average class day was 377 minutes (or about 6 hours and 20 minutes), although it was slightly longer for the "F"-graded schools. Further, the average class size (not pupil-teacher ratio) in first and fourth grades was between 24-25 students with lower-graded schools having slightly smaller classes.

---

[38] These responses are based on the question "What special measures, if any, does this school take to try to improve the performance of low-performing students?"

[39] Beginning in 2003 Florida required schools to retain all third graders that did not meet a pre-determined threshold on the FCAT Sunshine State Standards test.

One strategy that some schools employ to improve achievement is to assign students to smaller "class units" for particular reasons. This does not mean that overall class sizes in a school are reduced, but they may, for instance, reorganize students either within classrooms or possibly even across classrooms for particular subjects. As an example, in our sample 61 percent of the "F"-graded schools used these smaller class "units" for reading compared to just over 40 percent of "A"-graded schools. Similarly, "F"-graded schools were more likely to employ this strategy for math, writing, and low-performing students as well.

While the results in Table 4 suggested some "teaching to the test," we also sought more direct evidence that might suggest this shift in focus, or at a minimum a narrowing of the curriculum. We therefore asked the principals if the school had a policy on the minimum amount of time that fourth grade students must spend on particular subjects (and among those that did, what that minimum was). In 2002, 86 percent of the "F"-graded schools had policies that specified a minimum amount of time on math and reading, and three-quarters had a specified minimum amount of time on writing; these are all tested subjects. In contrast, only 61 percent had a policy on the minimum time spent on social studies which is not tested. The higher-graded schools were less likely to have school policies on how time was to be divided between the subjects.

In terms of how schools attempt to schedule the day and organize how teachers work together, we observe that nearly all schools attempted to organize teachers into teams or tried to schedule common prep periods to provide teachers with an opportunity to collaborate. Similarly, a majority also employed subject-matter specialist teachers to assist other teachers. Finally, 15 percent of "F"-graded schools (and fewer of the higher-graded schools) employed some other way of structuring schedules and staff. Examples of these other strategies include: inclusion, a

school-wide common reading or language arts block (for example, 90 minutes of uninterrupted reading each day)[40], ability grouping, organizing even lower grades into "departments," or employing technology specialists to help teachers.

Given the centrality of teachers to the learning process, it is not surprising that schools employ a variety of strategies to improve the performance of teachers. As of the spring of 2002, nearly all schools supervised such teachers more closely, assigned an aide or mentor to them, and provided additional professional development and/or an improvement plan. However, in terms of the resources available to teachers, we find that the "F"-graded schools provided less time for collaborative planning and class preparation than higher-graded schools. They did, however, require more days per year for professional development and had more money available for professional development activities each year (and more per teacher). Finally, while schools provided rewards for teacher performance (independent of any incentives used by the school district), the vast majority used comp or release time, attendance at conferences and workshops, or giving the teachers special leadership positions or assignments; fewer than one-third were found to use monetary rewards.

We also asked a number of questions about the level of control that teachers, principals and superintendents had over important decisions at the school (1 suggests no influence whereas 5 suggests complete control). Across the board, principals had the most say in the hiring of new full-time teachers, how the school's budget would be spent, and over the evaluation of teachers. The school superintendents had the most control over the curriculum.

Finally, while not necessarily the direct result of school policy, we were interested in gauging the climate at the school. The response indicates the extent to which the principal

[40] Implementing a common reading block is one of the corrective actions commonly adopted according to the *Assistance Plus* reports.

thought the statement was not at all accurate ("1") or was accurate ("5") of his or her school. We observe that in higher-graded schools the staff morale was higher, the teachers – new and experienced – were judged excellent, student disruption did not interfere with learning, violence was not a concern among parents, and parents closely monitored the academic progress of their children.

Given this portrait of school policy and the environment as of the spring of 2002, the question is whether schools, the low-performing schools in particular, *changed* any of these policies or procedures in response to their assigned performance grade.

2. What Policies Changed Between 2002 and 2004

Table 6 presents descriptive statistics of the change in school policy between the springs of 2002 and 2004 by the school's grade in the 2001-02 school year.[41] Again the individual variables are organized by their "domain." The "F"-graded schools appear to have been more likely to require grade retention and other forms of supplemental instruction for low-performing students, at least compared to higher-graded schools. Similarly, they were more likely to adopt summer school classes, extend the school year, and sponsor after school tutoring. In addition, while they lengthened the school day by, on average, about 2 minutes, the higher-rated schools shortened the length of the school day. Although the F-schools were not more likely to institute minimum time requirements in the tested subjects, they were more likely to relax minimum time

[41] One concern is that policies may appear to have changed because different people with different levels of information about what is going on in a school completed our survey. Unfortunately, we cannot readily assess in a systematic manner who completed the survey in each year. We have, however, pulled 100 random school surveys from 2001-02 and compared the name of the person completing the survey in 2001-02 to the name of the person completing the survey in 2003-04. Out of 100 surveys, 88 had the same respondent in the two rounds and 97 had the same position title. This informal finding, combined with Chiang's (2007) result that school grades did not influence principal turnover, makes us less concerned that changes in our findings are due to different people completing the survey.

requirements on the untested subjects – such as social studies and art – which is consistent with an incentive to narrow the curriculum in response to the accountability system.

In terms of teachers, the "F"-graded schools increased the amount of time available for collaborative planning and class preparation by about 80 minutes compared to only an 11 minute increase for the "A"-graded schools and a nearly one hour decrease for the "D"-graded schools. In general, teachers lost some autonomy over important decisions made at the school.

Finally, it is notable that all schools appear to have reduced class sizes between the two years by between 2-3 students, particularly among the lower-rated schools.

While these means are suggestive of an association between a school receiving an "F" grade and the policies and practices adopted by the school to improve, they do not necessarily indicate causation. We attempt to discern the causal impact of receiving an "F" grade in the models whose results are presented in Table 7. (We present OLS results for selected individual variables in Table 8.)[42]

The first four columns of Table 7 use SUR to aggregate the responses within each domain; the final column simply creates an "index" of the individual components. In addition, in the first two columns, we present the impact by comparing the "F"-graded schools to all higher graded schools; in the last three columns (3-5) we compare the "F"-graded schools to "D"-graded schools.

The first column does not include any other covariates and we estimate that schools that received an "F"-grade were more likely to adopt policies to improve low-performing students, lengthened instruction time, assigned students to smaller class "units" for particular reasons,

[42] The results are similar if we cluster the standard errors at the district level. We also note that we do not report results adjusting the standard errors for multiple testing primarily because we do not expect that schools will adopt policies in all of the domains and do not have a prior as to what fraction of outcomes should be significantly affected.

narrowed the curriculum, changed their scheduling systems, adopted policies to improve low-performing teachers, increased resources available to teachers, and had a worsening climate.

The estimates in column (2) improve on this specification in a number of ways. First, we control for the school's 2002 "simulated grade" such that we take into account whether or not the school "expected" to receive an "F" grade. In addition, we add the 2002 level of the 2004 policy (and an indicator for a missing response in 2002) and other school characteristics.[43] The coefficient estimates, conditional on these covariates, generally decrease and in some cases reverse sign. For example, while, on average, we estimate a positive (although insignificant) impact of receiving an "F"-grade on the principal's autonomy, conditional on some basic covariates we estimate that with an "F"-grade, school principals lose some autonomy as a result.

Column (3) employs a basic version of the regression discontinuity design that we exploited when analyzing the impacts on student test scores. We do so by controlling for a cubic in the total number of grade points earned by the school in 2002. Note that, by and large, the estimates do not change substantially. Next, we narrow the focus even further by comparing the impact of receiving an "F"-grade to that of receiving a "D"-grade while continuing to control for the cubic in the grade points (column (4)). Again, the estimates do not change substantially compared to those in columns (2) and (3). Finally, we compare the results using a SUR analysis to those generated by simply computing an index of the associated items in the domain. Again, the results do not generally change much although they are less precise. We primarily show these estimates as they facilitate the analysis in the last section.

[43] These variables are described in Section IV.2, above. We have also included the lagged variables (and indicators for these variables being missing) for all of the items in the domain with similar results. We only report results controlling for the individual the 2002 outcome for parsimony. In addition, as noted earlier, we also obtain similar results if we also include the 2000 level of the variable. We also include flags for missing expenditures or characteristics.

Table 8 shows OLS results of individual items within the domains that were statistically significant at the 10 percent level (or slightly above) as well as of a few items that we could not "easily" fit into a domain. In the first column we show the impact on "F"-graded schools compared to all schools and we do not control for any covariates. In the second column we continue to show the impact on "F"-graded schools compared to all schools but also control for all of the covariates also employed in column (3) of Table 7 (thereby adopting a regression discontinuity design). The third column is similar to column (4) of Table 7 in that it compares the "F"-graded schools to the "D"-graded schools.

We find that the schools facing the increased pressure adopt block scheduling, re-organize the school scheduling structure through "other" measures, and increase time for collaborative planning and class preparation for teachers; we also see some evidence that they provide a common prep period for teachers, and increase the district's control over how the budget is spent. And while it may appear that "F"-graded schools are more likely to reduce class sizes to improve student achievement, this response substantially decreases and becomes statistically insignificant when we add covariates.

In sum, we find that schools receiving an "F" grade are more likely to focus on low-performing students, lengthen the amount of time devoted to instruction, adopt different ways to organize the day and learning environment of the students and teachers, increase resources available to teachers, and decrease principal control, as was expected given the increased oversight built into the A+ Plan. While not statistically significant, the point estimates suggest that schools facing such pressures also respond to the parameters of the accountability system by possibly narrowing the curriculum to focus on the tested subjects and possibly focusing less on

gifted students (as we see *some* evidence that while these schools employ reduced class sizes for low-performing students they do not do so for the gifted students).[44]

C. Do the Policies Appear to Explain the Change in Test Scores?

Finally, we attempt to determine whether the adoption of these policies explains part of the test score gains observed by students attending the "F"-graded schools. We emphasize that we cannot generate causal estimates of the effectiveness of the individual policies; rather we are interested in gaining slightly more insight into the mechanics behind the achievement gains.

Table 9 compares three different types of models – (1) those that control for the set of background and school variables included in the specifications from Table 4; (2) those that further include controls for the policy indices presented in Table 7 as well as the class size variable described in Table 8; and (3) those that include just the five domains that we independently found to have the strongest cross-sectional relationship with student test score gains in models that include all of the domains.[45] To provide continuity with our previous analysis, we restrict the comparison to the "F" versus "D" grades, though "F"-versus-all school comparisons yield substantively and statistically similar results. We present evidence for the four different test scores for seventh graders, conditional on lagged fourth grade test scores, although the results are comparable when we look instead at fifth graders or sixth graders.

In the table we present p-values of tests of joint significance, depending on model specification, either of the five domains typically found to most strongly independently relate to

[44] Results, available on request, are generally similar when we use all schools.

[45] These domains are policies to improve low student performance; policies to improve low teacher performance; policies regarding teacher autonomy; reflections of school climate; and class size in general. All five domains typically are statistically significantly related to student test scores and consistently have the same signs across the various cross-sectional regressions. Policies to improve teacher performance and school climate are positively related to test score gains; those regarding improving low-performing students and teacher autonomy are negatively related to test score gains; and increased class sizes are negatively related to test score gains.

student test score gains in cross-section; or (in the case of the model in which we include all policy domains) of two different sets of policy variables – those that require additional resources (i.e., policies to improve low-performing students; policies to improve low-performing teachers; lengthening instructional time; reducing class sizes for subjects; and reducing class sizes in general) and those that are intended to improve efficiency, either by making existing resources more productive or by rearranging existing resources (i.e., teacher resources; teacher incentives; teacher/district/principal control; minimum time on testing and untested subjects; and scheduling systems.) In all specifications, the set of policy variables with the strongest cross-sectional relationships with student test scores are, unsurprisingly, strongly jointly significant. In addition, the set of school policy variables that are resource-related are also consistently jointly statistically significantly related to student test scores across the four test score measures. The set of policy variables aimed at increased efficiency is not statistically significantly related to student gains in reading, but is in mathematics.

More to the point of the analysis is the percentage reduction in the estimated "F"-grade test score effects when these policy variables are included as control variables. Across the specifications (including in other grades not reported in the table), the estimated effect of "F" grade receipt decreases with the inclusion of these policy variables, with percentage reductions that range from very modest to very large. The share of the test score gain associated with "F" grade receipt is at least 15 percent with regard to reading and at least 38 percent with regard to mathematics. Moreover, virtually the entire explained portion of the test score gains associated with an "F" grade is apparently due to the five policy domains that we found to have the strongest cross-sectional relationship with student test score gains.

While we do not ascribe a causal interpretation to these findings, the results indicate that the school policy variables captured in our surveys may explain a sizeable percentage of the differences in test score gains between "F"-graded schools and "D"-graded schools. Put differently, the evidence suggests that some combination of the policies and practices that the "F"-graded schools have put into place in apparent response to accountability pressures have contributed to the relative test score gains of the "F"-graded schools.

VI. Conclusion

This project presents an attempt to systematically get inside the "black box" of the effects of school accountability systems on student performance. In this paper we find that schools that received a grade of "F" in summer 2002 immediately improved the test scores of the next cohort of students, and that these test score improvements were not transitory, but rather remained in the longer term. We also find that "F"-graded schools engaged in systematically different changes in instructional policies and practices as a consequence of school accountability pressure, and that these policy changes may explain a significant share of the test score improvements (in some subject areas) associated with "F"-grade receipt.

We note also that the remaining estimated "F"-grade effect may be due to other unobserved changes in policies or programs, or additionally, to an increase in the productivity of school resources due to the "F"-label (e.g., through an increase in effort on the part of the school superintendent, principals and teachers). Further, we observe that the "F" schools adopted different policies than did "D" (or higher-graded) schools. And while we think that these policies reflect decisions made by the school principal and/or superintendent in an attempt to

improve school productivity, we do not strictly observe who may have instigated them and under the *Assistance Plus* Program there is scope for others to have much influence.

While suggesting that school accountability systems structured similarly to the A+ Plan can spur school improvement, it is premature to outline a prescription for the improvement of low-performing schools based on these findings, particularly since we do not observe student performance along all relevant dimensions. A comprehensive study of school efficiency would also include untested domains, such as student performance in social studies and the arts. That said, we find that accountability pressures have the potential to improve student test scores in low-performing schools, and that such pressures can induce school administrators to change their behavior in educationally beneficial ways.

References

Akerlof, George A. and Rachel Kranton. 2005. "Identify and the Economics of Organizations," *Journal of Economic Perspectives*, 19:1, pp 9-32.

Akerlof, George A. and Rachel Kranton. June 2007. *More than Money: Economics and Identity*, manuscript.

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114(2): 533-575.

Carnoy, Martin and Susanna Loeb (2002). "Does External Accountability Affect Student Outcomes? A Cross State Analysis." *Education Evaluation and Policy Analysis,* 24(4): 305-331.

Chakrabarti, Rajashri (2006) "Vouchers, Public School Response and the Role of Incentives: Evidence from Florida." Working Paper, Federal Reserve Bank of New York.

Chiang, Hanley (2007) "How Accountability Pressure on Failing Schools Affects Student Achievement," Harvard University mimeo (October).

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

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

Cullen, Julie Berry and Randall Reback (forthcoming). "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." in Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics, T. Gronberg and D. Jansen, eds., 14 (Amsterdam: Elsevier Science).

Figlio, David. (2006) "Testing, Crime and Punishment." *Journal of Public Economics* 90(4): 837-851.

Figlio, David and Lawrence Getzler (forthcoming). "Accountability, Ability and Disability: Gaming the System?" in Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics, T. Gronberg and D. Jansen, eds., 14 (Amsterdam: Elsevier Science).

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

Figlio, David and Cecilia Rouse (2006). "Do Accountability and Voucher Threats Improve Low-Performing Schools?" *Journal of Public Economics* 90(1-2): 239-255.

Figlio, David and Joshua Winicki (2005). "Food for Thought? The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89(1-2): 381-394.

Goldhaber, Dan and Jane Hannaway (2004). "Accountability with a Kicker: Preliminary Observations on the Florida A+ Accountability Plan." *Phi Delta Kappan* 85(8): 598-605.

Haney, Walt. (2002) "Lake Woebeguaranteed: Misuse of Test Scores in Massachusetts, Part I." *Education Policy Analysis Archives* 10(24).

Haney, Walt. (2000) "The Myth of the Texas Miracle in Education." *Education Policy Analysis Archives* 8(41).

Hannaway, Jane and Kristi Kimball. (2001) "Big Isn't Always Bad: School District Size, Poverty, and Standards-Based Reform" In Susan Fuhrman, *From the Capitol to the Classroom: Stadards-based Reform in the States*. National Society for the Study of Education: 99-123.

Hannaway, Jane and Laura Hamilton (2007) "Effects of Accountability Policies on Classroom Practices". Washington, DC: The Urban Institute.

Hanushek. Eric et al. (1994) *Making Schools Work: Improving Performance and Controlling Costs*. Washington, DC: The Brookings Institution.

Hanushek, Eric and Dale W. Jorgenson (eds.). (1996) *Improving America's Schools: The Role of Incentives*. National Academy Press.

Hanushek, Eric and Margaret Raymond (2005). "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24(2): 297-327.

Howell, William, and Paul Peterson (with Patrick Wolf and David Campbell). *The Education Gap: Vouchers and Urban Schools*. (Washington, DC: Brookings Institution, 2002).

Jacob, Brian (2004). "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics*.

Jacob, Brian and Lars Lefgren (2004). "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86(1): 226-244.

Jacob, Brian and Steven Levitt (2003). "Rotten Apples: An Investigation of the Prevalance and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118(3): 843-877.

Kling, Jeffrey R. and Jeffrey B. Liebman (2004). "Experimental Analyses of Neighborhood Effects on Youth." Industrial Relations Section (Princeton University) Working Paper Number 483 (March).

Koretz, Daniel (2003). "Using Multiple Measures to Address Perverse Incentives and Score Inflation." *Educational Measurement: Issues and Practice* 22, no. 2: 18-26).

Krieg, John (forthcoming). "Are Students Left Behind? The Distributional Effects of the No Child Left Behind Act." *Education Finance and Policy*.

Krueger, Alan (1999). "Experimental Estimates of Education Production Functions," *Quarterly Journal of Economics* 114(2): 497-531.

Krueger, Alan and Diane Whitmore (2001). "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, 1-28.

Krueger, Alan B. and Pei Zhu. (2004) "Another Look at the New York City Voucher Experiment," *American Behavioral Scientist*, 47, no. 5 (January): 658-698.

Linn, Robert. (2005) "Alignment, High Stakes, and the Inflation of Test Scores," in *Uses and Misuses of Data for Educational and Accountability Improvement* (ed. by Joan L. Herman and Edward H. Haertel) (Malden, Massachusetts: Blackwell Publishing), pp. 99-118.

Neal, Derek and Diane Whitmore Schanzenbach (2007) "Left Behind by Design: Proficiency Counts and Test-Based Accountability." Working paper, National Bureau of Economic Research, August.

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

Simon, Herbert A. (1957). Administrative Behavior: A Study of Decision-Making Processes in Administrative Organizations (2nd edition). New York: Free Press.

West, Martin and Paul Peterson (2006). "The Efficacy of Choice Threats Within School Accountability Systems: Results from Legislatively Induced Experiments," *The Economic Journal* 116: C46-C62.

Table 1: The Distribution of School Grades, by Year

School Grade	School Year					
	Summer 1999	Summer 2000	Summer 2001	Summer 2002	Summer 2003	Summer 2004
	All Schools					
A	183	552	570	887	1235	1203
B	299	255	399	549	565	515
C	1180	1115	1074	723	533	568
D	565	363	287	180	135	170
F	70	4	0	60	31	34
N	0	0	76	102	2	0
Total	2297	2289	2330	2501	2501	2490
	Elementary Schools					
A	119	485	389	623	928	974
B	214	180	324	368	360	333
C	713	614	636	452	299	284
D	448	260	215	124	63	67
F	61	4	0	35	18	9
N	0	0	46	68	2	0
Total	1555	1543	1610	1670	1670	1667

Source: Authors' calculations from state data.

Table 2: Transition Matrix in Predicted Grades Based on 2002 Grade Change, Elementary Schools (row percentages)

		Grade in 2002 based on new (summer 2002) grading system				
		A	B	C	D	F
Simulated grade in 2002 based on old (summer 2001) grading system	A	0.89 [264]	0.11 [33]	0.00 [0]	0.00 [0]	0.00 [0]
	B	0.48 [188]	0.38 [149]	0.13 [50]	0.26 [1]	0.00 [0]
	C	0.23 [167]	0.25 [181]	0.46 [331]	0.05 [37]	0.28 [2]
	D	0.02 [3]	0.01 [2]	0.38 [69]	0.44 [79]	0.15 [28]
	F	0.00 [0]	0.00 [0]	0.00 [0]	0.29 [2]	0.71 [5]

Notes: All row percentages (in parentheses) are student-weighted. The number of schools is in brackets. Simulated grade changes are generated by applying both the old grading system and the new grading system to 2002 student test scores. They are therefore generated based on precisely the same student tests; the only differences in the calculations are the formulas used to convert these same tests into school grades.

Table 3: Average School Characteristics by the School's Accountability Grade in 2002

	School Grade in 2002		
Variable	**A/B/C**	**D**	**F**
Number of Students	728 [258]	624 [219]	515 [269]
Proportion Black	0.22 [0.22]	0.58 [0.29]	0.74 [0.27]
Proportion Asian	0.02 [0.02]	0.01 [0.2]	0.004 [0.01]
Proportion Hispanic	0.18 [0.22]	0.18 [0.21]	0.12 [0.19]
Proportion Native American	0.003 [0.005]	0.002 [0.007]	0.001 [0.002
Proportion Mixed Race	0.02 [0.02]	0.01 [0.01]	0.01 [0.01]
Proportion Free or Reduced-Price Lunch Eligible	0.51 [0.24]	0.81 [0.15]	0.79 [0.19]
Proportion Gifted	0.04 [0.02]	0.02 [0.02]	0.01 [0.02]
Proportion Limited English Proficiency	0.09 [0.11]	0.11 [0.12]	0.09 [0.12]
Proportion Classified Disabled	0.16 [0.06]	0.15 [0.06]	0.17 [0.07]
Stability Rate	0.94 [0.03]	0.91 [0.03]	0.90 [0.04]
Special Education Expenditures per Pupil	$8,432 [2,098]	$9,651 [2,169]	$9,604 [2,097]
Regular Education Expenditures per Pupil	$4,411 [893]	$5,316 [1,164]	$5,720 [1,113]
At-risk Students Expenditures per Pupil	$5,518 [3,780]	$5,703 [3,736]	$4,929 [4,233]
Vocational Education Expenditures per Pupil	$102 [828]	$216 [1,295]	$465 [2,714]

Notes: Standard deviations in brackets. All means are based on data from the 2001-2002 school year. The means for the racial composition of the schools represent the average across grades among test takers; they were calculated from administrative data on individual students. The other characteristics are based administrative data from the FDOE.

Table 4: Difference-in-Difference Regression-Discontinuity Estimates of the Effect of Receiving an "F" Grade in Summer 2002 on Student Performance: School Fixed Effects Evidence

Model specification	Standardized test score			
	High-stakes reading	High-stakes math	Low-stakes reading	Low-stakes math
Performance of 2002-03 fifth-grade cohort in the year following policy change				
2002-03 cohort only	0.099	0.141	0.069	0.076
	(0.029)	(0.044)	(0.021)	(0.028)
First time "F" schools	0.097	0.091	0.067	0.062
	(0.033)	(0.050)	(0.023)	(0.031)
Additional effect of repeated "F"	0.041	0.237	0.047	0.123
	(0.034)	(0.067)	(0.033)	(0.058)
2002-03 cohort compared with 2001-02 cohort	0.140	0.212	0.074	0.122
	(0.036)	(0.047)	(0.026)	(0.029)
Performance of 2002-03 fifth-grade cohort in third year (seventh grade or equivalent) following policy change (as compared with the preceding cohort)				
2002-03 cohort compared with 2001-02 cohort	0.088	0.118	0.081	0.078
	(0.028)	(0.034)	(0.021)	(0.024)
Comparing "F" schools to "D" schools in summer 2002 only	0.079	0.150	0.151	0.135
	(0.045)	(0.064)	(0.037)	(0.039)
Restricting sample to those predicted to earn "D" grade under old system	0.120	0.139	0.117	0.127
	(0.034)	(0.044)	(0.035)	(0.037)
Alternative specifications: Considering mean reversion and selection				
Falsification test: Successive cohorts in **fourth grade**	-0.022	n/a	0.028	0.018
	(0.029)		(0.029)	(0.028)
Falsification test: School grades calculated based on **NRT scores**	-0.001	0.012	-0.020	-0.006
	(0.021)	(0.022)	(0.019)	(0.019)
Selection test: Differential **fourth grade** test scores of students remaining in same school versus those who left	-0.006	-0.005	0.008	-0.005
	(0.070)	(0.076)	(0.066)	(0.067)

Notes: Standard errors adjusted for school-level clustering are in parentheses. Dependent variables are test scores standardized to the Florida average by grade level. Regressions control for fourth grade FCAT-NRT scores in cubics (as well as indicators for missed scores), grade dummies, year dummies, indicators for current and historical disability status, race, ethnicity, English language learner status, free lunch eligibility, reduced lunch eligibility, subsidized lunch ineligibility conditional on application, prior year's school grade, predicted school grade, a cubic in the number of points the school received according to the 2002 accountability grading system, and elementary school fixed effects (or elementary-middle combination fixed effects in the case of sixth and seventh grade tests.) Cohorts compared are the fifth grade class of 2001-02 (pre-policy change) and 2002-03 (post-policy change). In the fourth grade falsification tests, both cohorts' test scores are observed before the policy shock; fourth grade falsification tests control for third grade NRT scores in cubics. The selection test does not include further lagged test scores.

Table 5: 2001-02 School Response Variable Means

	School Grade in 2002		
Domain/Variable	**A/B/C**	**D**	**F**
Policies to Improve Low-Performing Students			
Require grade retention	0.76	0.82	0.79
Require summer school	0.40	0.57	0.36
Require before/after school tutoring	0.44	0.62	0.68
Require in-school supplemental instruction	0.79	0.86	0.89
Require tutoring	0.61	0.72	0.82
Require Saturday classes	0.05	0.14	0.14
Require other policy	0.30	0.29	0.50
Lengthening Instructional Time			
Sponsor summer school	0.53	0.60	0.67
Sponsor year-round classes	0.01	0.01	0.00
Sponsor extended school year	0.19	0.24	0.25
Sponsor Saturday school	0.10	0.21	0.31
Sponsor after-school tutoring	0.78	0.88	0.86
Sponsor other school services	0.33	0.27	0.41
Average length of school day 1st and 4th grade (in minutes)	376.41	376.09	382.31
Reduced Class Size for Subject			
Math	0.23	0.27	0.43
Reading	0.43	0.56	0.61
Writing	0.28	0.40	0.36
Low academic performance	0.44	0.55	0.68
Narrowing of Curriculum			
Minimum time spent on math	0.67	0.81	0.86
Minimum time spent on reading	0.71	0.87	0.86
Minimum time spent on writing	0.62	0.81	0.75
Minimum time spent on social studies	0.43	0.45	0.61
Minimum time spent on art/music	0.59	0.64	0.71
Scheduling Systems			
Block scheduling	0.35	0.43	0.52
Common prep periods	0.90	0.92	0.92
Subject matter specialist teachers	0.59	0.75	0.85
Organize teachers into teams	0.95	0.95	0.96
Looping	0.43	0.41	0.33
Multi-age classrooms	0.29	0.37	0.46
Other schedule structure	0.11	0.07	0.15
Policies to Improve Low-Performing Teachers			
Closer teacher supervision	0.98	0.99	1.00
Assign aide to teachers	0.30	0.52	0.59
Assign mentor to teachers	0.89	0.87	0.92
Provide additional professional development	0.99	1.00	1.00
Provide development/improvement plan	0.97	0.96	1.00
Other improvement strategy	0.14	0.13	0.20

Teacher Resources			
Minutes per week for collaborative planning/class preparation	450.12	452.75	424.09
Days per year for individual professional development	3.24	3.94	5.08
Funds per student per year for professional development	$14.71	$28.48	$45.53
Teacher Incentives			
Monetary reward	0.29	0.22	0.29
Comp/release time	0.56	0.56	0.71
Choice of class	0.17	0.20	0.30
Attendance at conferences and workshops	0.64	0.65	0.71
Special leadership position/assignment	0.63	0.67	0.85
Other incentives	0.25	0.25	0.44
Teacher Control (1 = No Influence / 5 = Complete Control)			
Teacher control of establishing curriculum	3.39	3.40	3.33
Teacher control of hiring new full-time teachers	2.90	2.75	3.00
Teacher control of budget spending	3.20	2.97	3.15
Teacher control of teacher evaluation	1.84	1.74	2.07
District Control (1 = No Influence / 5 = Complete Control)			
District control of establishing curriculum	3.62	3.49	3.56
District control of hiring new full-time teachers	2.72	2.80	2.52
District control of budget spending	3.23	3.45	2.89
District control of teacher evaluation	2.28	2.41	2.19
Principal Control (1 = No Influence / 5 = Complete Control)			
Principal control of establishing curriculum	3.54	3.51	3.48
Principal control of hiring new full-time teachers	4.35	4.26	4.39
Principal control of budget spending	3.90	3.77	4.00
Principal control of teacher evaluation	4.64	4.65	4.79
School Climate (1 = Not at all accurate/ 5 = Very accurate)			
High staff morale	3.18	2.91	2.93
Staff support and encourage each other	4.22	4.04	4.32
Teacher understanding of expectations	4.40	4.14	4.57
New teacher excellence (three or fewer years)	3.89	3.62	3.61
Experienced teacher excellence (more than ten years)	4.07	3.74	3.69
Student disruption does not interfere with learning	3.07	2.21	1.75
Parents do not worry about violence in school	3.38	3.12	2.93
Parents closely monitor academic progress of child	3.29	2.68	2.46
Miscellaneous			
Reduced class size gifted academic performance	0.42	0.40	0.21
Average 1st and 4th grade class size	25.21	25.06	23.83
Minimum time spent on science	0.46	0.53	0.61
Whole school reform model	0.27	0.41	0.64

Table 6: Mean Change in Variables between 2001-02 and 2003-04

	School Grade in Summer 2002		
Domain/Variable	**A/B/C**	**D**	**F**
Policies to Improve Low-Performing Students			
Require grade retention	0.10	0.03	0.26
Require summer school	0.01	-0.11	-0.05
Require before/after school tutoring	0.06	-0.04	0.15
Require in-school supplemental instruction	0.05	0.07	0.11
Require tutoring	0.07	0.00	0.10
Require Saturday classes	0.02	0.07	0.11
Require other policy	-0.02	0.04	-0.43
Lengthening Instructional Time			
Sponsor summer school	-0.03	0.00	0.10
Sponsor year-round classes (/10)	-0.01	0.00	0.00
Sponsor extended school year	0.09	0.01	0.12
Sponsor Saturday school	0.04	0.09	0.05
Sponsor after-school tutoring	0.06	0.03	0.15
Sponsor other school services	0.06	-0.23	0.11
Average length of school day 1st and 4th grade (in minutes)	-0.98	-1.10	1.77
Reduced Class Size for Subject			
Math	-0.01	-0.04	-0.05
Reading	-0.04	-0.10	-0.10
Writing	-0.05	-0.17	0.11
Low academic performance	0.02	-0.10	-0.21
Narrowing of Curriculum			
Minimum time spent on math	0.08	0.07	0.05
Minimum time spent on reading	0.14	0.06	0.05
Minimum time spent on writing	0.02	-0.11	0.00
Minimum time spent on social studies	0.04	0.13	-0.25
Minimum time spent on art/music	0.03	0.13	-0.10
Scheduling Systems			
Block scheduling	-0.01	0.00	0.18
Common prep periods	0.01	-0.01	0.00
Subject matter specialist teachers	0.01	0.06	0.06
Organize teachers into teams	0.01	-0.01	0.00
Looping (/10)	0.01	-0.14	2.94
Multi-age classrooms (/10)	0.03	-0.92	-1.88
Other schedule structure	-0.03	-0.04	0.00
Policies to Improve Low-Performing Teachers			
Closer teacher supervision (/10)	0.03	-0.14	0.00
Assign aide to teachers	-0.08	-0.10	-0.06
Assign mentor to teachers	0.02	0.13	0.05
Provide additional professional development (/10)	-0.03	0.00	0.00
Provide development/improvement plan (/10)	0.04	-0.14	0.00
Other improvement strategy	0.03	0.06	0.00

Teacher Resources			
Minutes per week for collaborative planning/class preparation	11.10	-49.36	82.50
Days per year for individual professional development	0.49	-0.16	0.25
Funds per student per year for professional development	$33.13	$9.69	$42.33
Teacher Incentives			
Monetary reward	0.03	0.07	0.06
Comp/release time (/10)	-0.04	0.00	0.00
Choice of class	0.02	0.03	0.12
Attendance at conferences and workshops (/10)	0.07	0.82	0.53
Special leadership position/assignment (/10)	0.06	0.83	-1.11
Other incentives	-0.05	-0.08	0.13
Teacher Control (1 = No Influence / 5 = Complete Control)			
Teacher control of establishing curriculum	-0.11	-0.15	-0.10
Teacher control of hiring new full-time teachers	-0.02	-0.07	-0.20
Teacher control of budget spending	-0.08	-0.15	-0.05
Teacher control of teacher evaluation	-0.09	-0.04	-0.45
District/Superintendent Control (1 = No Influence / 5 = Complete Control)			
District control of establishing curriculum	0.14	0.18	0.20
District control of hiring new full-time teachers	0.00	-0.24	0.30
District control of budget spending	0.04	-0.29	0.44
District control of teacher evaluation	-0.01	-0.03	-0.30
Principal Control (1 = No Influence / 5 = Complete Control)			
Principal control of establishing curriculum	-0.03	0.26	0.15
Principal control of hiring new full-time teachers	0.01	0.25	0.00
Principal control of budget spending	0.04	0.18	-0.11
Principal control of teacher evaluation	0.03	0.07	0.00
School Climate (1 = Not at all accurate/ 5 = Very accurate)			
High staff morale	-0.08	-0.10	-0.21
Staff support and encourage each other	0.09	0.05	0.26
Teacher understanding of expectations	0.05	0.09	-0.26
New teacher excellence (three or fewer years)	-0.02	-0.07	0.00
Experienced teacher excellence (more than ten years)	0.05	-0.15	0.17
Student disruption does not interfere with learning	-0.04	-0.08	-0.37
Parents do not worry about violence in school	-0.11	-0.04	-0.21
Parents closely monitor academic progress of child	0.02	-0.11	-0.32
Miscellaneous			
Reduced class size gifted academic performance	-0.01	-0.08	0.05
Average 1st and 4th grade class size	-2.06	-3.41	-2.70
Minimum time spent on science	0.05	0.13	-0.10
Whole school reform model	0.48	0.28	0.00

Table 7: Seemingly-Unrelated Regression and OLS Results of the Effect of Receiving an F Grade in Summer 2002 on School Policy in 2003-04

	Seemingly-Unrelated Regression				Index
	F vs. All			F vs. D	
Domain	(1)	(2)	(3)	(4)	(5)
Policies to Improve Low-Performing Students	0.437	0.273	0.301	0.308	0.345
	(0.097)	(0.121)	(0.139)	(0.139)	(0.167)
Lengthening Instructional Time	0.279	0.174	0.262	0.264	0.237
	(0.095)	(0.104)	(0.123)	(0.123)	(0.139)
Reduced Class Size for Subject	0.557	0.459	0.292	0.299	0.364
	(0.190)	(0.212)	(0.241)	(0.240)	(0.258)
Narrowing of Curriculum	0.140	0.146	0.143	0.137	0.123
	(0.086)	(0.097)	(0.107)	(0.107)	(0.124)
Scheduling Systems	0.208	0.361	0.370	0.377	0.361
	(0.120)	(0.118)	(0.138)	(0.139)	(0.143)
Policies to Improve Low-Performing Teachers	0.215	0.130	0.148	0.154	0.204
	(0.080)	(0.092)	(0.096)	(0.097)	(0.156)
Teacher Resources	0.723	0.882	1.069	1.048	0.867
	(0.390)	(0.465)	(0.529)	(0.532)	(0.672)
Teacher Incentives	0.220	0.208	0.120	0.125	0.170
	(0.160)	(0.174)	(0.202)	(0.202)	(0.202)
Teacher Autonomy	-0.060	-0.166	-0.084	-0.091	-0.095
	(0.129)	(0.143)	(0.159)	(0.158)	(0.199)
District Control	-0.169	-0.098	0.072	0.082	0.066
	(0.172)	(0.175)	(0.203)	(0.203)	(0.201)
Principal Control	0.169	-0.087	-0.297	-0.294	-0.329
	(0.129)	(0.142)	(0.161)	(0.161)	(0.206)
School Climate	-0.454	0.035	0.124	0.121	0.116
	(0.124)	(0.139)	(0.158)	(0.157)	(0.160)
Specifications also control for:					
2002 "Simulated Grade"	N	Y	Y	Y	Y
Lagged dependent variable	N	Y	Y	Y	Y
Other school characteristics	N	Y	Y	Y	Y
2002 grade points	N	N	Y	Y	Y

Notes: Standard errors are in parentheses. The estimates in columns (1)-(4) are based on seemingly-unrelated regressions of variables in each of the "domains" as described in Tables 5 and 6; the estimates in column (5) use an "index" of the variables in each of the domains, as described in the text. The estimates in columns (1)-(3) compare the F-graded schools to all higher-grade schools; those in columns (4) and (5) compare the F-graded schools to the D-graded schools. The "2002 Simulated Grade" is a vector of dummy variables indicating the grade the school would have received in 2002 using the 2001 school grading formula. The "lagged dependent variable" is the school's value for the 2004 policy in 2002 an indicator for whether the 2002 response is missing. "Other school characteristics" include: racial and ethnic composition of the school, the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the racial and ethnic composition of the school, the percentage of the school's student body that is classified as disabled, eligible for free or reduced-price lunch, limited English proficient, and gifted, and the "stability rate" of the school. "2002 grade points" is a cubic in the number of points the school received according to the summer 2002 accountability grading system.

Table 8: OLS Results of the Impact of Receiving an F Grade in Summer 2002 on School Selected Individual Policies in 2003-04

	F vs. All		F vs. D
	No Covariates	With Covariates	
Variable	(1)	(2)	(3)
Average 1st and 4th grade class size	-2.946 (0.733)	-0.754 (0.907)	-0.747 (0.908)
Use block scheduling	0.250 (0.097)	0.247 (0.115)	0.247 (0.115)
Use common prep period	0.010 (0.054)	0.108 (0.071)	0.111 (0.071)
Use other scheduling structure	0.066 (0.077)	0.181 (0.105)	0.185 (0.106)
Minutes per week for collaborative planning/class preparation	21.791 (45.482)	124.604 (62.221)	123.702 (62.288)
Reduced class size gifted academic performance	-0.245 (0.099)	-0.065 (0.127)	-0.065 (0.127)
Whole school reform model	-0.056 (0.085)	-0.017 (0.110)	-0.019 (0.110)
Minimum time spent on science	0.101 (0.097)	-0.136 (0.119)	-0.127 (0.119)

Notes: Standard errors are in parentheses. The estimates in columns (1) and (2) compare the F-graded schools to all higher-grade schools; those in column (3) compare the F-graded schools to the D-graded schools. All specifications control for the "2002 Simulated Grade" (a vector of dummy variables indicating the grade the school would have received in 2002 using the 2001 school grading formula); the "lagged dependent variable" (the school's value for the 2004 policy in 2002 an indicator for whether the 2002 response is missing); and "Other school characteristics" (the school's estimated expenditures per student in 2002 for special education, vocational education, education for "at-risk" students, and "regular" education, the number of students in the school, the racial and ethnic composition of the school, the percentage of the school's student body that is classified as disabled, eligible for free or reduced-price lunch, limited English proficient, and gifted, and the "stability rate" of the school); and a cubic in the number of points the school received according to the summer 2002 accountability grading system.

Table 9: The Effect of Including School Policy/Practice Variables on Regression-Discontinuity Estimates of the Effect of Receiving an F vs. D Grade in Summer 2002 on Seventh-Grade Student Performance: Fifth-Grade Cohort of 2002-03

Model specification	Standardized test score			
	High-stakes reading	High-stakes math	Low-stakes reading	Low-stakes math
Models excluding school policy/practice variables	0.080 (0.030)	0.094 (0.039)	0.059 (0.025)	0.045 (0.027)
Models including all school policy/practice variables	0.068 (0.037)	0.058 (0.042)	0.042 (0.027)	0.022 (0.028)
p-value of joint significance of policy variables requiring resources	*0.002*	*0.003*	*0.001*	*0.008*
p-value of joint significance of policy variables aimed at greater efficiency	*0.330*	*0.001*	*0.458*	*0.096*
Models including only the five school policy/practice domains with the strongest cross-sectional relationship with test scores	0.068 (0.037)	0.059 (0.043)	0.044 (0.026)	0.015 (0.029)
p-value of joint significance of policy variables	*0.000*	*0.005*	*0.005*	*0.008*

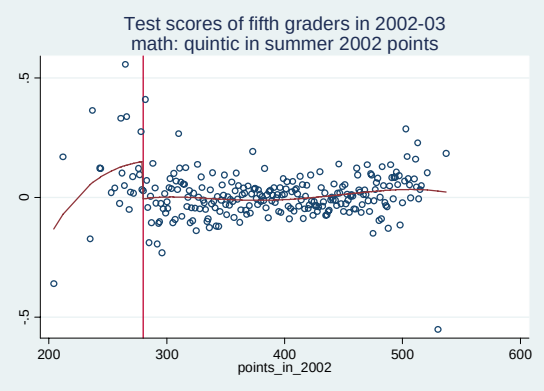
Notes: Standard errors adjusted for school-level clustering are in parentheses. Dependent variables are test scores standardized to the Florida average by grade level. Regressions include all variables described in Table 4, as well as the twelve policy indices described in Table 8 as well as the class size variable which itself is individually statistically significantly related to F grade receipt in summer 2002. The domains with the strongest cross-sectional relationship with test scores are policies to improve low student performance; policies to improve low teacher performance; policies regarding teacher autonomy; reflections of school climate; and class size in general.

Appendix Table 1: Difference in School Characteristics between Respondents and Non-Respondents to the 2004 School Survey Sample, by the School's Accountability Grade in 2001-02

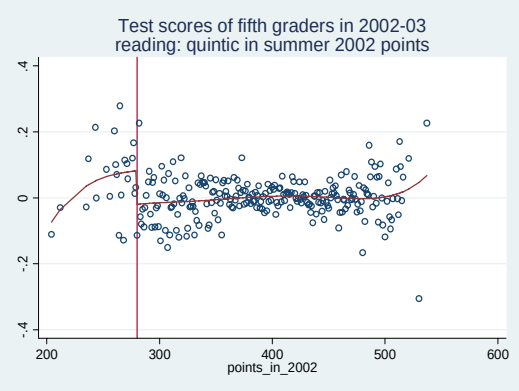
	School Grade in 2002		
Variable	**A/B/C**	**D**	**F**
Number of Students	1.356 (15.956)	39.510 (41.448)	-74.378 (86.885)
Proportion Black	-0.041 (0.014)	-0.055 (0.055)	-0.085 (0.085)
Proportion Asian	0.000 (0.001)	0.005 (0.003)	-0.002 (0.002)
Proportion Hispanic	-0.029 (0.014)	-0.028 (0.039)	0.021 (0.060)
Proportion Native American	0.000 (0.000)	-0.001 (0.001)	0.000 (0.001)
Proportion Mixed Race	-0.002 (0.001)	0.002 (0.002)	0.000 (0.004)
Proportion Free or Reduced-Price Lunch Eligible	-1.902 (1.453)	1.176 (2.787)	-0.622 (6.072)
Proportion Gifted	0.222 (0.337)	0.062 (0.360)	-0.421 (0.568)
Proportion Limited English Proficiency	-1.404 (0.704)	-1.664 (2.169)	1.222 (3.983)
Proportion Classified Disabled	1.255 (0.351)	0.499 (1.206)	3.159 (2.157)
Stability Rate	0.144 (0.177)	-0.071 (0.594)	2.899 (1.215)
Special Education Expenditures per Pupil	-$168.387 (131.352)	-$28.921 (430.789)	$575.636 (757.596)
Regular Education Expenditures per Pupil	-$118.849 (52.823)	-$313.196 (228.127)	$528.114 (394.682)
At-risk Students Expenditures per Pupil	-$119.886 (236.349)	-$2.533 (741.812)	$1,421.659 (1,521.961)
Vocational Education Expenditures per Pupil	-$4.938 (51.924)	-$61.475 (257.009)	-$1,318.667 (961.135)

Notes: Standard errors in parentheses. All school characteristics are based on data from the 2001-02 school year. The data include elementary schools in both the 2002 and 2004 sampling frames.

Appendix Figure 1a: Regression Discontinuity Estimates of Effects of "F" Grade on Fifth Graders' Test Score Gain Residuals

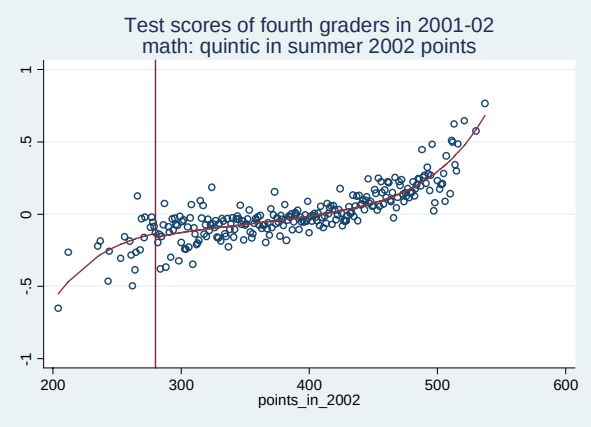


p-value of discontinuity: 0.000

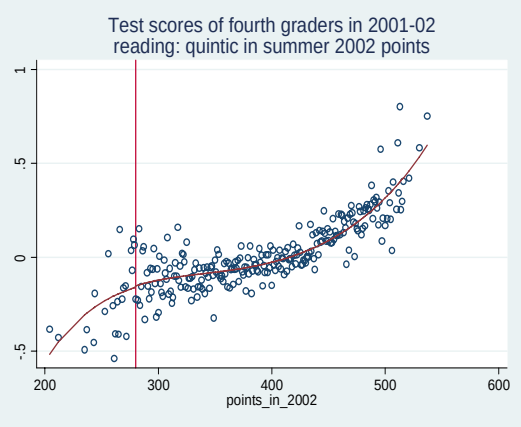


p-value of discontinuity: 0.001

Appendix Figure 1b: Regression Discontinuity Estimates of Effects of "F" Grade on Previous Year's Test Score Residuals



p-value of discontinuity: 0.564



p-value of discontinuity: 0.955

Note: The vertical line represents the 280 point threshold that separates a grade of "F" (to the left of the threshold) from a grade of "D" (to the right of the threshold.) The regression line plotted on the graphs represents the relationship between grade points (in quintics) and residual test scores.


CALDER