Journal of Public Economics 93 (2009) 1045–1057

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

How accountability pressure on failing schools affects student achievement

Hanley Chiang *

Mathematica Policy Research, Inc., P.O. Box 2393, Princeton, NJ 08543-2393, United States

ARTICLE INFO

Article history:
Received 5 December 2007
Received in revised form 15 June 2009
Accepted 15 June 2009
Available online 24 June 2009

JEL classification:
H75
I20
I28

Keywords:
School accountability
Sanction threats
Low-performing schools
School expenditures

ABSTRACT

Although an emerging body of evidence has shown that the threat of sanctions on low-performing schools can raise student test scores in the short run, the extent to which these test score improvements are due to schools' manipulation of the accountability system has remained uncertain. In this paper, I provide two new strands of evidence to evaluate the relative importance of educational reforms and gaming behavior in generating test score gains by threatened schools. First, using a regression discontinuity design that exploits Florida's system of imposing sanction threats on the basis of a cutoff level of performance, I estimate medium-run effects on student test scores from having attended a threatened elementary school. Threat-induced math improvements from elementary school largely persist at least through the first 1 to 2 years of middle school, while evidence for persistence of reading improvements is less consistent. Second, I analyze the effects of sanction threats on various features of educational production, and I find that sanction threats raise school spending on instructional technology, curricular development, and teacher training. Both strands of evidence are consistent with a predominant role for educational reforms in generating test score gains by threatened schools.

1. Introduction

The threat of sanctions on low-performing public schools has become a dominant form of incentive by which policymakers seek to raise student achievement. In the United States, penalties for repeated failure to meet performance standards have become universally established in all states as a result of the *No Child Left Behind Act of 2001* (NCLB). Beyond NCLB-mandated penalties, two-thirds of state education systems have their own, additional policies to penalize persistently low-performing schools (Bausell, 2007). Insofar as repeated failure is meaningfully penalized, struggling schools face a powerful incentive to raise their performance ratings. However, schools may have at their disposal a range of mechanisms for improving their ratings. The mechanisms consistent with policymakers' intent are those that reform the inputs and processes of educational production within failing schools, but schools may also choose to "game" or manipulate the accountability system in ways that raise test scores without contributing to students' knowledge and skills. In this paper, I use a rich collection of student-level and school-level data from Florida to evaluate the relative importance of educationally oriented reforms and gaming behavior in generating test score improvements by schools facing sanction threats. My analyses focus on two central questions. First, do sanction threats on elementary schools have persistent, positive impacts on student test scores even after the affected students have progressed into middle school? Second, do sanction threats induce significant changes in the quantity, allocation, or use of educational inputs within threatened schools? I gauge the centrality of educational reforms as a mechanism for improvement by the extent to which the answers are affirmative.

The importance of evaluating the sources of test score improvements in threatened schools has been underscored by two bodies of existing research on school accountability: one demonstrating positive, short-run effects of sanction threats on test scores, and another documenting the existence of school gaming behavior under accountability pressure. A number of studies on Florida's accountability system have found that sanction threats have raised the observed test scores of students during the time that they are attending the threatened schools (Greene, 2001; Figlio and Rouse, 2006; West and Peterson, 2006; Chakrabarti, 2007; Chakrabarti, 2008). The introduction of school accountability in Chicago has been shown to raise test scores by a greater extent within schools at higher risk of being placed on probation (Jacob, 2005).[1] Despite this encouraging evidence on test score gains, a second strand of research has found that schools under accountability pressure may alter testing conditions or student classifications to boost observed test scores in

* Tel.: +1 609 945 3310.
E-mail address: HChiang@mathematica-mpr.com.

0047-2727/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2009.06.002

[1] Other studies have examined overall effects of school accountability systems on test scores without focusing specifically on sanction threats. Carnoy and Loeb (2002) and Hanushek and Raymond (2005) provide cross-state evidence on the effects of accountability; Ladd (1999) examines a particular district's implementation of accountability; Grissmer and Flanagan (1998), Klein et al. (2000), and Koretz and Barron (1998) examine the time series of test scores within particular states that implement accountability.

ways unintended by policymakers. For instance, pressured schools have been shown to remove low-achieving students from school rating calculations by reclassifying them into special education (Deere and Strayer, 2001; Figlio and Getzler, 2002; Jacob, 2005; Cullen and Reback, 2006) or by subjecting them to longer disciplinary suspensions near testing dates (Figlio, 2006). Schools under pressure have also been observed to manipulate testing conditions through teacher cheating (Jacob and Levitt, 2003) and through inflating the caloric content of school lunches during the testing week (Figlio and Winicki, 2005). Collectively, these two strands of research raise — and leave open — the question of whether the observed test score gains from accountability pressure are *primarily* the result of gaming behavior or of educational reforms.

To evaluate the importance of educational reforms, I first estimate the causal effect of attending a threatened elementary school on students' test scores both *during* and *after* their elementary years. A common feature of school gaming behaviors is that either they should have minimal impacts on individual students' test scores if they merely change the pool of accountable students, or their impacts on any particular student's scores should be a purely transitory rise that disappears when the student exits the manipulated testing condition, such as the classroom of a cheating teacher. A persistent, threat-induced test score improvement would therefore be attributable to educational reforms. To estimate the causal effects of sanction threats, I use a regression discontinuity design that exploits the discontinuous structure of sanction threats in Florida's accountability system: schools whose prior-year performance lies below a threshold level must exhibit a sufficiently high performance in the current year to avoid sanctions. Short-run threat effects in the current year can thus be estimated from test score discontinuities at the threshold, and test score discontinuities exhibited by the same students in subsequent years yield estimates for medium-run effects.

Using the same regression discontinuity framework, a second empirical strategy assesses the impact of sanction threats on various features of educational production in threatened schools. If educational reforms are truly the main source of test score gains in threatened schools, then these schools should be observed to make changes to the inputs or technology of educational production. Motivated by open-ended interviews in which principals of threatened schools report undertaking changes to instructional practices and curricular content, my empirical analyses use school expenditure data to test for threat-induced resource shifts toward instructional and curricular reform.

Evidence from both approaches is generally consistent with the claim that educational reforms are the primary means by which threatened schools raise test scores. I find that most of the initial improvement in high-stakes math scores from attending a threatened elementary school persists for at least 1 to 2 years after entry into middle school; however, evidence for medium-run persistence of initial reading gains is mixed and not robust. I also find that sanction threats trigger increases in school expenditures on instructional and curricular development, teacher training, and technology surmised to support assessment-driven instruction, all of which are consistent with a reform-driven response to sanction threats.

This paper contributes to the existing literature on school accountability in a number of ways. First, to my knowledge, no prior study has examined the degree to which accountability pressure on schools persistently impacts all affected students beyond their time in those schools.[2] Second, this study extends a very sparse literature on the ways in which accountability pressure affects educational production. In the only other study on this topic of which I am aware, Jacob (2003) examines changes in school-level budgetary line items over a five-year period spanning the introduction of accountability in Chicago and finds declines in fine arts spending, decreases in the ratio of aides to teachers, and increases in the ratio of supervisors to teachers within lower-performing schools relative to higher-performing schools. My analyses provide the first estimates of immediate expenditure responses to the onset of sanction threats.

Simultaneously and independently of my paper, Rouse et al. (2007) have estimated the persistence of test score gains stemming from the same sanction threats examined by my paper. Whereas I find robust evidence of persistence only in math, they find persistence in both reading and math as a result of limiting their analyses to students initially affected by sanction threats in fifth grade.[3] In addition, Rouse et al. conduct surveys of Florida's principals and find that sanction threats induce schools to lengthen instructional time, change school-wide schedules, and increase the planning time and professional development with which teachers can improve instruction. Thus, their findings from survey data, similarly to my findings from administrative expenditure data, are consistent with the claim that schools facing accountability pressure are likely to implement pedagogical and curricular changes in their attempts to raise student achievement.

The remainder of this paper proceeds as follows. Section 2 gives an overview of the Florida accountability system. Section 3 describes the method for estimating initial effects of sanction threats on test scores, and section 4 presents results for these initial effects. Section 5 examines the persistence of test score gains from sanction threats. Section 6 analyses potential changes to educational production by which sanction threats affect achievement. Section 7 concludes.

2. Overview of the Florida accountability system

Florida's accountability system exemplifies a prototypical sanctioning regime in which schools failing once to meet a performance standard face an immediate threat of sanctions because a second failure within a short time span brings actual sanctions. The test scores used for calculating school ratings and threat status come from a high-stakes set of tests in reading, math, and writing known as the Florida Comprehensive Assessment Tests (FCAT), which are aligned to state curriculum standards. On the reading and math tests, which students take annually beginning in third grade during my analysis period, an individual student's FCAT achievement can be expressed as either a continuous scale score or a discrete achievement level ranging from 1 (worst) to 5 (best); roughly, levels 1, 2, and 3 correspond to below basic, basic, and proficient levels of achievement, respectively.[4,5] In addition, students also take the reading and math subtests of the Stanford Achievement Test, a low-stakes test for which scores are compared to a national norm group but are not used for accountability ratings.[6]

[2] One previous study by Donovan et al. (2006) uses data from a single university to estimate effects on collegiate performance from having attended a threatened high school.

[3] In section 5, I further discuss the sensitivity of the reading results to sample definition.

[4] The Florida Department of Education opts not to emphasize the labels of "basic" and "proficient" and only began using these labels to fulfill NCLB requirements.

[5] On the FCAT writing test, a student's performance is only expressed as an achievement level ranging from 1 to 6 in increments of 0.5, and standards for determining adequate writing achievement have evolved over time.

[6] In this paper, "FCAT" refers to the set of high-stakes tests that Florida calls the FCAT Sunshine State Standards (SSS) tests. Florida calls the Stanford tests "FCAT norm-referenced tests," but I will only refer to these tests as "low-stakes tests" to avoid confusion. Up through spring 2000, students took FCAT reading tests in grades 4, 8, and 10 and took FCAT math tests in grades 5, 8, and 10, and these scores were reported for accountability purposes. In spring 2001, only scores from these same subject-grade combinations were reported for accountability purposes, but all students in grades 3 through 10 took FCAT reading and math tests. In spring 2002, FCAT reading and math scores in all grades from 3 to 10 began to be reported for accountability purposes. The FCAT writing test has always been administered to and reported for grades 4, 8, and 10. FCAT science tests have recently been added to the state accountability system but were not relevant to school rating calculations in the period of my analysis. Since 2000, students have taken the Stanford Achievement Tests annually in grades 3 through 10.

In each summer since 1999, the Florida Department of Education (FLDOE) has assigned "school grades" of A, B, C, D, or F (ordered from best to worst) to public schools on the basis of FCAT scores from the recently completed spring. With an F grade denoting failure, schools receiving two Fs within a four-year period are subjected to both formal and informal penalties. Formally, students from double-F schools are permitted to transfer to another public school graded C or above[7] and, during my analysis period, also have the option of accepting a state-funded voucher to enroll in a private school.[8] School choice options serve as sanctions for double-F schools since a sufficiently large outflow of students can lead to cutbacks in staff and even school closure. In practice, the voucher take-up rate among eligible students was about 5% in 2002 (Peterson, 2006); while statewide data on the public school transfer option are not available, media reports from one large district indicate that the take-up rate was about 3% in 2002 (Grech, 2002). Thus, double-F status triggers a nontrivial but still modest outflow of students.

Beyond school choice penalties, an unofficial sanction also accompanies a second F: replacement of the incumbent principal. An examination of principal names for elementary schools receiving two Fs in the 2002 to 2005 time period reveals that in all five schools receiving both Fs under the same principal, a new principal was in place immediately after the second F. This rate of principal turnover far exceeds the 15% turnover rate in schools with just one F. Nevertheless, a web search for the names of the replaced principals indicates that many of these principals were demoted rather than fired from their district. In sum, a school that has just recently received its first F faces the threat of moderately — but not extremely — severe informal and formal sanctions attached to subsequent failure.[9]

Performance criteria for avoiding F grades have changed over time. From 1999 to 2001, an elementary school could get a D or better if, in *any one* subject, the percent of students attaining an acceptable achievement level was sufficiently high.[10] However, performance grades issued from 2002 to 2006 were based on very different criteria. These grades were determined by a continuous point index, known as "grade points," equal to the raw sum of six components, three of which reflected percentages of the school's students attaining acceptable achievement levels in the tested subjects, and three of which reflected percentages of students demonstrating acceptable achievement *growth* between the previous and current year.[11] The major innovation of this index was the inclusion of the latter three growth criteria. A school's letter grade was then determined by the interval of the grade point scale in which its grade point value lay; in particular, schools earning fewer than 280 grade points were designated as F schools. Notably, since the grade point index summed performance measures from multiple subjects, schools could not avoid F grades merely by focusing on one subject.

Table 1 shows the range of the post-2001 grade point scale corresponding to each letter grade as well as the 1999 through 2006 letter grade distribution for elementary schools in Florida.[12] As shown in Table 1, the number of F schools rose sharply with the implementation of the new grading system in 2002: although no schools received Fs in 2001, 38 elementary schools were designated as failing in 2002. It is advantageous to study the impact of the 2002 F grades on subsequent school performance because these F grades resulted almost exclusively from the implementation of the new grading scheme rather than from an idiosyncratic dip in performance that would lead to subsequent mean reversion. Indeed, my calculations indicate that only one of the 38 elementary schools receiving an F in 2002 would have received an F if the pre-2002 grading scheme had been in place in 2002.[13,14]

Finally, it is worth noting that a school's receipt of a 2002 F grade did trigger a small amount of assistance from the state. Principals of F schools attended a summit to connect with subject specialists who would volunteer to provide technical assistance for a few hours each month (Hegarty, 2002). Each F school also received state funding to hire one reading specialist. Thus, the effect of an F-triggered sanction threat cannot be disentangled from the effect of this assistance. Nevertheless, local educators regarded this assistance as quite modest (Flannery, 2002), and in nearly all states sanction threats are conjoined with some form of state assistance. Schools' responses to the receipt of a 2002 F grade are thus informative of the inducement effects of a typical sanction threat scenario.

3. Method for estimating initial threat effects on test scores

I first estimate the extent to which a school's receipt of a 2002 F grade impacts its test scores in 2003. The structure of school ratings in Florida is highly suited for empirical identification of threat effects through a regression discontinuity (RD) design. Whereas the unobservable factors influencing students' test scores, such as mean reversion and persistent school quality, are likely to be smoothly dependent on their schools' prior performance, a school's immediate threat status in 2003 is discontinuously determined by whether its 2002 grade point value lies below the F/D threshold of 280 points. Any observed, discontinuous difference in 2003 achievement between "F" and "D" schools arbitrarily close to the threshold can be attributed to the short-run impact of sanction threats. This approach follows a "sharp RD design" (Hahn et al., 2001) since a school's 2002 grade point value in relation to the threshold completely determines whether it receives an F and its concomitant sanction threat.[15]

To implement this approach, I designate 2002 as "year 0" and 2003 as "year 1" for notational convenience, and I let y_{is1} be the test score of student i attending school s in year 1, F_{s0} be an indicator for whether

[7] The student's district must offer her a place in at least one other school within the district. If she transfers to the offered school, then the district incurs all transportation costs; the student may also transfer to a public school outside of the district with available space, but the student's family incurs all transportation costs in this case.

[8] The voucher option was ruled unconstitutional by the Florida Supreme Court in January 2006.

[9] In addition, F schools face the stigma of having received a failing designation. Figlio and Rouse (2006) argue that test score improvements by F schools in the early years of the A through F grading system may have been driven more by stigma than by the threat of school choice sanctions. However, stigma and sanction threats are conceptually quite related: an F grade is stigmatizing at least partially because the state has chosen to signal the serious negative connotations of this grade through its sanctioning regime.

[10] Specifically, in *either* reading *or* math at least 60% of students must score at level 2 or above on the FCAT tests, *or* at least 50% of students must score at level 3 or above on the FCAT writing test.

[11] Specifically, the six components are: (i) percent of the school's students scoring at level 3 or above in reading; (ii) percent scoring at level 3 or above in math; (iii) average of the percent scoring at level 3 or above in writing and the percent scoring at level 3.5 or above in writing; (iv) percent demonstrating substantial growth in reading; (v) percent demonstrating substantial growth in math; and (vi) percent of the school's bottom quartile achievers in reading demonstrating substantial growth in reading. Since each percentage ranges from 0 to 100, the grade point index ranges from 0 to 600. Substantial growth in a given subject is defined as follows: a student scoring at level 3 or above in the prior year is regarded as making substantial growth if she achieves the same level or higher in the current year. A student scoring at level 1 or 2 in the prior year is regarded as making substantial growth if her achievement level increases by at least one level in the current year, or if the difference in scale scores between the current and prior year exceeds one year's worth of growth. The definition of one year's worth of growth is rather elaborate; see Florida Department of Education (2003) for details.

[12] I only consider schools classified by the FLDOE as having a solely "elementary" grade span and do not consider schools with a "combined" elementary–middle or elementary–middle–high school grade span.

[13] Calculations are available from the author upon request.

[14] It is also important to note that during the 2001–2002 school year schools were unlikely to have been able to anticipate their 2002 grades. Although the details of the new grading system were unveiled in November 2001 and approved in December 2001 (Harrison and Stepp, 2001; Stepp, 2001), the formulas for determining these new grades were just too complex for any school to form a reasonably accurate prediction of its grade.

[15] West and Peterson (2006) also estimate the short-run effects of the 2002 F grades on achievement in 2003. Although some of their analyses resemble a regression discontinuity design, they do not implement the more formal regression discontinuity methodology described in this section, and their estimates for the F effects are smaller than those found by my analyses.

Table 1
Grade point intervals for post-2001 performance grades and empirical distribution of 1999–2006 performance grades.

Letter grade	Grade point interval corresponding to letter grade from 2002 to 2006	Number of elementary schools receiving letter grade in:							
		1999	2000	2001	2002	2003	2004	2005	2006
A	410–600	117	475	362	597	870	962	946	943
B	380–409	199	170	309	350	338	319	324	360
C	320–379	670	573	601	422	272	259	283	289
D	280–319	436	254	212	121	52	62	79	36
F	0–279	63	4	0	38	16	9	18	7

Source: Author's tabulations of data from the Florida Department of Education. Counts do not include schools classified by the Florida Department of Education as having "combined" elementary–middle or elementary–middle–high grade spans.

school s received an F at the end of year 0, and GrdPts_{s0} be the grade points earned by school s at the end of year 0. First, I estimate the expected value of year 1 achievement for a student attending a (hypothetical) F school with a year 0 grade point value precisely at the F/D threshold; this quantity is the intercept $\hat{\alpha}_F$ from a regression of students' year 1 scale scores on their schools' year 0 grade points (net of 280), estimated using F schools with grade point values lying below 280 points by less than a distance or "bandwidth" h:

$$y_{is1} = \alpha_F + \beta_F(GrdPts_{s0} - 280) + \varepsilon_{is1} \quad \text{s.t.} \quad -h < GrdPts_{s0} - 280 < 0. \quad (1)$$

Likewise, to estimate expected year 1 achievement in a D school at the F/D threshold, I obtain the intercept $\hat{\alpha}_D$ from a comparable regression using D schools with grade point values exceeding 280 points by less than h:

$$y_{is1} = \alpha_D + \beta_D(GrdPts_{s0} - 280) + \varepsilon_{is1} \quad \text{s.t.} \quad 0 \le GrdPts_{s0} - 280 < h. \quad (2)$$

The estimated difference $\hat{\alpha}_F - \hat{\alpha}_D$ in year 1 scores between an F-graded and D-graded school precisely at the F/D threshold measures the effect of the F grade. This local linear regression approach essentially follows Imbens and Lemieux's (2008) guidelines for RD design. In practice, I combine estimation of the F-sided and D-sided local linear functions into one regression and control for a vector $\mathbf{X}$ of covariates:

$$y_{is1} = \beta_0 + \beta_1 F_{s0} + \beta_2(GrdPts_{s0} - 280) + \beta_3(F_{s0} \times (GrdPts_{s0} - 280)) + \mathbf{X}_{is1}\gamma + \varepsilon_{is1}. \quad (3)$$

In Eq. (3), the estimated coefficient $\hat{\beta}_1$ on the F grade indicator captures the short-run effect of the F grade on test scores. Each student's scale score is standardized by the statewide mean and standard deviation of scores in the student's expected grade level from an initial year; therefore, effect sizes are expressed in terms of standard deviations (s.d.) of student-level scores.[16]

Although the RD design is able to identify the causal effect of F grade receipt, the estimand, $E(y_{is1}|F_{s0}=1,\ \text{GrdPts}_{s0}=280) - E(y_{is1}|F_{s0}=0,\ \text{GrdPts}_{s0}=280)$, must be interpreted carefully. The estimated average effect is conditional on a fairly low grade point value of 280, so nothing inherent in the design permits generalization to higher-performing schools. Moreover, the treatment group, consisting of hypothetical F schools precisely at the threshold, must improve performance only by an arbitrarily small amount to avoid sanctions. Thus, the RD design mimics a randomized experiment in which the treatment group faces consequences if its performance is minutely lower in the following year, whereas the control group will only face such consequences after 2 years of minutely lower performance. Essentially, the RD design captures responses to the *immediacy* of sanction threats and not merely to the *existence* of a sanction component in the accountability regime. Immediacy of sanction threats is of considerable relevance to current policy debates: whereas the existence of sanctions was common even prior to NCLB, strict sanction timelines prescribed by the federal law have subjected more schools to immediate sanction threats.[17] Moreover, insofar as the stakes are higher when threats are more immediate, schools under immediate threat have the greatest incentives to raise test scores.

When estimating Eq. (3), I use regular, non-charter elementary schools operational in both 2002 and 2003.[18] In order for the F grade dummy to indicate the presence of sanction *threats* rather than actual sanctions, I remove from the estimation sample seven F schools and fifteen D schools that received F grades in any year from 1999 to 2001, leaving 28 F schools and 100 D schools in the bandwidth-unrestricted sample. Fig. 1 shows the number of in-sample schools observed within each five-point bin of 2002 grade points. Reassuringly, the densities of schools immediately below and above the threshold are similar, indicating that it is unlikely for schools in the proximity of the threshold to have actively manipulated their letter grade status.[19]

Within the schools of interest, since students only begin taking state tests in third grade, I limit the sample to fourth, fifth, and (the very few) sixth graders enrolled in 2003 in order to observe prior-year test scores.[20] Furthermore, in most regressions I restrict attention to "accountable students," or the three-fourths of students included in school grade calculations. This group excludes students with disabilities,[21] those in the first or second year of English for Speakers of Other Languages (ESOL),[22] and those not continuously enrolled at the

[16] The initial year, or the year in which the test is first given to all grade levels, is 2001 for the FCAT and 2000 for the low-stakes test. "Expected grade level" is the grade level in which the student would be observed if she were not retained in any year after her first year of being observed in the dataset.

[17] Under NCLB, penalties must be imposed on schools aided by the federal Title I program after two successive years of failure to meet a performance standard known as "adequate yearly progress."

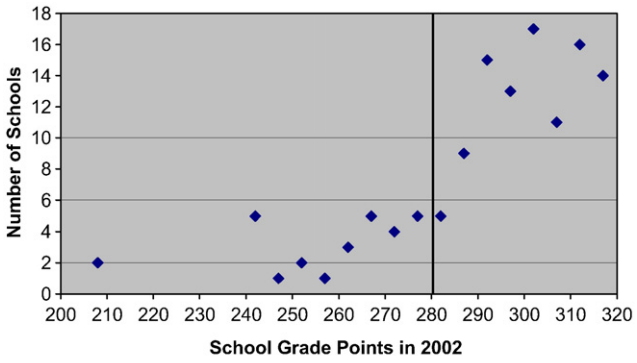
[18] Among the elementary schools receiving F or D grades in 2002, one F school and one D school are charter schools, and one F school and two D schools are no longer in existence in 2003. Moreover, one F school and three D schools were classified by the FLDOE as "elementary" in 2002 but were classified as having a "combined" (e.g. elementary and middle) grade span in 2003; although I only use schools classified as "elementary" in 2003, this sample criterion does not impact the results.

[19] This observation is in the spirit of the test of McCrary (2008) for agents' manipulation of the treatment-determining variable in an RD design. However, the number of bins in my analysis does not yield enough statistical power to conduct a formal test of manipulation.

[20] In Florida, a typical elementary school spans kindergarten through fifth grade (although a few elementary schools also have sixth grade), and a typical middle school spans sixth through eighth grade.

[21] Students are excluded if they are: educable mentally handicapped, trainable mentally handicapped, orthopedically impaired, language impaired, deaf or hard of hearing, visually impaired, emotionally handicapped, specific learning disabled, profoundly mentally handicapped, dual-sensory impaired, autistic, severely emotionally disturbed, traumatic brain injured, developmentally delayed, or "other health impaired."

[22] I do not observe how long a student has been enrolled in ESOL, so I count all students in ESOL as "excluded." Since only 5 to 6% of students in F and D schools are in ESOL, this procedure will only slightly inflate the number of excluded students.



Note: Each diamond represents the number of schools in a five-point bin of 2002 school grade points.

Fig. 1. Number of schools in sample, by school grade points in 2002.

same school from the fall to the spring. Since excluded students still take all tests, I do present additional results for threat effects on excludible test scores and on the probability of student reclassification.

Student-level longitudinal data on the population of students in Florida's public schools were provided by the Florida Department of Education (FLDOE). I have access to data on every student's reading and math scores, demographic characteristics, disciplinary information, and attendance rates from the 1998–1999 to 2004–2005 school years. The data do not include writing scores.

Before proceeding to the formal regression results, it is instructive to examine graphical depictions of the effects of the F grade. Fig. 2A and B plot average FCAT reading and math scores, respectively, achieved by schools in 2003 (or year 1) against their grade points in 2002 (or year 0). Data points represent actual values of the dependent variable averaged within five-point bins of 2002 grade points, and the 2002 grade point scale is divided by a vertical line representing the F/D threshold. On either side of the threshold, I have also drawn lines representing local regressions of the dependent variable on 2002 grade points, estimated using various bandwidths but without controlling for any covariates. For a given bandwidth, the difference between the intercepts of the F-sided and D-sided regression lines at the F/D threshold depicts the estimated F effect. In each figure, the depicted F effect is clearly positive at all bandwidths; in fact, school grade tabulations (not shown) reveal that all except one of the schools graded F in 2002 managed to earn a D grade or better in 2003. However, the figures also show that observed F effects appear to expand as bandwidths diminish. It is therefore of importance to select the preferred bandwidth carefully.

Bandwidth selection is informed by minimizing a statistical quantity known as the cross-validation criterion (Ludwig and Miller, 2005; Imbens and Lemieux, 2008). The underlying concept is as follows: since the local linear regressions to the left and right of the F/D threshold are aimed at predicting y_{is1} precisely at the discontinuity, the chosen bandwidth on either side of the threshold should minimize the expected squared prediction error at the threshold. The cross-validation criterion is an estimate of this expected squared error for each candidate bandwidth h. To construct this estimate, I identify two groups of schools, labeled G^F and G^D, consisting of the 50% of F-graded and D-graded schools, respectively, that are closest to the threshold in 2002. For each school $k \in G^F$, I run a *separate* regression of school-averaged scores in year 1, $\bar{y}_{s1}$, on year 0 grade points using only F-graded schools s to the left of (i.e., below) school k on the grade point scale by less than a bandwidth h:

$$\bar{y}_{s1} = \beta_0 + \beta_1 GrdPts_{s0} + \varepsilon_{s1} \quad \text{s.t.} \quad GrdPts_{k0} - h < GrdPts_{s0} < GrdPts_{k0}. \tag{4}$$

For a fixed value of h, each separate estimate of Eq. (4) produces a fitted value $\hat{\bar{y}}^h_{k1}$ for a different school k, and the *F-sided cross-validation criterion* is the mean squared prediction error

$$CV_F(h) = \frac{1}{|G^F|} \sum_{k \in G^F} \left(\bar{y}_{k1} - \hat{\bar{y}}^h_{k1} \right)^2 \tag{5}$$

over all schools k in G^F. Likewise for D schools, I estimate regressions of the form[23]

$$\bar{y}_{s1} = \beta_0 + \beta_1 GrdPts_{s0} + \varepsilon_{s1} \quad \text{s.t.} \quad GrdPts_{k0} < GrdPts_{s0} < GrdPts_{k0} + h \tag{6}$$

separately to the right of each $k \in G^D$ to obtain $\hat{\bar{y}}^h_{k1}$, and I calculate the *D-sided cross-validation criterion*

$$CV_D(h) = \frac{1}{|G^D|} \sum_{k \in G^D} \left(\bar{y}_{k1} - \hat{\bar{y}}^h_{k1} \right)^2. \tag{7}$$

On the F and D sides of the threshold, I thus give preference to bandwidths h^* that come close to minimizing $CV_F(h)$ and $CV_D(h)$, respectively.

Appendix Fig. A.1 plots CV_F and CV_D against bandwidth for each of the considered FCAT subjects.[24] The analyses indicate that a 28-point bandwidth is the smallest at which all cross-validation criteria are near their minima. To ensure that estimates for the F grade effects are not unduly influenced by observations far away from the threshold, I use this smallest justifiable bandwidth in the main analyses that follow.

4. Results for initial threat effects on test scores

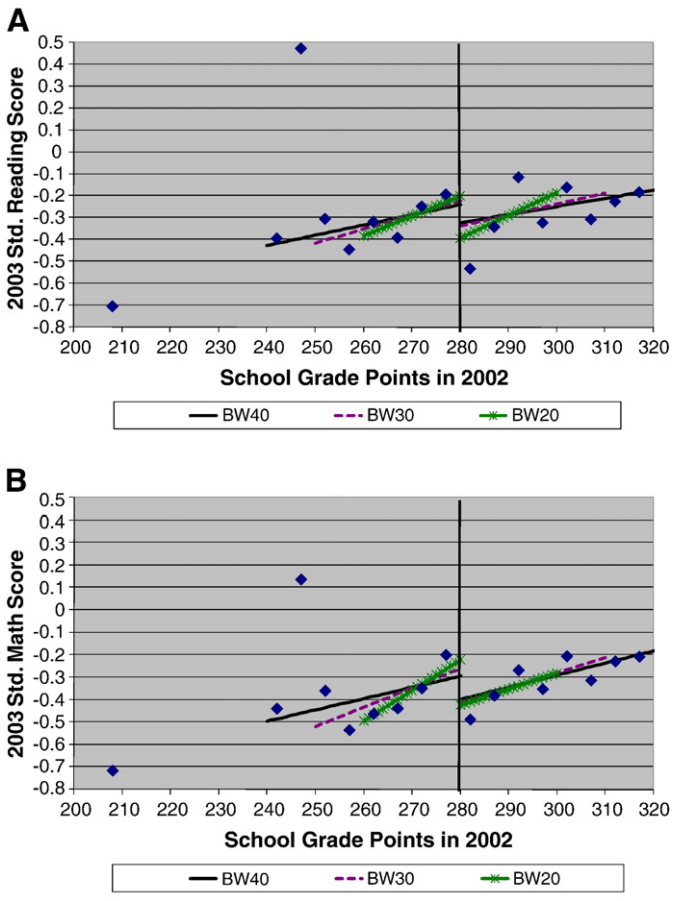
Given that Fig. 2A and B visually indicate a positive effect of 2002 F grades on test scores in 2003, I turn to more formal results based on estimation of Eq. (3) in the presence of covariates. Covariates have little influence on the size, but greatly improve the precision, of the estimated F effects. Table 2 lists the control variables used in Eq. (3) and provides sample means. Controls include variables for each student's demographic characteristics and prior achievement, for the average characteristics of the student's school-by-grade peers, for various measures of prior-year school resources, and for prior-year instructional quality (gauged by test score value added) in the student's current school-by-grade.[25,26] Table 2 also shows the

[23] In estimation of Eqs. (4) and (6), schools are weighted by the number of accountable students. Also, estimation samples for Eq. (4) exclude an outlying school located at a 2002 grade point value of 248; this school is evident in Figures 2A and B.

[24] All appendix material is available online at the website of the *Journal of Public Economics*.

[25] Specifically, with a student's 2003 FCAT score in a given subject as the dependent variable, the set of control variables consists of: indicators for blacks, Hispanics, females, students on free or reduced-price lunch, and students new to their school in 2003; mean values of the preceding indicators in the student's school-by-grade peer group; cubic polynomials of the student's 2002 scores in FCAT reading and FCAT math; linear terms of the student's 2002 scores in low-stakes reading and low-stakes math; mean values of 2002 (averaged FCAT and low-stakes) reading scores and 2002 (averaged FCAT and low-stakes) math scores in the student's current school-by-grade peer group; 2003 grade level indicators; indicator for the student's school being located within or on the fringe of a large city, defined by the Common Core of Data; operating costs per pupil (logged), pupil-to-teacher ratio, and average teacher experience in 2002 for the student's current school (obtained from the online Florida School Indicator Reports); and mean 2002 value added on the given test generated by the student's current school-by-grade instructional team, where value added is the residual from a statewide regression of 2002 scores on all individual test score covariates in 2001. When low-stakes scores are the dependent variable, all control variables remain the same, except that the set of individual prior-year achievement variables consists of cubic polynomials of 2002 low-stakes scores and linear terms of 2002 FCAT scores, and value-added measures are based on residuals of low-stakes scores.

[26] Even though the 2002 grade point index already incorporates some value-added elements, these elements take the form of *percentages* of students demonstrating acceptably high achievement growth, rather than a simple mean of value added across students. Thus, adding controls for mean 2002 value added generated by the student's current school-by-grade instructional team still improves the precision of the estimates.



Notes: Each diamond represents the average value of the dependent variable within five-point bins of 2002 school grade points. All local linear regression lines are drawn with the slopes and intercepts prevailing at the 2002 grade point value of 280 points. "BWy" denotes that the bandwidth is equal to y points. The BW40 line is estimated without the outlying observation located at a 2002 grade point value of 248 points, which corresponds to one school.

Fig. 2. A. Average standardized FCAT reading score in 2003, by school grade points in 2002. B. Average standardized FCAT math score in 2003, by school grade points in 2002.

predicted values of each covariate at the discontinuity from F-sided and D-sided local linear regressions of the covariate on 2002 grade points. In an ideal RD design, covariates should be balanced between treatment and control units at the discontinuity. Table 2 indicates that the F-sided and D-sided predicted values do not significantly differ for most covariates. Race/ethnicity composition and prior reading instructional quality are not balanced between F and D schools, but the differences are not significant at the 5% level, and inclusion or exclusion of these variables does not alter the basic conclusions described below.

Table 3 presents estimates for the short-run effects of a school's 2002 F grade on the test scores of its accountable students in 2003. Estimates based on the preferred 28-point bandwidth indicate that the onset of sanction threats from being graded as failing induced threatened schools to raise FCAT scores in the following year by 0.11 standard deviations (s.d.) in reading and 0.12 s.d. in math.[27] In reading, the estimated F effect on FCAT scores does not vary considerably with bandwidth; in math, however, estimated effects rise appreciably as bandwidths diminish.

One possible threat to internal validity is the potentially confounding influence of mean reversion. Since annual, school-level aggregate test scores are noisy measures of true school performance (Kane and Staiger, 2002a; Kane and Staiger, 2002b; Chay et al., 2005), schools receiving F grades in 2002 may have experienced transitory negative shocks that ceased to be present in 2003. If so, then inclusion of controls for historical school performance in 2001, 2000, and 1999[28] would have tended to diminish the size of the estimated F effects, which is not borne out by results in the penultimate row of Table 3.

Results from a "placebo test" also confirm that mean reversion is unlikely to be driving the main results. Suppose that the estimated effects of the 2002 F grades were due to a mean-reverting tendency for the worst-performing schools in a given year to bounce upward in performance in the following year. In such a scenario, schools that *would have* earned Fs in 2001 on the new grade point scale, had the new scale been in place, should exhibit a significantly greater test score improvement from 2001 to 2002 than schools that would have scored in the D range of the scale, even though *in actuality* the scale was not in place and no F grades were issued in 2001. As detailed in Appendix C, I thus assign "2001 pseudo-grades" to schools on the basis of a grade point index constructed analogously to the actual 2002 index. I find that pseudo-F schools improved by insignificant increments of 0.038 s.d. in FCAT reading and 0.018 s.d. in FCAT math relative to pseudo-D schools from 2001 to 2002,[29] indicating that the significant effects of the actual 2002 F grades are unlikely to be generated by mean reversion.

Another potential source of bias stems from possible student self-selection into or out of F and D schools immediately after the 2002 grades were issued, in such a way that unobservable student characteristics are not balanced across the discontinuity in 2003. Such a scenario is unlikely, given the absence of explicit school choice options for students of single-F and D schools. Nevertheless, I conduct a robustness check by switching to an estimation sample of students who are *predicted* to attend an F or D school in 2003 due to being observed in a non-terminal grade level of that school in 2002. In Eq. (3), the indicator for actually attending an F school in 2003, F_{s0}, is then instrumented by an indicator for being predicted to attend an F school.[30] The IV estimates, shown in the final row of Table 3, yield F grade effects of 0.12 s.d. and 0.09 s.d. on FCAT reading and math scores,

[27] Without covariates, the estimated F effects on FCAT reading scores and FCAT math scores are 0.110 (with standard error 0.106) and 0.145 (with standard error 0.104), respectively, at a 28-point bandwidth. Thus, as mentioned previously, inclusion of covariates leads to little change in point estimates but considerable improvement in precision.

[28] The timing of test introduction is the main constraint for constructing historical measures of school performance. Recall that the FCAT tests and low-stakes tests are administered to all third through tenth graders beginning in 2001 and 2000, respectively; up through 2000, FCAT tests are administered to selected grades only, and I only possess students' FCAT achievement levels, but not their scale scores, from the pre-2001 period. Thus, when FCAT (respectively, low-stakes) scores are the dependent variable in Eq. (3), I construct the measure of 2001 school performance to be the school-average residual from a statewide regression of students' 2001 FCAT (low-stakes) scores in the given subject on grade level dummies and cubic polynomials of the students' 2000 low-stakes reading and math scores. When FCAT scores are the dependent variable, the measure of 2000 school performance is the fraction of students attaining level 2 or above on the FCAT in the given subject; when low-stakes scores are the dependent variable, the measure of 2000 school performance is the school's average low-stakes score (adjusted for grade level) in the given subject. For all dependent variables, the measure of 1999 school performance is the fraction of students attaining level 2 or above on the FCAT in the given subject.

[29] These relative improvements are estimated from regressions using a pooled sample of accountable students attending pseudo-F or pseudo-D schools in either 2001 or 2002. I regress 2002 FCAT scores on school dummies, a year 2002 dummy, and an interaction term between the year 2002 dummy and a dummy for a 2001 pseudo-F grade; the relative improvement is the coefficient on the interaction term. I also control for the same covariates used in Eq. (3), with the exception that lagged FCAT scores are unavailable and are thus replaced by a cubic polynomial of lagged low-stakes scores.

[30] The first-stage coefficient on the instrument is 0.793 (with standard error 0.033). The variables multiplying β_2 and β_3 in Eq. (3) form a spline in the 2002 grade points of students' *predicted* schools. This framework represents a fuzzy RD design in which the grade point threshold determining the 2002 letter grades of students' predicted schools generates discontinuous changes in 2003 test scores and in the probability of attending an F school in 2003; the IV estimator is the ratio of these discontinuous changes.

Table 2
Sample means and predicted values at F/D threshold for control variables and exclusion variables.

Variable	Sample means for students in:		Predicted values at F/D threshold for:	
	F schools	D schools	F schools	D schools
	Control variables for accountable students			
1{black}	0.811	0.608	0.901	0.705*
1{Hispanic}	0.097	0.148	0.056	0.188*
1{female}	0.537	0.516	0.510	0.512
1{free/reduced-price lunch}	0.909	0.839	0.908	0.937
1{new student}	0.168	0.163	0.137	0.101
Fraction in school-by-grade:				
Black	0.784	0.590	0.869	0.679*
Hispanic	0.121	0.170	0.083	0.218
Female	0.495	0.488	0.478	0.481
Free/reduced-price lunch	0.926	0.853	0.920	0.945
New student	0.244	0.214	0.197	0.166
FCAT reading score in 2002	−0.491	−0.371	−0.425	−0.488
FCAT math score in 2002	−0.593	−0.463	−0.503	−0.570
Low-stakes reading score in 2002	−0.455	−0.351	−0.406	−0.468
Low-stakes math score in 2002	−0.482	−0.369	−0.364	−0.512
Mean in school-by-grade:				
2002 reading score (avg of FCAT and low-stakes)	−0.677	−0.556	−0.614	−0.715
2002 math score (avg of FCAT and low-stakes)	−0.735	−0.596	−0.634	−0.753
1{grade 4}	0.533	0.494	0.526	0.497
1{grade 5}	0.449	0.488	0.453	0.470
1{grade 6}	0.018	0.018	0.022	0.033
1{large city}	0.458	0.604	0.338	0.582
ln(operating costs per pupil) in 2002	8.72	8.65	8.71	8.64
Pupil-to-teacher ratio in 2002	14.9	15.9	14.8	15.0
Avg teacher experience in 2002	10.5	10.5	11.5	10.2
Mean value added generated by current school-by-grade instructional team in 2002:				
FCAT reading	−0.122	−0.079	−0.156	−0.063*
FCAT math	−0.165	−0.103	−0.159	−0.108
	Exclusion variables for all students			
1{excluded due to disability}	0.188	0.145	0.182	0.166
1{ESOL}	0.052	0.058	0.061	0.093
1{not enrolled in school in October}	0.064	0.046	0.054	0.053
1{excluded from accountability for any reason}	0.292	0.237	0.277	0.289
	Total sample size			
Number of accountable students	1897	8770		
Number of schools	20	68		

Notes: Sample means and sample sizes pertain to schools with 2002 grade point values within 28 points of the F/D threshold. Predicted values at the F/D threshold are based on a local linear regression of the indicated variable on 2002 grade points using a bandwidth of 28 points. Asterisks indicate that the D-sided predicted value differs significantly from the F-sided predicted value at a: *10% level, **5% level, and ***1% level. Tests of significant differences are based on standard errors clustered by school.

respectively, with the math effect losing statistical significance due to greater estimator imprecision. Overall, the IV estimates remain fairly proximate in magnitude to the estimates from the main specification.

In the last two columns of Table 3, I show the estimated effects of F grades on low-stakes scores. I find no evidence that low-stakes reading scores rise as a result of sanction threats. In math, the

Table 3
Effect of 2002 school F grade on test scores of accountable students in 2003.

Bandwidth/specification	Dependent var: standardized test score in 2003			
	FCAT reading	FCAT math	Low-stakes reading	Low-stakes math
40 points	0.092	0.071	0.000	0.014
	(0.052)*	(0.053)	(0.039)	(0.043)
40 points without outlier	0.110	0.078	0.005	0.015
	(0.050)*	(0.053)	(0.040)	(0.044)
28 points (preferred bandwidth)	0.112	0.118	0.025	0.065
	(0.056)*	(0.052)**	(0.045)	(0.046)
20 points	0.118	0.173	0.038	0.124
	(0.084)	(0.074)**	(0.053)	(0.050)**
28 points with controls for 1999–2001 school performance	0.107	0.123	0.028	0.061
	(0.057)*	(0.050)**	(0.046)	(0.045)
28 points with attendance in F school instrumented by predicted attendance	0.122	0.087	0.042	0.040
	(0.067)*	(0.056)	(0.051)	(0.054)

Notes: Each cell, taken from a separate regression, is the estimated coefficient on an indicator for attending a school with a 2002 F grade in 2003. Standard errors clustered by school are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and *** 1% level. Sample sizes by bandwidth/specification (FCAT sample size, low-stakes sample size) are: 40 points (15,743; 15,734), 40 points without outlier (15,682; 15,673), 28 points (10,667; 10,670), 20 points (6500, 6495), 28 points with 1999–2001 controls (10,667; 10,670), and 28 points with instrument (11,016; 11,028). All regressions include the control variables listed in Table 2 as well as: 2002 school grade points, 2002 grade points interacted with F grade indicator, and quadratic and cubic terms in student's 2002 reading and math scores on the indicated test (FCAT or low-stakes). School-level and school-by-grade level control variables in the instrumental variables specifications refer to the student's predicted school. "Outlier" refers to a single school with a 2002 grade point value of 248; see Fig. 2A and B.

estimated gain in low-stakes scores from attending a threatened school is about 0.06 s.d. at the preferred bandwidth but is insignificant and highly sensitive to bandwidth. The larger effects of accountability pressure on high-stakes scores relative to low-stakes scores can be due to various factors, as discussed by Jacob (2007). One possibility is that schools manipulate testing conditions more actively on high-stakes testing days, but the analyses in section 5 will show the unimportance of these gaming behaviors in accounting for high-stakes gains. Other possible explanations, which I do not have the data to evaluate, include differential student effort and differences in test format or content between the high-stakes and low-stakes exams.

Table 4 examines the extent to which the test scores of different student subpopulations rise as a result of attending threatened schools. In both reading and math, sanction threats have the greatest positive effect on the test scores of students who performed most poorly during the previous year. The greater magnitude of effects on the lowest-performers is unsurprising since these students have the largest potential for growth, for which schools are rewarded under Florida's grade point system; however, these patterns differ from those found in school rating regimes based on pass rates, which yield the largest effects on students near the margin of passing (Jacob, 2005; Reback, 2008; Neal and Schanzenbach, forthcoming). Table 4 also shows that a school's receipt of an F has negative effects on the test scores of students excluded from school rating calculations; although imprecise, these estimates nevertheless suggest that the types of responses undertaken by threatened schools do not appear to benefit excluded students.

Since the test scores of accountable and excluded students show strikingly different responses to sanction threats, it is possible that the positive effects observed for accountable students are the spurious artifacts of strategic efforts by threatened schools to exclude students who are unlikely to improve. Although I do not observe *ex ante* the likelihood that any particular student will improve, I test whether receipt of an F grade induces schools to increase the overall rate of exclusion, using the main RD specification with a student-level exclusion indicator or disability indicator as the dependent variable.[31] The results, displayed in Table 5, provide no support for the hypothesis that sanction threats increase the overall rate of exclusion. The F grade is observed to have a modest effect on the disability rates of students already proficient in math, but this subpopulation accounts for very little of the test score gains from accountability pressure. In all, Table 5 confirms that the observed F grade effects on the test scores of accountable students are not driven by increased student reclassification.

In sum, I find that the onset of sanction threats leads to a moderately sizable rise in the high-stakes test scores of students included in accountability calculations during the first year of a school's threatened status. Sanction threats have little to no positive effects on test scores irrelevant to school ratings. These short-run effects highlight the importance of discerning the sources of observed achievement gains by accountable students.

Table 4
Effect of 2002 school F grade on test scores in 2003: by subpopulation.

Subpopulation	Dependent var: standardized test score in 2003			
	FCAT reading	FCAT math	Low-stakes reading	Low-stakes math
All accountable students	0.112	0.118	0.025	0.065
	(0.056)**	(0.052)**	(0.045)	(0.046)
Level 1 (below basic) in subject in 2002	0.156	0.175	0.036	0.127
	(0.069)**	(0.083)**	(0.052)	(0.052)**
Level 2 (basic) in subject in 2002	0.115	0.039	0.023	0.028
	(0.064)*	(0.060)	(0.064)	(0.056)
Level 3 (proficient) or above in subject in 2002	0.056	0.097	0.030	0.032
	(0.056)	(0.057)*	(0.048)	(0.066)
Grade 4 in 2003	0.104	0.117	0.009	0.045
	(0.084)	(0.080)	(0.066)	(0.061)
Grade 5 in 2003	0.091	0.165	0.030	0.096
	(0.061)	(0.075)**	(0.048)	(0.077)
Excluded students	−0.088	−0.046	−0.023	−0.059
	(0.065)	(0.099)	(0.062)	(0.066)

Notes: Each cell, taken from a separate regression, is the estimated coefficient on an indicator for attending a school with a 2002 F grade. Standard errors clustered by school are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and ***1% level. Sample sizes by subpopulation (FCAT reading sample size, FCAT math sample size, low-stakes reading sample size, low-stakes math sample size) are: all accountable (10,667; 10,667; 10,670; 10,670), level 1 (4120, 3775, 4124, 3785), level 2 (2130, 3191, 2130, 3193), level 3 or above (4417, 3701, 4416, 3692), grade 4 (5346, 5346, 5345, 5345), grade 5 (5132, 5132, 5139, 5139), and excluded (3508, 3508, 3506, 3506). All regressions use a 28-point bandwidth and include the control variables listed in Table 2, as well as: 2002 school grade points, 2002 grade points interacted with F grade indicator, and quadratic and cubic terms in student's 2002 reading and math scores on the indicated test (FCAT or low-stakes).

Table 5
Effect of 2002 school F grade on exclusion status in 2003: by subpopulation.

Subpopulation	Dependent var: indicator for being in exclusion category in 2003	
	Disability	Any exclusion
All students	0.001	0.013
	(0.010)	(0.019)
Level 1 (below basic) in math in 2002	−0.002	0.019
	(0.016)	(0.028)
Level 2 (basic) in math in 2002	−0.008	0.015
	(0.012)	(0.025)
Level 3 (proficient) or above in math in 2002	0.019	0.003
	(0.010)*	(0.026)
Mean of dependent var in D schools	0.145	0.237

Notes: Each cell, taken from a separate regression, is the estimated coefficient on an indicator for attending a school with a 2002 F grade. Standard errors clustered by school are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and ***1% level. Sample sizes by subpopulation are: all accountable (14,175), level 1 (6102), level 2 (3892), and level 3 or above (4181). All regressions use a 28-point bandwidth and include the control variables listed in Table 2, as well as: 2002 school grade points, 2002 grade points interacted with the F grade indicator, quadratic and cubic terms in the student's 2002 reading and math FCAT scores, lagged dependent variable, mean of the lagged dependent variable in the student's 2003 school-by-grade level group, and mean 2002 value added in the dependent variable generated by the student's current school-by-grade instructional team.

5. Medium-run threat effects on test scores

To determine whether educational reforms are primarily responsible for short-run test score gains from sanction threats, I analyse whether sanction threats have persistent impacts on students who have attended a school in its first year of threatened status. Recall that students from the estimation sample are observed in fourth through sixth grades in year 1 (2003), and I now consider their test scores in the medium-run period encompassing years 2 and 3 (2004 and 2005), by the end of which most students will have completed 1 to 2 years of middle school. Specifically, I examine whether students' test scores in years 2 and 3 are discontinuous at the 280-point threshold that determined whether they attended a threatened school *in year 1*. To establish additional notation, let s continue to index a student's *current* school, but let k index the school that the student attended in year 1. For $t=2$ and 3, I run local linear regressions of y_{ist}, the test score of student i currently attending school s in year t, on the grade points GrdPts_{k0} earned in year 0 by the school k that she attended in year 1:

$$y_{ist} = \beta_0 + \beta_1 F_{k0} + \beta_2(GrdPts_{k0} - 280) + \beta_3(F_{k0} \times (GrdPts_{k0} - 280)) + \mathbf{X}_{ist}\gamma + \varepsilon_{ist} \qquad t = 2, 3. \tag{8}$$

[31] In addition to the covariates used in Eq. (3), I also control for the lagged dependent variable, the mean of the lagged dependent variable in the student's school-by-grade peer group, and the mean 2002 "value added" in the dependent variable generated by the student's current school-by-grade instructional team, where value added is the residual from a statewide regression of the variable in 2002 on the variable in 2001.

Table 6
Effect of attending F-graded school in year 1 on FCAT reading scores in years 1 through 3: by subpopulation.

	Dependent var: standardized FCAT reading score				
		With same controls as those used in year 1		With additional controls	
	Year 1	Year 2	Year 3	Year 2	Year 3
All accountable students	0.112	−0.013	0.050	0.002	0.050
	(0.056)**	(0.049)	(0.060)	(0.047)	(0.060)
Level 1 (below basic) in reading in year 0	0.156	0.049	0.099	0.064	0.086
	(0.069)**	(0.064)	(0.074)	(0.061)	(0.080)
Level 2 (basic) in reading in year 0	0.115	−0.020	0.068	−0.013	0.070
	(0.064)*	(0.050)	(0.074)	(0.055)	(0.078)
Level 3 (proficient) or above in reading in year 0	0.056	−0.069	−0.009	−0.045	0.021
	(0.056)	(0.056)	(0.060)	(0.054)	(0.053)
Grade 4 in year 1	0.104	−0.067	0.055	−0.066	0.044
	(0.084)	(0.060)	(0.065)	(0.063)	(0.065)
Grade 5 in year 1	0.091	0.018	0.068	0.047	0.083
	(0.061)	(0.072)	(0.077)	(0.073)	(0.071)

Notes: Each cell, taken from a separate regression estimated on accountable students, is the estimated coefficient on a dummy variable indicating that the student's year 1 school received an F in year 0. Standard errors clustered by year 1 school are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and ***1% level. All regressions use a 28-point bandwidth. Sample sizes by subpopulation (year 1 sample size, year 2 sample size, year 3 sample size) are: all accountable (10,667; 10,203; 9602), level 1 (4120, 3933, 3658), level 2 (2130, 2041, 1923), level 3 or above (4417, 4229, 4021), grade 4 (5346, 5125, 4848), and grade 5 (5132, 4898, 4587). Control variables used in year 1 are the same as those used in the regressions represented in Table 4. For a given year $t>1$, "additional controls" consist of the following factors observed in each of the years 2 through t: fraction black, fraction Hispanic, fraction female, and fraction on free or reduced-price lunch in the student's school-by-grade peer group; mean values of year 0 (averaged FCAT and low-stakes) reading scores and year 0 (averaged FCAT and low-stakes) math scores in the student's school-by-grade peer group; mean year 0 value added in FCAT reading generated by the student's school-by-grade instructional team; school grade points in year 0; and indicators for missing values of the value added and grade point variables.

The total, cumulative effect of accountability pressure in the medium run is captured by the coefficient on F_{k0}, the indicator for whether a student's year 1 school received an F in year 0. I continue to use a bandwidth of 28 points for the RD design. In addition to the controls used in Eq. (3) for the estimation of initial F effects, Eq. (8) also includes a temporally expanding set of controls for three features of the student's educational environment in each of the years 2 through t: characteristics of the student's school-by-grade peer group, quality of the student's school-by-grade instructional team as measured by value added generated in year 0, and school quality as measured by school grade points in year 0.[32] As before, covariates improve precision but do not materially change the point estimates.

Table 6 shows the estimated effect of attending a threatened elementary school in year 1 on a student's FCAT reading score in that year and in the subsequent 2 years. The first column simply restates the initial-year effects shown in Table 4. The next two columns give the estimated F effects on reading scores in years 2 and 3 without controlling for any variables beyond the covariates used in estimation of the initial-year effects. The preferred medium-run estimates are shown in the final two columns, based on regressions controlling for features of the educational environment that each student has experienced up through the given year. According to estimates for the whole sample of accountable students, the immediate reading improvement from attending a threatened elementary school diminishes by more than half in the 2 years after exposure to the initial year of pressure; the final, three-year cumulative improvement of 0.05 s.d. is both practically and statistically insignificant. Point estimates do suggest some degree of threat-induced reading gain in the medium run for students originally below reading proficiency and for students initially exposed in fifth grade, but these estimates lack precision. In sum, evidence for a persistent impact of sanction threats on reading scores is, at best, weak and sensitive to sample definition.

However, I find stronger evidence of medium-run gains in math scores due to sanction threats, as shown in Table 7. Again, I focus on the final two columns, which display estimated F effects that control for middle school characteristics. For the whole sample of accountable students, attending an elementary school in its first year of threatened status cumulatively raises FCAT math scores by 0.11 s.d. over a three-year period encompassing 1 to 2 years of middle school. This medium-run effect size is nearly as large as the initial effect size from year 1. Notably, much of the persistent effect is driven by students who attend fifth grade in year 1 and who thus experience just 1 year of their elementary school's threatened status. Although the medium-run effect for students initially exposed in fourth grade is statistically insignificant, the effect for those initially exposed in fifth grade is more than strong enough to render the effect in the combined sample both significant and moderately sizable, unlike in the case of reading.

In Appendix Table B.1, I show the estimated medium-run effects of the F grade on FCAT scores using a wide variety of specifications for the

Table 7
Effect of attending F-graded school in year 1 on FCAT math scores in years 1 through 3: by subpopulation.

	Dependent var: standardized FCAT math score				
		With same controls as those used in year 1		With additional controls	
	Year 1	Year 2	Year 3	Year 2	Year 3
All accountable students	0.118	0.044	0.099	0.072	0.109
	(0.052)**	(0.068)	(0.068)	(0.057)	(0.054)**
Level 1 (below basic) in math in year 0	0.175	0.141	0.190	0.179	0.180
	(0.083)**	(0.076)*	(0.088)**	(0.067)***	(0.078)**
Level 2 (basic) in math in year 0	0.039	−0.025	0.114	0.024	0.134
	(0.060)	(0.089)	(0.077)	(0.076)	(0.067)**
Level 3 (proficient) or above in math in year 0	0.097	0.017	−0.011	0.018	0.026
	(0.057)*	(0.070)	(0.070)	(0.067)	(0.060)
Grade 4 in year 1	0.117	−0.042	0.050	−0.034	0.075
	(0.080)	(0.082)	(0.075)	(0.073)	(0.069)
Grade 5 in year 1	0.165	0.114	0.172	0.153	0.174
	(0.075)**	(0.097)	(0.083)**	(0.078)*	(0.065)***

Notes: Each cell, taken from a separate regression estimated on accountable students, is the estimated coefficient on a dummy variable indicating that the student's year 1 school received an F in year 0. Standard errors clustered by year 1 school are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and ***1% level. All regressions use a 28-point bandwidth. Sample sizes by subpopulation (year 1 sample size, year 2 sample size, year 3 sample size) are: all accountable (10,667; 10,203; 9601), level 1 (3775, 3608, 3349), level 2 (3191, 3059, 2880), level 3 or above (3701, 3536, 3372), grade 4 (5346, 5125, 4848), and grade 5 (5132, 4898, 4586). Control variables used in year 1 are the same as those used in the regressions represented in Table 4. For a given year $t>1$, "additional controls" consist of the following factors observed in each of the years 2 through t: fraction black, fraction Hispanic, fraction female, and fraction on free or reduced-price lunch in the student's school-by-grade peer group; mean values of year 0 (averaged FCAT and low-stakes) reading scores and year 0 (averaged FCAT and low-stakes) math scores in the student's school-by-grade peer group; mean year 0 value added in FCAT math generated by the student's school-by-grade instructional team; school grade points in year 0; and indicators for missing values of the value added and grade point variables.

[32] Specifically, with a student's FCAT score in year t ($t=2, 3$) on a given subject as the dependent variable, I include additional controls for the following factors observed in each of the years 2 through t: fraction black, fraction Hispanic, fraction female, and fraction on free or reduced-price lunch in the student's school-by-grade peer group; mean values of year 0 (averaged FCAT and low-stakes) reading scores and year 0 (averaged FCAT and low-stakes) math scores in the student's school-by-grade peer group; mean value added on the given test in year 0 generated by the student's current school-by-grade instructional team, where value added is the residual from a statewide regression of 2002 scores on all individual test score covariates in 2001; school grade points in year 0; and indicators for missing values of the value added and grade point variables. The value added and grade point variables are measured in year 0 so that they are not a function of achievement in years 1 and beyond; since my analyses are aimed at assessing the *total cumulative* effect of attending a threatened school in year 1 on test scores in year t, a control variable that is a function of achievement in years 1 and beyond may inappropriately absorb some of the cumulative effects of sanction threats.

RD design, including different bandwidths and different forms for the control function of the running variable $GrdPts_{k0}$. I also show alternative standard errors clustered at the level of the running variable values, as proposed by Lee and Card (2008), rather than at the level of year 1 school.[33] The conclusion that attending a threatened school raises math scores in the medium run is quite robust to different RD specifications and clustering approaches.[34] In reading, the estimated medium-run effects are generally insignificant and small in magnitude, with a few exceptions. However, Rouse et al. (2007) restrict attention to students initially affected in fifth grade and estimate significant medium-run F effects on reading scores. Appendix Table B.2, which shows reading estimates stratified by cohort for various RD specifications, confirms that estimated reading gains in the medium run are sensitive to choice of cohort. The medium-run reading gain by initial fourth-graders is close to zero in every RD specification, whereas some RD specifications, including Rouse and colleagues' choice of a cubic control function estimated on all elementary schools, yield significant medium-run reading gains for initial fifth-graders. Given the sensitivity of the reading estimates, I can draw no firm conclusions on the persistence of reading improvements attributable to sanction threats.

In math, however, the finding that initial threat-induced improvements are preserved in the medium run indicates that educational reforms, rather than school gaming behavior, are likely to be the source of the initial gain. Notably, these findings apply only to achievement on the high-stakes FCAT test, given that initial F effects on low-stakes scores are weak. Nevertheless, the persistence of FCAT math gains provides further motivation to identify specific reforms to educational production that may account for these achievement gains.

6. Changes to educational production by threatened schools

Open-ended phone interviews with school principals served as a starting point for identifying possible reforms by which threatened schools sought to raise test scores and for guiding subsequent econometric analysis. In the summer and fall of 2007, I interviewed five principals who, in the 2002–2003 school year, led elementary schools that had received F grades in 2002.[35] Despite constituting a small sample, the interviews conveyed a surprisingly consistent portrait of the changes or attempted reforms that the principals' schools underwent soon after the onset of sanction threats.

First, all of the interviewees reported intensifying efforts to promote a school culture conducive to learning shortly after 2002. Four of the principals discussed the importance of raising and maintaining high expectations for performance among staff members and students, and at least three interviewees worked to enhance a sense of order within their schools by emphasizing behavioral intervention, conflict resolution, and structured daily routines. In all of the interviews, the principals conveyed a strong belief that a positive school culture was an essential prerequisite for school improvement.

A second reform consistently highlighted by the interviews was an increased use of assessments to guide instruction. According to the principals, regularly scheduled assessments provided teachers with data by which they could monitor student progress and modify instruction to address areas of weak performance. In at least three schools, the principals hired outside consultants to guide the implementation of these and other pedagogical changes, and most of the principals also purchased computer-based programs to give students individualized tutorials and practice questions that generated additional data for tracking student progress. Changes in pedagogical methods were accompanied by curricular changes in the expected direction: greater alignment with FCAT-tested content.

Two other types of changes catalyzed by the F grade were mentioned in multiple interviews. The receipt of an F triggered abnormally high teacher turnover in at least three of the interviewees' schools, partially as a result of involuntary transfers of poorly performing teachers out of the F-graded schools. Finally, all five principals indicated that their schools received additional financial resources from the district or state as a result of the F grade; some specific funds were intended for hiring subject specialists, or "coaches," to help classroom teachers improve instruction and for financing after-school or weekend supplemental instruction for low-achieving students.[36] However, most interviewees downplayed the importance of these funds, which were known to be transitory allocations.

Although many reforms reported by the interviewees revolve around changes to the *technology* of educational production that are not readily quantifiable, some of the improvement strategies may be reflected in observable allocations of financial inputs. In particular, the interviews suggest that activities related to the reform of pedagogy and curricula may be likely candidates for increases in funding. I thus begin the formal analysis of threat-induced changes in educational production by analyzing responses of school spending.

From the FLDOE's Office of Funding and Financial Reporting, I have obtained data on annual school-level costs in various categories. Given a spending or resource measure R_{sd1} for school s within district d in year 1, I estimate threat effects on spending with the standard RD specification based on a 28-point bandwidth:

$$R_{sd1} = \beta_0 + \beta_1 F_{sd0} + \beta_2 (GrdPts_{sd0} - 280) + \beta_3 (F_{sd0} \times (GrdPts_{sd0} - 280)) + + \beta_4 R_{sd0} + \beta_5 \overline{R}_{-s,d0} + \mathbf{X}_{sd1} + \varepsilon_{sd1}. \quad (9)$$

The F grade indicator F_{sd0} is again the independent variable of interest. I control for the lagged dependent variable R_{sd0}, the mean $\overline{R}_{-s,d0}$ of the lagged dependent variable for other elementary schools in the same district, and a vector $\mathbf{X}$ of school characteristics that may influence funding in year 1.[37] In estimation of Eq. (9), all schools are weighted by full-time equivalent enrollment in year 1, and all expenditures are expressed in 2001 dollars and then logged.

[33] Within the preferred bandwidth, while there are 88 schools, there are only 38 distinct running variable values. In this case, it is not clear whether the finite sample distribution of the standard errors clustered by running variable values is sufficiently close to the asymptotic distribution. Nevertheless, the two types of clustering approaches yield very similar standard errors, as shown in Appendix Table B.1.

[34] Moreover, although 10.0% of students in the original year 1 sample are no longer observed in year 3, simple back-of-the-envelope calculations can rule out the possibility that attrition is spuriously generating the main findings on persistence. For example, even if the average latent F effect on the year 3 math scores of attriters were equal to zero, the average F effect on year 3 math scores for the combined population of attriters and non-attriters would still be 0.098 s.d. ($=0.9*0.109+0.1*0$) under the preferred specification.

[35] The average demographic characteristics of accountable fourth through sixth graders attending the interviewees' schools in 2002–2003 are very similar to the average characteristics of F-school students shown in the first column of Table 2. In the interviewees' schools, 78% of students are black, 9% are Hispanic, 52% are female, 90% are on free or reduced-price lunch, and 18% are new to their school in 2002–2003. However, three of the five schools are from the same district, and two of the five schools do not have a 2002 grade point value within 28 points of the F/D threshold. One of the interviewees was her school's new principal in 2002–2003, while the other four interviewees were incumbent principals in 2002–2003. Also, one of the principals opted to write responses by e-mail rather than to speak over the phone.

[36] Case studies by Goldhaber and Hannaway (2004) of two F schools in the early years of the grading system also indicate that F schools received an influx of additional resources from their districts.

[37] I control for the district mean of the lagged dependent variable because the change in a school's allocation of a particular resource may depend on the school's prior allocation *relative* to that received by other schools in the district. The vector X consists of: fraction of students on free or reduced-price lunch, fraction with a disability, fraction who are Limited English Proficient, and average FCAT reading and math scores in year 0. The demographic characteristics are obtained from the online Florida School Indicator Reports, and the average FCAT scores are generated from the student-level microdata.

Table 8
Effect of 2002 school F grade on school expenditures in the 2002–2003 school year.

Dependent var	Estimated effect of F grade	2003 variable mean in D schools	2003 mean of exponentiated variable in D schools	Calculated effect of F grade ($ per pupil)
ln(total school costs per pupil)	0.073	8.74	$6373	$485
	(0.061)			
ln(total instructional costs per pupil)	0.071	8.33	$4252	$314
	(0.078)			
ln(total non-instructional costs per pupil)	0.088	7.63	$2121	$195
	(0.068)			
Share of costs devoted to instruction	−0.007	0.667	–	
	(0.019)			
Ratio of FTE students to FTE staff units	−1.52	15.95	–	
	(1.19)			
Categories of instruc. costs (non-exhaustive):				
ln(teacher salaries and benefits per pupil)	0.064	8.22	$3790	$250
	(0.053)			
ln(contracted service costs per pupil)	0.181	4.11	$77	$15
	(0.159)			
ln(instruc. materials costs per pupil)	0.100	5.37	$231	$24
	(0.147)			
ln(instruc. equipment costs per pupil)	0.938	4.19	$97	$151
	(0.430)**			
Categories of non-instruc. costs (non-exhaustive):				
ln(pupil support costs per pupil)	0.102	5.78	$339	$36
	(0.107)			
ln(media center costs per pupil)	−0.087	4.92	$155	−$13
	(0.128)			
ln(instructional and curricular development costs per pupil)	0.565	5.03	$249	$189
	(0.235)**			
ln(teacher training costs per pupil)	0.317	4.49	$111	$42
	(0.255)			
ln(school administration costs per pupil)	0.079	6.00	$421	$34
	(0.061)			
ln(plant operation costs per pupil)	0.066	6.22	$526	$36
	(0.050)			
ln(plant maintenance costs per pupil)	0.143	5.16	$186	$29
	(0.189)			

Notes: In the column labeled "Estimated effect of F grade," each cell, taken from a separate regression for which the dependent variable is the indicated expenditure variable in the 2002–2003 school year, is the estimated coefficient on an indicator for receiving a 2002 F grade. Huber–White standard errors are in parentheses. Asterisks indicate significance at: *10% level, **5% level, and ***1% level. Sample size is 88 schools. All regressions use a 28-point bandwidth and control for: 2002 school grade points, 2002 grade points interacted with F grade indicator, lagged dependent variable, district mean of lagged dependent variable (excluding the given school), fraction of school's students on free or reduced-price lunch, fraction with disability, fraction who are Limited English Proficient, and lagged average FCAT reading and math scores in the school. Each school is weighted by the number of full-time equivalent students in the 2002–2003 school year. "Calculated effect" is equal to: [exp("Estimated effect") − 1]*[2003 mean of exponentiated variable in D schools].

Throughout this analysis, I do not distinguish whether expenditure changes are due to the decisions of district or school administrators.

The first column of Table 8 presents the estimated impacts of an F grade on various types of school expenditures during the first year of threatened status. Although many of the effects are imprecisely estimated, their signs point to an expansion of available resources for threatened schools. Within the broad category of direct instruction, the only specific item category for which the F grade is found to have a statistically significant effect on spending is instructional equipment; the moderate increment of $151 per pupil attributable to the F grade likely reflects, in part, the principals' reported purchases of computer-based tutorial programs for data-driven instruction.

The estimated F effects on expenditure categories outside of direct classroom instruction show further hints that threatened schools attempt to modify pedagogy and curricula. Receipt of an F grade raises a school's expenditures on instructional and curricular development by 76%, or about $189 per pupil; these additional expenditures partially stem from state financing for reading coaches but may also be due to the hiring of consultants who are reported to have helped with curricular and pedagogical reforms. Sanction threats also trigger a proportionally sizable, but not quite statistically significant, rise in spending on teacher training. Overall, findings from school expenditure data, when combined with principals' responses in interviews, indicate that threatened schools spend more on the technology, activities, and human resources that facilitate changes to pedagogy and curricula.

It is unclear, however, whether threat-induced spending changes are actually responsible for the observed F effects on achievement. As shown in Appendix Table B.3, estimated F effects on achievement do not diminish even after controlling for the expenditure variables responsive to sanction threats. This accounting approach is uninformative, however, because it exploits expenditure variation that is potentially endogenous to achievement, such as the preferences of school administrators. Prior literature also provides uncertain guidance for determining whether expenditure changes are likely to be responsible for the test score gains of F-graded schools. The sparse literature on the effectiveness of expenditure categories responsive to the F grade, such as teacher training, has not yielded consistent findings (Angrist and Lavy, 2001; Jacob and Lefgren, 2004). Despite the uncertain link between threat-induced expenditure shifts and achievement gains, the spending patterns are at least *consistent* with a major role for pedagogical and curricular reforms in driving the positive threat effect on achievement.

In additional analyses, I do not find evidence of F-triggered changes in various non-financial inputs or processes of educational production. Measures of student absenteeism and disciplinary incidents, potentially indicative of school culture, are unresponsive to a school's entry into threatened status (Appendix Table B.4).[38] Although principal

[38] This finding is based on the same RD framework as that in Eq. (3), but such that the dependent variable is any of the four following student-level variables: number of absences, an indicator for ten or more absences, an indicator for receiving any serious punishment, and the number of incidents for which the student receives serious punishments.

quality has often been regarded as a key ingredient for school improvement (Purkey and Smith, 1983), I do not find that a school's receipt of a first F grade in 2002 is associated with any higher departure rate for incumbent principals between the 2001–2002 and 2002–2003 school years (Appendix Table B.5). From this finding, it appears that districts do not regard principal removal as a desired strategy for reforming schools that have only failed once.

7. Conclusion

The threat of sanctions from school accountability systems provides powerful incentives for low-performing schools to raise test scores, but there is the potential for observed test score gains to stem from non-educational manipulation of testing conditions. In this paper, I have provided two strands of evidence indicating that test score gains from sanction threats are likely to arise from educationally oriented reforms. First, exploiting the discontinuous structure of threat determination in Florida, I show that students' initial improvement in high-stakes math scores from attending threatened elementary schools largely persists for at least 1 to 2 years after their entry into middle school. Second, I show that a school's entry into threatened status raises its spending on inputs and activities thematically linked with curricular and pedagogical reform. Both phenomena would have been unlikely to occur if threatened schools had relied primarily on gaming strategies to raise test scores.

These findings from Florida are likely to be relevant for gauging the effects of sanction threats in various other accountability systems. School choice sanctions similar to those in Florida (but without vouchers) are becoming increasingly popular: as of 2006–2007, seventeen other states had the authority to permit student transfers from all low-performing schools (Bausell, 2007). However, authority does not necessarily imply that the sanction is imposed with regularity; the findings of this paper are applicable insofar as failing schools face a real prospect of sanctions attached to further failure. One system that does have a firm timeline for the imposition of penalties, including public school choice, is the NCLB-mandated sanction regime for schools receiving Title I aid. My results are potentially informative of responses by poorly performing Title I schools to being below the NCLB performance standard; however, as the performance standard has been increasing over time, schools on the margin of being threatened have increasingly been higher-performing schools, to which my analyses cannot necessarily be generalized.

Although the persistence of threat-induced math improvements may be an encouraging finding for policymakers, the types of skills retained by affected students deserve further investigation. The responses of interviewed principals, who indicate that instruction within pressured schools is increasingly oriented toward the high-stakes assessment, raise two questions regarding the productive value of "teaching to the test." First, it is important to determine whether persistence of observed improvements arises from retained knowledge of *subject content* or greater familiarity with the *format* of test questions; the latter type of familiarity is arguably of less value. Future analysis of specific test questions should examine whether threat-induced achievement gains on high-stakes test questions are reflected in gains on low-stakes test questions of similar content but dissimilar format, along a line of inquiry similar to the analysis of aggregate test score inflation by Jacob (2007). Second, even for gains stemming from content mastery, it is of interest to determine the types of skills and knowledge that students have retained. For instance, Jacob (2005) has found that math achievement gains attributable to high-stakes testing in Chicago were primarily concentrated in basic skills that are relatively easy to teach, whereas gains in reading, a subject arguably less amenable to formulaic test preparation, were more evenly distributed across test questions. In my paper, the stronger evidence for persistence of math gains relative to reading gains raises the possibility that the retained skills may indeed be those easiest to teach, but a rigorous analysis of this question again requires data on responses to specific test items.

Finally, the modest expenditure increases but moderately sizable test score gains that my analyses have attributed to sanction threats raise new questions for the longstanding debate over the effects of school expenditures. Previous literature on school expenditures either concerns types of expenditures far different than those affected by sanction threats or is too sparse and inconclusive for drawing inferences on whether threat-induced expenditure changes can account for the observed test score gains. Moreover, much of the existing literature examines contexts in which there are weak to nonexistent incentives for using marginal expenditures to raise student achievement. My findings leave open the possibility that small expenditure increases on inputs related to pedagogical and curricular reform may impact achievement within schools facing explicit pressures to improve. Future research should aim to find exogenous variation for identifying the effects of these inputs on student achievement, conditional on the presence of immediate sanction threats. Regardless of whether the expenditures *per se* or the reforms that underlie them account for the test score gains from sanction threats, the expenditure and test score responses estimated by this paper provide evidence for the efficacy of accountability pressure in triggering educationally oriented changes and persistent achievement gains.

Acknowledgements

I am grateful for the feedback from two anonymous referees, Melissa Clark, Daniel Fetter, Alexander Gelber, Guido Imbens, Anna Mastri, Josh Mitchell, Ebonya Washington, and seminar participants at Abt Associates, the College of William and Mary, Harvard University, Mathematica Policy Research, RAND, and the Urban Institute. I am especially grateful to David Cutler, Caroline Hoxby, and Lawrence Katz for their essential guidance. I thank Jeff Sellers, Hanna Skandera, John Winn, and Cheryl Yecke at the Florida Department of Education for providing student-level data, and I thank Richard Harbin, Susan Klos, and Matt Chingos for providing various pieces of school-level data used in this paper. This paper is based on a dissertation chapter written while I was at Harvard University. This research was supported by the Harvard University Multidisciplinary Program in Inequality and Social Policy (National Science Foundation IGERT grant 0333403). All errors and interpretations are mine.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jpubeco.2009.06.002.

References

Angrist, J., Lavy, V., 2001. Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. Journal of Labor Economics 19, 343–369.
Bausell, C., 2007. State of the states. Education Week 26, 86–93.
Carnoy, M., Loeb, S., 2002. Does external accountability affect student outcomes? A cross-state analysis. Educational Evaluation and Policy Analysis 24, 305–331.
Chakrabarti, R., 2007. Vouchers, Public School Response and the Role of Incentives: Evidence From Florida. Staff Report Number 306, Federal Reserve Bank of New York.
Chakrabarti, R., 2008. Impact of Voucher Design on Public School Performance: Evidence From Florida and Milwaukee Voucher Programs. Staff Report Number 315, Federal Reserve Bank of New York.
Chay, K., McEwan, P., Urquiola, M., 2005. The central role of noise in evaluating interventions that use test scores to rank schools. American Economic Review 95, 1237–1258.
Cullen, J., Reback, R., 2006. Tinkering toward accolades: school gaming under a performance accountability system. In: Gronberg, T., Jansen, D. (Eds.), Improving School Accountability: Check-Ups or Choice, Advances in Applied Microeconomics, 14. Elsevier Science, Amsterdam, pp. 1–34.
Deere, D., Strayer, W., 2001. Competitive Incentives: School Accountability and Student Outcomes in Texas. Mimeo, Texas A&M University.

Donovan, C., Figlio, D., Rush, M., 2006. Cramming: The Effects of School Accountability on College-Bound Students. Working Paper 12628, National Bureau of Economic Research.
Figlio, D., 2006. Testing, crime, and punishment. Journal of Public Economics 90, 837–851.
Figlio, D., Getzler, L., 2002. Accountability, Ability, and Disability: Gaming the System. Working Paper 9307, National Bureau of Economic Research.
Figlio, D., Rouse, C., 2006. Do accountability and voucher threats improve low-performing schools? Journal of Public Economics 90, 239–255.
Figlio, D., Winicki, J., 2005. Food for thought: the effects of school accountability plans on school nutrition. Journal of Public Economics 89, 381–394.
Flannery, M., June 16, 2002. State offers "F" schools help, advice — but little cash. Palm Beach Post 1A.
Florida Department of Education, 2003. 2003 Guide to Calculating School Grades: Technical Assistance Paper. Office of Assessment and School Performance, Division of Accountability, Research, and Measurement, Tallahassee, FL.
Goldhaber, D., Hannaway, J., 2004. Accountability with a kicker: observations on the Florida A+ accountability plan. Phi Delta Kappan 85, 598–605.
Grech, D., July 2, 2002. Dade students using vouchers for first time. The Miami Herald 1A.
Greene, J., 2001. An Evaluation of the Florida A-Plus Accountability and School Choice Program. Manhattan Institute for Policy Research, New York.
Grissmer, D., Flanagan, A., 1998. Exploring Rapid Achievement Gains in North Carolina and Texas. National Education Goals Panel, Washington, DC.
Hahn, J., Todd, P., van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69, 201–209.
Hanushek, E., Raymond, M., 2005. Does school accountability lead to improved student performance? Journal of Policy Analysis and Management 24, 297–327.
Harrison, S., Stepp, H., 2001. FCAT grading to get tougher. The Miami Herald, November 9, 1B.
Hegarty, S., July 12, 2002. State: "F" Means First for Help, Funds. St. Petersburg Times.
Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. Journal of Econometrics 142, 615–635.
Jacob, B., 2003. Getting inside accountability: lessons from Chicago. In: Gale, W., Pack, J. (Eds.), Brookings–Wharton Papers on Urban Affairs 2003. Brookings Institution Press, Washington, DC, pp. 41–70.
Jacob, B., 2005. Accountability, incentives, and behavior: the impact of high-stakes testing in the Chicago public schools. Journal of Public Economics 89, 761–796.
Jacob, B., 2007. Test-Based Accountability and Student Achievement: An Investigation of Differential Performance on NAEP and State Assessments. Working Paper 12817, National Bureau of Economic Research.
Jacob, B., Lefgren, L., 2004. The impact of teacher training on student achievement: quasi-experimental evidence from school reform efforts in Chicago. Journal of Human Resources 39, 50–79.
Jacob, B., Levitt, S., 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating. Quarterly Journal of Economics 118, 843–877.
Kane, T., Staiger, D., 2002a. Volatility in school test scores: implications for test-based accountability systems. In: Ravitch, D. (Ed.), Brookings Papers on Education Policy 2002. Brookings Institution Press, Washington, DC, pp. 235–283.
Kane, T., Staiger, D., 2002b. The promise and pitfalls of using imprecise school accountability measures. Journal of Economic Perspectives 16, 91–114.
Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., 2000. What Do Test Scores in Texas Tell Us? RAND, Santa Monica, CA.
Koretz, D., Barron, S., 1998. The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS). RAND, Santa Monica, CA.
Ladd, H., 1999. The Dallas school accountability and incentive program: an evaluation of its impacts on student outcomes. Economics of Education Review 18, 1–16.
Lee, D., Card, D., 2008. Regression discontinuity inference with specification error. Journal of Econometrics 142, 655–674.
Ludwig, J., Miller, D., 2005. Does Head Start Improve Children's Life Chances? Evidence From a Regression Discontinuity Design. Working Paper 11702, National Bureau of Economic Research.
McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. Journal of Econometrics 142, 698–714.
Neal, D., Schanzenbach, D., forthcoming. Left Behind by Design: Proficiency Counts and Test-Based Accountability. The Review of Economics and Statistics.
Peterson, P., 2006. Opportunity scholarships. In: Peterson, P. (Ed.), Reforming Education in Florida: A Study Prepared by the Koret Task Force on K-12 Education, Hoover Institution 2006. Hoover Press, Stanford, CA.
Purkey, S., Smith, M., 1983. Effective schools: a review. Elementary School Journal 83, 427–452.
Reback, R., 2008. Teaching to the rating: school accountability and the distribution of student achievement. Journal of Public Economics 92, 1394–1415.
Rouse, C., Hannaway, J., Goldhaber, D., Figlio, D., 2007. Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher And Accountability Pressure. Working Paper 13681, National Bureau of Economic Research.
Stepp, H., 2001. State overhauls school grading. The Miami Herald, December 19, 1B.
West, M., Peterson, P., 2006. The efficacy of choice threats within school accountability systems: results from legislatively induced experiments. Economic Journal 116, C46–C62.