Rajashri Chakrabarti and Noah Schwartz

Unintended Consequences of School Accountability Policies: Evidence from Florida and Implications for New York

- A key question for educators is whether accountability policies linked to measurable performance outcomes induce schools to "game the system," rather than make genuine improvements.

- This study of an influential Florida program allowing students from failing schools to transfer to better ones suggests that the failing schools engaged in differential classifications of students into exempt categories to artificially boost accountability.

- The finding that schools resort to strategic classifications offers lessons for the design of accountability programs elsewhere, including New York City's Progress Reports program and New York's implementation of the federal No Child Left Behind Act.

1. Introduction

Over the past two decades, state and federal education policies have increasingly emphasized school accountability. This approach focuses on the assignment of rewards and sanctions for schools based on measurable outcomes, usually student performance on standardized tests. A common criticism of accountability policies is that they may induce schools to "game the system" along with—or instead of—making genuine educational improvements. This article investigates whether schools resorted to such strategic behavior in response to the Florida Opportunity Scholarship Program (FOSP), an influential accountability policy that made students from low-performing schools eligible for vouchers to transfer to better ones. Our findings have important implications for New York City's Progress Reports program and New York's implementation of the federal No Child Left Behind (NCLB) Act, which were modeled on the Florida program but contain crucial design changes.

Rajashri Chakrabarti is an economist at the Federal Reserve Bank of New York; Noah Schwartz is a former assistant economist at the Bank.
Correspondence: rajashri.chakrabarti@ny.frb.org

The authors thank David Figlio, Sarah Turner, and participants at Duke University, Harvard University, the Massachusetts Institute of Technology, the American Economic Association, the American Education Finance Association, the Econometric Society, and the Society of Labor Economists conferences for helpful discussions, the Florida Department of Education for data, and Brandi Coates for excellent research assistance. The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

Starting in the 1998-99 school year,[1] Florida began assigning letter grades to schools on a scale of A to F based on student performance on statewide standardized tests.[2] The Florida Opportunity Scholarship Program, introduced in June 1999, embedded a voucher program within this accountability system. It made students from low-performing schools eligible for vouchers to transfer to private schools and higher-performing public schools. Specifically, students from any school receiving two F grades in four years were made eligible for vouchers. These vouchers were funded by public school revenue, with funds following students to their new schools. Thus, FOSP can be viewed as a "threat of vouchers" program—schools receiving an F grade for the first time were at risk of being subjected to vouchers, but vouchers were actually issued only if the school received another F grade in the next three years.

> *Did the exemptions for certain LEP [limited-English-proficient] and ESE [exceptional student education] students induce schools to classify some weaker students into these excluded categories to remove them from school-grade calculations and artificially boost scores?*

Consider the incentives faced by a school threatened by vouchers after receiving its first F grade. As the lowest grade, that mark was associated with stigma, especially because of the publicity and visibility these grades drew. In addition, vouchers were associated with a loss of revenue and shame. As a result, threatened schools had strong incentives to avoid receiving another F grade. This article studies how schools may have responded to this risk, given the features of the program.

Under Florida rules, the test scores of certain high-needs students were excluded from the calculation of school grades, presumably to avoid penalizing schools with large numbers of such students. One exempted category was limited-English-proficient (LEP) students who were in an English-for-speakers-of-other-languages (ESOL) program for less than two years. Several types of special-education (exceptional student education, or ESE) students were also exempted, as we discuss. The features of this program motivate an important question: Did the exemptions for certain LEP and ESE students induce schools to classify some weaker students into these excluded categories to remove them from school-grade calculations and artificially boost scores?

Using data from the Florida Department of Education and a regression-discontinuity estimation strategy, we look for any evidence of increased classifications of students into these excluded categories after the introduction of the program. The regression-discontinuity approach essentially entails comparing schools that just barely avoided an F with ones that just barely received an F. Arguably, these two groups are very similar, and only differ in that the first was not threatened by the program while the second was. So, a comparison is expected to yield a causal estimate of the effect of FOSP. Employing this technique, we find that the program led to increased classification of students into the excluded LEP category in the high-stakes grade 4 and in grade 3, the entry grade for that high-stakes year, following the program's inception. Specifically, schools threatened by the program elected to classify as excluded LEP an additional 0.31 percent of students in grade 4 and an additional 0.36 percent of students in grade 3 in the year after the program was implemented. In contrast, we find no evidence that the threatened schools resorted to increased classification into excluded ESE categories in that school year. As we discuss, ESE classification was associated with substantial costs during this period,[3] which might have discouraged this form of classification. These findings suggest the use of strategic classifications into excluded categories by the failing schools after the inception of the program.[4]

> *[Our] findings suggest the use of strategic classifications into excluded categories by the failing schools after the inception of the [Florida Opportunity Scholarship Program (FOSP)].*

This article is related to two strands of literature. The first studies the effect on public school performance of voucher programs, "threat of voucher" programs, and programs that incorporate threat of vouchers and stigma. This literature generally finds positive effects of school accountability

[1] Going forward, we refer to school years by the calendar year of the spring semester.

[2] Florida had a different accountability system in place before 1999. This system assigned numeric grades of I-IV (I-lowest, IV-highest) to schools based on test scores.

[3] We argue that Florida's McKay Scholarship program for students with disabilities acted as a major disincentive to such classification. Since it made every student with a disability in Florida public schools eligible for vouchers, schools that classified students into ESE categories risked losing these students and the corresponding per-pupil funding.

programs on public school performance in the United States.[5] The second strand investigates whether schools facing accountability systems respond by gaming the system. Researchers have presented evidence of various types of strategic behaviors: reclassification of weaker students into exempted disability categories, suspensions of weaker students during the testing period, teacher cheating, increased focus on high-stakes marginal students, and even strategic boosting of the caloric content of school lunches on testing days.[6]

[Our findings] from Florida have important implications for other programs, including the major school accountability policies in the New York region.

Despite the wealth of literature on gaming behaviors of public schools facing accountability systems, it is not immediately obvious that schools facing accountability-tied sanctions will behave in a similar way. Understanding the incentives and behaviors of public schools in such systems is becoming more relevant in today's world due to the shift toward education policies incorporating sanctions as their centerpiece. This article diverges from and advances this literature by analyzing whether accountability-tied sanctions (specifically vouchers) induce schools to behave in similar strategic ways.[7]

Our findings from Florida have important implications for other programs, including the major school accountability policies in the New York region. New York City's Progress Reports program and New York's implementation of the federal No Child Left Behind Act were both modeled in part on the Florida program, tying sanctions (including school choice) and rewards to student test scores and other measurable outcomes. Importantly, though, both policies contain design differences that should discourage the type of gaming that might have occurred in Florida. These programs incorporate into accountability measures the performance of all students, including limited-English-proficient, special education, and other subgroups. In fact, New York City even gives "extra credit" to schools for achieving progress with English-language learners, special education students, and other high-needs groups. Therefore, schools have no adverse incentives to resort to strategic reclassification of low-performing students into special education and limited-English-proficient categories. We do note, though, that these rules can cause their own type of gaming, perhaps inducing schools to classify their higher-performing students into these groups in an effort to artificially boost their scores and grades.

2. Program Details

The Florida Opportunity Scholarship Program, introduced in June 1999, made students from the worst-performing public schools eligible for vouchers ("opportunity scholarships") to attend private schools and higher-performing public schools. Under the program, all students of a public school became eligible for vouchers if the school received two F grades in a period of four years. A school receiving an F grade for the first time was exposed to the threat of vouchers, but vouchers were not implemented unless and until it received a second F within the next three years. Vouchers resulted in loss of revenue and negative publicity. Moreover, the F grade, being the lowest-performing grade, was associated with stigma and shame.

The Florida Opportunity Scholarship Program . . . made students from the worst-performing public schools eligible for vouchers . . . to attend private schools and higher-performing public schools.

School grades were based on student performance on the Florida Comprehensive Assessment Test (FCAT). The FCAT writing test was first administered in 1993. Following a field test in 1997, the FCAT reading and math tests were first administered in 1998. The reading and writing tests were given in grades 4, 8, and 10, and the math tests in grades 5, 8, and 10.

[4] It is worth considering how such classification might affect the students involved. One the one hand, strategic placements into LEP categories can potentially have a demoralizing effect on students and might expose them to weaker student groups. On the other hand, such placements might expose them to more resources with a positive effect on learning. Hanushek, Kain, and Rivkin (2002) study the effect of placement of students with disabilities into special education programs. They find that the programs led to significant gains in math achievement, especially for learning-disabled and emotionally handicapped students. But they do not look at the effect of placement into LEP categories, nor the impact of strategic placement into these categories. Unfortunately, there is virtually no literature on the impact of such strategic placement into exempt categories, making this question an avenue for important future research.

[5] See Greene (2001), Hoxby (2003a, 2003b), Greene and Winters (2003), Figlio and Rouse (2006), West and Peterson (2006), Rouse et al. (2007), Chakrabarti (2008a, 2008b), Chiang (2009), and Figlio and Hart (2010).

[6] See Jacob and Levitt (2003), Jacob (2005), Figlio and Winicki (2005), Cullen and Reback (2006), Figlio and Getzler (2006), Figlio (2006), Reback (2008), Neal and Schanzenbach (2010), and Chakrabarti (2013).

[7] The only exception is Chakrabarti (2013), who studies the behavior of public schools facing accountability-tied vouchers on other types of strategic behaviors, such as whether threatened schools focus more on high-stakes marginal students and subject areas.

The system of assigning letter grades to schools on a scale of A through F started in 1999. The state assigned a school an F grade if it failed to achieve the minimum criteria in all three FCAT subjects (reading, math, and writing), a D grade if it failed the minimum criteria in only one or two subject areas, and a C grade if it passed the minimum criteria in all three. To pass the minimum criteria in reading and math, a school needed to have at least 60 percent of its students score at level 2 or above in the respective subject; to pass the minimum criteria in writing, at least 50 percent had to score at level 3 or above.[8]

While the test scores of all regular students were included in the calculation of school grades, the scores of students in some limited-English-proficient and exceptional student education categories were excluded. Specifically, scores of LEP students who were in an ESOL program for less than two years were not included in the computation of grades, nor were scores of ESE students in eighteen ESE categories. Only LEP students with two or more years in an ESOL program and ESE students in speech-impaired, gifted, and hospital/homebound categories were included in school grade computations.[9]

Henceforth, we refer to the less than two years in an ESOL program category as the "excluded" LEP category and the two years or more in an ESOL program category as the "included" LEP category. Similarly, we refer to the speech-impaired, gifted, and hospital/homebound categories as "included" ESE categories and to the other ESE categories as "excluded" ESE categories.

3. Data

We obtained all data for this study from the Florida Department of Education. The information includes grade-level data on LEP enrollment in grades 2, 3, 4, and 5 for 1999 and 2000 as of February in each year (just before the tests were administered). We also know the number of students in an ESOL program for less than two years and the number of students in an ESOL program for two years or more in each of these grades in each year under consideration.

School-level data on enrollment in the various ESE categories were also obtained. In addition to total ESE enrollment, these data report enrollment in each of the ESE categories in each Florida school for 1999 and 2000.

The third type of data we retrieved was the distribution of students across grades K-12 in each Florida school in 1999 and 2000. We also had access to data on various socioeconomic characteristics of schools, including gender composition, racial composition, and the percentage of students eligible for free or reduced-price lunch. Finally, we obtained several measures of school-level and district-level per-pupil expenditures for both years under consideration.

4. Empirical Strategy

Under the Florida Opportunity Scholarship Program, schools that received an F grade in 1999 were directly threatened with stigma and vouchers since all of their students would be eligible for vouchers if the school received another F grade in the next three years. We refer to these schools as "F" schools. The schools that received a D grade in 1999 were closest to the "F" schools in terms of grade, but were not directly threatened by the program. We refer to them as "D" schools. Our empirical strategy essentially compares schools that barely received an F to those that barely received a D, as we explain below.

Because grades were not randomly assigned to schools, the schools that received an F grade in 1999 were likely to be quite different from those that did not, both in terms of observable and unobservable characteristics. These differences may themselves affect the outcome of interest—whether schools engage in strategic ESE or LEP classification. Thus, simply comparing the outcomes of "F" schools to those of "D" schools will not yield a causal estimate of the effect of the program; there are many confounding variables besides the program that could explain any differences we observe.

By comparing the schools that fell just below the cutoff ("F" schools) with those just above ("D" schools), we get an estimate of the effect of the [FOSP].

[8] We mainly focus on the responses of the schools that just received an F versus those that just received a D in 1999. In Section 6.4, we study the response of the "D" schools relative to the "C" schools as well. While the "D" schools did not face any direct threat of vouchers, they may have faced an indirect threat as they were close to an F grade and might have also faced stigma by being one of the lowest-performing groups. Correspondingly, we focus on the criteria for F, D, and C grades. Detailed descriptions of the criteria for the other grades are available at schoolgrades.fldoe.org.

[9] Florida classified ESE students into twenty-one ESE categories in total: educable mentally handicapped, trainable mentally handicapped, orthopedically handicapped, occupational therapy, physical therapy, speech-impaired, language-impaired, deaf or hard of hearing, visually impaired, emotionally handicapped, specific learning disabled, gifted, hospital/homebound, profoundly mentally handicapped, dual-sensory-impaired, autistic, severely emotionally disturbed, traumatic brain injured, developmentally delayed, established conditions, and other health-impaired.

To minimize the influence of confounding variables, we use a regression-discontinuity strategy (Hahn, Todd, and van der Klaauw 2001; van der Klaauw 2002; Imbens and Lemieux 2008) to analyze the effect of the program. The analysis essentially entails comparing the response of schools that barely failed to that of schools that barely passed. The institutional structure of the Florida program allows us to follow this strategy. We exploit the fact that there was a sharp discontinuity in how the F grade was assigned. Schools that scored below a fixed cutoff received an F, and thus the threat, while schools that scored above the cutoff did not. By comparing the schools that fell just below the cutoff ("F" schools) with those just above ("D" schools), we get an estimate of the effect of the program. Presumably, these two groups of schools were nearly identical in terms of socioeconomic and demographic characteristics (a testable assumption that we examine later), and the only difference between them was that one group was subjected to stigma and the threat of vouchers while the other was not.

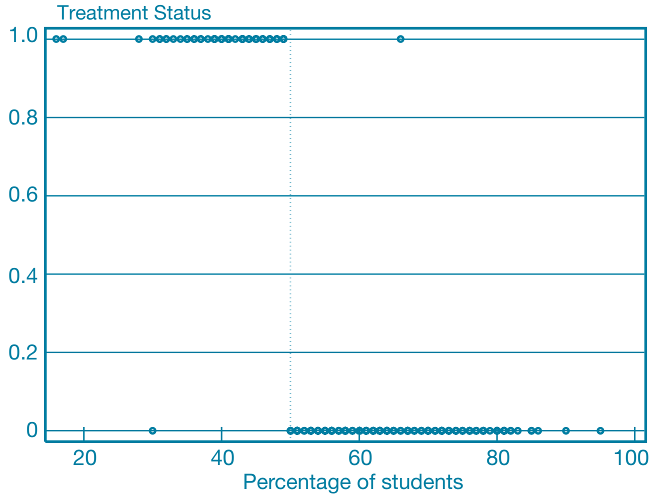
We focus on the sample of "F" and "D" schools that failed both reading and math in 1999. In this sample, according to the Florida grading rules, only the schools that also failed writing would receive an F, while the schools that passed writing would receive a D. Therefore, in this sample, schools that had less than 50 percent of their students pass the 1999 writing FCAT would receive an F and face a direct threat, while schools at or above 50 percent on the writing portion would not.

In the rest of this article, we refer to schools receiving an F grade in 1999 as being in the "treatment" group. Treated schools were exposed to the threat of vouchers and sanctions. Using the sample of "F" and "D" schools that failed both reading and math in 1999, we illustrate in Chart 1 the relationship between treatment status (those receiving an F in 1999) and the schools' percentages of students scoring at or above level 3 in FCAT writing, or the "running variable" (r_i) in the regression-discontinuity literature. There are 269 schools in this sample, with 65 falling below the cutoff of 50 percent on the writing portion and 204 schools at or above the cutoff. The chart shows that all but one of the schools in this sample that had less than 50 percent of their students scoring at or above level 3 actually received an F grade. Similarly, all but one that had 50 percent or more of their students scoring at or above level 3 were assigned a D grade. The result demonstrates that, in this sample, the percentage of students scoring at or above level 3 in writing indeed uniquely predicts assignment to treatment for all but two schools, and there is a sharp increase in the probability of treatment at the 50 percent mark. In fact, the estimated discontinuity is 1 and highly significant; there was a perfect correlation between falling below 50 percent and receiving an F. Using this sample ("F" and "D" schools that failed in reading and math in 1999), we rank schools in terms of the percentage of students scoring at or above level 3 in FCAT writing and then pick schools that are close to the cutoff. Our analysis uses this set of schools.

> *The percentage of students scoring at or above level 3 in writing indeed uniquely predicts assignment to treatment for all but two schools, and there is a sharp increase in the probability of treatment at the 50 percent mark.*

Chart 1

Relationship between Treatment Status and Percentage of Students Scoring at or above Level 3 in 1999 FCAT Writing



Source: Authors' calculations.

Notes: Treatment status is 1 if a school received a grade of "F" and 0 if it received a grade of "D." FCAT is the Florida Comprehensive Assessment Test.

We also consider two alternate samples in which both "F" and "D" schools fail reading and writing or math and writing. (According to the Florida rules, "F" schools would also fail math [reading], unlike "D" schools.) We find that indeed in these samples, the probability of treatment increases sharply when less than 60 percent of a school's students scored at or above level 2 in math (reading). The sizes of these samples, however, are considerably smaller than those of the first sample we described, and these samples are considerably less dense in the vicinity of the cutoff. So, we focus on the first sample above, in which the "D" schools passed the writing cutoff and the "F" schools missed it, and both groups of schools missed the cutoffs in the other two subject areas. Note, though, that the results from the alternate samples are qualitatively similar. Also, as a robustness check, we present in section 6.2 estimates from a combined sample in which we pool the three samples.

Consider the following model, where Y_i is school i's outcome,[10] T_i equals 1 if school i received an F grade in 1999 and $f(r_i)$ is a function of the running variable r_i. Recall that the running variable here is the percentage of students scoring at or above level 3 in FCAT writing:

$$Y_i = \gamma_0 + \gamma_1 T_i + f(r_i) + \varepsilon_i \, . \tag{1}$$

Hahn, Todd, and van der Klaauw (2001) show that γ_1 is identified by the difference in average outcomes of schools that just missed the cutoff and those that just made it, provided that the conditional expectations of the other determinants of Y are smooth through the cutoff. Here, γ_1 identifies the local average treatment effect (LATE) or the effect of getting an F at the cutoff.

The estimation can be done in many ways. We use local linear regressions with a triangular kernel and a rule-of-thumb bandwidth, as suggested by Silverman (1986). We also allow for flexibility on both sides of the cutoff by using a linear spline functional form that enables us to include an interaction term between the running variable and a dummy indicating whether or not the school falls below the cutoff (see equation 2 below). We estimate alternate specifications that do not include controls as well as those that use them.[11] Assuming the covariates are balanced on both sides of the cutoff (we formally test this assumption below), the purpose of including covariates is variance reduction. They are not required for the consistency of γ_1. Thus, our preferred specification is:

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 r_i + \alpha_3 (T_i \times r_i) + (\Sigma_k \alpha_{4k} X_{ik}) + \varepsilon_i \, , \tag{2}$$

where $f(r_t) = r_i + (T_i \times r_i)$ denotes the linear spline functional form; $\Sigma_k X_{ik}$ denotes the set of covariates (or controls) and includes racial composition (percentage black, Hispanic, Asian, American Indian, multiracial; percentage white serves as the excluded category), gender composition (percentage male), percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditures.

To test the robustness of our results, we also experiment with alternative bandwidths. The results remain qualitatively similar, and are available on request. We also conduct a parametric estimation in which we include a third-order polynomial in the percentage of students scoring at or above level 3 in writing and interactions of the polynomial with a dummy indicating whether or not the school falls below the cutoff. We also estimate alternative functional forms that include a fifth-order polynomial instead of a third-order polynomial and the corresponding interactions.[12] The results are very similar in each case, and are available on request.

An advantage of a regression-discontinuity analysis is that identification relies on a discontinuous jump in the probability of treatment at the cutoff. Consequently, mean reversion—a potentially confounding factor in other settings—is not apt to be important here, as it likely varies continuously with the running variable (r_i) at the cutoff. Also, regression-discontinuity analysis essentially entails comparison of schools that are very similar, even virtually identical, except that the schools to the left of the cutoff faced a discrete increase in the probability of treatment. As a result, another potentially confounding factor—existence of differential preprogram trends—is not likely to be important here.

4.1 Testing the Validity of the Regression-Discontinuity Analysis

We now investigate whether the underlying assumptions governing the validity of the regression-discontinuity design are satisfied in this context. First, we check whether schools just below the cutoff differed from those just above it in terms of preprogram characteristics. Recall that any such differences would confound our attempt to attribute a difference in outcomes to the program. There is not much reason to expect any differences between these groups. For such differences to arise, certain types of schools would need to strategically manipulate their test scores in an effort to fall on one side of the cutoff. However, the program was announced in June 1999, while the tests were given a few months before (in January and February), making it unlikely that Florida's schools had the necessary information and time to resort to such manipulation.

Nevertheless, we check for discontinuities in predetermined characteristics of schools at the cutoff. For the regression-discontinuity strategy to be valid, preexisting characteristics should vary continuously through the cutoff. The only factor that should vary discontinuously is the probability of treatment. In such a case, any discontinuity in student

[10] In most of this article, Y_i refers to schools' percentages of students in various ESE and LEP categories. Exceptions are in sections 4.1 and 6.1, where Y_i also refers to various demographic and socioeconomic characteristics of the schools. See those sections for more details.

[11] Covariates used as controls include racial composition of schools, gender composition of schools, percentage of students eligible for free or reduced-price lunches, and real per-pupil expenditures.

[12] We use odd-order polynomials because they are more efficient (Fan and Gijbels 1996) and are not subject to boundary bias problems, as even-order polynomials are.

Table 1

Testing Validity of Regression-Discontinuity Analysis: Looking for Discontinuities in Preprogram Characteristics at Cutoff

	Percentage				
Panel A	(1) White	(2) Black	(3) Hispanic	(4) Asian	(5) American Indian
	2.92	-5.06	2.43	0.09	-0.16
	(7.24)	(11.39)	(6.73)	(0.28)	(0.06)
Panel B	Percentage Multiracial	Percentage Male	Percentage Free/ Reduced-Price Lunch	Enrollment	Real Per-Pupil Expenditure
	-0.23	-1.21	-5.97	-14.45	-1.97
	(0.26)	(1.44)	(5.36)	(60.32)	(2.29)
	Percentage				
Panel C	Exceptional Student Education (ESE)	Excluded ESE	Included ESE	Learning-Disabled	Emotionally Handicapped
	-2.92	-2.89	-0.03	0.05	-0.63
	(1.87)	(1.83)	(0.78)	(0.79)	(0.56)
	Percentage Excluded Limited-English-Proficient (LEP)				
Panel D	Grade 2	Grade 3	Grade 4	Grade 5	
	0.03	0.30	0.24	0.30	
	(0.18)	(0.20)	(0.22)	(0.18)	
	Percentage Included LEP				
Panel E	Grade 2	Grade 3	Grade 4	Grade 5	
	-0.54	0.06	-0.09	0.26	
	(0.51)	(0.56)	(0.28)	(0.41)	

Source: Authors' calculations.

Note: Robust standard errors adjusted for clustering using the running variable are in parentheses.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

classification (into excluded or included ESE and LEP categories) at the cutoff can be attributed to the discontinuity in the probability of treatment, or, in other words, to the program. The discontinuity estimates for preprogram characteristics (using the regression-discontinuity strategy described above) are presented in Table 1. As expected, they are small and never statistically distinguishable from zero.

Following McCrary (2008), we also use a density test to investigate whether there is selection at the cutoff. The idea is that if schools strategically placed themselves on one side of the cutoff, we would expect to see a clustering close to it, and consequently an unusual spike in the density of the running variable (the percentage of students at or above level 3 in writing). However, as Table 2 shows, we find no evidence of discontinuity in the density of the running variable at the cutoff.

5. Results

Having established that a regression-discontinuity approach in this setting is valid, we now look at the program's behavioral

Table 2

Testing Validity of Regression-Discontinuity Analysis: Looking for Discontinuities in Density of Running Variable

	1999
Difference	-0.01 (0.01)

Source: Authors' calculations.

Notes: The table shows the percentage of students at or above FCAT level 3 in writing. Standard error is in parentheses and is clustered using the running variable (percentage of students at or above writing cutoff). FCAT is the Florida Comprehensive Assessment Test.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

effect on threatened schools. We focus on the elementary grades; grades 4 and 5 were the tested grades during this period in Florida.

For reference, we first look at the behavior of the schools in our sample in the preprogram period. Table 1 (panels C-E) shows classification into excluded and included LEP and ESE categories in 1998-99, the school year just before the program started. Each entry shows the average difference between the soon-to-be-threatened and the nonthreatened schools. There is no evidence that the schools that would be threatened the next year behaved any differently than the nonthreatened schools in terms of excluded or included LEP classification in any of the high- or low-stakes grades. We also see no evidence of differential classification into excluded or included ESE categories in 1999. The picture in the post-program period, however, is very different.

Table 3 examines the effect of the FOSP on the percentage of students classified into the excluded and included LEP categories in grades 2-5 in 1999-2000, the first school year after the program went into effect.[13] Again, each entry in the table shows the difference between the LEP percentages of threatened versus nonthreatened schools.

Consider the excluded LEP category in the top panel. In the year after the program's inception, there was a positive and statistically significant difference between threatened and nonthreatened schools in terms of the percentage of students classified as excluded LEP in the high-stakes grade 4 and the entry grade 3. In contrast, there is no evidence of a statistically significant difference in the low-stakes grade 2 or the high-stakes grade 5. Of note, though, is that while the grade 2 and grade 5 effects are not statistically significant, they are positive and not statistically different from the grade 3 or grade 4 effects.

Table 3

Effect of Program on Classification in Excluded and Included Limited-English-Proficient Categories

	(1) Grade 2	(2) Grade 3	(3) Grade 4	(4) Grade 5
Percentage excluded	0.29	0.36**	0.31**	0.27
	(0.23)	(0.18)	(0.12)	(0.25)
Observations	123	121	119	116
R^2	0.53	0.54	0.40	0.43
Percentage included	0.11	-0.42	0.04	0.01
	(0.30)	(0.48)	(0.31)	(0.39)
Observations	123	121	119	116
R^2	0.66	0.57	0.53	0.33

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, and real per-pupil expenditure.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

The estimates suggest that in the first year of the program, schools facing stigma and the threat of vouchers classified an additional 0.31 percent of students into the excluded LEP category in grade 4 and an additional 0.36 percent in grade 3.

> *In the year after the [FOSP's] inception, there was a positive and statistically significant difference between threatened and nonthreatened schools in terms of the percentage of students classified as excluded LEP in the high-stakes grade 4 and the entry grade 3.*

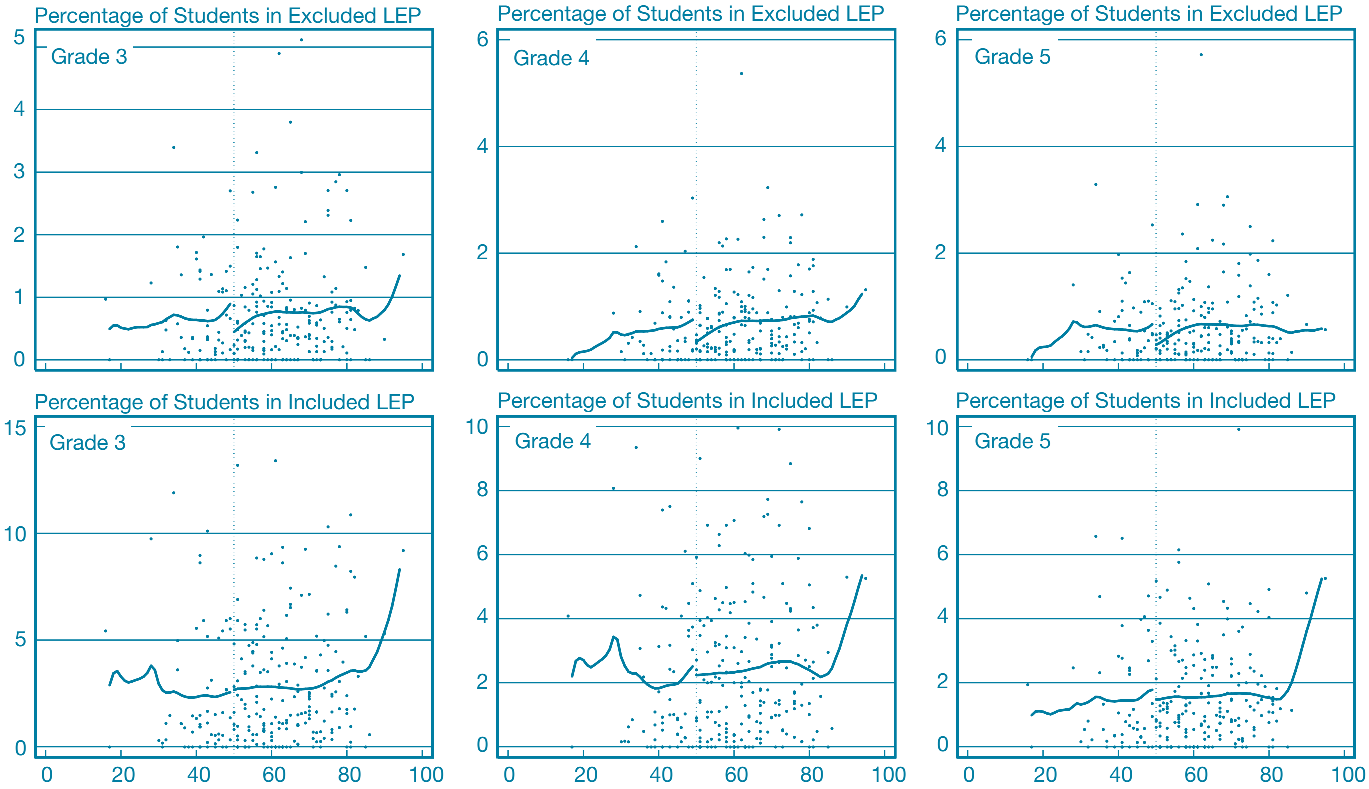
To put these numbers in perspective, we note that the average enrollment of these schools in the immediate preprogram period was approximately 713 students. Thus, the threatened schools classified an additional 53 percent of their excluded LEP students in grade 4 and an additional 55 percent of their excluded LEP students in grade 3. The results are, in turn,

[13] These variables are defined as enrollment in excluded and included LEP categories in each grade as a percentage of total school enrollment.

Chart 2

Effect of Program on Classification in Excluded and Included Limited-English-Proficient (LEP) Categories

Regression-Discontinuity Estimates; February 2000 Survey



Source: Authors' calculations.

Notes: The x-axis in each panel depicts the percentage of students at or above level 3 in FCAT (Florida Comprehensive Assessment Test) writing.

equivalent to classification of an additional 2.6 students in grade 4 and 2.3 students in grade 3 into the excluded LEP category.

The lower half of Table 3 presents the program's effects on the percentage of students in the included LEP category. There is no evidence that the program led to differential classification into included LEP in any grade in the first year after the program; the discontinuities are small and statistically insignificant.[14] Chart 2 illustrates the impact on classifications into excluded and included categories.[15] Consistent with the above findings, the chart provides evidence in favor of increased classifications into excluded LEP categories in grades 3 and 4 (and these discontinuities are statistically significant). There is evidence of a smaller (statistically insignificant) discontinuity in grade 5, but none in favor of any differential classification into included LEP categories.

Tables 4 and 5 examine the effect of the program on ESE classification. Table 4, column 1, shows the effect on total ESE classification. The dependent variable for this analysis is percentage ESE enrollment (total ESE enrollment as a share of total enrollment). The estimates show no evidence of any differential classification in the threatened schools at the cutoff.

While trends in total ESE classification provide a summary picture, they are unlikely to provide a conclusive look at whether the "F" schools resorted to such classification. Yet in our view, the absence of shifts in total ESE classification does not rule out the possibility of shifts in certain ESE categories.

[14] Of note here is that neither the excluded LEP effects nor the included LEP effects are statistically different across grades.

[15] While the regression-discontinuity estimates in the tables were obtained from specifications that included all covariates mentioned above, the estimates in the charts were obtained from specifications that did not include any covariate. The similarity of the two sets of estimates attests to the robustness of the estimates.

Table 4

Effect of Program on Classification in Exceptional Student Education (ESE) Categories

	Percentage		
	(1) Students in ESE	(2) Students in Excluded ESE	(3) Students in Included ESE
	0.44 (0.40)	0.70 (0.56)	-0.24 (0.29)
Observations	130	130	130
R^2	0.92	0.92	0.84

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and preprogram (1999) percentage of students in All (Column 1), Excluded (Column 2), or Included (Column 3) ESE categories.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

Table 5

Effect of Program on Classification in Learning-Disabled and Emotionally Handicapped Categories

	Percentage	
	(1) Students in Learning-Disabled	(2) Students in Emotionally Handicapped
	-0.18 (0.26)	0.08 (0.16)
Observations	130	130
R^2	0.80	0.93

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and preprogram (1999) percentage of students in All (Column 1), Excluded (Column 2), or Included (Column 3) ESE categories.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

To offer a closer look, Table 4 also displays the effect of the program on classification into excluded and included ESE categories. The dependent variables here are the percentages of total enrollment classified into excluded (column 2) and included (column 3) categories. The estimates show no evidence that the threatened schools resorted to greater classification into excluded ESE categories in the first year of the program. The effects are not at all statistically significant, nor are they economically significant. There is also no statistically or economically significant evidence of differential classification out of (or into) the included categories.[16] Consistent with this evidence, Chart 3 offers no evidence of (statistically significant) differential classification into excluded or included ESE categories.

The various ESE categorizations differ in the extent of their severity, and consequently it may be easier to reclassify students into some categories than others. While some categories such as those involving observable or severe disabilities or physical handicaps are comparatively nonmutable, others such as learning disabled and emotionally handicapped are often mild and comparatively mutable. Classification into these latter categories often has a large subjective element and, as such, could be prone to manipulation. While the above analysis does not find evidence of differential classification into excluded categories as a whole, it does not rule out the possibility of increased classification into certain categories that are more easily manipulated on the spectrum of special needs.

To investigate this possibility, we examine the effect of the program on classification into two mutable excluded categories: learning disabled (column 1) and emotionally handicapped (column 2). We find no evidence that the threatened schools tended to differentially classify students into either of these categories; the discontinuities are small and not statistically significant.

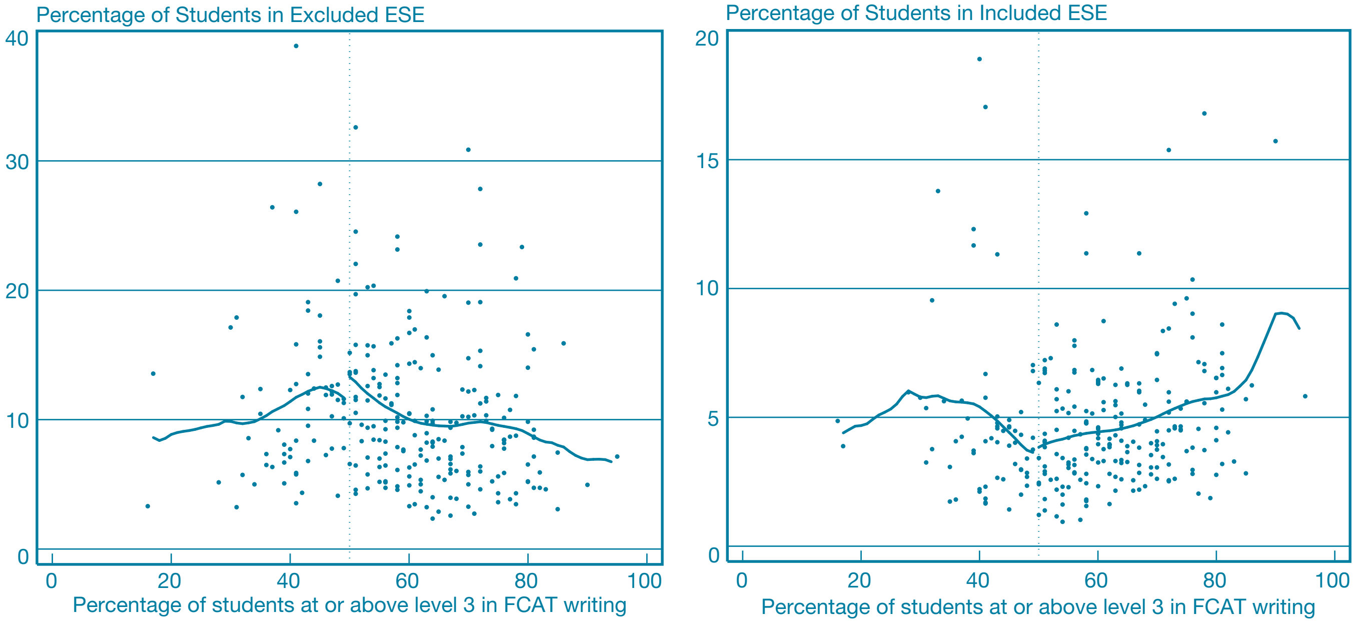
Our next step is to ask what might be driving these classification patterns that we do see. It is worth considering two explanations: 1) the "wake-up-call" hypothesis and 2) the "strategic-classifications" hypothesis.

[16] Recall that these are school-level effects, unlike grade-level effects for LEP. Also of note here is that the excluded LEP effect (computed from data aggregated over the available grades to generate a school-level measure for easier comparison) is both economically and statistically different from the excluded ESE effect. However, the included LEP effect is not statistically different from the included ESE effect.

Chart 3

Effect of Program on Classification in Excluded and Included Exceptional Student Education (ESE) Categories

Regression-Discontinuity Estimates, 2000



Source: Authors' calculations.

Note: FCAT is the Florida Comprehensive Assessment Test.

To summarize, we observe that the program led to statistically significant increased classifications into excluded LEP categories in high-stakes grade 4 and entry grade 3 in the threatened schools. Yet we find no evidence of any difference in classifications into included LEP categories. Neither do we find evidence of any difference in classification into ESE categories (excluded or included) in the threatened schools. Our next step is to ask what might be driving these classification patterns that we do see. It is worth considering two potential explanations: 1) the "wake-up-call" hypothesis and 2) the "strategic-classifications" hypothesis.

Under a wake-up-call hypothesis, one might argue that the F grade served as a wake-up call for these schools and led them to proactively classify their low-performing students into LEP or ESE groups to ensure greater and more specialized support for these students. Under a strategic-classifications hypothesis, an opposing argument can be made that the threatened schools tended to classify their low-performing students into excluded categories in a strategic effort to boost their scores and grades.

While the data do not allow us to pinpoint the exact cause of such classifications, there seems to be somewhat more evidence that strategic classifications are the more likely driver of the results. One would expect the wake-up call to manifest itself in increased classifications in all grades symmetrically, with a school acting on a genuine desire to help weaker students in each grade. It is not clear why such classification into an LEP track would be more prominent in high-stakes grade 4, and the entry to that high-stakes year, grade 3. Also the wake-up-call hypothesis would predict classifications into both ESE and LEP categories, perhaps more into ESE, as ESE categories provide more resources as well as more specialized help to students.

While the data do not allow us to pinpoint the exact cause of such classifications, there seems to be somewhat more evidence that strategic classifications are the more likely driver of the results.

In contrast, a strategic-classifications hypothesis would point to schools classifying students into excluded LEP in the high-stakes grades or entry grades. Specifically, students classified into the excluded LEP category in grade 4 would not count toward school grades either in the current year or in the following year, when these students would advance to grade 5, another high-stakes grade. Note, though, that doing the additional classification all at once may have been difficult, which is why the administrators may have chosen to spread out the process to the entry grade 3.

Strategic classifications would also tend to result in classifications only into excluded LEP, but not excluded ESE categories, since there were considerable costs associated with reclassification into ESE categories. ESE designations had to be approved by the parents and a group of experts (such as physicians and psychologists). But the steepest cost to ESE classification was posed by the McKay Scholarship program. Created in 1999 and fully implemented statewide in the 2000-01 school year,[17] this program made every student with disabilities in Florida public schools eligible for vouchers to move to a private school or to another public school. Thus, reclassification of students into special education categories was associated with a risk of losing the students and their corresponding per-pupil funding. Moreover, because special education students were more expensive to educate than regular students, McKay vouchers cost more than Opportunity Scholarships—approximately $7,000 versus $3,500 per student on average. This fact meant that schools were likely to lose more funding with the departure of an ESE student under the McKay program than with the loss of a regular student under the FOSP. Consequently, the McKay Scholarship program acted as a strong disincentive to this sort of reclassification. The strategic-classifications view, therefore, seems to be more compelling in this scenario, as it matches better the patterns observed in the data.[18] However, the implication that strategic classifications play a role should only be taken as suggestive, and not conclusive. A further caveat is worth mentioning here. As with any regression-discontinuity analysis, the estimates obtained above are all local average treatment effects, meaning that the effects obtained are local to the cutoff only. These results should not be generalized to the whole sample.

> *The strategic-classifications view . . . seems to be more compelling in this scenario, as it matches better the patterns observed in the data.*

6. Sensitivity Checks

6.1 Compositional Changes of Schools and Sorting

If differential student sorting or compositional changes occurred in the treated schools, then the above effects may in part be driven by those changes.[19] To investigate this issue, we examine whether the FOSP led to a differential change in the demographic composition in the treated schools. We use the same regression-discontinuity strategy outlined above, but the dependent variables are now demographic (the percentages of students that are white, black, Hispanic, Asian, American Indian, multiracial, male, eligible for free or reduced-price lunch, as well as enrollment). We find no evidence of differential shifts in the treated schools in these characteristics after the introduction of the program. (These results are not reported here, but are available on request.) Thus, it seems safe to conclude that the results described above are not being driven by differential changes in the composition of schools or student sorting.

6.2 Does Combining the Three Discontinuity Samples Affect Results?

To broaden our analysis, we also apply an alternative regression-discontinuity strategy in which we combine the three samples described in section 4: the sample that failed in reading and math, but just passed or failed in writing (F/D writing sample); the sample that failed in reading and writing, but just passed or failed in math (F/D math sample); and the sample of schools that failed in math and writing, but just passed or failed in reading (F/D reading sample). In the F/D reading (math) sample, according to Florida rules, schools with

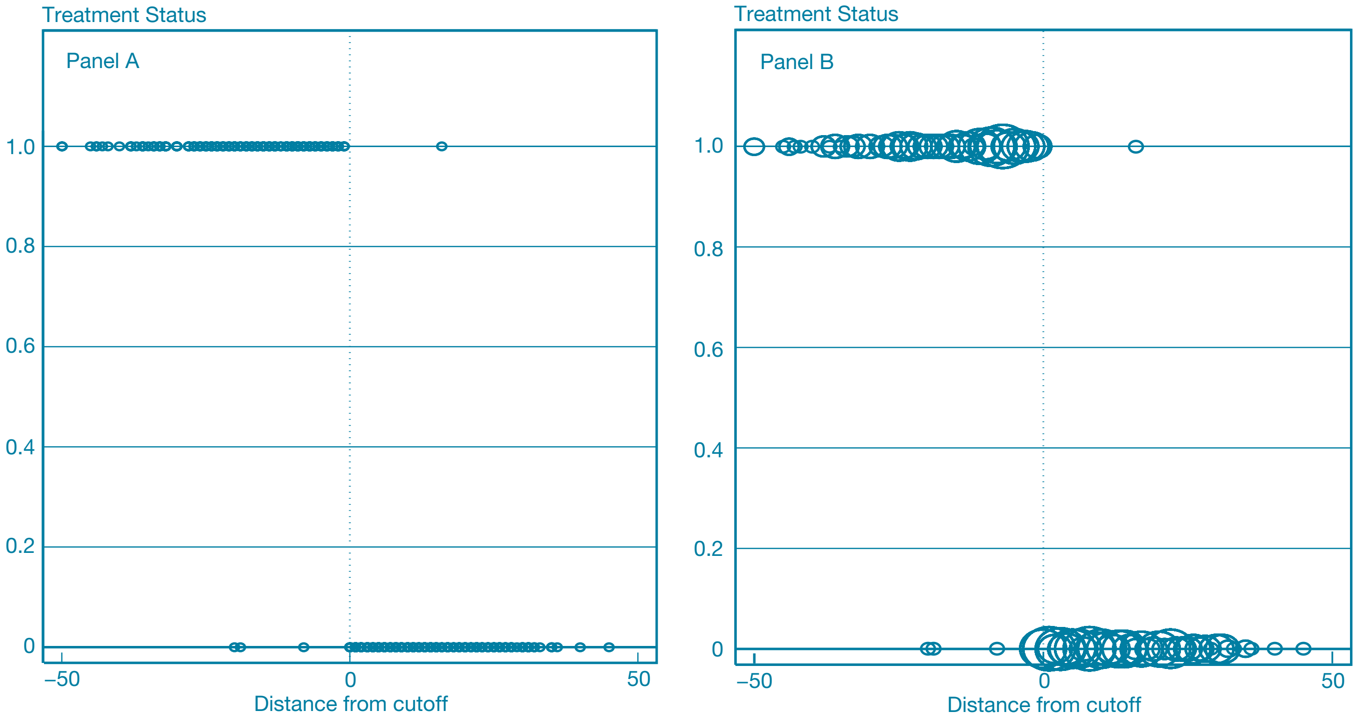
[17] The McKay program was run as a small pilot in the 1999-2000 school year with only one school and two students participating in the program.

[18] A question worth considering here is whether such classification was enough for an "F" school to escape an F grade in the near future. Note that the percentages of students classified into the excluded LEP category were not small (53 percent and 55 percent). The additional classification in terms of numbers of students of between two and three in grade 3 and grade 4 does not appear to be big. However, these were marginal schools located close to the cutoff that only barely failed to make the cutoff. So, for such schools, even such a small classification could potentially make a difference. Also, consider that schools may not respond in only one margin. Such classifications along with responses along other margins could together make a difference in terms of grade.

[19] None of the threatened schools was subjected to vouchers in the 1999-2000 school year, so the concern about vouchers leading to sorting is not applicable here. However, the F and D grades alone (exposing schools to the threat of vouchers) could lead to differential sorting of students in these two types of schools. Figlio and Lucas (2004) find that following the first assignment of school grades in Florida, the better students differentially entered schools receiving A grades, though this differential sorting tapered off over time.

Chart 4

Relationship between Treatment Status and Distance from Cutoff (Combining the Three Discontinuity Samples)



Source: Authors' calculations.

just under 60 percent of their students scoring at or above level 2 in FCAT reading (math) should receive an F, while schools with just above (or exactly) 60 percent should receive a D. In the F/D writing sample, schools with just under 50 percent of their students scoring at or above level 3 in FCAT writing should receive an F, while schools with just above (or exactly) 50 percent of their students scoring at or above level 3 should receive a D. Centering these running variables at their respective cutoffs (60 percent or 50 percent), we pool the three samples to improve efficiency. We first examine the relationship between treatment status and the running variable in each of these samples as well as in the pooled sample. Chart 4 illustrates this relationship for the pooled sample—specifically, between probability of treatment and the respective running variable centered at the cutoff (marking essentially the distance from the relevant cutoff). In Chart 4, panel B is the same as panel A, except that the sizes of the bubbles are proportional to the number of schools at that point. In each of the individual samples (Chart 1 for the writing sample; others available on request) as well as in the pooled sample (Chart 4), there is a sharp discontinuity at the cutoff, with an estimated discontinuity size of 1. The underlying validity assumptions (continuity of preexisting observables and continuity of density) are also satisfied for each of the individual samples and the pooled sample (estimates available on request).

There is no evidence of any increased classification into either the total ESE or excluded/included ESE categories, nor is there evidence of any change in classification into learning-disabled or emotionally handicapped categories.

The results for the LEP categories using the combined sample are reported in Table 6. The picture depicted in the table is very similar to that obtained above, both quantitatively

Table 6

Effect of Program on Classification in Excluded and Included Limited-English-Proficient Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples

	(1) Grade 2	(2) Grade 3	(3) Grade 4	(4) Grade 5
Percentage excluded	0.19	0.34**	0.30**	0.26
	(0.26)	(0.16)	(0.12)	(0.23)
Observations	215	216	213	205
R^2	0.03	0.05	0.03	0.04
Percentage included	0.12	-0.04	0.18	0.08
	(0.95)	(0.66)	(0.57)	(0.52)
Observations	215	216	213	205
R^2	0.02	0.05	0.02	0.02

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, and real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

and qualitatively. The estimates suggest that the "F" schools tended to classify an additional 0.34 percent of their total students into the excluded LEP category in the entry grade 3 and an additional 0.30 percent of their total students into the excluded LEP category in the high-stakes grade 4. These effects are statistically significant and equivalent to classifying as LEP an additional 2.37 students in grade 3 and an additional 2.1 students in grade 4. There is no statistically significant evidence of any change in classification in either the low-stakes grade 2 or high-stakes grade 5.

The results for ESE using the combined sample are reported in Tables 7 and 8. Like before, there is no evidence of any increased classification into either the total ESE or excluded/included ESE categories, nor is there evidence of any change in classification into learning-disabled or emotionally handicapped categories.

Table 7

Effect of Program on Classification in Exceptional Student Education (ESE) Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples

	Percentage		
	(1) Students in ESE	(2) Students in Excluded ESE	(3) Students in Included ESE
	-0.94	-1.01	0.34
	(1.40)	(1.61)	(0.77)
Observations	241	241	241
R^2	0.04	0.02	0.06

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

6.3 Are the Results Robust to Expressing the LEP Share as Percentage of Grade Enrollment?

Recall from footnote 13 that the various LEP or ESE shares (or percentages) are computed as percentages of total school enrollment. Since all ESE data are available at the school level, it is natural to divide ESE enrollment by total school enrollment to get the corresponding ESE percentage. However, since LEP data are available at the grade level, there are two alternatives: expressing excluded and included LEP as percentages of grade enrollment or as percentages of total school enrollment. In the above analysis, we take the latter route to be consistent with the definitions of various ESE percentages and to facilitate comparison with the ESE results. One disadvantage of using this definition, though, is that grade-specific LEP shares are also affected by enrollment changes in other grades.[20]

[20] Note, though, that when one divides by grade enrollment, grade-level LEP shares will change if non-LEP enrollment of that grade changes, even though LEP enrollment does not. Such a change will also be reflected in the first definition, in which we divide by total school enrollment, but dividing by total enrollment will dampen the effect of the change of the non-LEP share of the grade. Each measure, therefore, has its advantages and disadvantages.

Table 8

Program Effects on Classification in Learning-Disabled and Emotionally Handicapped Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples

	Percentage	
	(1) Students in Learning-Disabled	(2) Students in Emotionally Handicapped
	-0.23	-0.38
	(0.60)	(0.46)
Observations	241	241
R^2	0.06	0.03

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

Table 9

Program Effects on Classification in Excluded and Included Limited-English-Proficient (LEP) Categories: A Regression-Discontinuity Analysis Using Excluded and Included LEP as Percentages of Grade-Level Enrollment

	(1) Grade 2	(2) Grade 3	(3) Grade 4	(4) Grade 5
Percentage excluded	1.91	2.48**	1.62***	1.39
	(1.34)	(1.18)	(0.55)	(1.76)
Observations	123	121	119	116
R^2	0.53	0.51	0.42	0.43
Percentage included	0.28	-3.25	-1.13	-1.98
	(2.18)	(2.80)	(1.60)	(2.71)
Observations	123	121	119	116
R^2	0.66	0.57	0.55	0.35

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, and real per-pupil expenditure.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

To ensure that changes in enrollment in other grades are not driving the results above, and that they are robust to the definition of percentage (or share) used, we reestimate the above regression-discontinuity specifications for LEP using the alternative definition. In this section, percentage LEP is defined as LEP enrollment in that grade divided by total enrollment in the same grade.

The results for LEP are presented in Table 9 and are similar to those obtained above. There is evidence of increased classification into excluded LEP in both the entry grade 3 and high-stakes grade 4. To put the effects below in perspective, we note that in the immediate preprogram period (1999), average grade 3 and grade 4 enrollments of the schools under consideration were 125 and 124, respectively. Facing the threat of vouchers and stigma, the "F" schools resorted to an additional classification of 2.48 percent of their grade 3 students into the excluded LEP category in that grade and 1.62 percent of their grade 4 students into the excluded LEP category in grade 4. We observed that the coefficients here are bigger than earlier because of the difference in the definition of LEP share (excluded LEP expressed as a percentage of grade enrollment rather than school enrollment). These figures are equivalent to an increase of 2.87 students in grade 3 and 2.0 students in grade 4 and are similar to those obtained above. Moreover, there is no statistically significant evidence of a change in classification into either excluded categories in low-stakes grade 3 or high-stakes grade 5 nor is there evidence of any change in classification into included categories in any of the grades.

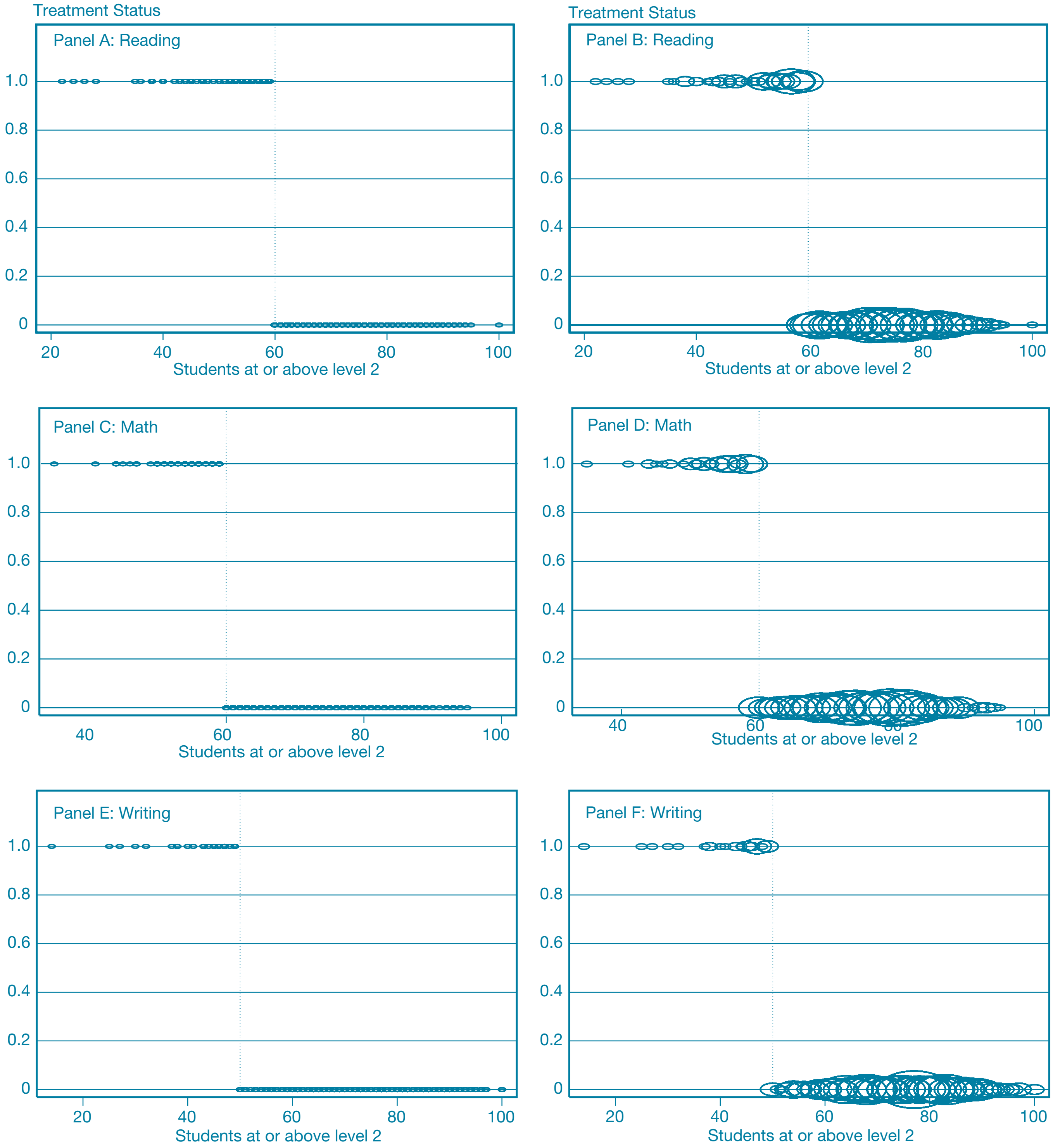
6.4 How "D" Schools Responded Relative to "C" Schools: A Regression-Discontinuity Analysis at the C/D Cutoff

A related question is whether the "D" schools exhibited any strategic behavior in terms of additional classification into excluded LEP and ESE categories. "D" schools did not face any direct threat of vouchers or stigma, but they were close to getting an F. Moreover, while they were not the lowest-performing schools, they were one of the lower-performing groups, and hence might have felt stigma to some extent. In this section, we

Chart 5

Relationship between Treatment Status (D) and Running Variable in Reading, Math, and Writing Samples

Regression-Discontinuity Estimates, 2000



Source: Authors' calculations.

Notes: The x-axis in all panels depicts percentages. In this chart, treatment status is 1 if a school received a grade of "D" and 0 if it received a grade of "C."

investigate whether the "D" schools responded differently than the "C" schools, ranking higher in the grade scale.

Once again, we use a regression-discontinuity strategy to study this response. Recall from section 2 that according to Florida rules, a school was assigned a D grade if it passed the minimum criteria in one or two of the three subject areas, while it got a C grade if it passed the minimum criteria in all three subject areas. Consider the group of schools that passed in two of the three subject areas. In this sample of schools, those that failed the third subject area should have received a D, while those that passed the third subject area should have received a C. There are three such possible samples: schools that passed in math and writing, but just passed or failed in reading (reading sample); schools that failed in reading and writing, but just passed or failed in math (math sample); and schools that passed in reading and math, but just passed or failed in writing (writing sample). According to Florida rules, the minimum criteria of each subject area yielded a sharp cutoff. In each of these samples, schools that were just below the cutoff in the third subject area should have received a D, and schools just above should have gotten a C.

Chart 5 illustrates the relationship between treatment status (for the purposes of this section, receiving a D rather than a C)[21] and the running variable for each of the three samples. Panels A and B show the relationship in the reading sample, where the running variable is the percentage of students at or above level 2; panels C and D illustrate the relationship in the math sample, where the running variable is the percentage of students at or above level 2; panels E and F depict the relationship in the writing sample, where the running variable is the percentage of students at or above level 3. For each sample, the second panel (B, D, and F) is the same as the first one (A, B, and C), except that each dot is weighted by the number of schools at that time. The smallest bubble corresponds to one school, while bigger bubbles correspond to larger numbers of schools. Indeed, we find that in the first two samples (Chart 5, panels A-B and panels C-D, respectively), the probability of treatment (getting a D) increases discontinuously at 60 percent as a function of the percentage of students scoring at or above level 2 in reading (math). In the third sample, the probability of treatment increases discontinuously at 50 percent as a function of the percentage of students scoring at or above level 3 in writing. As can perhaps be anticipated from the panels, each of these samples yields an estimated discontinuity of size 1 at the respective cutoffs.

> *[W]hile the "D" schools may have faced an indirect threat and some stigma since they were close to F status, those issues were not enough to lead to any strategic classifications into ... excluded categories.*

Table 10

Effect of Program on Classification in Excluded and Included Limited-English-Proficient Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples for Schools at the C/D Cutoff

	(1) Grade 2	(2) Grade 3	(3) Grade 4	(4) Grade 5
Percentage excluded	-0.09	-0.09	-0.02	-0.22
	(0.06)	(0.06)	(0.06)	(0.14)
Observations	331	327	333	321
R^2	0.45	0.40	0.57	0.42
Percentage included	0.27	0.30	0.20	-0.07
	(0.17)	(0.24)	(0.12)	(0.13)
Observations	311	311	306	294
R^2	0.92	0.90	0.85	0.76

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, and real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained; regressions in the last three rows also include the lagged dependent variable as an additional covariate (see footnote 20).

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

To leverage efficiency gains and to build power, we pool these three samples together, centering the running variables at the respective cutoffs. First, we check whether the standard assumptions that govern the validity of regression-discontinuity techniques are satisfied in this context. Specifically, we find that for each of these samples as well as the combined sample, observable preprogram characteristics are indeed smooth through the cutoff. The preprogram results for the reading sample are presented in the appendix;[22] results for the other samples are not reported for lack of space, but are

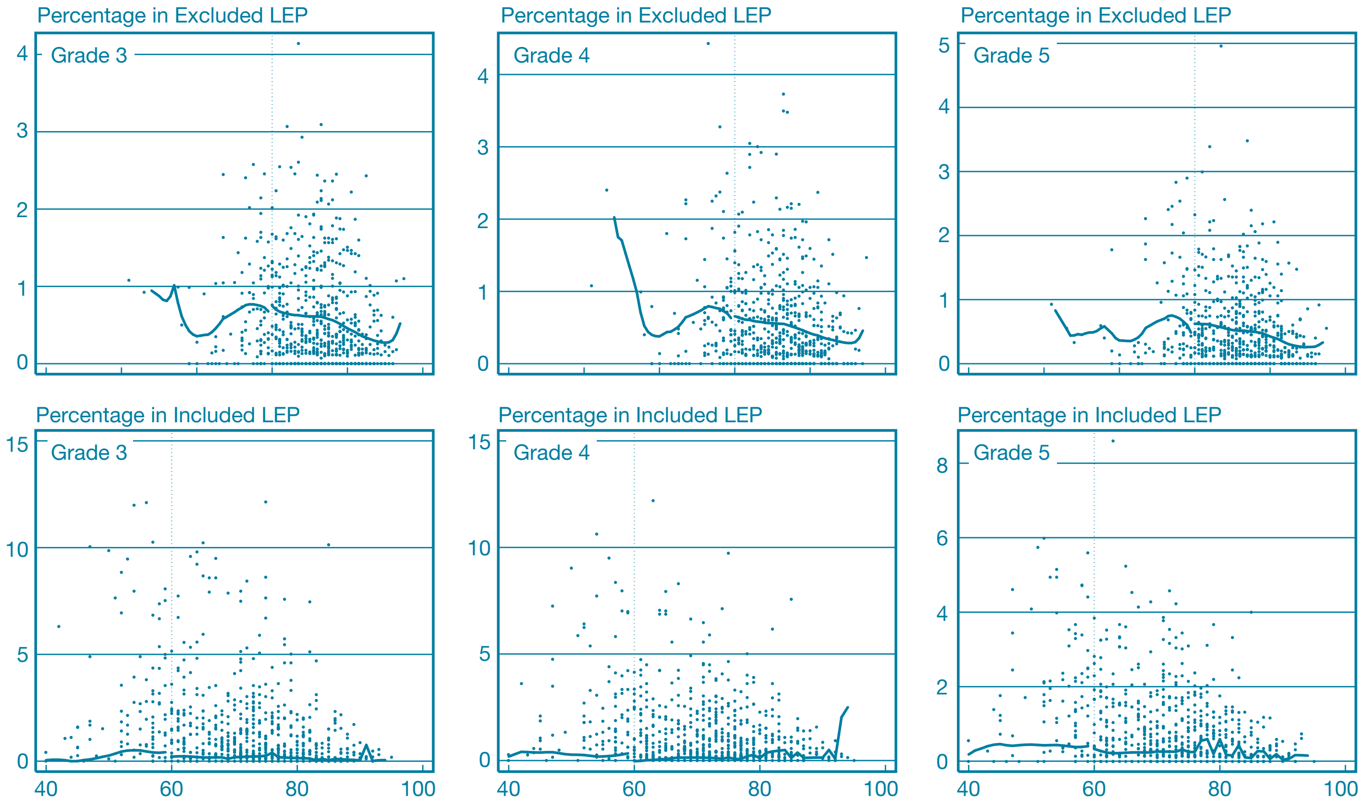
[21] Here, receiving a D in the immediate preprogram year (1999) is considered to be the treatment. In the rest of the article, getting an F in 1999 is the treatment.

[22] One exception is the estimate for included LEP percentage in grade 5, which is statistically different from zero. However, with a large number of differences, it is natural to have a few statistically different from zero, even if by random variation. Still, we observe that even though the coefficients for percentage LEP in the other grades are not statistically different from zero, they are not small. Therefore, in the estimations for included LEP in this subsection, we include the lagged dependent variable as an additional covariate.

Chart 6

Effect of Program on Classification in Excluded and Included Limited-English-Proficient (LEP) Categories on Schools at the C/D Cutoff

Regression-Discontinuity Estimates at C/D Cutoff; February 2000 Survey



Source: Authors' calculations.

Note: The x-axis in each panel depicts the percentage of students at or above level 2 in FCAT (Florida Comprehensive Assessment Test) reading in 1999.

available on request. We also find no evidence of discontinuity in the density of any of the running variables at the cutoff. (These results are also not reported here, but are available on request.)

Having established the validity of regression-discontinuity design in this context, and using the combined sample, we investigate in Table 10 and Chart 6 the effect of the program on classification into excluded and included LEP categories in "D" schools at the cutoff (relative to "C" schools). Interestingly, there is no evidence of any differential classification in the "D" schools at the cutoff into either excluded or included LEP categories in any of the low- or high-stakes grades.

We also look at the effect of getting a D on classification into total ESE, excluded ESE, and included ESE (Table 11 and Chart 7) as well as into more mutable learning-disabled and emotionally handicapped categories (Table 12). Once again, there is no evidence of any differential classification into any of these categories at the cutoff. These results imply that while the "D" schools may have faced an indirect threat and some stigma since they were close to F status, those issues were not enough to lead to any strategic classifications into any of the excluded categories. In contrast, the direct threat of vouchers and the stigma effect associated with the lowest grade led to additional classifications by the "F" schools (at the cutoff) into excluded LEP categories in high-stakes grade 4 and entry grade 3.

Table 11

Effect of Program on Classification in Exceptional Student Education (ESE) Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples for Schools at the C/D Cutoff

	Percentage		
	(1) Students in ESE	(2) Students in Excluded ESE	(3) Students in Included ESE
	-0.001	-0.001	0.000
	(0.008)	(0.006)	(0.004)
Observations	383	383	383
R^2	0.17	0.20	0.05

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

Table 12

Effect of Program on Classification in Learning-Disabled and Emotionally Handicapped Categories: A Regression-Discontinuity Analysis Combining the Three Discontinuity Samples of Schools at the C/D Cutoff

	Percentage	
	(1) Students in Learning-Disabled	(2) Students in Emotionally Handicapped
	0.001	-0.001
	(0.003)	(0.003)
Observations	383	383
R^2	0.16	0.07

Source: Authors' calculations.

Notes: Robust standard errors adjusted for clustering using the running variable are in parentheses. All regressions control for racial composition, gender composition, percentage of students eligible for free/reduced-price lunch, real per-pupil expenditure, and include sample dummies to control for the respective sample from which the observation is obtained.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

7. Implications for Education Policies in New York

The Florida experience yields important lessons for school accountability programs elsewhere. These policies include New York City's accountability framework, known as the Progress Reports program, and the federal No Child Left Behind Act as implemented by New York State.

In 2007, the New York City Department of Education introduced a new accountability system centered on annual school progress reports. These publicly available school "report cards" assign each school a letter grade ranging from A to F based on three separate components: school environment, student performance, and student progress (accounting for 15 percent, 30 percent, and 55 percent of the overall score, respectively). The school environment score is based on responses to surveys given to teachers, parents, and students in grade 6 and above. The student-performance and progress scores are based on student performance on statewide standardized math and English language arts tests. The performance score is based on the level of test scores in the current year, while the progress score is based on improvements or declines in test scores compared to previous years.

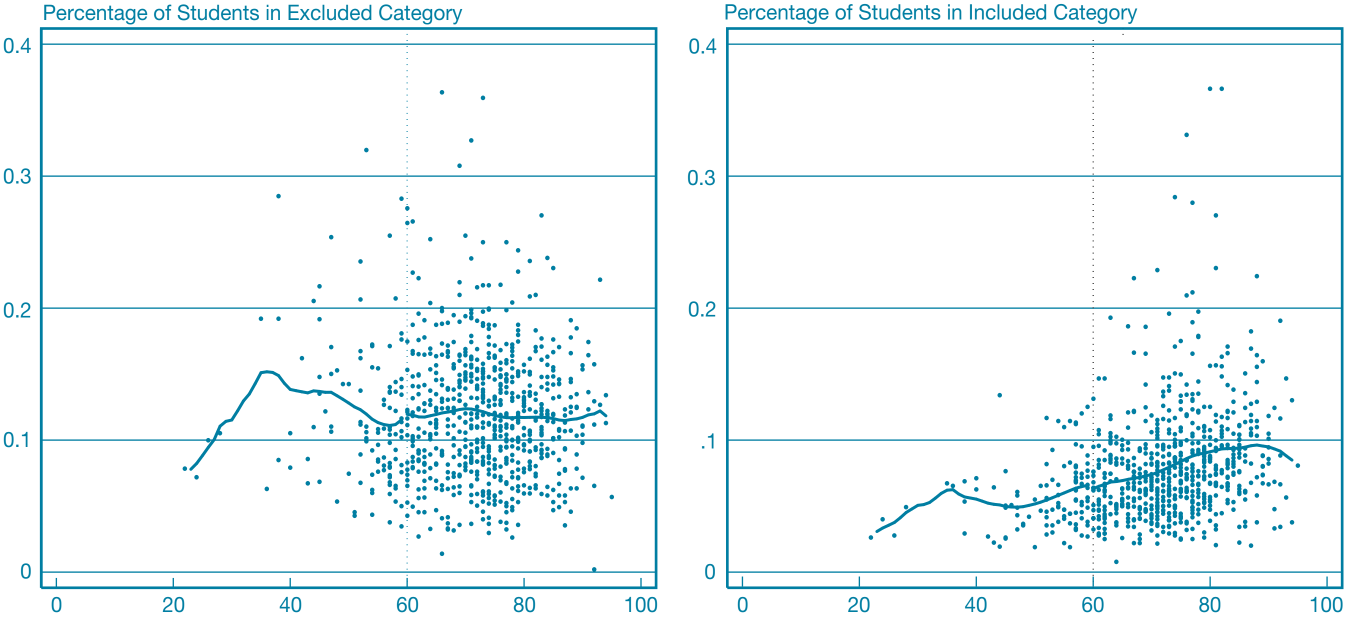
In contrast to the Florida program, New York City's accountability program includes not only high-needs students in grade calculations, but also gives schools additional credit for making achievement gains with particular high-needs groups: English language learners (ELL), special education students, and students performing in the lowest third of all students citywide. Overall scores are calculated as a weighted sum of the scores in each component plus any additional credit received. Letter grades from A to F correspond to specific thresholds on the overall score scale. Thus, additional credit can (and has already often) allowed schools to receive a higher grade.

The Florida experience yields important lessons for school accountability programs elsewhere . . . [including] New York City's accountability framework, known as the Progress Reports program, and the federal No Child Left Behind Act as implemented by New York State.

Chart 7

Effect of Program on Classification in Excluded and Included Exceptional Student Education (ESE) Categories on Schools at the C/D Cutoff

Regression-Discontinuity Estimates, 2000



Source: Authors' calculations.

Note: The x-axis in each panel depicts the percentage of students at or above level 2 in FCAT (Florida Comprehensive Assessment Test) reading in 1999.

This approach attaches clear rewards for high scores and clear sanctions for low scores. Schools receiving high grades are eligible for increases in per-pupil funding, and their principals are eligible for bonuses ranging from $7,000 to $25,000. In contrast, schools receiving low grades (F or D) are threatened with principal dismissal, restructuring, or even closure. This threat is credible and has often been implemented in practice.[23] In addition to the possibility of leadership change or closure, all schools receiving F and D grades (or a C grade three years in a row) are required to implement school improvement measures and target-setting. Finally, students in "F" schools are eligible to transfer to better-performing public schools.

Although the Progress Reports program does not include a voucher element, it is in many ways similar to the Florida voucher program. For example, it assigns schools letter grades based in part on student performance on standardized tests and imposes sanctions on low-performing schools, including allowing students to transfer out of failing schools.[24] But a key difference is that the New York City program includes the test scores of all ELL and special education students in the computation of school grades. In fact, it gives schools extra credit for achieving progress with ELL and special education students as well as other high-needs groups (such as students in the lowest third citywide). This additional credit can be substantial—in 2007, 161 schools received a higher grade due to additional credit (Rockoff and Turner 2010). Consequently, the strategic classification we describe earlier in the Florida context would not be expected to take place in New York City. However, the New York City program rules can generate other adverse incentives for classification. Since the failing schools there can earn additional credit for demonstrating progress of ELL and special education students, they might have an impulse to classify their higher-performing students in these categories in an effort to artificially boost scores.[25] Whether or not this behavior actually happened is a topic of future research.

[23] In December 2007, the New York City Department of Education announced that seven of the forty-two schools receiving F grades and two of the eighty-seven schools receiving D grades would be closed or phased out in the following year (Rockoff and Turner 2010); this sent a clear signal to other low-performing schools that the threat of closure was credible.

[24] Students are eligible to transfer to public schools but do not receive vouchers to transfer to private schools, as they do in Florida.

We now turn to the federal education law—the No Child Left Behind Act—as implemented in New York. Like New York City's Progress Reports program, NCLB establishes an accountability framework modeled on the Florida program, though with important differences.

NCLB, a major reform of the Elementary and Secondary Education Act, was signed into law on January 8, 2002. The states, including New York, implemented it soon thereafter. In compliance with the law, New York established targets for adequate yearly progress (AYP). AYP is determined based on each school's progress toward meeting target proficiency levels for all students in English-language arts, mathematics, and science. Schools must achieve these proficiency targets for the student body as a whole, and also for particular subgroups of students. Schools must also have an average of 95 percent of students participating in state tests over two years. Finally, schools must meet a target for attendance rate or, in the case of high schools, for graduation rate. If a school does not meet requirements in any one of these categories, it is said to miss AYP.

> *In all, the features of both New York City's Progress Reports program and the federal No Child Left Behind Act (as implemented in New York) represent important steps forward in eliminating adverse incentives for the type of strategic reclassification that appears to have taken place in Florida.*

Schools that receive Title I federal funds are subject to NCLB sanctions if they miss AYP for two consecutive years. A Title I school missing AYP for two consecutive years is required to provide public school choice to its students. That rule permits students to transfer to better-performing public schools, with per-pupil funding following the students to their new schools. If a school misses AYP for three consecutive years, it is required to provide (and finance) supplemental educational services (such as tutoring) in addition to public school choice. Missing AYP for four consecutive years leads to corrective action in addition to the above sanctions; missing it for five consecutive years leads to restructuring in addition to the sanctions.

Recall that schools must meet AYP not only for the student body as a whole, but for particular subgroups: white, black, Hispanic, Asian, and American Indian students; students with disabilities; students with limited English proficiency; and students from low-income families. If a school fails to meet the target for any subgroup, it is deemed to have missed AYP. Thus, LEP students, students with disabilities, and other subgroups are not only included in the calculation of scores for the "All Students" group, they also separately count toward AYP formation.[26] Therefore, the potential incentives to reclassify weak students into ungraded groups are not present here.

In all, the features of both New York City's Progress Reports program and the federal No Child Left Behind Act (as implemented in New York) represent important steps forward in eliminating adverse incentives for the type of strategic reclassification that appears to have taken place in Florida. These two programs do not permit high-needs students to be excluded from the calculation of school grades.[27] All students count toward grade formation, and, in the case of the New York City program, the weaker categories carry more weight. While this program design can potentially ward off the gaming of the system seen in Florida, it introduces an incentive to move stronger students into high-needs categories as a way to boost scores.

8. Conclusion

This article analyzes the responses of public schools to the Florida Opportunity Scholarship Program, an influential school accountability policy employing vouchers as a sanction for low school achievement. Looking closely at the institutional details of the program, we identify the incentives it establishes and the behavior of public schools responding to it. Under the program, two types of students were excluded from the calculation of school grades: limited-English-proficient students in an ESOL program for less than two years and several categories of special education students. As a result, threatened schools may have had an incentive to reclassify their low-performing students into these exempted categories in order to remove them from school grade calculations and thereby artificially inflate their marks. Did this actually happen in practice?

Using data obtained from the Florida Department of Education and a regression-discontinuity approach, we compare LEP and ESE classification in schools that barely

[25] It is important to note, though, that students have to test into the special education categories. Consequently, it can be relatively difficult to have higher-performing students test into these categories since they are more likely to pass the diagnostic tests.

[26] The only exemption is for any subgroup with less than forty students in a school (less than fifty for the students with disabilities subgroup). Subgroups with small numbers of students are not evaluated separately, but students in these groups are still included in the evaluation of the "All Students" group.

[27] An exception should be noted here: If a school's total enrollment is less than forty, and even a summing of total enrollment over three years does not yield a total of forty, then that school and its students are exempted from AYP determination. But, as might be expected, this is a very rare occurrence.

avoided the threat of vouchers with such classification in schools that barely received the threat. We find robust evidence that the threatened schools classified a greater percentage of their students into the excluded LEP category in high-stakes grade 4 and entry grade 3. We find no evidence of any differential classification into the included LEP category in any of the grades. For reference, there was no evidence of a difference in behavior between threatened versus non-threatened schools before the program. These findings suggest that schools threatened with vouchers and stigma tended to reclassify students into the excluded LEP category in an effort to remove them from the effective test-taking pool in both the current year and the following year.

In contrast, we find no evidence that the program led to greater classification into excluded (or included) ESE categories by the threatened schools. This result is not surprising given the substantial costs associated with ESE classification. The main disincentive to this form of classification was posed by Florida's McKay Scholarship program, which made any student with disabilities in Florida public schools eligible for vouchers to move to a private school or another public school. Under the McKay program, schools that classified students into excluded ESE categories faced losing them and their corresponding per-pupil funding. Since McKay vouchers cost about twice as much on average as FOSP vouchers, schools actually risked losing more funding with a move of an ESE student under the McKay program than with the departure of a regular student under the Florida program. It is likely that threatened schools weighed the costs and benefits of their options and chose to respond in the least costly ways.

These findings have important implications for school accountability policies in the New York region. New York City's Progress Reports program and New York's implementation of the federal No Child Left Behind Act were modeled in part on the Florida program, though both have avoided the types of exemptions that incentivized gaming of the system in Florida. Because the policies hold schools accountable for the performance of all students—including limited-English-proficient and special education students—New York schools do not have adverse incentives to classify weaker students into these categories. Moreover, schools have the motivation to improve the performance of these and other historically low-performing groups since such improvements are tied to better school grades and concomitant rewards. The New York City program rules, however, have the potential to induce schools to classify their high-performing students into these high-needs groups in an effort to earn extra credit and better grades. Whether or not this kind of sorting actually happened is a topic of future research.

The general lesson to take from examining the Florida and New York accountability policies is that policymakers must be careful when designing exemptions, special allowances, or credits for certain groups of students since these accommodations can create adverse incentives and unintended consequences. While accountability policies must acknowledge the challenges schools face in educating students with limited English proficiency, disabilities, and other special needs, excluding them entirely from accountability measures may induce struggling schools to reclassify low-performing students into exempted categories. The danger is that such an approach can lead to strategic sorting rather than genuine improvements to the quality of education for the students whom the programs aimed to help.

Appendix

Testing Validity of 1999 Regression-Discontinuity Analysis: Looking for Discontinuities in Preprogram Characteristics at the C/D Cutoff (Reading Sample)

	Percentage				
	(1)	(2)	(3)	(4)	(5)
Panel A	White	Black	Hispanic	Asian	American Indian
	5.99	-6.51	3.12	-0.51	-0.18
	(4.074)	(3.959)	(5.560)	(0.310)	(0.126)
Panel B	Percentage Multiracial	Percentage Male	Percentage Free/Reduced-Price Lunch	Enrollment	Real Per-Pupil Expenditure
	0.20	1.67	-1.19	18.66	0.61
	(0.137)	(0.809)	(1.294)	(42.168)	(0.426)
	Percentage				
Panel C	Exceptional Student Education (ESE)	Excluded ESE	Included ESE	Learning-Disabled	Emotionally Handicapped
	-0.002	-0.004	0.002	-0.004	0.001
	(0.008)	(0.008)	(0.005)	(0.004)	(0.004)
	Percentage Excluded Limited-English-Proficient (LEP)				
Panel D	Grade 2	Grade 3	Grade 4	Grade 5	
	0.075	-0.051	-0.197	-0.058	
	(0.084)	(0.094)	(0.115)	(0.196)	
	Percentage Included LEP				
Panel E	Grade 2	Grade 3	Grade 4	Grade 5	
	0.852	0.952	0.442	0.908***	
	(0.531)	(0.608)	(0.456)	(0.289)	

Source: Authors' calculations.

Note: Robust standard errors adjusted for clustering using the running variable are in parentheses.

***Statistically significant at the 1 percent level.
**Statistically significant at the 5 percent level.
*Statistically significant at the 10 percent level.

References

Chakrabarti, R. 2008a. "Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs." Federal Reserve Bank of New York Staff Reports, no. 315, January.

———. 2008b. "Can Increasing Private School Participation and Monetary Loss in a Voucher Program Affect Public School Performance? Evidence from Milwaukee." Journal of Public Economics 92, nos. 5-6 (June): 1371-93.

———. 2013. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida." Economic Inquiry 51, no. 1 (January): 500-26.

Chiang, H. 2009. "How Accountability Pressure on Failing Schools Affects Student Achievement." Journal of Public Economics 93, nos. 9-10 (October): 1045-57.

Cullen, J., and R. Reback. 2006. "Tinkering towards Accolades: School Gaming under a Performance Accountability System." In T. J. Gronberg and D. W. Jansen, eds., Improving School Accountability: Check-Ups or Choice. Advances in Applied Microeconomics 14. Amsterdam: Elsevier.

Fan, J., and I. Gijbels. 1996. "Local Polynomial Modeling and Its Applications." Monographs on Statistics and Applied Probability 66. London: Chapman and Hall.

Figlio, D. 2006. "Testing, Crime, and Punishment." Journal of Public Economics 90, nos. 4-5 (May): 837-51.

Figlio, D., and L. Getzler. 2006. "Accountability, Ability, and Disability: Gaming the System?" In T. J. Gronberg and D. W. Jansen, eds., Improving School Accountability: Check-Ups or Choice. Advances in Applied Microeconomics 14. Amsterdam: Elsevier.

Figlio, D., and C. Hart. 2010. "Competitive Effects of Means-Tested Vouchers." NBER Working Paper no. 16056, June.

Figlio, D., and M. Lucas. 2004. "What's in a Grade? School Report Cards and the Housing Market." American Economic Review 94, no. 3 (June): 591-604.

Figlio, D., and C. Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" Journal of Public Economics 90, nos. 1-2 (January): 239-55.

Figlio, D., and J. Winicki. 2005. "Food for Thought? The Effects of School Accountability Plans on School Nutrition." Journal of Public Economics 89, nos. 2-3 (February): 381-94.

Greene, J. 2001. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." Manhattan Institute for Policy Research civic report, February.

Greene, J., and M. Winters. 2003. "When Schools Compete: The Effects of Vouchers on Florida Public School Achievement." Manhattan Institute for Policy Research Education Working Paper no. 2, August.

Hahn, J., P. Todd, and W. van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design." Econometrica 69, no. 1 (January): 201-9.

Hanushek, E. A., J. F. Kain, and S. G. Rivkin. 2002. "Inferring Program Effects for Special Populations: Does Special Education Raise Achievement for Students with Disabilities?" Review of Economics and Statistics 84, no. 4 (November): 584-99.

Hoxby, C. 2003a. "School Choice and School Productivity: Could School Choice Be a Tide that Lifts All Boats?" In C. Hoxby, ed., The Economics of School Choice. Chicago: University of Chicago Press.

———. 2003b. "School Choice and School Competition: Evidence from the United States." Swedish Economic Policy Review 10: 9-65.

Imbens, G. W., and T. Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." Journal of Econometrics 142, no. 2 (May): 615-35.

Jacob, B. 2005. "Accountability, Incentives, and Behavior: The Impacts of High-Stakes Testing in the Chicago Public Schools." Journal of Public Economics 89, nos. 5-6 (June): 761-96.

Jacob, B., and S. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Quarterly Journal of Economics 118, no. 3 (August): 843-77.

McCrary, J. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." Journal of Econometrics 142, no. 2 (February): 698-714.

References (Continued)

Neal, D., and D. W. Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability." Review of Economics and Statistics 92, no. 2 (May): 263-83.

Reback, R. 2008. "Teaching to the Rating: School Accountability and Distribution of Student Achievement." Journal of Public Economics 92, nos. 5-6 (June): 1394-415.

Rockoff, J. E., and L. J. Turner. 2010. "Short-Run Impacts of Accountability on School Quality." American Economic Journal: Economic Policy 2, no. 4 (November): 119-47.

Rouse, C. E., J. Hannaway, D. Figlio, and D. Goldhaber. 2007. "Feeling the Florida Heat: How Low-Performing Schools Respond to Voucher and Accountability Pressure." National Center for Analysis of Longitudinal Data in Education Research Working Paper no. 13, November.

Silverman, B. W. 1986. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.

van der Klaauw, W. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." International Economic Review 43, no. 4 (November): 1249-87.

West, M., and P. Peterson. 2006. "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments." Economic Journal 116, no. 510 (March): 46-62.

The views expressed are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. The Federal Reserve Bank of New York provides no warranty, express or implied, as to the accuracy, timeliness, completeness, merchantability, or fitness for any particular purpose of any information contained in documents produced and provided by the Federal Reserve Bank of New York in any form or manner whatsoever.