ACCOUNTABILITY WITH VOUCHER THREATS, RESPONSES, AND THE TEST-TAKING POPULATION: REGRESSION DISCONTINUITY EVIDENCE FROM FLORIDA

Rajashri Chakrabarti
Federal Reserve Bank
of New York
33 Liberty Street
New York, NY 10045
Rajashri.Chakrabarti@ny.frb.org

Abstract

Florida's 1999 A-plus program was a consequential accountability program that embedded vouchers in an accountability regime. Under Florida rules, scores of students in several special education (ESE) and limited English proficient (LEP) categories were not included in the computation of school grades. One might expect these rules to induce F schools (that faced stigma and threat of vouchers) to strategically classify their weaker students into these excluded categories. The interplay of these rules with those of the McKay program for disabled students, however, created an interesting divergence of incentives as far as classifications into excluded LEP and ESE were concerned. Because classifying students into ESE made them eligible for McKay vouchers that were funded by public school revenue, the McKay program acted as a strong disincentive to such classification. Using a regression discontinuity strategy, I investigate whether the differences in incentives led the F schools to exhibit different behaviors as far as classifications into excluded ESE and LEP were concerned. Indeed, I find robust evidence in favor of classification into excluded LEP in high-stakes grade 4 and entry grade 3. In contrast, I do not find evidence of classification into excluded ESE.

1. INTRODUCTION

Continued concerns over public school performance after the publication of *A Nation at Risk* in 1983 have pushed public school reform to the forefront of policy debates in the United States (National Commission on Excellence in Education 1983). Various reforms have been debated, and school accountability and school choice have been among the foremost of these. This paper analyzes the effect of a consequential accountability system in Florida on public school incentives and behavior. Understanding the behavior and responses of public schools facing alternative school reform initiatives is paramount to an effective policy design and this paper takes a step forward in that direction. Moreover, the federal No Child Left Behind Act is similar to and largely modeled after the Florida program, which makes understanding the impact of the Florida program all the more interesting and relevant.

Written into law in June 1999, the Florida A-plus program embedded a voucher program within a school accountability system. It graded public schools on a scale of A–F (A = highest, F = lowest) based primarily on Florida Comprehensive Assessment Test (FCAT) scores in reading, math, and writing. Unlike in other pure accountability systems, the A-plus program attached consequences to the lowest-performing grade, F. Specifically, it made all students of a Florida public school eligible for vouchers (opportunity scholarships) if the school received two F grades in a period of four years. The F grade, being the lowest-performing grade, exposed schools to shame and stigma, in the sense that all their students would be eligible for vouchers if the school received another F grade in the next three years. In addition, schools getting an F grade for the first time were directly threatened by vouchers. Vouchers were associated with a loss in revenue (equivalent to state aid per pupil for each student) and also negative media publicity and visibility. Therefore, the schools receiving the first F grade had strong incentives to try to avoid the second F, to escape stigma and threat of vouchers. This paper studies some alternative ways in which these schools might have responded facing the incentives built into this consequential accountability system.[1]

This study exploits the fact that the Florida rules created a key divergence of incentives as far as classification in special education and limited English proficient categories were concerned. Under Florida rules, scores of students in several special education categories—Exceptional Student Education (ESE) categories—and limited English proficient (LEP) categories were not included

1. For the 1999 F schools, the consequences (threat of vouchers) attached to the accountability program remained in effect for the next three years only. Therefore, the main focus of this study is the behavior of the 1999 F schools during these three years. I briefly study the responses to the 2002 accountability program later (section 8), however, to examine whether the responses to that program were consistent with responses observed in response to the 1999 program.

in the computation of grades. Given these rules, one might expect the threatened schools to strategically classify some of their weaker students into these "excluded" ESE and LEP categories so as to remove them from the relevant test-taking pool in an effort to boost scores.

Although this might have been a plausible response in the absence of other incentives, Florida had a scholarship program for disabled students that created an interesting difference in incentives for classification along these two margins. Created in 1999, and fully implemented in the 2000–1 school year, the McKay Scholarship program for disabled students made every disabled Florida public school student eligible for vouchers to move to a private school (religious or nonreligious) or to another public school. Thus classification into special education categories was associated with a risk of loss of the student to McKay vouchers. Like the opportunity scholarship vouchers, the McKay vouchers were also funded by public school revenue. The McKay scholarships were far more generous than the opportunity scholarships, however. They ranged between $4,500 and $20,000, and averaged around $7,000. In contrast, the opportunity scholarships during this period (1999–2000 through 2001–2) averaged at around $3,500. Thus the interaction of the rules of Florida's consequential accountability system and the McKay scholarship program created an interesting bifurcation of incentives as far as classifications in ESE and LEP were concerned. Whereas F schools trying to escape the second F grade still had incentives to classify their low-performing students into excluded LEP categories, such an incentive did not exist for ESE because of the potential cost posed by the McKay scholarship program. In this paper, I study whether the F schools behaved according to these incentives, and specifically, whether they exhibited a difference in response as far as classifications in LEP and ESE were concerned.

Using a regression discontinuity (RD) estimation strategy that exploits the institutional details of the Florida program, I find that the program led to increased classification into excluded LEP categories in high-stakes grade 4 and the entry grade to high-stakes grade 3[2] in the first year after the program. Specifically, threatened schools classified an additional 0.31 percent of their total students in the excluded LEP category in grade 4 and an additional 0.36 percent of their total students in this category in grade 3 in the first year after program. These figures amounted to 53 percent of the excluded LEP students in grade 4, and 55 percent of the excluded LEP students in grade 3, respectively, in that year. In terms of numbers of students, these were equivalent to classification of 2.3 additional students in excluded LEP in grade 4 and 2.6 additional students in grade 3.

2. This grade will be referred to as "entry grade 3" in the rest of the paper.

In contrast, I do not find any evidence that the threatened schools resorted to increased classification into "excluded" ESE categories in any of the three years after program. There is also no evidence of any change in classification in either included ESE or included LEP categories. These results are reasonably robust—they are not explained by student-sorting or changes in demographic and socioeconomic compositions of schools or schools' levels of spending, and withstand a variety of other sensitivity checks.

Exploiting further the differences in extents of McKay voucher competition across schools and the role of 2002 changes in Florida's accountability system yields some interesting insights.[3] Schools facing more McKay voucher competition tended to classify fewer students into excluded ESE categories, but more into excluded LEP categories. These differences in behaviors are again consistent with incentives for classification along these two dimensions. Schools facing more McKay competition had a higher probability of losing their ESE students and hence were likely less inclined to classify students into ESE. The presence of a larger concentration of McKay-accepting private schools also implied larger private school competition in general, however, and hence a large potential loss of students.[4] So, it is reasonable to expect these schools to resort to larger strategic classification into excluded LEP categories in an effort to artificially boost their scores, as a lower grade likely increased the chances of such loss.

Florida's accountability system underwent some major changes in 2002. The 1999 accountability system was relatively straightforward in that it required certain percentages of students to score at or above a cutoff to pass in that subject area, and schools could escape an F by satisfying the criterion in only one of the three subject areas. In such a scenario, removing a selected few low-performing students from the test-taking pool might have seemed promising to schools to escape an F grade. In contrast, the 2002 shock made the accountability system far more complicated. In addition to level scores, the system introduced points for gain scores and entailed aggregation of points over a number of criteria, and also made it impossible to avoid an F grade on the basis of a single test. As a result, one might expect the 2002 changes to have reduced the attractiveness and relative benefit of strategically classifying students into an excluded LEP category. Indeed, consistent with this, I find no evidence that the F schools increased classification into excluded LEP (or ESE) categories in response to the 2002 accountability shock.

This study is related to two strands of literature. The first strand investigates whether schools facing accountability systems and testing regimes

3. I would like to thank an anonymous referee for suggesting these strategies.
4. This is because McKay private schools were regular private schools that made themselves available to accept McKay students.

respond by gaming the system in various ways. Cullen and Reback (2006), Figlio and Getzler (2006), and Jacob (2005) find evidence of classification of low-performing students into excluded disabled categories. Jacob (2005) finds evidence of teaching to the test, preemptive retention of students, and substitution away from low-stakes subjects. Jacob and Levitt (2003) find evidence of teacher cheating, and Reback (2008), Ladd and Lauen (2010), and Neal and Schanzenbach (2010) find evidence in favor of differential focus on marginal students. Figlio (2006) finds that low-performing students were given longer suspensions during the testing period than higher-performing students for similar crimes. Figlio and Winicki (2005) find that schools faced with accountability systems increased the caloric content of school lunches on testing days in an attempt to boost performance.

The second strand of literature analyzes the effect of Florida's A-plus choice and accountability program on public school performance and behavior. This literature finds evidence in favor of improvement of the treated schools in response to the program (Greene 2001; Greene and Winters 2003; Chakrabarti 2008a; Figlio and Rouse 2006; West and Peterson 2006). Rouse et al. (2007) and Chiang (2009) find evidence in favor of persistence of achievement gains in the medium-run of students who attended F schools in Florida. Both studies also find evidence in favor of behavioral changes of these schools—such as more focus on instruction and teacher development. Chakrabarti (2012) finds that threatened schools facing the same program in Florida tended to focus more on students expected to score just below the minimum criteria cutoffs. Goldhaber and Hannaway (2004) and Chakrabarti (2012) also find evidence that F schools tended to overwhelmingly focus on writing, rather than reading and math (passing in one subject was sufficient to escape an F).[5]

Thus, whereas there is a rich literature that investigates whether accountability regimes led affected schools to reclassify their low-performing students into excluded categories, this study investigates whether schools facing a consequential accountability system (that embedded vouchers in a full-fledged accountability system) behaved in a similar way. What makes this study even more distinct and sets it apart from the existing literature is its ability to tap into the unique institutional details of Florida programs that generated very different incentives for classifications into excluded ESE and LEP categories. As discussed previously, the interaction of the A-plus and McKay program rules created incentives for increased classifications in excluded LEP categories, but

5. Also related to this study is Figlio and Hart (2010), who study the Florida tax credit scholarship program. They find evidence in favor of improvement of threatened schools facing voucher threats via the Florida tax credit scholarship program. For studies on the impacts of publicly funded means-tested voucher programs on public schools in the United States, see Hoxby (2003a, b) and Chakrabarti (2008b).

not in ESE. Exploiting these differences in incentives, I investigate whether the F schools responded differently in these two dimensions. In other words, the difference in incentives along two forms of exclusions allows me to examine in a more definitive way the role of incentives and responses facing such exclusions. The findings also have important policy implications. On the one hand they illustrate that presence of excluded categories may lead to strategic classifications into these excluded categories, on the other they illustrate that counter-incentives offered by alternative policy tools can go a long way in thwarting such gaming.

Finally, a recent study that is worth discussing here is Winters and Greene (2011). The authors study the effects of Florida's McKay scholarship program during 2002–5 and find that, on the one hand competition from the McKay program decreased the probability of a student to be diagnosed as learning disabled and, on the other the increased competition led to an improvement in performance of the public schools. The current study differs from that paper in that its focus is on the effect of Florida's A-plus program. Moreover, it studies not only classification in learning disabled categories but classification in different excluded and included ESE and LEP categories. It also relates to a different time period. The key difference, however, is that the current paper seeks to study how the differences in incentives for classifications in ESE and LEP created by the interplay of Florida's A-plus and McKay programs led schools to respond along these two margins.

2. INSTITUTIONAL DETAILS

Florida's A-plus choice and accountability program, signed into law in June 1999, embedded vouchers in an accountability system. Under this program, all students of a public school became eligible for vouchers or "opportunity scholarships" 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 and stigma, but its students did not become eligible for vouchers unless and until it got a second F within the next three years.

Following a field test in 1997, the FCAT reading and math tests were first administered in 1998. The FCAT writing test was first administered in 1993. The reading and writing tests were given in grades 4, 8, and 10 and math tests in grades 5, 8, and 10.

The system of assigning letter grades to schools started in the year 1999,[6] and they were based on the FCAT reading, math, and writing tests. The state designated a school an F if it failed to attain the minimum criteria in all three FCAT subjects, and a D if it failed the minimum criteria in only one or

6. Before 1999, schools were graded by a numeric system of grades, I–IV (I = lowest, IV = highest).

two of the three subject areas. To pass the minimum criteria in reading and math, at least 60 percent of the students had to score at level 2 and above in the respective subject, and to pass the minimum criteria in writing, at least 50 percent had to score 3 or above.[7]

Scores of all regular students were included in the computation of school grades. Scores of students in only some ESE and LEP categories, however, were included in the calculation of grades. Specifically, ESE students belonging to the three categories of speech impaired, gifted, and hospital/homebound as well as LEP students with more than two years in an English for Speakers of Other Languages (ESOL) program were included in school grade computations. In contrast, 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. 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. From now on, I will refer to the "less than two years in an ESOL program" category as the "excluded" LEP category, and "two years or more in an ESOL program" category as the "included" LEP category. Similarly, I will refer to the speech impaired, gifted, and hospital/homebound categories as "included" ESE categories, and the other ESE categories as "excluded" ESE categories.

To understand the incentives built into the system, it is important to understand the rules and procedures governing placement into ESE and LEP categories in Florida. Every student entering a Florida public school was offered a survey that elicited the student's exposure to English (questions included whether a language other than English was spoken at home, whether the first language was other than English, or whether the student frequently spoke a language other than English). Students answering any of these questions in the affirmative were given an eligibility assessment, and, conditional on performance in this assessment (scoring within the limited English proficient range), were classified as an LEP student. Classification into LEP could happen in other ways as well, however. Upon request of a teacher or school administrator or parent, a student who was previously not an English language

7. Because I will investigate the responses of the schools that just received an F in 1999 versus those that just received a D in 1999, I will focus on the criteria for F and D grades. Detailed descriptions of the criteria for the other grades are available at http://schoolgrades.fldoe.org.

learner could be referred to an "ELL Committee."[8] The ELL Committee could determine a student to be an English language learner based on any two of the following criteria (1) extent and nature of prior educational or academic experience, social experience, and a student interview; (2) written recommendation and observation by current and previous instructional and supportive services staff; (3) level of mastery of basic competencies or skills in English and heritage language according to local, state, or national criterion-referenced standards; (4) grades from current or previous years; and (5) test results other than results from ELL eligibility assessments. Thus, there was considerable flexibility into classification into LEP and teachers and school administrators played an important role in this classification decision.

Classification into ESE also afforded considerable flexibility. A child starting in a regular education category could transition into a disability category in various ways. Accidents or sickness could lead to physical or mental disabilities that could warrant classification into some ESE categories. Even apart from that, there was another relevant way in which such classification could take place. As the curriculum grew more rigorous, a child could face a challenge that could be identified by a teacher, school administrator, or parent. In such a case, the child would be evaluated by a committee consisting of teachers, school administrators, parents, developmental specialists (who were often part of the school) and psychologists, and could be placed into ESE based on deliberations and recommendation of this committee. The basic takeaway from this discussion is that there was considerable flexibility in placement into ESE and LEP categories and the school (school administrators and teachers) could play a key role in such placements.

3. DATA

The data for this study were primarily obtained from the Florida Department of Education. I focus on elementary schools in this paper and the data include grade-level data on enrollment in LEP categories in each of the grades 2, 3, 4, and 5 for the years 1999–2002 as of February of the corresponding year (just before the tests were administered). These data report the number of students in an ESOL program for less than two years (excluded category) and the number of students in an ESOL program for two years or more (included category) in each of these grades in the years under consideration.

School-level data were also obtained on the distribution of students in the various ESE categories. In addition to information on total ESE enrollment,

8. The ELL Committee typically consists of teachers, school administrators (e.g., principal, assistant principal, guidance counselor), developmental specialist (often part of the school), and parents.

these data also report enrollment in each of the ESE categories in each Florida school for the years 1999–2002.

Data on socioeconomic characteristics include data on gender composition, race composition, and percent of students eligible for free or reduced-price lunches. School finance data consist of school-level per-pupil expenditures data and are available for the years under consideration.

In addition to this data, this study has benefited from data shared by Marcus Winters and Jay Greene.[9] These data include grade-level LEP enrollment in both excluded and included categories in each of the grades 4 and 5 for the years 2002–5, school-level ESE enrollment in each of the ESE categories for the years 2002–5, and McKay private school competition data. The latter include data on number of elementary McKay private schools within a 5-mile radius of each elementary public school in 2001.

I also supplement these data sets with private school location data for the 1997–98 school year. These data are obtained from the Private School Surveys (PSS) conducted by the National Center for Education Statistics, an arm of the U.S. Department of Education. The PSS have been conducted biennially since 1989–90 and I use private school location (address) data for 1997–98 to get pre-program distribution of private schools.[10] I geocode every elementary public and elementary private school in the state of Florida and compute the number of elementary private schools within 1-, 2-, and 5-mile radii of each public elementary school. These counts serve as measures of pre-program competition.

4. EMPIRICAL STRATEGY

Under the Florida A-plus program, schools that received a grade of F in 1999 directly faced stigma and threat of vouchers. I will refer to these schools as "F schools" from now on. The schools that received a D in 1999 were closest to the F schools in terms of grade, but were not directly threatened by the program. I will refer to them as "D schools" in the rest of the paper. Given the nature of the Florida program, the threat of vouchers faced by the 1999 F schools would be applicable for the next three years only. Therefore, I study the behavior of the F schools (relative to the D schools) during the first three years of the program (that is, up to 2002). I focus on elementary schools in this paper, because only a few middle and high schools received an F grade in 1999.

9. Many thanks are due to Marcus Winters and Jay Greene for graciously sharing part of their data with me that enabled some of the analysis in this paper.
10. I use data for 1997–98 because the surveys are done biennially and data are not available for the immediate pre-program year 1998–99.

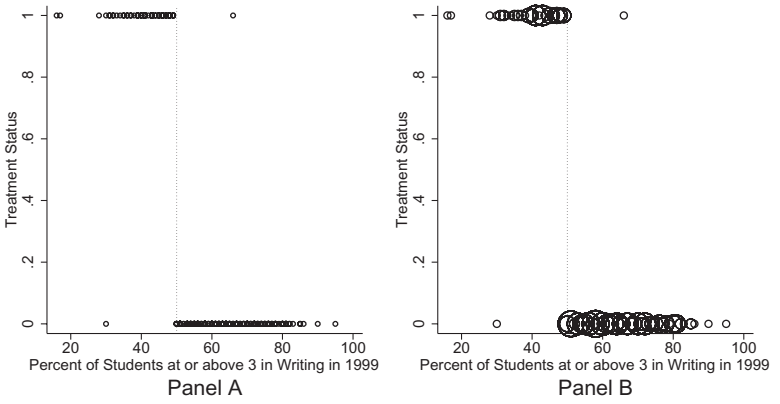

Figure 1. Regression Discontinuity Analysis: Relationship between Percent of Students at or above 3 in Writing and Treatment Status

I use a regression discontinuity (RD) analysis to analyze the effect of the program. The analysis essentially entails comparing the response of schools that barely missed D and received an F with schools that barely got a D. The institutional structure of the Florida program allows me to follow this strategy. The program created a highly nonlinear and discontinuous relationship between the percentage of students scoring above a predesignated threshold and the probability that the school's students would become eligible for vouchers in the near future, which enables the use of such a strategy.

Consider the sample of F and D schools that failed to meet the minimum criteria in both reading and math in 1999. In this sample, according to the Florida grading rules, only F schools would fail the minimum criteria in writing also, whereas D schools would pass it. Therefore, in this sample the probability of treatment would vary discontinuously as a function of the percentage of students scoring at or above 3 in 1999 FCAT writing (p_i). There would exist a sharp cutoff at 50 percent—whereas schools below 50 percent would face a direct threat, those above 50 percent would not face any such direct threat.

Using the sample of F and D schools that failed the minimum criteria in both reading and math in 1999, figure 1, panel A, illustrates the relationship between assignment to treatment (i.e., facing stigma and the threat of vouchers) and the schools' percentages of students scoring at or above 3 in FCAT writing. The figure shows that all but one of the schools in this sample that had less than 50 percent of their students scoring at or above 3 actually received an F grade. Similarly, all schools (except one) in this sample that had 50 percent or a larger percentage of their students scoring at or above 3 were assigned a D grade. Note that many of the dots correspond to more than one school. Figure 1, panel B, illustrates the same relationship where the sizes of the dots

are proportional to the number of schools at that point. The smallest dot in this figure corresponds to one school. These two panels show that in this sample, the percentage of students scoring at or above 3 in writing, indeed, uniquely predicts (except two schools) assignment to treatment and there is a discrete change in the probability of treatment at the 50 percent mark.[11]

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

Consider the following model, where Y_i is school i's outcome, F_i equals 1 if school i received an F grade in 1999 and $f(p_i)$ is a function representing other determinants of outcome Y_i expressed as a function of p_i:

$$Y_i = \alpha_0 + \alpha_1 F_i + f(p_i) + \epsilon_i \tag{1}$$

Hahn, Todd, and Van der Klaauw (2001) show that α_1 is identified by the difference in average outcomes of schools that just missed the cutoff and those that just made the cutoff, provided the conditional expectations of the other determinants of Y are smooth through the cutoff. Here, α_1 identifies the local average treatment effect at the cutoff.

The estimation can be done in multiple ways. In this paper, I use local linear regressions with a triangular kernel and a rule-of-thumb bandwidth suggested by Silverman (1986). I also allow for flexibility on both sides of the cutoff by including an interaction term between the running variable and a dummy indicating whether or not the school falls below the cutoff. I estimate alternate specifications that do not include controls as well as those that do use

11. I also consider two corresponding samples where both F and D schools failed the minimum criteria in reading and writing (math and writing). According to the Florida rules, F schools would fail the minimum criteria in math (reading) also, unlike D schools. I find that, indeed in these samples, the probability of treatment changes discontinuously as a function of the percentage of students scoring at or above level 2 in math (reading) and there is a sharp cutoff at 60 percent. The sizes of these samples are considerably smaller than the first one, however, and the samples just around the cutoff are considerably less dense. So I focus on the first sample, where 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. The results reported in this paper are from this sample. Note, though, that the results from the other two samples are qualitatively similar.

controls.[12,13] Assuming the covariates are balanced on both sides of the cutoff (I later test this restriction), the purpose of including covariates is variance reduction. They are not required for the consistency of α_1.

To test the robustness of the results, I also experiment with alternative bandwidths. The results remain qualitatively similar and are available on request. In addition, I also do a parametric estimation where I include a third-order polynomial in the percentage of students scoring at or above 3 in writing and interactions of the polynomial with a dummy indicating whether or not the school falls below the cutoff. I also estimate alternative functional forms that include fifth-order polynomial instead of a third-order polynomial and the corresponding interactions.[14] The results remain very similar in each case and are available on request.

Testing Validity of the Regression Discontinuity Analysis

Using the described local linear regression technique, I first investigate whether there is a discontinuity in the probability of receiving an F as a function of the assignment or running variable (percentage of students scoring at or above 3 in 1999 FCAT writing) in the sample reported in this paper. As could be perhaps anticipated from figure 1, I indeed find a sharp discontinuity at 50. The estimated discontinuity is 1 and it is highly statistically significant.

Next, I examine whether the use of an RD strategy is valid here. As discussed previously, identification of α_1 requires that the conditional expectations of various preprogram characteristics be smooth through the cutoff. Using the strategy outlined here, I test if that were indeed the case. I also test for any selection of schools around the cutoff. Note, though, that there is not much reason to expect strategic manipulation or selection in this particular situation. The program was announced in June 1999 and the tests were given a few months before, in January and February of 1999. Also, any form of strategic response with the objective of precise manipulation of test scores likely takes quite some time. It is unlikely the schools had the time or information to manipulate the percentage of students above certain cutoffs before the tests.

Nevertheless, I check for both continuity of predetermined characteristics and density of the running variable at the cutoff, using the strategy outlined earlier. The graphs corresponding to the test of continuity of predetermined

12. Unless otherwise noted, 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 expenditure.
13. As is customary in the literature, I cluster these standard errors by the running variable to account for common components of variance that can be induced if the functional form of the estimated conditional expectations function deviates from the actual.
14. I use odd order polynomials because they have better efficiency (Fan and Gijbels 1996) and are not subject to boundary bias problems, unlike even order polynomials.

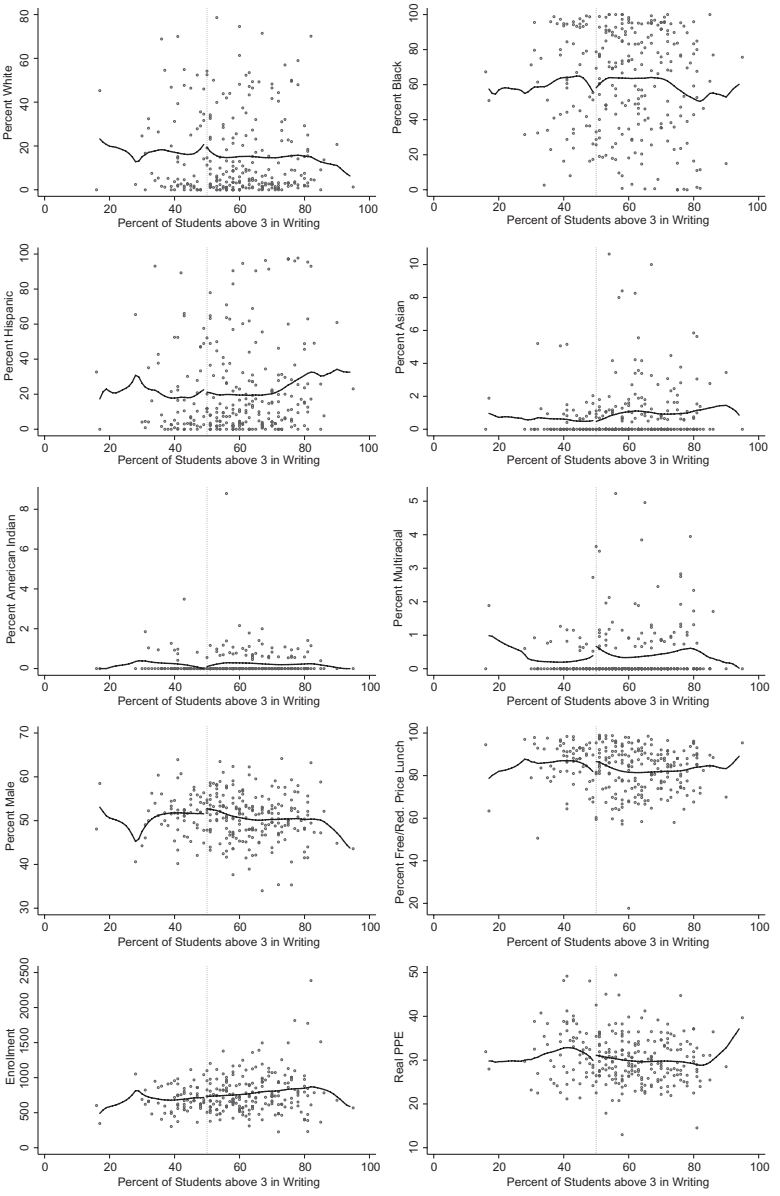

Figure 2A. Testing Validity of Regression Discontinuity Design: Pre-Program Characteristics Relative to the Cutoff

characteristics are presented in figures 2A and 2B and the discontinuity estimates in table 1. Figure 2A considers preprogram (1999) demographic and socioeconomic characteristics, and figure 2B considers classification in excluded and included LEP and ESE categories in the pre-program (1999) period. The discontinuity estimates are never statistically distinguishable from zero.

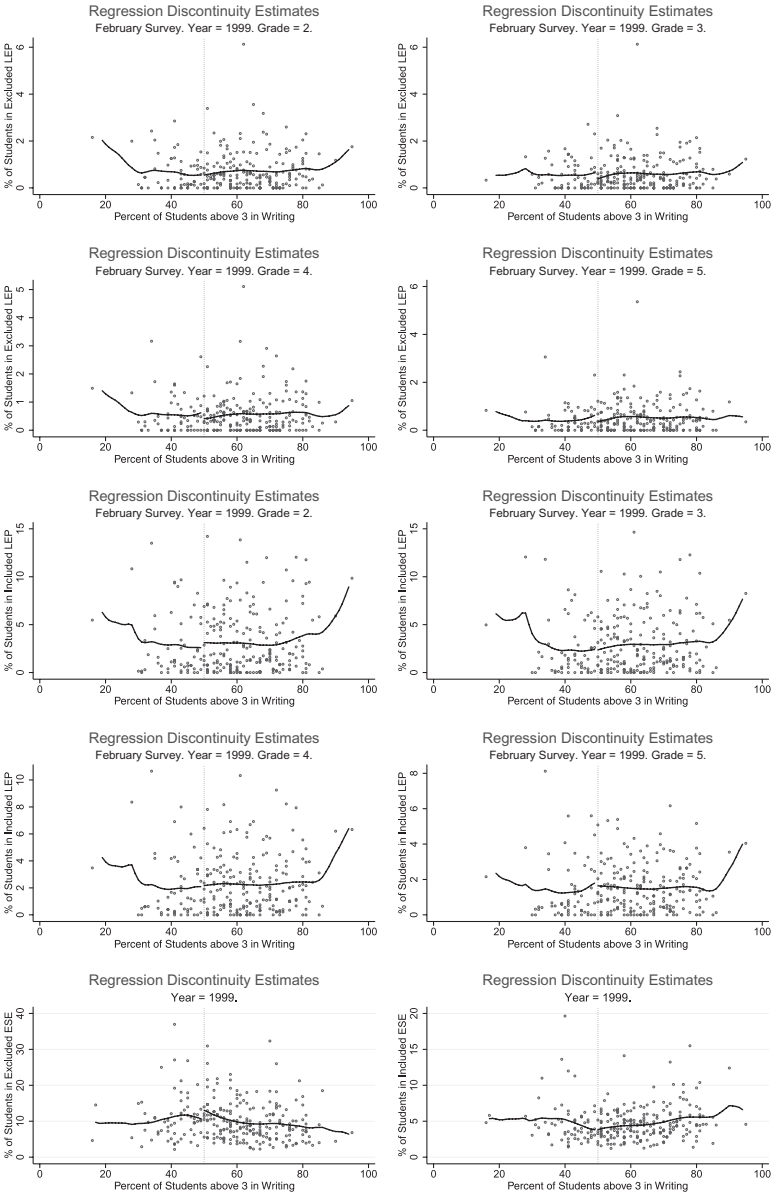

Figure 2B. Testing Validity of Regression Discontinuity Design: Examining Classification in Excluded and Included LEP and ESE Categories Relative to Cutoff in Pre-Program Period

Visually examining the graphs, it seems that, unlike in the cases of the other predetermined characteristics, there is a small discontinuity in the variable "percentage of school students eligible for free or reduced-price lunches." But the discontinuity is small and not statistically significant (with a *p*-value of .28). Also, note that even if it was statistically significant, with a large number of

Table 1. Testing Validity of Regression Discontinuity Analysis: Looking for Discontinuities in Pre-Program Characteristics at the Cutoff

Panel A	**% White (1)**	**% Black (2)**	**% Hispanic (3)**	**% Asian (4)**	**% American Indian (5)**
	2.92	−5.06	2.43	0.09	−0.16
	(7.24)	(11.39)	(6.73)	(0.28)	(0.06)
Panel B	**% Multiracial (6)**	**% Male (7)**	**% Free/Reduced Price Lunch (8)**	**Enrollment (9)**	**Real PPE (10)**
	−0.23	−1.21	−5.97	−14.45	−1.97
	(0.26)	(1.44)	(5.36)	(60.32)	(2.29)
Panel C	**% ESE (11)**	**% Excluded ESE (12)**	**% Included ESE (13)**	**% Learning Disabled (14)**	**% Emotionally Handicapped (15)**
	−2.918	−2.891	−0.026	0.052	−0.633
	(1.874)	(1.827)	(0.779)	(0.795)	(0.563)
Panel D	**% Excluded LEP**				
	Grade 2 (16)	**Grade 3 (17)**	**Grade 4 (18)**	**Grade 5 (19)**	
	0.027	0.304	0.244	0.299	
	(0.183)	(0.199)	(0.222)	(0.182)	
Panel E	**% Included LEP**				
	Grade 2 (20)	**Grade 3 (21)**	**Grade 4 (22)**	**Grade 5 (23)**	
	−0.544	0.057	−0.086	0.260	
	(0.510)	(0.557)	(0.280)	(0.410)	

Note: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses.
*$p < .10$, **$p < .05$, ***$p < .01$.

comparisons one might expect a few to be statistically different from zero by sheer random variation. So, from this discussion, it seems reasonable to say that this case passes the test of smoothness of predetermined characteristics through the cutoff.

Following McCrary (2008), I next test whether there is unusual bunching at the cutoff. Using the density of the running variable (percentage of students at or above 3 in writing in 1999) and the given strategy, I test for a discontinuity in the density of the running variable at the cutoff. As can be seen from table 2, there is no evidence of a statistically significant discontinuity in the density function at the cutoff in 1999.

Table 2. Testing Validity of Regression Discontinuity Analysis: Looking for Discontinuities in the Density of the Running Variable

	1999
Difference	−0.01
	(0.01)

Note: Standard errors are in parentheses and are clustered by the running variable (% of school's students at or above the writing cutoff).
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

5. RESULTS

Appendix table A.1 presents summary statistics for the discontinuity sample of F and D schools that fell within the Silverman bandwidth, pooled for the years under consideration (1999–2002).[15] Panel A shows the racial composition of students in these schools—about 64 percent were black, followed by Hispanic at 21 percent and white at 14 percent. These schools only served a small number of Asian and American Indian students. Male students constituted a slight majority, and most students came from low-income families, being eligible for free or reduced-price lunches (panel B). The average school had an enrollment of 713 students. Panel C shows that about 16 percent of the students were ESE students. The large majority of them, about three-quarters, were in excluded ESE categories, and the rest were in included ESE categories. A little over 4 percent of the students were classified as Learning Disabled (LD), while about 1 percent was classified as Emotionally Handicapped (EH). Panels D and E report the percentages of excluded LEP and included LEP, respectively, across grades 2–5. Pooling grades 2–5 together, it can be seen that the excluded LEP students in these grades constituted about 2.6 percent of the average school's enrollment, and included LEP students constituted 9.6 percent of its enrollment.

Having established that the use of an RD strategy in this setting is valid, I next look at the effect of the program on the behavior of threatened schools. For reference, let's first look at the behavior of these same schools in the

15. Two of the 1999 F schools became eligible for vouchers in 1999. They were in the state's "critically low-performing schools" list in 1998 and were grandfathered into the program. Consistent with the previous literature (Chiang 2009; West and Peterson 2006), I exclude them from the analysis because they faced different incentives. Note, though, that results do not change if they are included in the analysis. One of these F schools falls outside the bandwidth and hence does not affect estimation. The other one falls within the bandwidth, but it falls very close to the left end of the bandwidth and hence only gets a relatively small weight in the estimation. Its inclusion or exclusion does not affect results. None of the other F schools received a second F in either 2000 or 2001. Four schools received an F in 2000 and all of them were D schools. No other D school received an F either in 2000 or 2001.

pre-program period. Figure 2B and table 1 (panels C–E) look at the LEP and ESE classification in excluded and included categories in 1999, the year just before the program. There is no evidence that the schools that would be threatened the next year behaved any differently from the nonthreatened schools in excluded or included LEP classification in any of the high-stakes or low-stakes grades. Nor is there any statistically significant evidence of any differential classification in excluded or included ESE categorization in 1999.[16] The picture in the post-program period is very different, as seen subsequently.

Table 3 looks at the effect of the program on percentage of students in excluded (columns 1–3) and included (columns 4–6) LEP categories in various grades. These variables are defined as enrollment in excluded or included LEP categories in various grades as a percentage of total school enrollment.

First, consider the excluded category. In the first year after the program, the table finds that the program led to a statistically significant increase in the percentage of students classified in excluded LEP categories in the high-stakes grade 4 and the entry grade 3. In contrast, there is no evidence of an increase in the low-stakes grade 2 or the high-stakes grade 5. The estimates suggest that in the first year after the program, F schools classified an additional 0.31 percent of their total students in the excluded LEP category in grade 4 and an additional 0.36 percent of their students in grade 3. Because it might have been difficult to do the classification all at once, the administrators might have chosen to phase out the process to the entry grade 3. These figures are equivalent to an additional classification of 53 percent of their excluded LEP students in grade 4 and an additional classification of 55 percent of their excluded LEP students in grade 3. In terms of numbers of students, this is equivalent to classification of an additional 2.3 students in grade 3 and 2.6 students in grade 4. In the second year after the program (column 2), there is evidence of positive and statistically significant shifts in the excluded LEP category in grades 4 and 5 in the threatened schools. Compared with the effects in the first year, it seems that the increase in grade 5 (grade 4) in the second year was generated by the increased classification in grade 4 (grade 3) in the first year after the program. There does not seem to have been any new classification in the second year after the program. Similarly, there is no evidence of any new classification in the third year after the program (column 3).

Columns 4–6 present the effects of the program on the percentage of students in the included LEP category. There is no evidence that the program led to differential classification in any of the three years after the program.

16. Note that while the 1999 ESE estimates are not statistically significant, the magnitudes of some of the estimates are not small. So in the ESE analysis that follows, I include the lagged dependent variable as an additional covariate in addition to the usual set of covariates used in this paper (see footnote 12). I discuss this in more detail toward the end of this section.

Table 3. Effect of the Program on Classification into Excluded and Included LEP Categories: A Regression Discontinuity Analysis

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	2 Years After (2)	3 Years After (3)	1 Year After (4)	2 Years After (5)	3 Years After (6)
Grade 2	0.292	0.170	0.024	0.115	0.477	−0.594
	(0.228)	(0.268)	(0.232)	(0.300)	(0.685)	(0.534)
Observations	123	124	120	123	124	120
R^2	0.532	0.573	0.537	0.664	0.652	0.735
Grade 3	0.362*	0.295	0.261	−0.422	−0.212	0.811
	(0.181)	(0.278)	(0.264)	(0.484)	(0.436)	(0.835)
Observations	121	122	124	121	122	124
R^2	0.540	0.483	0.579	0.568	0.603	0.605
Grade 4	0.314**	0.365**	0.144	0.039	−0.220	0.532
	(0.118)	(0.218)	(0.295)	(0.310)	(0.360)	(0.399)
Observations	119	124	121	119	124	121
R^2	0.403	0.523	0.398	0.531	0.435	0.546
Grade 5	0.270	0.317***	0.449	0.011	0.065	0.276
	(0.253)	(0.097)	(0.380)	(0.391)	(0.455)	(0.259)
Observations	116	117	122	116	117	122
R^2	0.430	0.544	0.444	0.325	0.365	0.397

Notes: Robust standard errors adjusted for clustering by the running variable are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and real per pupil expenditure.
*$p < .10$, **$p < .05$, ***$p < .01$.

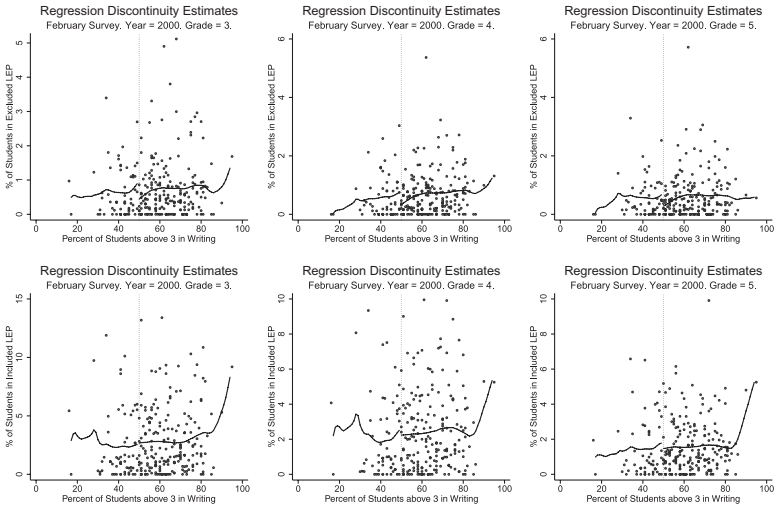

Figure 3A. Examining the Effects on Classification in Excluded and Included LEP Categories, 2000

The fact that there is no evidence of any additional classification in included LEP categories, unlike that in the excluded LEP categories, is informative. Recall that the included LEP category consists of students who are in an ESOL program for two years or more—LEP students move from excluded to included categories after two years. The absence of increased memberships in the included categories suggests that increased classifications did not take place in the excluded categories in the earlier low-stakes grades in the pre-program as well as post-program years. The absence of any additional classification in the included categories is comforting and adds more confidence that the increased classifications in the excluded categories indeed indicate strategic behavior.

Figures 3A and 3B display the effects of the program on classification in excluded and included LEP categories graphically. Although the estimates presented in the table include controls, the graphs display results of estimations without controls. As can be seen, the patterns are similar and do not depend on inclusion of controls.

These results can be summarized as follows. In the pre-program period, there is no evidence that the would-be threatened schools behaved any differently than the would-be nonthreatened schools in terms of categorization of students in excluded or included LEP categories in any of the high-stakes or low-stakes grades. In contrast, the program led to increased classification of students into the excluded LEP category in the high-stakes grade 4 and the entry grade 3 in the first year after the program. There is no evidence of any new classification in this category either in the second or third years after the program. Nor is there any evidence of differential classification in the included category in any of the three years after the program. Students classified into the

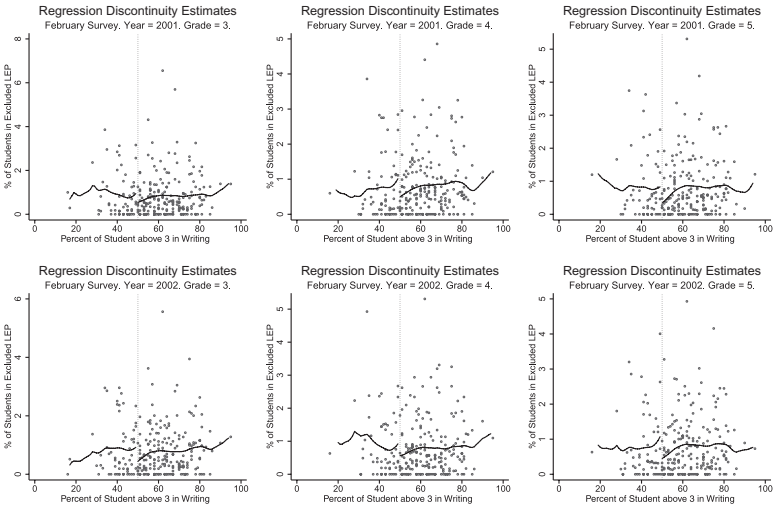

Figure 3B. Examining the Effects on Classification in Excluded LEP Categories, 2001 and 2002

excluded LEP category in grade 4 in the first year after the program would not count in school grades either in the current year or in the following year (that is, in both high-stakes grades 4 and 5). Students classified into the excluded LEP category in grade 3 would not count the following year when they would be in the high-stakes grade 4. So these findings suggest that the threatened schools attempted to remove certain students from the effective test-taking pool, both in the current year and in the following year, by classifying them into the excluded LEP category.

Table 4 looks at the effect of the program on ESE classification. Panel A looks at the effect on total ESE classification. The dependent variable for this analysis is percentage ESE enrollment—that is, total ESE enrollment as a percentage of total enrollment. The estimates show that there is no evidence in favor of any differential classification in the threatened schools at the cutoff.

Although trends in total ESE classification provide a summary picture, they are unlikely to provide a conclusive picture in terms of whether the F schools resorted to such classification of students. For example, the absence of shifts in total ESE classification does not rule out the possibility that relative classification in excluded categories took place in the F schools.

To have a closer look, table 4, panels B and C, look at the effect of the program on classification in excluded (panel B) and included (panel C) ESE categories. The dependent variable here is percentage of total enrollment classified in excluded (panel B) and included (panel C) categories. The estimates show no evidence that the threatened schools resorted to relative classification into excluded categories in any of the three years after the program. Nor is there any evidence of differential classification in the included categories.

Table 4. Effect of the Program on ESE Classifications: A Regression Discontinuity Analysis

Panel A	**% of Students in ESE Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	0.437	−1.142	−0.648
	(0.403)	(0.830)	(0.873)
R^2	0.920	0.838	0.733
Panel B	**% in Excluded ESE Categories**		
	1 Year After Program (4)	**2 Years After Program (5)**	**3 Years After Program (6)**
	0.699	−0.433	−0.086
	(0.565)	(1.062)	(1.060)
R^2	0.920	0.852	0.740
Panel C	**% in Included ESE Categories**		
	1 Year After Program (7)	**2 Years After Program (8)**	**3 Years After Program (9)**
	−0.236	−0.633	−0.493
	(0.286)	(0.413)	(0.379)
R^2	0.837	0.706	0.570
Panel D	**% in Learning Disabled Category**		
	1 Year After Program (10)	**2 Years After Program (11)**	**3 Years After Program (12)**
	−0.178	−0.445	0.347
	(0.261)	(0.478)	(0.407)
R^2	0.801	0.727	0.643
Panel E	**% in Emotionally Handicapped Category**		
	1 Year After Program (13)	**2 Years After Program (14)**	**3 Years After Program (15)**
	0.083	−0.138	0.038
	(0.158)	(0.179)	(0.235)
R^2	0.925	0.886	0.788
Observations	130	132	132

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, real per pupil expenditure, and pre-program (1999) percentage of students in All ESE Categories (panel A), Excluded (panel B), Included (panel C), Learning Disabled (panel D), or Emotionally Handicapped (panel E) category.

$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

The ESE categories vary in the extents of their severities. Whereas some categories, such as those with observable or severe disabilities or physical handicaps, are comparatively nonmutable, others, such as learning disabled and emotionally handicapped, are much milder and comparatively mutable categories.[17] Classification in these latter categories often has a large amount of subjective element to it and hence could be easily manipulated. This analysis does not find much evidence in favor of relative classification into excluded categories in F schools. This does not rule out the possibility that this kind of behavior took place in the F schools, however. Increased classification may have taken place in some specific categories that are more mutable and hence more amenable to manipulation, and consideration of all excluded categories together masks this kind of behavior. If such classification did take place, it is most likely to have taken place in such mutable categories.

Table 4, panels D and E, investigate the effect of the program on relative classification in mutable excluded categories—learning disabled (panel D) and emotionally handicapped (panel E). There is no evidence the threatened schools tended to differentially classify students into either learning disabled or emotionally handicapped categories.[18]

Figure 4, panels A–D, looks at the effect of the program on classification in total excluded, included, emotionally handicapped, and learning disabled categories, respectively. As earlier, the graphs display results from RD estimations that do not include controls whereas those in the tables include controls. The graphical patterns in figure 4 mirror closely the results obtained in table 4. The discontinuities are either small or indistinguishable from zero and they are never statistically significant. Thus, to summarize, I find no evidence that the treated schools resorted to strategic classification into excluded ESE categories.

To summarize, the program led the F schools to relatively over-classify students in the excluded LEP category in the high-stakes grade 4 and the entry grade to the high-stakes grades, grade 3. In contrast, there is no evidence of any differential classification in included LEP categories. Nor is there any evidence of relative classification in either included ESE or included LEP. These patterns

17. See Cullen (2003), Singer et al. (1989), and Figlio and Getzler (2006).
18. One point to note here is that the number of observations differs somewhat between the ESE analysis and the LEP analysis. The number of observations for the LEP analysis varies between 116 and 124 (table 3), whereas that for ESE analysis varies between 130 and 132. This is because the former is a grade-level analysis and the latter is a school-level analysis, and the grade distributions vary across schools. Whereas school-level analysis includes all elementary schools within the bandwidth, not all schools have all grades between grades 2–5. Correspondingly, the school-level analysis has a slightly larger number of observations than the grade-level analysis. Also of note here is that, consistent with this explanation, the number of observations in the aggregated "school-level" LEP analysis (section 6) where I pool grades 2–5 has 129–132 observations, more similar to the school-level ESE analysis.

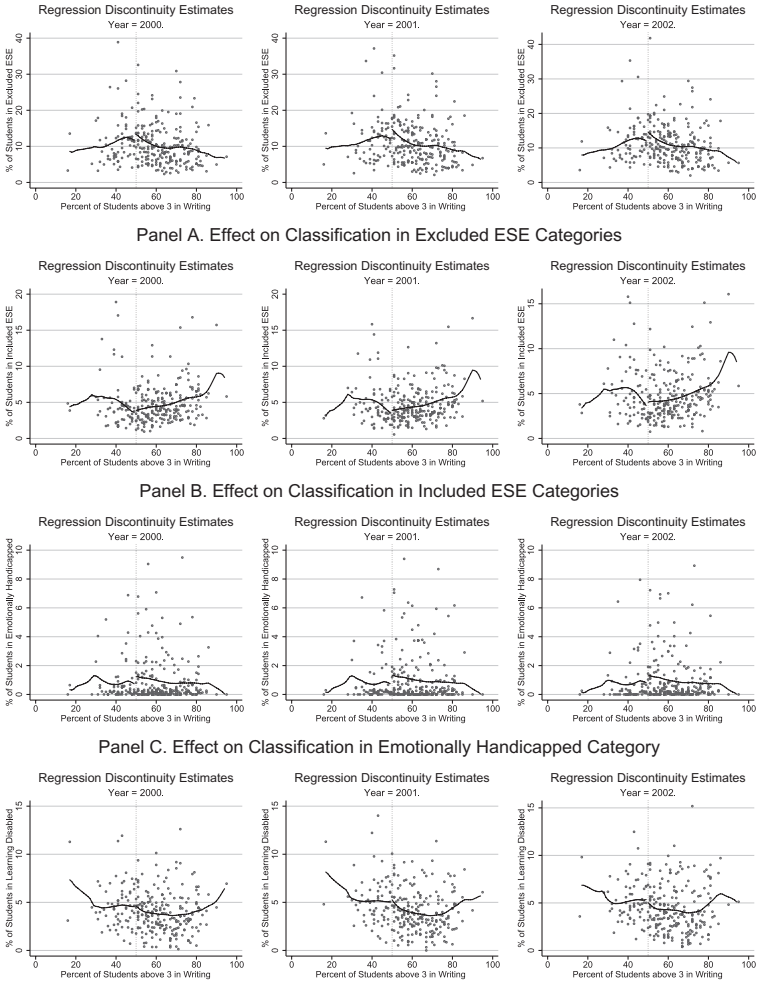

Figure 4. Examining the Effect of the Program on Classification in Special Education (ESE) Categories

suggest that the different incentives created by the interplay of the A-plus and McKay rules encouraged the F schools to respond very differently along the ESE and LEP margins. Whereas the impending threat of vouchers and stigma increased the attractiveness and benefit of strategic classifications into excluded ESE and LEP categories, categorization into ESE was associated with a direct cost, unlike categorization into LEP. Classification into ESE exposed the schools to the threat of loss of those ESE students (and the corresponding revenue) to McKay vouchers. This discouraged classification into ESE. There was no such counterincentive for LEP classification, however, encouraging strategic classification into excluded LEP categories.

Some points are worth noting here before moving on to the next section. The first relates to the set of covariates used in the regressions. As noted in footnote 12, the set of covariates generally used in this study include racial composition of schools, gender composition of schools, percentage of students eligible for free or reduced-price lunches, and real per-pupil expenditures. The RD estimates for LEP reported in table 3 are obtained from regressions that control for this set of covariates. On the other hand, the results for ESE reported in table 4 are obtained from regressions that include the pre-program value of the dependent variable in addition to these covariates. The decision to include the latter follows from the pre-program patterns seen in table 1. Although there is no evidence of any statistically significant discontinuity in the pre-program ESE variables at the cutoffs (table 1, panel C), magnitudes of the estimates in some cases are not small. Consequently, I control for the pre-program value of the dependent variable in the ESE analysis.

It is important to note here that the differences in post-program patterns seen earlier between LEP and ESE classification cannot be attributed to this difference in covariates. In appendix table A.2, I present estimates for LEP where I control for the one-year lagged value of the dependent variable in addition to the usual set of covariates. As can be seen, the results are qualitatively similar to those in table 3, which also speaks to the robustness of the estimates.

Second, one of the control variables—real per-pupil expenditure—deserves some special attention. One might argue that this variable is potentially endogenous as ESE and LEP counts determine school funding. Although ESE and LEP counts do determine school funding, it is the previous year's count and not the current year's count that determines school funding. In contrast, both the dependent variable (percentage count variable) and the real per-pupil expenditure covariate relate to the current year, and hence inclusion of the latter is likely not a problem. Nevertheless, to test for robustness of the estimates to inclusion of real per-pupil expenditure, I estimate RD specifications that exclude real per-pupil expenditure as a control variable. The corresponding estimates for LEP are reported in table A.3 and those for ESE are reported in table A.4. Once again, the estimates remain qualitatively similar to those reported in tables 3 and 4, so inclusion of real per-pupil expenditure is not driving results.

Third, it should be noted here that, as in any other RD analysis, effects obtained in this study are local, average treatment effects. As a result, the effects obtained are local to the cutoff and could be underestimates of the treatment effect. Whereas D schools did not directly face the threat of vouchers or stigma (associated with the lowest-performing grade), they were close to getting an F and hence likely faced an indirect threat. In fact, there was a 5 percent

probability that a D school might receive an F grade in the next year.[19] In such a case, the program effects shown here could be underestimates. But the extent of underestimation is not expected to be large as the probability of treatment (receiving an F) of the D schools was not large.

6. ROBUSTNESS CHECKS

Compositional Changes of Schools and Sorting

If there is differential student sorting or compositional changes in the treated schools, then the effects we have seen can be in part driven by those changes. None of the threatened schools received a second F grade in 2000 or 2001, and therefore none of their students became eligible for vouchers. Thus, the concern about vouchers leading to sorting is not applicable here. A valid question here though is whether the McKay program led to sorting of ESE students that affected F schools differently. But any such differential sorting will be reflected in impacts on ESE analyzed previously. The absence of impacts on any of the ESE categories analyzed in this paper—total ESE, excluded ESE, included ESE, mutable categories (LD and EH)—indicates that the relative sorting of ESE students was not a driving factor. For the sake of completeness, I also investigate whether the program generated shifts in immutable categories in F schools. I find no evidence of such differential shifts. The results are not reported here for lack of space, but are available on request.

Note that just the grades themselves (F and D) could lead to a differential sorting of students in these two types of schools.[20] To investigate this issue further, I examine whether the demographic composition of the treated schools saw a relative shift after the program. I use the same RD strategy outlined previously, but the dependent variables are now various socioeconomic variables.

The results of this analysis are presented in table 5.[21] As can be seen, there is no evidence of any differential shift in the treated schools in any of the characteristics in any of the three years after program, except for percent

19. Recall that 1999 was the first year when Florida graded its schools on a scale of A–F. But using the 1999 state grading criteria and the percentages of students scoring below the minimum criteria in the three subjects (reading, math, and writing) in 1998, I was able to assign F and D grades in 1998. Five percent of these 1998 D schools received an F in 1999.

20. Figlio and Lucas (2004) find that following the first assignment of school grades in Florida, the better students differentially selected into schools receiving grades of A, though this differential sorting tapered off over time.

21. All estimates reported in table 5 are obtained from RD specifications that control for racial composition of schools (percentage in racial groups other than that represented by the dependent variable and the omitted group), gender composition of schools, percentage of students eligible for free or reduced-price lunches, and real per-pupil expenditure to put them on an equal footing with the estimates in the rest of the paper. Percent white is treated as the omitted category for demographic composition covariate set except when the dependent variable is percent white. Percent Hispanic is treated as the omitted category in this case.

Table 5. Are Compositional Changes or Sorting Driving Results? Investigating Demographic Shifts Using a Regression Discontinuity Analysis

	1 Year After Program (1)	2 Years After Program (2)	3 Years After Program (3)
% White	−0.460	−0.325	−4.321
	(2.130)	(2.860)	(2.784)
R^2	0.612	0.642	0.535
% Black	−0.517	−0.699	2.873
	(2.974)	(3.598)	(3.868)
R^2	0.845	0.849	0.806
% Hispanic	0.460	0.325	4.321
	(2.130)	(2.860)	(2.784)
R^2	0.806	0.819	0.782
% Asian	0.409	0.450*	0.322
	(0.294)	(0.238)	(0.212)
R^2	0.295	0.302	0.319
% American Indian	−0.144	−0.152	−0.136
	(0.093)	(0.153)	(0.118)
R^2	0.167	0.100	0.085
% Multiracial	0.000	−0.000	0.000
	(0.000)	(0.000)	(0.000)
R^2	0.999	0.999	0.999
% Male	−0.577	1.013	0.557
	(0.683)	(0.730)	(0.801)
R^2	0.095	0.084	0.168
% Free/Reduced Price Lunch	0.528	−1.493	0.782
	(2.289)	(1.855)	(2.332)
R^2	0.547	0.635	0.461
Total Enrollment	−15.697	−19.473	−21.334
	(65.387)	(57.167)	(57.942)
R^2	0.256	0.325	0.411
Real PPE	0.316	2.960	−1.735
	(1.463)	(1.931)	(1.077)
R^2	0.231	0.431	0.420
Observations	130	129	128

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch (except when dependent variable is % free or reduced price lunch), and real per pupil expenditure (except when dependent variable is real per pupil expenditure). See footnote 20 for details.
*$p < .10$, **$p < .05$, ***$p < .01$.

Table 6A. Testing Validity of Regression Discontinuity Analysis: Looking for Discontinuities in Pre-Program LEP Classification using Aggregated Data

	% Excluded LEP (Grades 2–5)	% Included LEP (Grades 2–5)
	(1)	(2)
	0.565	−0.052
	(0.531)	(1.753)
Observations	132	132
R^2	0.508	0.595

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. Controls include racial composition, gender composition, percent of students eligible for free or reduced price lunch, and real per pupil expenditure.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Asian in the second year after the program. So from this analysis, it seems safe to conclude that the results obtained previously are not driven by differential changes in composition of schools or student sorting.

Are Differences in Levels of Aggregation Driving Results?

Recall that whereas the LEP data are available and analyzed at the grade level, ESE data are available only at the school level, leading to a corresponding school-level analysis for ESE. One might argue that the differences in the levels of aggregation are driving the differences in the ESE and LEP patterns and doubt whether the LEP patterns will survive similar aggregation of the data.

As mentioned earlier, I focus on elementary schools in this study. An overwhelming 80 percent of the elementary schools were either PK–5 or K–5, 18 percent of the schools were PK–6 or K–6, and the remaining very small proportion of schools were either PK–3, PK–4, 1–5, 3–5, or 4–5.

One thing to note here is that whereas the ESE analysis is based on these elementary schools, the LEP analysis includes data on most of the key elementary grades. To assess the role of aggregation in generating the given patterns, I aggregate the LEP data for the available grades 2–5 and look for any discontinuity in LEP classification using this aggregated data. To set the stage, table 6A looks at the aggregate LEP patterns in the pre-program period. There is no evidence of any discontinuity in either percent excluded LEP or percent included LEP students at the cutoff in the pre-program period. In contrast, table 6B looks at the patterns in the post-program period using aggregated data. Consistent with the previous grade-level results for LEP (table 3), there is evidence once again of increased classification in the first year after the program. In response to the program, in grades 2–4 taken together, the F schools classified an additional 1.2 percent of their students into excluded LEP in the first year

Table 6B. Effect of the Program on Classification into Excluded and Included LEP Categories: A Regression Discontinuity Analysis Using Aggregated Data

Panel A	**% in Excluded LEP Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	1.198**	1.089*	0.841
	(0.612)	(0.627)	(1.080)
Observations	129	130	129
R^2	0.574	0.607	0.576
Panel B	**% in Included LEP Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	−0.339	0.144	1.044
	(1.006)	(1.496)	(1.674)
Observations	129	130	129
R^2	0.601	0.595	0.659

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. Controls include racial composition, gender composition, percent of students eligible for free or reduced price lunch, and real per pupil expenditure. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

after the program. This figure is equivalent to 52 percent of their excluded LEP students in these grades. There is also evidence of a positive shift in the second year after the program. But comparing the magnitude of this effect with that in the first year indicates this shift is likely driven by additional classification in the first year after the program. Thus, although there is evidence of classification in the first year after the program, there is no evidence of any added classification in the later years. To summarize, the results obtained from this aggregate LEP analysis are qualitatively similar to those obtained from the grade-level analysis, and they continue to show evidence of increased classifications into excluded LEP. In other words, differences in levels of aggregation are not driving the differences in results between LEP and ESE.

Are the LEP Effects Statistically Different from the ESE Effects?

Because, based on data availability, the unit of analysis is different between ESE and LEP (LEP analysis uses grade-level data, whereas ESE analysis uses school-level data), I have used separate RD analysis to examine the effects of the program on ESE and LEP classifications (see section 5). A natural question to ask, however, is whether the LEP effects statistically differ from the ESE effects. To address this question, I compare the ESE effects with the LEP effects

obtained from the aggregated data analysis (to bring them to a comparatively equal footing) statistically.

For this purpose, using school-level aggregated data, I integrate the ESE and LEP estimations in a single model and conduct an RD difference-in-differences analysis. I estimate the following specification.

$$Y_i = \beta_0 + \beta_1 F_i + \beta_2 LEP + \beta_3 (F_i * LEP) + f(p_i) + \epsilon_i \quad (2)$$

where *LEP* is a dummy variable that takes a value of 1 for *LEP* and 0 for *ESE*, $Y = \{$percentage of students in excluded categories, percentage of students in included categories$\}$. I continue to use local linear regressions with a triangular kernel, flexible functional forms on both sides of the cutoff, and the Silverman bandwidth for the RD estimation. The interpretations of the coefficients are as follows. Any differential classification made by the F schools in ESE would be captured by β_1; β_2 captures any differential classification in LEP relative to ESE that is common to both F and D schools; β_3 captures any differential classification in LEP in F schools (relative to D schools) in comparison with any differential classification in ESE in F schools (relative to D schools). In other words, β_3 indicates if the F-school LEP effects (relative to D schools) are statistically (and economically) different from the corresponding ESE effects in F schools (relative to D schools).

The results of this analysis are presented in table 7. Panel A presents results for excluded categories and panel B for included categories. Let's focus on panel A first. As expected, the first row shows no evidence of any differential classification into excluded ESE categories in F schools relative to D schools in any of the years. The second row (coefficient of LEP) is also expected—an artifact of the definition of the excluded LEP and ESE categories. Although excluded LEP category only includes LEP students who are in an ESOL program for less than two years, excluded ESE categories include the eighteen categories outlined in section 2, and is considerably larger in size. This can be seen from summary statistics table A.1. Percent excluded ESE exceeds the pooled percent excluded LEP in grades 2–5 by 9.035 percentage points, which essentially is reflected in this coefficient (second row). The interaction coefficient (third row) is the key coefficient of interest. It shows that the F-school LEP effects were indeed *statistically and economically* larger than the F-school ESE effects.

In contrast, the picture in panel B is different. There is no evidence of any differential classification in included ESE in F versus D schools (first row), nor in included LEP in F schools (relative to D schools) in comparison with included ESE in F schools (relative to D schools) as seen in the third row. The positive significant coefficients of LEP (second row) are again artifacts of the construction of the ESE and LEP groups. Included ESE consisted of

Table 7. Directly Comparing Program Effects in LEP and ESE: Are There Statistical Differences? A Regression Discontinuity Difference-in-Differences Analysis

Panel A	**% of Students in Excluded Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
F	−0.485	−0.724	−0.653
	(1.471)	(1.141)	(1.123)
LEP	−9.866***	−10.741***	−10.571***
	(1.166)	(1.143)	(0.880)
F * LEP	1.280*	1.296*	1.208
	(0.690)	(0.777)	(1.705)
Observations	240	244	242
R^2	0.567	0.552	0.568
Panel B	**% of Students in Included Categories**		
F	0.189	0.012	0.428
	(1.076)	(1.078)	(1.306)
LEP	5.018***	4.884***	6.068***
	(0.600)	(0.721)	(0.907)
F * LEP	−0.614	−0.042	0.108
	(1.457)	(1.791)	(2.157)
Observations	240	244	242
R^2	0.368	0.365	0.394

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and real per pupil expenditure.
*$p < .10$, **$p < .05$, ***$p < .01$.

only three groups (learning disabled, hospital/homebound, gifted), whereas included LEP constitutes the bulk of the LEP students (who are in an ESOL program for two years or more) and is larger in size. This difference in sizes of the included LEP and ESE groups (percent included LEP in grades 2–5 versus percent included ESE) can also be seen from appendix table A.1.

7. ASSESSING THE ROLE OF MCKAY COMPETITION: DIFFERENTIATING BETWEEN SCHOOLS FACING DIFFERENT LEVELS OF COMPETITION

In this section, I assess the role of McKay voucher competition. Specifically, I differentiate between schools facing different extents of McKay competition and investigate whether there were differences between ESE and LEP classification patterns in schools facing more versus less McKay competition.

I use two measures of McKay competition. First, I start with a measure that gives the number of McKay-accepting elementary private schools within a 5-mile radius of each elementary public school in 2001. Although the advantage of this measure is that it exploits the count of private schools that actually made themselves available for McKay vouchers, this metric has an important disadvantage. Because it exploits the post-program distribution of schools, and private school decision to opt in is likely endogenous to the A-plus program (example F/D grades), this count measure likely suffers from an endogeneity problem.

The ideal metric would be to use the distribution of McKay private schools in the pre-program period. But because there was no McKay program during this period, it is impossible to get this metric. The correlation between the distribution of elementary McKay private schools and elementary private schools, however, is very high, 0.895.[22] This implies that the number of elementary private schools in the near vicinity of a private school is a good proxy of McKay voucher competition. Exploiting this fact, I use a second set of measures of competition (*count*)—the number of elementary private schools within 1-, 2-, and 5-mile radii of each elementary public school in the pre-program period (1998). The latter is my preferred measure of McKay private competition (because it allows me to get around the endogeneity problem). Results reported in the paper pertain to this count. The results corresponding to the 2001 count are qualitatively similar, however, and available on request. I estimate the following specification using the RD technique outlined in section 4:

$$Y_i = \gamma_0 + \gamma_1 F_i + \gamma_2 count + \gamma_3 (F_i * count) + f(p_i) + \epsilon_i \quad (3)$$

The coefficient γ_1 captures any differential classification made by F schools (relative to D schools); γ_2 captures the common effect of McKay competition on F and D schools; and γ_3 captures any additional effect of McKay competition on F schools (relative to D schools).

Tables 8A–8B present the results for ESE classification. Table 8A looks at the impact on total ESE classification (panel A), classification in excluded ESE (panel B), and included ESE classification (panel C). Table 8B looks at the impact on classification in mutable excluded categories: learning disabled (panel A) and emotionally handicapped (panel B). Consistent with the patterns obtained in section 5, there is no evidence of any increased classification in F schools relative to D schools in any of the ESE categories (first row of each

22. Specifically, correlation between the two counts within a 5-mile radius in 2001 is 0.895. The 2001 count obtained from Marcus Winters and Jay Greene relate to a 5-mile radius. Consequently, the correlation relates to this distance. The count measures I use for 1998 relate to 1-, 2-, and 5-mile radii. Only the results for 5 miles are reported in this paper to save space. Results for the other radii are available on request.

Table 8A. Did McKay Voucher Competition Affect Classification into Special Education Categories?

Panel A	**% of Students in ESE Categories**		
	Using Pre-Program Count		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
F	−0.200	−1.407	−2.751
	(0.609)	(1.018)	(1.631)
Count	−0.007*	−0.012	−0.032
	(0.004)	(0.022)	(0.038)
F * Count	0.024	0.053	0.065
	(0.017)	(0.040)	(0.040)
R^2	0.922	0.844	0.743
Panel B	**% in Excluded ESE Categories**		
	Using Pre-Program Count		
	1 Year After Program (4)	**2 Years After Program (5)**	**3 Years After Program (6)**
F	−0.031	−1.907	−1.555
	(0.655)	(1.186)	(1.600)
Count	−0.011	−0.023*	−0.031
	(0.008)	(0.012)	(0.031)
F * Count	0.028	0.062	0.057
	(0.018)	(0.044)	(0.034)
R^2	0.923	0.859	0.747
Panel C	**% in Included ESE Categories**		
	Using Pre-Program Count		
	1 Year After Program (7)	**2 Years After Program (8)**	**3 Years After Program (9)**
F	−0.151	−0.409	−1.132
	(0.377)	(0.413)	(0.661)
Count	0.004	0.013	0.001
	(0.008)	(0.009)	(0.014)
F * Count	−0.004	−0.011	−0.023
	(0.010)	(0.015)	(0.019)
R^2	0.839	0.711	0.577
Observations	128	130	130

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, real per pupil expenditure, and pre-program (1999) percentage of students in All ESE Categories (panel A), Excluded (panel B), Included (panel C), Learning Disabled (panel D), or Emotionally Handicapped (panel E) category. Count refers to the number of private schools within 5 miles of a public school in 1998.
*$p < .10$, **$p < .05$, ***$p < .01$.

Table 8B. Did McKay Voucher Competition Affect Classification into Special Education Categories?

Panel A	**% in Learning Disabled Category**		
	Using Pre-Program Count		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
F	−0.568	−0.963	−0.320
	(0.420)	(0.686)	(0.741)
Count	−0.006	−0.017*	−0.012
	(0.007)	(0.010)	(0.010)
F * Count	0.015	0.023	0.026
	(0.011)	(0.018)	(0.020)
R^2	0.809	0.739	0.649
Panel B	**% in Emotionally Handicapped Category**		
	Using Pre-Program Count		
	1 Year After Program (4)	**2 Years After Program (5)**	**3 Years After Program (6)**
F	0.163	0.055	0.103
	(0.247)	(0.284)	(0.417)
Count	0.002	−0.004	−0.004
	(0.002)	(0.003)	(0.007)
F * Count	−0.003	−0.009	−0.002
	(0.004)	(0.006)	(0.009)
R^2	0.928	0.892	0.795
Observations	128	130	130

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, real per pupil expenditure, and pre-program (1999) percentage of students in Learning Disabled (panel A) or Emotionally Handicapped (panel B) category. Count refers to the number of private schools within 5 miles of a public school in 1998.
*$p < .10$, **$p < .05$, ***$p < .01$.

panel, tables 8A and 8B). In contrast, the coefficient of "count" is almost always negative and often statistically significant. This implies schools facing greater McKay competition responded by lowering classifications into special education. This pattern is seen for total ESE classification, classification in excluded ESE categories, and LD and EH categories. The results in table 8B, panel A, are consistent with those obtained in Winters and Greene (although for different time periods)[23] exhibiting decreased classifications in learning disabled categories in schools facing larger McKay competition. In contrast,

23. Winters and Greene (2011) relate to 2002–2005, whereas the focus of this study is 1998–2002.

the coefficient of the interaction term shows no evidence of any differential effect of McKay competition on F schools' classification into ESE. To conclude, the McKay scholarship program for disabilities was faced by both F and D schools, and they responded by decreasing classifications into ESE categories, but there was no differential effect of higher extents of McKay competition on F schools.[24]

The tables for LEP present an interesting contrast (table 8C). Consistent with results in section 5, there is evidence of increased classification into excluded LEP categories in grades 3 and 4 in the first year after the program. These effects are also quantitatively similar to those obtained in table 3. There is also evidence of positive shifts in excluded categories in grades 4 and 5 in the second year after the program, but these patterns suggest that these are generated by the increased classification taking place in the year before, in grades 3 and 4. What is interesting is that facing McKay competition (or general competition, recall that the measure is the number of elementary private schools in the vicinity), both F and D schools respond with increased classification into excluded LEP categories and these effects are often statistically significant. This behavior is consistent with incentives. Schools facing more competition face a larger threat of loss of students, and, because a lower grade may increase the chances of such losses, respond by strategically classifying into excluded LEP in an effort to manipulate their grade and make themselves more attractive (and hence potentially avert loss). Although the coefficients of the interaction terms are in most cases positive—indicating F schools facing larger competition tended to respond more strongly with added classifications—these effects are never statistically significant from zero. There is no evidence of any effect on included LEP categories.

8. ASSESSING THE IMPACT OF THE 2002 PROGRAM ON ESE AND LEP CLASSIFICATIONS

Florida's accountability program underwent some drastic changes in 2002. It became far more complicated and introduced points for gain scores, in addition to the level scores in the earlier system. Points for a number of metrics were to be added to achieve the total number of points, which in

24. It might be worthwhile to think what difference in incentives F schools might face (relative to D schools) toward ESE classification, when facing increased McKay competition. Because an F grade carries a larger shame effect, these schools may be more likely to lose ESE students to McKay vouchers relative to D schools (even though they face the same competition). This would induce F schools to classify students into ESE even less often than D schools. F schools also face the incentives posed by A-plus, however, and would have incentives to classify more low-performing students into excluded ESE. Because these two incentives work against each other, it is not clear whether there should be a differential F effect. This is consistent with the findings in tables 8A and 8B.

Table 8C. Did McKay Voucher Competition Affect Classification into Limited English Proficient Categories? Using Pre-Program Competition Measure

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	2 Years After (2)	3 Years After (3)	1 Year After (4)	2 Years After (5)	3 Years After (6)
			Grade 2			
F	0.339	0.340	0.167	0.225	1.062	0.064
	(0.245)	(0.308)	(0.244)	(0.789)	(0.691)	(0.563)
Count	−0.002	0.005	0.002	0.016	0.002	0.017
	(0.004)	(0.003)	(0.003)	(0.017)	(0.013)	(0.012)
F * Count	−0.001	−0.008	−0.006	−0.007	−0.027	−0.029
	(0.005)	(0.006)	(0.006)	(0.025)	(0.020)	(0.017)
Observations	121	122	118	121	122	118
R^2	0.536	0.584	0.548	0.668	0.658	0.741
			Grade 3			
F	0.453**	0.425	0.330	0.330	0.049	1.326
	(0.173)	(0.366)	(0.326)	(0.711)	(0.589)	(0.986)
Count	0.003	0.004	0.002	0.016	0.006	0.005
	(0.003)	(0.004)	(0.004)	(0.010)	(0.013)	(0.013)
F * Count	0.004	0.006	−0.003	−0.032	−0.012	−0.020
	(0.004)	(0.008)	(0.004)	(0.026)	(0.024)	(0.022)
Observations	119	120	122	119	120	122
R^2	0.544	0.494	0.586	0.584	0.605	0.609
			Grade 4			
F	0.442***	0.414*	0.336	0.411	0.540	0.141
	(0.127)	(0.225)	(0.229)	(0.454)	(0.376)	(0.743)
Count	0.008*	0.007**	0.000	0.009	0.009	−0.000
	(0.004)	(0.003)	(0.005)	(0.008)	(0.009)	(0.009)
F * Count	0.006	0.003	0.007	−0.017	−0.032	0.014
	(0.005)	(0.008)	(0.007)	(0.011)	(0.027)	(0.019)
Observations	117	122	119	117	122	119
R^2	0.430	0.539	0.411	0.538	0.447	0.551
			Grade 5			
F	0.403	0.304**	0.625	0.161	0.278	0.280
	(0.237)	(0.151)	(0.412)	(0.552)	(0.596)	(0.379)
Count	0.004	0.008*	0.009***	0.008	−0.005	−0.000
	(0.003)	(0.004)	(0.003)	(0.011)	(0.007)	(0.006)

Table 8C. Continued.

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	**2 Years After (2)**	**3 Years After (3)**	**1 Year After (4)**	**2 Years After (5)**	**3 Years After (6)**
	Grade 5					
F * Count	0.005	0.007	0.007	−0.007	−0.008	−0.000
	(0.005)	(0.007)	(0.007)	(0.011)	(0.010)	(0.010)
Observations	114	115	120	114	115	120
R^2	0.444	0.598	0.461	0.330	0.375	0.399

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, real per pupil expenditure. Count refers to the number of private schools within 5 miles of a public school in 1998.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

turn determined the grade of the school. Most importantly, the new system made it completely impossible to escape an F grade on the basis of a single test, unlike that in the earlier system. Under the 1999 accountability program, schools could escape an F by making the cutoff in any one of the three subject areas of reading, math, and writing. In contrast, even getting the maximum possible number of points in one of the subjects in the newer accountability program would not deliver the number of points needed to escape an F. Under the old program, targeted removal of specific students from the test-taking pool could go a long way in averting an F grade, unlike that under the new program. So, one would expect the new program to have reduced the relative attractiveness of classification into excluded LEP categories. Still another point is worth noting here. Although it is difficult for low-performing students to make the proficiency cutoff (the requirement under the 1999 program), it is often easier for low-performing students to have larger gains merely because of mean reversion (which, in turn, would contribute to school points under the new program). The new system therefore had built-in incentives that, to some extent, discouraged removal of low-performing students from the test-taking pool. Taking advantage of the differences in incentives across the 1999 and 2002 programs, I investigate whether the 2002 program led the F schools to behave in ways different from under the 1999 program (relative to the D schools) in terms of classification into LEP and ESE.

I estimate the impact of the 2002 accountability program on the 2002 F schools (relative to the 2002 D schools) using an RD design. The rules of the new accountability program created a highly nonlinear relationship

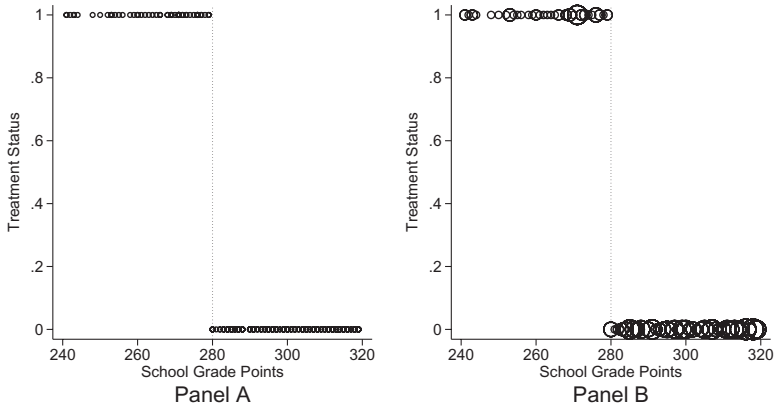

Figure 5. 2002 Regression Discontinuity Analysis: Relationship between School Grade Points and Treatment Status

between the schools' points and the probability of receiving a certain grade. Specifically, there were cutoffs on the score point range that determined the grade of the school. Schools that scored below the threshold of 280 points received an F grade, whereas those at or above 280 received a D. Indeed, as figure 5 shows, there was a strict discontinuity at 280 in the probability of getting an F—schools scoring below 280 received an F grade with probability one, whereas those at or above 280 received an F with probability zero.

Exploiting the institutional structure, using data on ESE and LEP classifications for 2002–5, and utilizing the cutoff of 280 and the RD design described in section 4, I estimate the impact of the 2002 shock on classifications in these categories. Table 9A presents tests for validity of this RD strategy by investigating whether the pre-existing characteristics of the schools were smooth through the cutoff. Except percent excluded LEP in grade 5 (which is barely significant even at 10 percent) and percent included LEP in grade 4, none of the other coefficients are statistically different from zero. As mentioned earlier, with a large number of coefficients, one might expect some to be statistically different from zero even by random variation. There is also no evidence of any discontinuity in the density of the running variable at the cutoff, with a discontinuity size of 0.005 (which is not statistically significant).

Having established the validity of the RD strategy, tables 9B and 9C present the impact of the 2002 shock on classification in LEP and ESE, respectively. Table 9B finds no evidence of any relative classification in either excluded or included LEP categories. Table 9C, panels A–E, respectively, look at the impacts on total ESE classification, excluded ESE classification, included ESE classification, and classification in the two mutable categories (learning disabled and emotionally handicapped). In neither category is there any evidence of relative classifications in F schools.

Table 9A. Testing Validity of 2002 Regression Discontinuity Analysis: Looking for Discontinuities in Pre-Program Characteristics at the Cutoff

Panel A	**% White (1)**	**% Black (2)**	**% Hispanic (3)**	**% Asian (4)**	
	−3.033	−2.537	6.352	−0.745	
	(6.122)	(11.996)	(9.038)	(0.462)	
Panel B	**% American Indian (5)**	**% Male (6)**	**% Free/Reduced Price Lunch (7)**	**Enrollment (8)**	
	−0.037	−3.839	−9.785	−99.913	
	(0.123)	(2.868)	(8.212)	(320.313)	
Panel C	**% ESE (9)**	**% Excluded ESE (10)**	**% Included ESE (11)**	**% Learning Disabled (12)**	**% Emotionally Handicapped (13)**
	−0.587	−0.709	0.121	0.180	−0.822
	(1.050)	(1.262)	(1.301)	(1.313)	(0.679)
Panel D	**% Excluded LEP**				
	Grade 4 (14)	**Grade 5 (15)**			
	−0.033	−0.047*			
	(0.022)	(0.027)			
Panel E	**% Included LEP**				
	Grade 4 (16)	**Grade 5 (17)**			
	−0.837**	−0.662			
	(0.395)	(0.459)			

Note: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses.
*$p < .10$, **$p < .05$, ***$p < .01$.

The results are consistent with the earlier discussion and the incentives that prevailed. The specific changes in the 2002 rules reduced the relative benefit and attractiveness of classification into excluded LEP and hence did not lead to any such classification, in sharp contrast to its precursor, the 1999 program. The incentives relating to McKay scholarship still prevailed, which discouraged additional classification into ESE. The contrasting results between the impacts of the 2002 program and the 1999 program for ESE versus LEP are interesting. The continuity of incentives for ESE led to very similar patterns for ESE under both the 1999 and 2002 programs. In contrast, the break in incentives for LEP classification brought about by the 2002 shock led to very different patterns for LEP classifications under the two systems—increased

Table 9B. Effect of the 2002 Program on Classification into Excluded and Included LEP Categories: A Regression Discontinuity Analysis

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	2 Years After (2)	3 Years After (3)	1 Year After (4)	2 Years After (5)	3 Years After (6)
Grade 4	0.030	−0.039	−0.014	−0.418	0.229	0.333
	(0.051)	(0.035)	(0.061)	(0.270)	(0.303)	(0.380)
Observations	78	77	77	78	77	77
R^2	0.231	0.472	0.174	0.845	0.790	0.694
Grade 5	−0.010	−0.020	0.049	−0.270	−0.246	0.299
	(0.040)	(0.019)	(0.049)	(0.254)	(0.206)	(0.244)
Observations	77	76	76	77	76	75
R^2	0.253	0.360	0.428	0.814	0.701	0.601

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and pre-program (2002) percentage of students in Excluded (columns 1–3) or Included (columns 4–6) LEP categories. $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

classifications under the former, but no evidence of such behaviors under the latter.

9. CONCLUSION

This paper analyzes the behavior of public schools facing a consequential accountability program in Florida. Florida's A-plus program graded schools on a scale of A–F, and made all students of a school eligible for vouchers if the school received two F grades in a period of four years. Consequently, schools receiving their first F were exposed to stigma and threat of vouchers. Utilizing the institutional details of the program, I analyze some of the incentives and responses of the F schools facing the program.

Under the A-plus program, scores of students in some ESE and LEP categories were excluded from grade computations. In the absence of other incentives, this might induce F schools to strategically classify some of their weaker students in these excluded categories to remove them from the effective test-taking pool. But the interplay of the rules of Florida's McKay program for disabled students and the A-plus program led to some interesting divergence of incentives between classification into LEP and ESE categories. The McKay program made all public school ESE students eligible for publicly funded McKay vouchers to move to private schools. Consequently, it discouraged

Table 9C. Effect of the 2002 Program on ESE Classifications: A Regression Discontinuity Analysis

Panel A	**% of Students in ESE Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	−1.373	0.204	0.539
	(2.268)	(1.872)	(1.220)
R^2	0.771	0.708	0.497
Panel B	**% in Excluded ESE Categories**		
	1 Year After Program (4)	**2 Years After Program (5)**	**3 Years After Program (6)**
	−0.512	0.334	1.348
	(1.215)	(1.684)	(1.118)
R^2	0.830	0.670	0.402
Panel C	**% in Included ESE Categories**		
	1 Year After Program (7)	**2 Years After Program (8)**	**3 Years After Program (9)**
	−0.756	−0.123	−0.811
	(1.070)	(0.586)	(0.511)
R^2	0.775	0.700	0.577
Panel D	**% in Learning Disabled Category**		
	1 Year After Program (10)	**2 Years After Program (11)**	**3 Years After Program (12)**
	−0.825	−0.167	−0.842
	(1.075)	(0.579)	(0.594)
R^2	0.775	0.709	0.583
Panel E	**% in Emotionally Handicapped Category**		
	1 Year After Program (13)	**2 Years After Program (14)**	**3 Years After Program (15)**
	0.375	1.104	1.320
	(0.420)	(1.381)	(1.327)
R^2	0.909	0.886	0.764
Observations	79	78	77

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and pre-program (2002) percentage of students in All ESE Categories (panel A), Excluded (panel B), Included (panel C), Learning Disabled (panel D), or Emotionally Handicapped (panel E) category.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

classification into ESE categories because this directly exposed the schools to loss of that student and the corresponding revenue. But no such disincentive was present for classification into LEP. Consistent with these incentives, I indeed find robust evidence of additional classification into excluded LEP categories in high-stakes grade 4 and the entry grade to high-stakes grades

(grade 3) in F schools. In contrast, there is no evidence of any relative classification into excluded ESE categories. Nor is there any evidence of increased classification into included ESE or included LEP categories.

A closer look at the role of McKay competition yields some interesting results. Again the interplay of incentives of the A-plus and McKay programs encouraged schools to behave in strikingly different ways as far as classifications into ESE and LEP were concerned. Consistent with incentives, schools facing a higher concentration of McKay competition responded by classifying fewer students into excluded ESE categories but more students into excluded LEP categories.

Contrasting the rules, incentives, and responses of the 2002 accountability shock with its precursor, the 1999 A-plus program, yields some new insights. The 2002 accountability shock made the system much more complicated, introduced additional points for gains, and made it impossible for a school to escape an F grade on the basis of a single subject, unlike the 1999 system. These blunted the incentives for strategic classification into excluded LEP categories—strategic removal of a few weaker students from the test-taking pool was no longer as promising. Consistent with these incentives, there is no evidence that the F schools resorted to additional classification into excluded LEP categories (relative to D schools), in stark contrast with the responses to the 1999 system. On the other hand, the continuation of similar incentives for ESE led to very similar patterns for ESE across both the 1999 and 2002 programs.

These findings have important policy implications. They imply that, while facing exclusion rules, schools have an inclination to respond by strategically increasing classifications into excluded categories—appropriate counter-incentives can go a long way in averting these unintended behaviors. It follows that when designing accountability or consequential accountability policies, policy makers should be wary of creating exemptions for certain groups of students as they might create adverse incentives to game the system. Embedding appropriate counter-incentives in these programs, however, can potentially thwart such strategic behaviors.

I thank David Figlio, Brian Jacob, Joydeep Roy, Sarah Turner; seminar participants at Columbia University, Duke University, University of Florida, Harvard University, University of Maryland, MIT, Northwestern University, American Economic Association Conference, Association for Education Finance and Policy Conference, Econometric Society Conference, Association for Public Policy Analysis and Management Conference, Society of Labor Economists Conference; the editors and two anonymous referees for valuable comments, Jay Greene and Marcus Winters for sharing their data, and the Florida Department of Education for most of the data used in this analysis. Brandi Coates and Noah Schwartz provided excellent research assistance. The views expressed in this paper are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System. All errors are my own.

REFERENCES

Chakrabarti, Rajashri. 2008a. Impact of voucher design on public school performance: Evidence from Florida and Milwaukee voucher programs. Federal Reserve Bank of New York Staff Paper No. 315.

Chakrabarti, Rajashri. 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(5–6): 1371–93. doi:10.1016/j.jpubeco.2007.06.009

Chakrabarti, Rajashri. 2012. Vouchers, public school response and the role of incentives: Evidence from Florida. *Economic Inquiry*. Available http://onlinelibrary.wiley.com/doi/10.1111/j.1465-7295.2012.00455.x/pdf. Accessed 25 September 2012.

Chiang, Hanley. 2009. How accountability pressures on failing schools affects student achievement. *Journal of Public Economics* 93(9–10): 1045–57. doi:10.1016/j.jpubeco.2009.06.002

Cullen, Julie. 2003. The impact of fiscal incentives on student disability rates. *Journal of Public Economics* 87(7–8): 1557–89. doi:10.1016/S0047-2727(01)00203-1

Cullen, Julie, and Randall Reback. 2006. Tinkering towards accolades: School gaming under a performance accountability system. In *Advances in applied microeconomics, vol. 14: Improving school accountability: Check-ups or choice*, edited by Timothy J. Gronberg and Dennis W. Jansen, pp. 1–35. Amsterdam: Elsevier Science. doi:10.3386/w12286

Fan, Jianqing, and Irene Gijbels. 1996. *Local polynomial modeling and its applications*. London: Chapman and Hall.

Figlio, David. 2006. Testing, crime and punishment. *Journal of Public Economics* 90(4–5): 837–51. doi:10.1016/j.jpubeco.2005.01.003

Figlio, David, and Lawrence Getzler. 2006. Accountability, ability and disability: Gaming the system? In *Advances in applied microeconomics, vol. 14: Improving school accountability: Check-ups or choice*, edited by Timothy J. Gronberg and Dennis W. Jansen, pp. 35–50. Amsterdam: Elsevier Science.

Figlio, David, and Cassandra Hart. 2010. Competitive effects of means-tested vouchers. NBER Working Paper No. 16056.

Figlio, David, and Maurice Lucas. 2004. What's in a grade? School report cards and the housing market. *American Economic Review* 94(3): 591–604. doi:10.1257/0002828041464489

Figlio, David, and Cecilia Rouse. 2006. Do accountability and voucher threats improve low-performing schools? *Journal of Public Economics* 90(1–2): 239–55. doi:10.1016/j.jpubeco.2005.08.005

Figlio, David, and Joshua Winicki. 2005. Food for thought? The effects of school accountability plans on school nutrition. *Journal of Public Economics* 89(2–3): 381–94. doi:10.1016/j.jpubeco.2003.10.007

Goldhaber, Dan, and Jane Hannaway. 2004. Accountability with a kicker: Observations on the Florida A+ accountability plan. *Phi Delta Kappan* 85(8): 598–605.

Greene, Jay. 2001. *An evaluation of the Florida A-plus accountability and school choice program*. New York: Manhattan Institute for Policy Research, Civic Report.

Greene, Jay, and Marcus Winters. 2003. When schools compete: The effects of vouchers on Florida public school achievement. Manhattan Institute for Policy Research Education Working Paper No. 2.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. Identification and estimation of treatment effects with a regression discontinuity design. *Econometrica* 69(1): 201–09. doi:10.1111/1468-0262.00183

Hoxby, Caroline. 2003a. School Choice and school productivity: Could school choice be the tide that lifts all boats? In *The economics of school choice*, edited by Caroline Hoxby, pp. 287–323. Chicago: University of Chicago Press.

Hoxby, Caroline. 2003b. School choice and school competition: Evidence from the United States. *Swedish Economic Policy Review* 10(2): 11–67.

Jacob, Brian. 2005. Accountability, incentives and behavior: The impacts of high-stakes testing in the Chicago public schools. *Journal of Public Economics* 89(5–6): 761–96. doi:10.1016/j.jpubeco.2004.08.004

Jacob, Brian, and Steven Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118(3): 843–78. doi:10.1162/00335530360698441

Ladd, Helen F., and Douglas L. Lauen. 2010. Status versus growth: The distributional effects of school accountability policies. *Journal of Policy Analysis and Management* 29(3): 426–50. doi:10.1002/pam.20504

McCrary, Justin. 2008. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142(2): 698–714. doi:10.1016/j.jeconom.2007.05.005

National Commission on Excellence in Education. 1983. *A Nation at Risk: The Imperative for Educational Reform*. Washington, DC: U.S. Government Printing Office.

Neal, Derek, and Diane W. Schanzenbach. 2010. Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics* 92(2): 263–83. doi:10.1162/rest.2010.12318

Reback, Randall. 2008. Teaching to the rating: School accountability and distribution of student achievement. *Journal of Public Economics* 92(5–6): 1394–1415. doi:10.1016/j.jpubeco.2007.05.003

Rouse, Cecilia E., Jane Hannaway, David Figlio, and Dan Goldhaber. 2007. Feeling the Florida heat: How low-performing schools respond to voucher and accountability pressure. CALDER Working Paper No. 13, Urban Institute.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Singer, Judith, Judith Palfrey, John Butler, and Deborah Walker. 1989. Variation in special education classification across school districts: How does where you live affect what you are labeled? *American Educational Research Journal* 26(2): 261–81.

West, Martin, and Paul Peterson. 2006. The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *Economic Journal* 116(510): C46–C62. doi:10.1111/j.1468-0297.2006.01075.x

Winters, Marcus A., and Jay P. Greene. 2011. Public school response to special education vouchers: The impact of Florida's McKay scholarship program on disability diagnosis and student achievement in public schools. *Educational Evaluation and Policy Analysis* 33(2): 138–58. doi:10.3102/0162373711404220

APPENDIX

Table A.1. Summary Statistics

Panel A	**% White (1)**	**% Black (2)**	**% Hispanic (3)**	**% Asian (4)**	**% American Indian (5)**
mean	13.841	64.439	20.866	0.641	0.213
sd	(17.783)	(28.888)	(24.235)	(1.057)	(0.598)

Panel B	**% Male (6)**	**% Free/Reduced Price Lunch (7)**	**Enrollment (8)**	**Real PPE (9)**
mean	51.402	84.640	713.193	32.289
sd	(2.228)	(11.699)	(217.018)	(7.698)

Panel C	**% ESE (10)**	**% Excluded ESE (11)**	**% Included ESE (12)**	**% Learning Disabled (13)**	**% Emotionally Handicapped (14)**
mean	16.213	11.694	4.519	4.317	1.003
sd	(6.186)	(6.041)	(2.669)	(2.250)	(1.638)

Panel D	**% Excluded LEP**			
	Grade 2 (15)	**Grade 3 (16)**	**Grade 4 (17)**	**Grade 5 (18)**
	0.753	0.665	0.617	0.584
	(0.810)	(0.747)	(0.684)	(0.676)

Panel E	**% Included LEP**			
	Grade 2 (19)	**Grade 3 (20)**	**Grade 4 (21)**	**Grade 5 (22)**
	3.004	2.776	2.229	1.553
	(3.206)	(2.905)	(2.280)	(1.578)

Notes: Means with standard deviations of the discontinuity sample within the Silverman bandwidth (pooled for the sample period 1999–2002) in parentheses. Consistent with data usage in this study (see section 3 for details), the summary statistics relating to ESE (panel C) pertain to school level data, and the summary stats relating to LEP (panels D and E) pertain to grade-school level data.

Table A.2. Effect of the Program on Classification into Excluded and Included LEP Categories: A Regression Discontinuity Analysis

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	2 Years After (2)	3 Years After (3)	1 Year After (4)	2 Years After (5)	3 Years After (6)
Grade 2	0.348	0.205	0.086	0.705	0.876	0.403
	(0.216)	(0.234)	(0.218)	(0.518)	(0.535)	(0.425)
Observations	117	119	114	117	119	114
R^2	0.613	0.712	0.656	0.920	0.910	0.908
Grade 3	0.318**	0.190	0.176	−0.309	0.422	0.808
	(0.111)	(0.204)	(0.217)	(0.202)	(0.433)	(0.622)
Observations	114	113	115	114	113	115
R^2	0.659	0.611	0.678	0.891	0.872	0.870
Grade 4	0.312**	0.364*	0.105	0.431	0.662	0.575
	(0.115)	(0.209)	(0.193)	(0.261)	(0.447)	(0.361)
Observations	111	114	112	111	114	112
R^2	0.590	0.572	0.493	0.900	0.797	0.832
Grade 5	0.173	0.307**	0.305	0.133	0.320	0.468
	(0.172)	(0.126)	(0.234)	(0.311)	(0.285)	(0.322)
Observations	108	107	110	108	107	110
R^2	0.641	0.633	0.645	0.811	0.716	0.756

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, real per pupil expenditure, and pre-program (1999) percentage of students in Excluded (columns 1–3) or Included (columns 4–6) LEP categories.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.

Table A.3. Effect of the Program on Classification into Excluded and Included LEP Categories: A Regression Discontinuity Analysis Excluding Real PPE

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After (1)	2 Years After (2)	3 Years After (3)	1 Year After (4)	2 Years After (5)	3 Years After (6)
Grade 2	0.268	0.193	0.004	0.021	0.756	−0.618
	(0.223)	(0.286)	(0.230)	(0.385)	(0.715)	(0.554)
Observations	124	124	120	124	124	120
R^2	0.528	0.571	0.527	0.633	0.639	0.734
Grade 3	0.359*	0.298	0.264	−0.404	0.133	0.792
	(0.181)	(0.311)	(0.258)	(0.459)	(0.384)	(0.849)
Observations	122	122	124	122	122	124
R^2	0.545	0.483	0.578	0.561	0.579	0.604
Grade 4	0.296**	0.353*	0.148	0.051	−0.011	0.490
	(0.120)	(0.210)	(0.286)	(0.347)	(0.412)	(0.380)
Observations	121	124	121	121	124	121
R^2	0.404	0.522	0.398	0.484	0.421	0.543
Grade 5	0.265	0.296***	0.443	0.093	0.260	0.270
	(0.258)	(0.097)	(0.372)	(0.484)	(0.454)	(0.247)
Observations	117	117	122	117	117	122
R^2	0.431	0.543	0.443	0.214	0.333	0.397

Notes: Robust standard errors adjusted for clustering by school district are in parentheses. All regressions control for racial composition, gender composition, and percent of students eligible for free or reduced price lunch.
*$p < .10$, **$p < .05$, ***$p < .01$.

Table A.4. Effect of the Program on ESE Classifications: A Regression Discontinuity Analysis Excluding Real PPE

Panel A	**% of Students in ESE Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	0.469	−0.821	−0.854
	(0.401)	(0.879)	(0.830)
R^2	0.920	0.833	0.720
Panel B	**% in Excluded ESE Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	0.618	−0.249	−0.230
	(0.601)	(1.062)	(1.037)
R^2	0.920	0.850	0.733
Panel C	**% in Included ESE Categories**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	−0.138	−0.507	−0.563
	(0.298)	(0.384)	(0.371)
R^2	0.834	0.700	0.564
Panel D	**% in Learning Disabled Category**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	−0.207	−0.283	0.252
	(0.268)	(0.424)	(0.409)
R^2	0.792	0.717	0.630
Panel E	**% in Emotionally Handicapped Category**		
	1 Year After Program (1)	**2 Years After Program (2)**	**3 Years After Program (3)**
	0.076	−0.098	0.039
	(0.168)	(0.180)	(0.224)
R^2	0.922	0.885	0.788
Observations	133	132	132

Notes: Robust standard errors adjusted for clustering by the running variable (% of school's students at or above the writing cutoff) are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and pre-program (1999) percentage of students in All ESE Categories (panel A), Excluded (panel B), Included (panel C), Learning Disabled (panel D), or Emotionally Handicapped (panel E) category.
$^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$.