Federal Reserve Bank of New York
Staff Reports

Vouchers, Public School Response, and the Role of Incentives

Rajashri Chakrabarti

Staff Report no. 306
October 2007
Revised November 2010

Vouchers, Public School Response, and the Role of Incentives

Rajashri Chakrabarti

Federal Reserve Bank of New York Staff Reports, no. 306
October 2007; revised November 2010.
JEL classification: H4, I21, I28

Abstract

This paper analyzes the incentives and responses of public schools in the context of an educational reform. The literature on the effect of voucher programs on public schools typically focuses on student and mean school scores. This paper tries to go inside the black box to investigate some of the ways in which schools facing the Florida voucher program behaved. The program embedded vouchers in an accountability regime. Schools getting an "F" grade for the first time were exposed to the threat of vouchers, but did not face vouchers unless and until they got a second "F" within the next three years. In addition, "F," being the lowest grade, exposed the threatened schools to stigma. Exploiting the institutional details of this program, I analyze the incentives built into the system and investigate the behavior of the threatened public schools facing these incentives. There is strong evidence that they did respond to incentives. Using highly disaggregated school-level data, a difference-in-differences estimation strategy, and a regression discontinuity analysis, I find that the threatened schools tended to focus more on students below the minimum criteria cutoffs rather than reading and math. These results are robust to controlling for differential preprogram trends, changes in demographic compositions, mean reversion, and sorting. The findings have important policy implications.

Key words: vouchers, incentives, regression discontinuity, mean reversion

---

Chakrabarti: Federal Reserve Bank of New York (e-mail: rajashri.chakrabarti@ny.frb.org). The author thanks Steve Coate, Sue Dynarski, Ron Ehrenberg, David Figlio, Ed Glaeser, Caroline Hoxby, Brian Jacob, Bridget Long, Paul Peterson, Miguel Urquiola, seminar participants at Duke University, the University of Florida, Harvard University, the University of Maryland, MIT, Northwestern University, the Econometric Society Conference, the Association for Public Policy Analysis and Management Conference, and the Society of Labor Economists Conference for helpful discussions. She also thanks the Florida Department of Education for data used in this analysis, and the Program on Education Policy and Governance at Harvard University for its postdoctoral support. Dan Greenwald 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.

1 Introduction

The concern over public school performance in the last two decades has pushed public school reform to the forefront of policy debate in the United States. School accountability and school choice, and especially vouchers, are among the most hotly debated instruments of public school reform. Understanding the behavior and response of public schools facing these initiatives is key to an effective policy design. This paper takes an important step forward in that direction by analyzing public school behavior under the Florida voucher program.

The Florida voucher program, known as the "opportunity scholarship" program, is unique in that it embeds a voucher program within a school accountability system. Moreover, the federal No Child Left Behind (NCLB) Act is similar to and largely modeled after the Florida program, which makes the latter all the more interesting and relevant. Most studies to date, studying the effect of voucher programs on public schools, have looked at the effect on student and mean school scores. In contrast, this study tries to go inside the black box to investigate some of the ways in which schools facing the voucher program behaved in the first three years after program.[1] Exploiting the institutional details of the Florida program during this period, it analyzes the incentives built into the system, and investigates public school behavior and response facing these incentives.

The Florida voucher program, written into law in June 1999, made all students of a school eligible for vouchers if the school got two "F" grades in a period of four years. Thus, the program can be looked upon as a "threat of voucher" program—schools getting an "F" grade for the first time were directly threatened by vouchers, but vouchers were implemented only if they got another "F" grade in the next three years. Vouchers were associated with a loss in revenue and also media publicity and visibility. Moreover, the "F" grade, being the lowest performing grade, was likely associated with shame and stigma. Therefore, the threatened schools had a strong incentive to try to avoid the second "F". This paper studies some alternative ways in which the threatened schools responded, facing the incentives built into the system.

Under the 1999 Florida grading criteria, certain percentages of a school's students had to score above some specified cutoffs on the score scale for it to escape the second "F". Therefore the threatened schools

[1] Under the Florida voucher program (described below), schools getting an "F" grade in 1999 were directly threatened by vouchers, but this threat remained valid for the next three years only. Therefore, I study the behavior of the 1999 threatened schools during these three years.

had an incentive to focus more on students expected to score just below these high stakes cutoffs rather than equally on all students. Did this take place in practice? Second, to escape an F grade, the schools needed to pass the minimum criteria in only one of the three subject areas of reading, math and writing. Did this induce the threatened schools to concentrate more on one subject, rather than equally on all? If so, which subject area did the schools choose to concentrate on? One alternative would be to concentrate on the subject area closest to the cutoff.[2] But subject areas differ in the extent of difficulties, so it is not immediately obvious that it is easiest to pass the cutoff in the subject area closest to the cutoff. Rather, schools are likely to weigh the extent of difficulties of the different subjects and their distances from the cutoffs, and choose the subject that is least costly to pass the cutoff. In addition to analyzing the above questions, this study also tries to look at a broader picture. If the threatened schools concentrated on students expected to score just below the high stakes cutoffs, did their improvements come at the expense of higher performing ones?

Using highly disaggregated school level Florida data from 1993 through 2002, and a difference-in-differences analysis as well as a regression discontinuity analysis, I investigate the above issues. There is strong evidence that public schools responded to the incentives built into the system. First, I find that the threatened schools concentrated more on students below and closer to the high stakes cutoffs, rather than equally on all students. Note that, as discussed in detail later, this improvement of the low performing students does not seem to have come at the expense of the higher performing students. Rather, there seems to have been a rightward shift of the entire score distribution, with improvement concentrated more in the score ranges just below the high stakes cutoff. This pattern holds in all the three subjects of reading, math and writing. Second, I find that the threatened schools indeed focused more on one subject area. They did not focus more on the subject area closest to the cutoff. Rather, they concentrated on writing, irrespective of the distances of the subject areas from the high stakes cutoffs. This is consistent with the perception among Florida administrators that writing scores were considerably easier to improve than scores in reading or math. These results are quite robust in that they withstand several sensitivity tests including controlling for pre-program trends, mean reversion, sorting, changes in demographic compositions and other observable characteristics of schools. Also, the results from the difference-in-differences analysis are qualitatively similar to those obtained from the regression discontinuity analysis.

[2] The cutoffs differ across subjects (as will be detailed below). Here "cutoff" refers to the cutoff in the corresponding subject area.

This study is related to two strands of literature. The first strand investigates whether schools facing accountability systems and testing regimes respond by gaming the system in various ways. This relates to the moral hazard problems associated with multidimensional tasks under incomplete observability, as pointed out by Holmstrom and Milgrom (1991). Cullen and Reback (2006), Figlio and Getzler (2006) and Jacob (2005) show that schools facing accountability systems tend to reclassify their low performing students as disabled in an effort to make them ineligible to contribute to the school's aggregate test scores, ratings or grades. Jacob (2005) also finds evidence in favor of teaching to the test, preemptive retention of students and substitution away from low-stakes subjects, while Jacob and Levitt (2003) find evidence in favor of teacher cheating. Reback (2005) finds that schools in Texas facing accountability ratings have tended to relatively improve the performance of students who are on the margin of passing. Studying Chicago public schools, Neal and Schanzenbach (2010) similarly find that introduction of *No Child Left Behind* and other previous accountability programs induced schools to focus more on the middle of their achievement distributions. Figlio (2006) finds that low performing students are given harsher punishments during the testing period than higher performing students for similar crimes, once again in an effort to manipulate the test taking pool. Figlio and Winicki (2005) find that schools faced with accountability systems increase the caloric content of school lunches on testing days in an attempt to boost performance.

While the above papers study the response of public schools facing accountability systems, the present paper studies public school response and behavior facing a voucher system,—a voucher system that ties vouchers to an accountability regime. Although there is considerable evidence relating to the response of public schools facing accountability regimes, it would be instructive to know how public schools behave facing such a voucher system, an alternative form of public school reform. Second, this study also uses a different estimation strategy than that used in the above literature. The above literature uses a difference-in-differences strategy. In contrast, this paper uses a regression discontinuity analysis in addition to a difference-in-differences strategy that can get rid of some potential confounding factors, such as mean reversion and existence of differential pre-program trends. Third, in addition to investigating whether the voucher program led the threatened schools to focus on marginal students, this paper also investigates whether the program induced these schools to focus more on a specific subject area. None of the above papers investigate this form of alternative behavior.

The second strand of literature that this paper is related to analyzes the effect of vouchers on public

school performance. Theoretical studies in this literature include McMillan (2004) and Nechyba (2003). Modeling public school behavior, McMillan (2004) shows that under certain circumstances, public schools facing vouchers may find it optimal to reduce productivity. Nechyba (2003) shows that while public school quality may show a small decline with vouchers under a pessimistic set of assumptions, it will improve under a more optimistic set of assumptions.

Combining both theoretical and empirical analysis, Chakrabarti (2008a) studies the impact of two alternative voucher designs—Florida and Milwaukee—on public school performance. She finds that voucher design matters—the "threat of voucher" design in the former has led to an unambiguous improvement of the treated public schools in Florida and this improvement is larger than that brought about by traditional vouchers in the latter. Other empirical studies in this literature include Greene (2001, 2003), Hoxby (2003a, 2003b), Figlio and Rouse (2006), Chakrabarti (2008b) and West and Peterson(2006).[3] Greene (2001, 2003) finds positive effects of the Florida program on the performance of treated schools. Figlio and Rouse (2006) find some evidence of improvement of the treated schools under the program in the high stakes state tests, but these effects diminish in the low stakes, nationally norm-referenced test. West and Peterson (2006) study the effects of the revised Florida program (after the 2002 changes) as well as the NCLB Act on test performance of students in Florida public schools. They find that the former program has had positive and significant impacts on student performance, but they find no such effect for the latter. Based on case studies from visits to five Florida schools (two "F" schools and three "A" schools), Goldhaber and Hannaway (2004) present evidence that F schools focused on writing because it was the easiest to improve.[4] Analyzing the Milwaukee voucher program, Hoxby (2003a, 2003b) find evidence of a positive productivity response to vouchers after the Wisconsin Supreme Court ruling of 1998. Following Hoxby (2003a, 2003b) in the treatment and control group classification strategy, and using data for 1987-2002, Chakrabarti (2008b) finds that the shifts in the Milwaukee voucher program in the late 1990's led to a higher improvement of the treated schools in the second phase of the Milwaukee program than that in the first phase.

Most of the above studies analyze the effect of different voucher programs on student and mean school scores and document an improvement in these measures. This study, on the other hand, tries to

[3] For a comprehensive review of this literature as well as other issues relating to vouchers, see Howell and Peterson (2005), Hoxby (2003b) and Rouse (1998).

[4] Schools that received a grade of "A" in 1999 are referred to as "A" schools. Schools that received a grade of "F" ("D") in 1999 will henceforth be referred to as "F" ("D") schools.

delve deeper so as to investigate where this improvement comes from. Analyzing the incentives built into the system, it seeks to investigate some of the alternative ways in which the threatened schools in Florida behaved. Chakrabarti (2008a) and Figlio and Rouse (2006) analyze the issue of teaching to the test, but they do not examine the forms of behavior that are of interest in this paper. Evidence on the alternative forms of behavior of public schools facing a voucher program is still sparse. This study seeks to fill this important gap.

2 The Program and its Institutional Details

The Florida Opportunity Scholarship Program was signed into law in June 1999. 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 getting 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.

To understand the incentives created by the program, it is important to understand the Florida testing system and school grading criteria.[5] In the remainder of the paper, I refer to school years by the calendar year of the spring semester. Following a field test in 1997, the FCAT (Florida Comprehensive Assessment Test) 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 FCAT reading and math scores were expressed in a scale of 100-500. The state categorized students into five achievement levels in reading and math that corresponded to specific ranges on this raw score scale.[6] The FCAT writing scores, on the other hand, were expressed in a scale of 1-6. The Florida Department of Education reports the percentages of students scoring at 1, 1.5, 2, 2.5, ..., 6 in FCAT writing. For simplicity, as well as symmetry with reading and math, I divide the writing scores into five categories and call them levels 1-5. Scores 1 and 1.5 will together constitute level 1; scores 2 and 2.5 level 2; 3 and 3.5 level 3; 4 and 4.5 level 4; 5, 5.5 and 6 level 5. (The results in this paper are not sensitive to the definitions of these categories.)[7] In the remainder of the paper, for

[5] Since I am interested in the incentives faced by the threatened schools and this mostly depends on the criteria for "F" grade and what it takes to move to a "D", 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.

[6] Levels 1, 2, 3, 4 and 5 in grade 4 reading corresponded to score ranges 100-274, 275-298, 299-338, 339-385 and 386-500 respectively. Levels 1, 2, 3, 4 and 5 in grade 5 math corresponded to score ranges of 100-287, 288-325, 326-354, 355-394 and 395-500 respectively.

[7] Defining the categories in alternative ways or considering the scores separately do not change the results.

writing, level 1 will refer to scores 1 and 1.5 together; level 2 scores 2 and 2.5 together etc.; while 1, 2, 3, ..., 6 will refer to the corresponding raw scores.

The system of assigning letter grades to schools started in the year 1999,[8] 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 the three subjects of FCAT reading, math and writing, and a "D" if it failed the minimum criteria in only one or two of the three subject areas. To pass the minimum criteria in reading and math, at least 60% of the students had to score at level 2 and above in the respective subject, while to pass the minimum criteria in writing, at least 50% had to score 3 and above.

3 Theoretical Discussion

This section and subsections 3.1-3.3 explore some alternative ways of response of public schools facing a Florida-type "threat of voucher" program and the 1999 grading system. Assume that there are n alternative ways in which a public school can apply its effort. Quality q of the public school is given by $q = q(e_1, e_2, ..., e_n)$ where $e_i, i = \{1, 2, ..., n\}$, represents the effort of the public school in alternative i. Assume that e_i is non-negative for all i and that the function q is increasing and concave in all its arguments. Any particular quality level q can be attained by multiple combinations of $\{e_1, e_2, ..., e_n\}$—the public school chooses the combination that optimizes its objective function. Public school cost is given by $C = C(e_1, e_2, ..., e_n)$, where C is increasing and convex in its arguments.

The Florida program designates a quality cutoff $\bar{q}$ such that the threatened schools get a second "F" and vouchers are implemented if and only if the school fails to meet the cutoff. A school deciding to meet the cutoff can do so in a variety of ways—its problem then is to choose the best possible way. More precisely, it faces the following problem:

$$\text{Minimize } \ C = C(e_1, e_2, ..., e_n) \ \text{ subject to } \ q(e_1, e_2, ..., e_n) \geq \bar{q}$$

The public school chooses effort level ${e_i}^*, i = \{1, 2, ..., n\}$ such that ${e_i}^*$ solves $\frac{\delta C({e_i}^*)}{\delta {e_i}^*} \geq \lambda \frac{\delta q({e_i}^*)}{\delta {e_i}^*}$ and ${e_i}^*[\frac{\delta C({e_i}^*)}{\delta {e_i}^*} - \lambda \frac{\delta q({e_i}^*)}{\delta {e_i}^*}] = 0$, where λ is the Lagrange multiplier and $q(e_1^*, e_2^*, ..., e_n^*) = \bar{q}$. If ${e_i}^*$ is strictly positive, ${e_i}^*$ solves $\frac{\delta C({e_i}^*)}{\delta {e_i}^*} = \lambda \frac{\delta q({e_i}^*)}{\delta {e_i}^*}$.

Thus the amounts of effort that the public school chooses to expend on the various alternatives depend on the marginal costs and marginal returns from the alternatives. While it delegates higher

[8] Before 1999, schools were graded by a numeric system of grades, I-IV (I-lowest, IV-highest).

efforts to alternatives with higher marginal returns and/or lower marginal costs, the effort levels in alternatives with lower marginal returns and higher marginal costs are lower. It can choose a single alternative l (if $\frac{\delta C(e_l{}^*)}{\delta e_l{}^*} - \lambda \frac{\delta q(e_l{}^*)}{\delta e_l{}^*} = 0 < \frac{\delta C(e_k{}^*)}{\delta e_k{}^*} - \lambda \frac{\delta q(e_k{}^*)}{\delta e_k{}^*}$ for all $k \neq l$) or it can choose a mix of alternatives. In the latter case the net marginal returns $(\frac{\delta q}{\delta e_i} - \frac{1}{\lambda}\frac{\delta C}{\delta e_i})$ from each of the alternatives in the mix are equal (and in turn equal to zero) at the chosen levels of effort. This paper empirically analyzes the behavior of public schools and investigates what alternatives the public schools actually chose when faced by the 1999 Florida "threat of voucher" program.

3.1 The Incentives Created by the System and Alternative Avenues of Public School Responses

3.1.1 Focusing on Students below the Minimum Criteria Cutoffs

Given the Florida grading system, threatened public schools striving to escape the second "F" would have an incentive to focus on students expected to score just below the minimum criteria cutoffs.[9] Marginal returns from focusing on such students would be expected to be higher than that on a student expected to score at a much higher level (say, level 4). If marginal costs were not too high, the threatened schools would be expected to resort to such a strategy.

If schools did indeed behave according to this incentive, then the percentage of students scoring at level 1 in reading and math would be expected to fall after the program as compared to the pre-program period. In writing, the cutoff level is 3 (rather than level 2 in reading and math). Therefore, while the threatened schools would have an incentive to focus on students expected to score below 3, they would be induced to focus more on students expected to score in level 2, since they were closer to the cutoff and hence easier to push over the cutoff. So while a downward trend would be expected in both the percentages of students scoring in levels 1 and 2, the fall would be more prominent in level 2.

3.1.2 Choosing between Subjects with Different Extents of Difficulties Versus Focusing on Subject Closer to the Cutoff

As per the Florida grading criteria, the threatened schools needed to pass the minimum criteria in only one of the three subjects to escape a second F grade. Therefore the schools had an incentive to focus more on one particular subject area, rather than equally on all. Note that it is unlikely that the

[9] Some ways to do this would be to target curriculum to low performing students, put more emphasis on the basic concepts rather than advanced topics in class or repeating material already covered rather than moving quickly to new topics.

concerned schools will focus exclusively on one subject area and completely neglect the others because there is an element of uncertainty inherent in student performance and scores, the degree of difficulty of the test, etc. and schools surely have to answer to parents for such extreme behavior. But if they behave according to incentives, it is likely that they will concentrate more on one subject area. The question that naturally arises in this case is: which subject area will the threatened schools focus on?

One possibility is to focus more on the subject area closest to the cutoff i.e. the subject area for which the difference between the percentage of students scoring below the cutoff in the previous year and the percentage required to pass the minimum criteria is the smallest.[10] However, the subject areas differ in terms of their extent of difficulties, and hence the schools may find it more worthwhile to focus on a subject area farther from the cutoff, which otherwise is easier to improve in. In other words, the distance from the cutoff has to be weighed against the extent of difficulty or ease in a subject area, and the effort that a school decides to put in will depend on both factors.

4 Data

The data for this study were obtained from the Florida Department of Education. These data include school-level data on mean test scores, grades, percentages of students scoring in different levels, grade distribution of schools, socio-economic characteristics of schools and school finances. In spite of being school level data, these data are highly disaggregated—in addition to data on mean school scores, data are available on percentages of students scoring in different ranges of the score scale for each of reading, math and writing.

School level data on the percentage of students scoring in each of the five levels are available from 1999 to 2002 for both FCAT grade 4 reading and grade 5 math. In addition, data are available on percentages of students scoring in levels 1 and 2 in 1998 for both reading and math. Data are also available on mean scale scores and number of students tested for each of reading and math from 1998-2002.

In grade 4 writing, data are available on the percentage of students scoring at the various score points. These data are available from 1994 to 1996 and again from 1999 to 2002. In addition, data on mean scale scores in writing and number of students tested are available from 1994-2002. Data on school grades are available from 1999 to 2002.

[10] As outlined earlier, the required percentage of students below cutoff that would allow the school to pass the minimum criteria in the respective subject is 40% in reading and math and 50% in writing.

School level data on grade distribution (K-12) of students are available from 1993-2002. Data on socio-economic characteristics include data on gender composition (1994-2002), race composition (1994-2002) and percent of students eligible for free or reduced-price lunches (1997-2002). School finance data consist of several measures of school level and district level per pupil expenditures and are available for the period 1993-2002.

5 Empirical Analysis

Under the Florida opportunity scholarship program, schools that received a grade of "F" in 1999 were directly threatened by "threat of vouchers" and stigma,–the former in the sense that all their students would be eligible for vouchers if the school received another "F" grade in the next three years. These schools will constitute my treated group of schools and will be referred to 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. They will constitute my control group of schools and will be referred to as "D schools" in the rest of the paper.[11] 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, upto 2002).

5.1 Did the Threatened Schools Focus on Students Expected to score Below the Minimum Criteria Cutoffs

As discussed above, if the treated schools tend to focus more on students anticipated to score below the minimum criteria cutoffs, the percentage of students scoring in level 1 in F schools in reading and math should exhibit a decline relative to D schools after the program. In FCAT writing, although relative declines are likely in both levels 1 and 2, the relative decline in level 2 would be larger than in level 1, if the treated schools responded to incentives.

Note that the underlying assumption here is that in the absence of the program, the score distribution of students (that is, percentage of students at various levels) in F schools (relative to D schools) would

[11] Two of the "F schools" became eligible for vouchers in 1999. They were in the state's "critically low-performing schools list" in 1998 and were grandfathered in the program. I exclude them from the analysis because they likely faced different incentives. None of the other F schools got a second "F" in either 2000 or 2001. Four schools got an F in 2000 and all of them were "D schools". I exclude these four "D schools" from the analysis. (Note though that results do not change qualitatively if I include them in the analysis.) No other D school received an "F" either in 2000 or 2001.

remain similar to that before. This does not seem to be an unreasonable assumption because as I show later, there is no evidence of any differences in pre-existing trends in various levels in F schools (relative to D schools). This implies that before the program the relative score distribution of students remained similar over the years.

To investigate whether the F schools tended to focus on marginal students, I look for shifts in the percentages of students scoring in the different levels (1-5) in the F schools relative to the D schools in the post-program period. Using data from 1999 to 2002, I estimate the following model:

$$P_{ijt} = \sum_{j=1}^{5} \alpha_{0j} L_j + \sum_{j=1}^{5} \alpha_{1j} (F * L_j) + \sum_{k=2000}^{2002} \sum_{j=1}^{5} \alpha_{2kj} (D_k * L_j) + \sum_{k=2000}^{2002} \sum_{j=1}^{5} \alpha_{3kj} (F * D_k * L_j) + \alpha_{4j} X_{ijt} + \varepsilon_{ijt} \quad (1)$$

where P_{ijt} denotes the percentage of students in school i scoring in level j in year t; F is a dummy variable taking the value of 1 for F schools and 0 for D schools; $L_j, j = \{1, 2, 3, 4, 5\}$ are level dummies that take a value of 1 for the corresponding level, 0 otherwise; $D_k, k = \{2000, 2001, 2002\}$ are year dummies for years 2000, 2001 and 2002 respectively. The variables $(D_k * L_j)$ control for post-program common year effects and X_{ijt} denote the set of control variables. Control variables include racial composition of schools, gender composition of schools, percentage of students eligible for free or reduced-price lunches, real per pupil expenditure and interaction of the level dummies with each of these variables. The coefficients on the interaction terms $(F * D_k * L_j)$ represent the program effects on the F schools in each of the five levels and in each of the three years after the program. I also run the fixed effects counterpart of this regression which includes school by level fixed effects (and hence does not have the level and level interacted with treated dummies). These regressions are run for each of the subject areas—reading, math and writing.

Figure 1 shows the distribution of percentage of students scoring below the minimum criteria cutoffs in F and D schools in 1999 and 2000 in the three subject areas of reading, math and writing. 1999 is the last pre-program year and 2000 the first post-program year. Panels A and B (C and D) look at the distributions in level 1 reading (level 1 math) in the two years respectively, while panels E and F look at the distributions in level 2 writing in 1999 and 2000 respectively. In each of reading, math and writing, the graphs show a relative leftward shift of the F school distribution in comparison to the D school distribution in 2000. This suggests that the F schools were characterized by a greater fall in the percentage of students scoring in level 1 reading, level 1 math and level 2 writing after the program.

Figure 2 shows the distribution of reading, math and writing scores by treatment status in 1999 and 2000. In each of reading and math, there is a fall in the percentage of students scoring in level 1 in F

schools relative to D schools in 2000. In writing, on the other hand, while there are relative falls in both levels 1 and 2 in F schools, the relative fall in level 2 is much more prominent than the fall in level 1. Another important feature—seen in all reading, math and writing—is that there is a general relative rightward shift in the F distribution in 2000, with changes most concentrated in the crucial levels.

Table 1 presents results on the effect of the program on percentages of students scoring in levels 1-5 in FCAT reading, math and writing. Using model 1, columns (1)-(2) look at the effect in reading, columns (3)-(4) in math and columns (5)-(6) in writing. For each set, the first column reports the results from OLS estimation and the second column from fixed effects estimation. All regressions are weighted by the number of students tested and control for racial composition and gender compositions of schools, percentage of students eligible for free or reduced-price lunches, real per pupil expenditure and interactions of each of these variables with level dummies.

In reading, both OLS and FE estimates show relative decline in percentage of students in level 1 in F schools in each of the three years after the program.[12] On the other hand, there are increases in the percentage of students scoring in levels 2, 3 and 4. The level 1, 2 and 3 effects are always statistically significant (except level 2 in first year), while level 4 effects never are. The level 5 percentages saw small, though statistically significant declines. Moreover, the changes in level 1 percentages always economically (and in most cases, statistically) exceed the changes in each of the other levels in each of the three years after the program. These patterns are consistent with the hypothesis that in reading schools chose to focus more on students they expected to score below the minimum criteria cutoff.

The results in math (columns (3)-(4)) are similar. There is a steep and statistically significant decline in the percentage of students scoring in level 1, in each of the three years after the program. Increases are seen in percentages of students in levels 2, 3 and 4, which are statistically significant in most cases. Percentages of students in level 5 on the other hand saw a small decline, though the effects are not statistically significant in most cases. Once again, the decline in the level 1 percentages exceed the changes in the other levels, both economically and statistically.

Columns (5)-(6) present the results for writing. The percentages of students scoring in both levels 1

[12] Although the state still continued to grade the Florida schools on a scale of A though F, the grading criteria underwent some changes in 2002. So a natural question that arises here is whether the 2002 effects (that is, the effects in the third year after program) were induced by the 1999 program or were also contaminated by the effect of the 2002 changes. However, these new grading rules were announced in December 2001 and were extremely complicated combining student learning gains in addition to level scores. Since the FCAT tests were held in February and March 2002, just a couple of months after the announcement, it is unlikely that the 2002 effects were contaminated by responses to the 2001 announcement. Moreover, the results are very similar if the year 2002 is dropped and the analysis is repeated with data through 2001.

and 2 saw a decline after the program. But interestingly, the decline in level 2 is larger (both economically and statistically) than in level 1. In writing, there is no evidence of a fall in the percentage of students scoring in level 5.

It should be noted here that the changes in table 1 in each of the levels are net changes. For example, it is possible that some students moved from higher levels to level 1. If this did happen, then the actual fall in level 1 in terms of movement of students from level 1 to the higher levels is even larger than that suggested by the estimate. Again, to the extent that there may have been movements from level 2 to upper levels, the actual increases in level 2 in reading and math are larger than that seen in the level 2 estimates above. Similarly, to the extent that there have been moves from level 1 to level 2 in writing, the actual fall in level 2 writing is larger than that seen in the above level 2 estimates. It is possible that some students moved from the upper levels to levels 1 and 2, but this does not seem to have been a major factor. This is because there is not much evidence of declines in the upper levels and the cumulative percentage changes of students in levels 3, 4 and 5 are always large and positive. This discussion suggests that the falls in the percentages of students just below the minimum criteria cutoff can be actually larger than that suggested by the estimates (but not smaller).

Figures 3, 4 and 5 show the trends in the percentages of students scoring in levels 1-5 in reading, math and writing respectively. Consistent with the results obtained above, there is a large decline in the percentage of students scoring in level 1 in each of reading and math which exceeds the changes in the other levels. In writing, on the other hand, the decline is considerably larger in level 2 than in level 1, once again in conformity with the above regression results.

The patterns in reading, math and writing support the hypothesis that the F schools chose to focus more on students expected to score just below the minimum criteria cutoffs. More importantly, consistent with the incentives created by the program, while the declines in reading and math were concentrated in level 1, the decline in writing was most prominent in level 2, rather than level 1. Recall that the cutoffs in reading and math were level 2, which justify the declines in level 1. On the other hand, the writing cutoff of 3 induced the F schools to concentrate more on students expected to score in level 2 (i.e. closer to the cutoff) than in level 1. These patterns strongly suggest that the threatened schools focused on students expected to score below and close to the high stakes cutoffs.

A question that naturally arises in this context is whether the improvements of the lower performing students came at the expense of the higher performing ones. There is no evidence of such a pattern in

math or writing (except in the first year after program in math). In reading and in first year math there is a statistically significant decline in the percentage of students in level 5, but the effects are small. I later investigate whether this pattern continues to hold under a regression discontinuity analysis as well.

The computation of treatment effects above assumes that the D schools are not treated by the program. Although D schools do not directly face the threat of vouchers, they are close to getting an "F" and hence are likely to face an indirect threat. In such a case, the program effects shown in this paper (both difference-in-differences and regression discontinuity estimates) would be underestimates, but not overestimates.

5.1.1 Existence of Pre-Program Trends

The above estimates of the program effects will be biased if there are differential pre-program trends between F and D schools in the various levels. Using pre-program data, I next investigate the presence of such pre-program trends. In FCAT writing, pre-program data on percentage of students scoring in each of the different levels are available for the years 1994-1996. In FCAT reading and math, data on percentage of students scoring in levels 1 and 2 are available for the pre-program years 1998 and 1999.[13] To investigate the issue of pre-existing trends, I estimate the following regression as well as its fixed effects counterpart (that includes school by level fixed effects) using pre-program data:

$$P_{ijt} = \sum_j \beta_{0j} L_j + \sum_j \beta_{1j}(F * L_j) + \sum_j \beta_{2j}(L_j * t) + \beta_{3j} \sum_j (F * L_j * t) + \beta_{4j} X_{ijt} + \varepsilon_{ijt} \qquad (2)$$

where t denotes time trend, $j = \{1, 2, 3, 4, 5\}$ for writing and $j = \{1, 2\}$ for reading and math. The coefficients of interest here are β_{3j}.

Table 2, columns (1)-(2) report the results in reading, (3)-(4) in math and (5)-(6) in writing. The first column in each set reports the results from OLS estimation, the second from fixed effects estimation. There is no evidence of any differential trends in F schools relative to D schools in any of the levels and in any of the subject areas. Therefore it is unlikely that the previous results are biased by pre-program trends.

5.1.2 Mean Reversion

Another concern here is mean reversion. Mean reversion is the statistical tendency whereby high and low scoring schools tend to score closer to the mean subsequently. Since the F schools were by definition

[13] Data on percentage of students in all the five levels are available only from 1999.

the lowest scoring schools in 1999, it is natural to think that any decrease in the percentage of students in these levels (level 1 in reading and math; levels 1 and 2 in writing) after the program is contaminated by mean reversion. However, since I do a difference-in-differences analysis, my estimates of the program effect will be contaminated only if the F schools revert to a greater extent towards the mean than the D schools.

I use the following strategy to check for mean reversion in level 1. The idea is to measure the extent of decline, if any, in the percentage of students scoring in level 1 (in reading and math) in the schools that received an F grade in 1998 relative to the schools that received a D grade in 1998, during the period 1998-99. Since this was the pre-program period, this gain can be taken as the mean-reversion effect in level 1 for F schools relative to the D schools, and can be subtracted from the program effects previously obtained to arrive at mean reversion corrected effects. A similar strategy can be used to check mean reversion in the other levels.

The system of assigning letter grades to schools started in Florida in 1999. However, 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. These schools will henceforth be referred to as 98F and 98D schools respectively.[14] Using this sample of 98F and 98D schools, I investigate the relative changes, if any, in the percentage of students scoring in levels 1 and 2 in the 98F schools (relative to the 98D schools) during 1998-99.[15]

Table 3 reports the results for mean reversion in reading (columns (1)-(3)) and math (columns (4)-(6)). Relative to the 98D schools, there is no evidence of mean reversion of the 98F schools in either reading or math and in either level 1 or level 2.

5.1.3 Compositional Changes of Schools and Sorting

School level data brings with it the hazards of potential compositional changes of schools. In the presence of such changes, the program effects will be biased if the F schools were characterized by different compositional changes than the D schools. I investigate this issue further by examining whether the F

[14] Note that the mean percentages of students in the different levels in F and D schools in 1999 are very similar respectively to the corresponding mean percentages in 98F and 98D schools in 1998, which attests to the validity of this approach.

[15] Note that mean reversion in only levels 1 and 2 (in reading and math) can be assessed using this method, since data on percentages in the other levels are not available for 1998. Data on percentages in the different levels in writing are not available for 1998, which precludes the use of this method in writing. While data are available for the pre-program years 1994-97 in writing, the FCAT reading and math tests were not given then. Therefore, there is no way to impute F and D grades to schools in those years using the 1999 grading criteria. However, I also do a regression discontinuity analysis which serves to get rid of this mean reversion problem (if any).

schools exhibited differential shifts in demographic compositions after the program.

Another related issue is student sorting which can, once again, bias the results. None of the threatened schools received a second "F" grade in 2000 or 2001, therefore none of their students became eligible for vouchers. Therefore the concern about vouchers leading to sorting is not applicable here. However, the F and D grades can lead to a differential sorting of students in these two types of schools.[16] The above decline in percentage of students in lower levels in F schools relative to D schools could be driven by sorting if the F schools faced a relative flight of low performing students and a relative influx of high performing students in comparison to the D schools. There is no a priori reason as to why low performing and high performing students respectively would choose to behave in this way. However, note that F schools may have an incentive to encourage the low performing students to leave. Chakrabarti (2010) finds no evidence that there was a differential movement of special education students away from F schools (relative to the D schools) after the program. Also, note that if indeed the F schools successfully induced the low performing students to leave, this would likely be captured in changes in student composition of the school after the program.

However, to investigate this issue further as well as to directly address the potential problem of changes in school composition, I examine whether the demographic composition of the F schools saw a relative shift after the program as compared to the pre-program period. Using data from 1994-2002, I estimate the following regression (as well as its fixed effects counterpart):

$$y_{it} = \phi_0 + \phi_1 F + \phi_2 t + \phi_3 (F * t) + \phi_4 v + \phi_5 (v * t) + \phi_6 (F * v) + \phi_7 (F * v * t) + \varepsilon_{it} \tag{3}$$

where y_{it} represents the demographic characteristic of school i in year t and v is the program dummy, $v = 1$ if year$>$ 1999 and 0 otherwise. This regression investigates whether there has been any relative shift in demographic composition of the F schools in the post-program period after controlling for pre-program trends and post-program common shocks. The coefficients in the interaction terms $(F * v)$ and $(F * v * t)$ capture the relative intercept and trend shifts of the F schools.

Table 4 presents the estimation results for specification (3). The results reported include school fixed effects, the corresponding results from OLS are very similar and hence omitted. There is no evidence of any shift in the various demographic variables except for a modest positive intercept shift for Hispanics. However, if anything, this would lead to underestimates of the program effects. Moreover, the regressions

[16] 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.

in this paper control for any change in demographic composition. To sum, it is unlikely that the patterns seen above are driven by sorting.

5.1.4 "Threat of Vouchers" Versus Stigma

As discussed earlier, while on the one hand, the threatened schools (F schools) faced the "threat of vouchers", on the other they faced the stigma associated with getting the lowest performing grade "F". In this section, I discuss whether it is possible to separate out the two effects and whether it is possible to say whether the above effects were caused by the "threat of vouchers" or "stigma". I use the following strategies to investigate this issue.

First, although the system of assigning letter grades to schools started in 1999, Florida had an accountability system in the pre-1999 period which categorized schools into four groups 1-4 (1-low, 4-high) based on FCAT writing, and reading and math norm referenced test scores. The rationale behind this strategy is that is that if there was a stigma effect of getting the lowest performing grade, group 1 schools should improve in comparison to the group 2 schools even in the pre-program period. Using FCAT writing data for two years (1997 and 1998), I investigate the performance of schools that were categorized in group 1 in 1997 relative to the 1997 group 2 schools during the period 1997-98. While data on percentage of students in the different levels separately are not available for these two years, data on mean scores as well as data on percentage of students scoring in levels 1 and 2 together (that is, % of students below 3) are available for both years. I investigate trends in both mean scores and percentage of students below 3 for group 1 schools (relative to group 2 schools) during 1997-98 and compare these patterns with the post-program patterns. It should be noted here that the minimum criteria for writing in the pre-program period was exactly the same as in the post-program period and the cutoff in the pre-program period was also 3 (same as in the post-program period). So stigma effect (if any) would induce similar responses in both pre- and post-program periods,—fall in percentage of students below 3.[17]

[17] I do not use the pre-1999 reading and math norm referenced test (NRT) scores because different districts used different NRTs during this period, which varied in content and norms. Also districts often chose different NRTs in different years. Thus these NRTs were not comparable across districts and across time. Moreover, since districts could choose the specific NRT to administer each year, the choice was likely related to time varying (and also time-invariant) district unobservable characteristics which also affected test scores. Also note that this discussion assumes that if there is a stigma effect associated with getting an F, this would induce a relative improvement of F schools relative to D schools. However, it is not clear that this would be the case in the first place. Stigma is the "bad" label that is associated with getting an F. Since D schools were very close to getting F (more so, in the regression discontinuity analysis), and if F grade carries a stigma, then D schools should be threatened by the stigma effect also. In fact, one might argue that since D schools were unscarred while F schools were already scarred, the former might have a larger inducement to improve to avoid the scar.

Table 5 investigates whether there is a stigma effect of getting the lowest performing grade using pre-program FCAT writing scores. Columns (1)-(3) find that there is no evidence of any improvement in mean scores. In contrast, there was a large improvement in FCAT writing mean scores in the post-program period (Chakrabarti 2008a). Columns (4)-(6) look at the effect on percentage of students scoring in levels 1 and 2. Once again, there is no evidence of any stigma effect. Not only are the effects not statistically significant, but the magnitudes are also very small compared to the post-program patterns (table 1). This implies that threat of vouchers rather than stigma was the driving factor behind the post-program patterns seen above.

Second, all schools that received an F in 1999 received higher grades (A,B,C,D) in the years 2000, 2001. Therefore although stigma effect on F schools might have been operative in 2000, this was not likely to have been the case in 2001 or 2002 since none of the F schools got an F in the preceding year. As seen above, the patterns in the different levels were not specific to 2000 only, but similar patterns prevailed in 2001 and 2002 also. Since F schools continued to face the threat of vouchers till 2002, this provides further evidence in favor of the threat of voucher effect and against the stigma effect.

Third, I also use another strategy to investigate this issue. This strategy exploits the relationship between private school distribution around threatened schools and its relationship with threatened school response.[18] F schools that had more private schools in their near vicinity would likely lose more students if vouchers were implemented, and hence would face a greater threat of vouchers than those that had less. However, since stigma was a "bad" label associated with F, these schools would face the same stigma. Therefore if the response was caused by "threat of vouchers", then one would expect to see a greater response from F schools that had more private schools in their near vicinity. This, however would not be the case if the response was driven by stigma. To investigate this issue, I exploit the pre-program distribution of private schools, and investigate whether threatened schools that had more private schools in their immediate vicinity showed a greater response.

Using data from 1999 to 2002, I run the following fixed effects regression and its corresponding OLS counterpart (which also includes relevant lower level interactions and variables). The variable *number* represents the number of private schools within a certain radius and f_{ij} denotes school by level fixed effects.[19] The coefficients of interest here are θ_{4j},—they show the differential effects on F schools of

[18] I would like to thank David Figlio for suggesting this strategy.

[19] The results presented here relate to a one mild radius. But I have also experimented with 2 mile, 3 mile and 5 mile radii,–the results remain qualitatively similar and are available on request.

having an additional private school in its near vicinity on the various levels.

$$P_{ijt} = f_{ij} + \sum_{k=2000}^{2002} \sum_{j=1}^{5} \theta_{1kj}(D_k * L_j) + \sum_{k=2000}^{2002} \sum_{j=1}^{5} \theta_{2kj}(F * D_k * L_j) + \sum_{j=1}^{5} \theta_{3j}(v * L_j * number) + \sum_{j=1}^{5} \theta_{4j}(F * v * L_j * number) + \theta_{5j} X_{ijt} + \varepsilon_{ijt} \quad (4)$$

Table 6 investigates whether F schools that had more private schools in their near vicinity responded more to the program. Both OLS and fixed effects results in each of reading, math and writing indicate that this indeed has been the case. In reading and math, F schools with greater private school presence showed a higher fall in their percentages of students scoring in level 1, while in writing these schools showed a higher fall in level 2. This indicates that F schools that had more private school presence around them tended to focus more on students expected to score just below the high stakes cutoffs. This indicates that the effects above were driven by threat of vouchers rather than stigma.

To summarize, the above three strategies strongly suggest that the F-school effects obtained above were driven by "threat of vouchers" rather than stigma. Note though that even otherwise, the effects obtained above captures the effects of the whole program on F-schools, that is the effects of a combination of threat of vouchers and stigma generated by a voucher system that embeds vouchers in an accountability framework. (As outlined in the introduction, the objective of the paper is to identify the effect of the whole voucher program, that is the effect of an accountability tied voucher program on the threatened schools.)

5.1.5 Using Regression Discontinuity Analysis to Examine the Differential Focus on Students below Minimum Criteria

I also use a regression discontinuity 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 non-linear and discontinuous relationship between the percentage of students scoring above a pre-designated 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 where both failed to meet the minimum criteria in 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, while 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%—while schools below 50% would

face a direct threat, those above 50% would not face any such direct threat.

Using the sample of F and D schools that fail minimum criteria in both reading and math in 1999, Figure 6 Panel A illustrates the relationship between assignment to treatment (i.e. facing the threat of vouchers) and the schools' percentages of students scoring at or above 3 in FCAT writing. The figure shows that except one, all schools in this sample that had less than 50% of their students scoring at or above 3 actually received an F grade. Similarly, all schools (except one) in this sample that had 50% 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 6, 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, 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% mark.

I also consider two corresponding samples where both F and D schools fail the minimum criteria in reading and writing (math and writing). According to the Florida rules, F schools 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%. However, the sample sizes in the case of these samples are considerably smaller than above and the samples just around the cutoff are considerably less dense. So I focus on the first sample above, 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.

An advantage of a regression discontinuity analysis is that identification relies on a discontinuous jump in the probability of treatment at the cutoff. Consequently, a potential confounding factor such as mean reversion that is important in a difference-in-differences setting is not likely to be important here, as it likely varies continuously with the run variable (p_i) at the cutoff. Also, regression discontinuity 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 in a difference-in-differences setting, existence of differential pre-program trends, is not likely to be important here.

Consider the following model, where Y_i is school $i's$ outcome, T_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.

$$Y_i = \gamma_0 + \gamma_1 T_i + f(p_i) + \epsilon_i$$

Hahn, Todd and Van Der Klaauw(2001) show that γ_1 is identified by the difference in average outcomes of schools that just missed the cutoff and those that just made the cutoff, provided the conditional expectations of the other determinants of Y are smooth through the cutoff. Note that the interpretation of the treatment effect here is different from that in the above difference-in-differences analysis. Here, γ identifies the local average treatment effect (LATE) at the cutoff while the difference-in-differences analysis identifies the average treatment effect on the treated (ATT).

The estimation can be done in multiple ways here. 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 run variable and a dummy indicating whether or not the school falls below the 50% cutoff. I estimate alternate specifications that do not include controls as well as those that use controls. 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. Assuming the covariates are balanced (I later test this restriction), the purpose of inclusion of covariates is variance reduction. They are not required for the consistency of γ_1.

To test 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 50% cutoff. I also estimate alternative functional forms that include fifth order and seventh order polynomials instead of a third order polynomial and the corresponding interactions.[20] The results remain very similar in each case and are available on request.

Using the above 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 run variable (percentage of students

[20] 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.

scoring at or above 3 in 1999 FCAT writing) in the sample reported in this paper. As could be perhaps anticipated from Figure 6, I indeed find a sharp discontinuity at 50. The estimated discontinuity is 1 and it is very highly significant.

Next, I examine whether the use of a regression discontinuity strategy is valid here. As discussed above, identification of γ_1 requires that the conditional expectations of various pre-program characteristics are smooth through the cutoff. Using the strategy outlined above, I test if that was indeed the case. Note though that there is not much reason to expect manipulation or selection in this particular situation. The program was announced in June 1999 while the tests were given 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. So, it is unlikely that the schools had the time and information to manipulate percentage of students above certain cutoffs before the tests. Nevertheless, I check for continuity of pre-determined characteristics at the cutoff, using the strategy outlined above. The corresponding graphs are presented in Figure 7 and the discontinuity estimates in table 7a. The discontinuity estimates are never statistically distinguishable from zero. Visually examining the graphs, it seems that unlike in the cases of the other pre-determined characteristics, there is a small discontinuity in the variable, "percentage of school's students eligible for free or reduced price lunches". But the discontinuity is small and not at all statistically significant (with a p-value of 0.28). Also, note that even if it was statistically significant, with a large number of comparisons, one might expect a few to be statistically different from zero even by sheer random variation. So, from the above discussion, it seems reasonable to say that this case passes the test of smoothness of predetermined characteristics through the cutoff. Following McCrary (2008), I also test whether there is unusual bunching at the cutoff. Using density of the run variable (percentage of students at or above 3 in writing in 1999) and the strategy above, I test for discontinuity in the density of the run variable at the cutoff. As can be seen from table 7b, there is no evidence of a statistically significant discontinuity in the density function at the cutoff in 1999.

Having established that the use of regression discontinuity strategy in this setting is valid, I next look at the effect of the program on the behavior of threatened schools. Figures 8a, 8b and 8c respectively look at the effect of the program on percentage of students scoring in levels 1-5 in reading, math and writing respectively in the three years after program. The first column presents effects in 2000, the second column in 2001 and the third column in 2002. The corresponding regression discontinuity estimates are

presented in table 8. In both reading and math and in each of the years, the percentage of students scoring in level 1 dropped sharply just to the left of the 50% cutoff and these effects are statistically significant. These imply that the program led to a decline in the percentage of students scoring in level 1 in the threatened schools. Upward shifts are also visible in levels 2, 3 and 4. These patterns support the hypothesis that the schools tended to concentrate on students expected to score just below the minimum criteria cutoff in reading and math. And there was a movement of students from level 1 to the upper levels. Recall that as noted in section 5.1, the falls and the increases seen in the various levels are net changes. For example, as suggested by the table, an increase in percentage of students in level 2 is generated by movements from level 1. To the extent that there may have been movements from level 2 to the upper levels, the actual movements into level 2 are larger than that seen in the level 2 estimates. Also note that there may have been movements from the upper levels (3, 4 or 5) to level 2, but this does not seem to have been important as the net changes in the upper levels (and the cumulative net changes in the upper levels) have always been positive. This intuition applies to levels above 2 as well.

In writing, while declines are visible in both levels 1 and 2 to the left of the 50% cutoff, the decline in level 2 is substantially larger than that in level 1 in both 2000 and 2002. In 2001, the fall in level 2 is less than in level 1, but note that the fall in level 2 is a net fall. Since, it is likely that a major chunk of the level 1 students moved to level 2, the actual fall in level 2 in writing in each year is considerably larger than that seen in the estimates/graphs. These patterns confirm the earlier results and provide additional evidence that the threatened public schools concentrated more on the students expected to score below and close to the minimum criteria cutoffs.

A related question here is whether the improvement of the low performing students came at the expense of the higher performing ones. The previous difference-in-differences analysis showed some evidence of a small decline in percentage of students scoring in level 5 in reading and first year math, although not in writing. But, as seen in table 8, in the RD analysis, there is no evidence of any effects in level 5 except an increase in 2002 writing. Thus, it does not seem that the improvements of the low performing students came at the expense of the higher performing ones.

To sum up, there is strong evidence that the threatened schools concentrated more on marginal students (i.e., students expected to score below and close to the minimum criteria cutoffs) and there have been perceptible statistically significant declines in the percentages of students just below the minimum criteria cutoffs. This pattern holds in all the three subjects—reading, math and writing. But there is

no evidence that the increased focus of attention on the marginal students adversely affected the higher performing ones. Rather, there seems to have been a rightward shift of the entire score distribution in each of reading, math and writing, although the improvements were concentrated in score ranges just below the respective minimum criteria cutoffs. (A possible explanation of the rightward shift of the entire distribution is that the program induced the schools to become more efficient in general.)

5.2 Choosing between Subjects with Different Extents of Difficulties Versus Focusing on Subjects Closer to the Cutoff

For each F school, I first rank the subject areas in terms of their distances from the respective subject cutoffs. Distance of a subject from the respective subject cutoff is defined as the difference between the percentage of students scoring below the cutoff in that subject in 1999 and the percentage required to pass the minimum criteria in that subject. Next, based on the ranks of the subjects, I generate three dummies, "low", "mid" and "high". "Low" takes a value of 1 if the subject is closest to the cutoff, 0 otherwise; "mid" takes a value of 1 if the subject is second in terms of distance from the cutoff, 0 otherwise; "high" takes a value of 1 if the subject is farthest from the cutoff, 0 otherwise. The analysis in this section will combine the reading, math and writing scores (percent scoring below minimum criteria) in a single model. Therefore, for purposes of analysis in this section, I standardize the reading, math and writing scores by grade, subject and year to have means of 0 and standard deviations of 1.

Using the sample of F schools and data from 1999 and 2000, I estimate the following model:

$$y_{ist} = \gamma_0 read + \gamma_1 math + \gamma_2 write + \gamma_3 low + \gamma_4 mid + \gamma_5(read * D00) + \gamma_6(math * D00) + \gamma_7(write * D00) + \gamma_8(low * D00) + \gamma_9(mid * D00) + \gamma_{10} X_{ist} + \varepsilon_{ist} \quad (5)$$

where y_{ist} represents the percentage of students below minimum criteria cutoff (standardized by grade, subject and year) in school i subject s in year t; $read$, $math$ and $write$ are subject dummies that take a value of 1 for the corresponding subject and 0 otherwise; and X_{ist} denotes the set of control variables. Control variables include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditure and interactions of the subject dummies with these variables. $High$ is taken to be the omitted category. The coefficients $\gamma_5 - \gamma_9$ capture the program effects. If the F schools focused on subject areas on the basis of their distances from the cutoff then γ_8, $\gamma_9 < 0$ and $|\gamma_8| > |\gamma_9|$. On the other hand, if the schools choose to focus on a certain subject area, then the coefficient of the interaction term between that subject and 2000 year dummy will be negative and larger in magnitude than the other corresponding interaction terms. I also estimate the fixed effects counterpart of this model that includes

school by subject fixed effects (and hence does not have subject and distance rank dummies which are absorbed).

Table 9 presents the results from estimation of model 5. While columns (1)-(2) present the results without controls, columns (3)-(4) present those with controls. Controls include racial composition of schools, gender composition of schools, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of the subject dummies with these variables. The first column of each set reports results from OLS estimation and the second column from fixed effects estimation.

There is no evidence that the threatened schools concentrated most on the subject closest to the cutoff. The coefficients of the relevant interaction terms are actually positive and are never different from zero statistically. Nor are they statistically different between themselves, as seen in the last row of table 9.

In each of the columns, the first three coefficients indicate a decline in the percentage of students scoring below the minimum criteria cutoffs in each of the three subjects. However, the decline in writing by far exceeds the corresponding declines in the other two subjects. As the p-values indicate, this decline in writing exceeds the declines in reading and math statistically also. To summarize, this table finds no evidence in favor of the hypothesis that the threatened schools concentrated most on the subject closest to the cutoff. Rather the schools seem to have disproportionately favored FCAT writing. While there are improvements in each of the three subject areas, the improvement in writing is substantially larger than that in the other two subject areas both economically and statistically.

I next explore these issues further by disaggregating the above effects. Did the F schools choose to focus on writing because of its relative ease, irrespective of its rank? To investigate this question, I estimate the following model as well as the fixed effects counterpart of it that includes school by subject fixed effects. The coefficients of interest here are $\delta_5 - \delta_{13}$.

$$y_{ist} = \delta_0 read + \delta_1 math + \delta_2 write + \delta_3 low + \delta_4 mid + \delta_5(low * D00 * read) + \delta_6(low * D00 * math) + \delta_7(low * D00 * write) + \delta_8(mid * D00 * read) + \delta_9(mid * D00 * math) + \delta_{10}(mid * D00 * write) + \delta_{11}(high * D00 * read) + \delta_{12}(high * D00 * math) + \delta_{13}(high * D00 * write) + \delta_{14} X_{ist} + \varepsilon_{ist} \quad (6)$$

Table 10a investigates whether the F schools chose to focus on writing irrespective of its distance from the cutoff (relative to the other subjects). It presents results from estimation of model 6. Columns (1)-(2) report results from specifications without controls, while columns (3)-(4) include controls. There are

declines in the percentage of students scoring below the cutoffs in each of the three subjects, irrespective of their distances from the cutoffs. However, these declines are largest in magnitude for writing and holds irrespective of whether writing has a rank of "low", "mid" or "high". For example, the decline in writing for "F" schools which were closest to the cutoff in writing ("low" in writing) exceeded the decline in reading (math) for schools that were "low" in reading (math), "mid" in reading (math) or "high" in reading (math). The scenario is exactly the same when writing ranks "mid" or "high". Note that these improvements are not only economically larger, but as table 10b shows, they are statistically so too. Moreover, the improvements in the different subjects do not have a definite hierarchy or a one-to-one relationship with distances from the cutoff.

Regression Discontinuity Analysis: A problem with the above analysis is that it cannot rule out the fact that focus on writing is due to a year specific shock to that subject area (or other subject areas). For example, one can argue that the F schools chose to concentrate on writing because it suddenly became easier. A way out of this problem is to do a regression discontinuity analysis that compares the effect on schools just below the cutoff to those just above the cutoff between F and D. I use the regression discontinuity strategy discussed above for this purpose and also use the above discontinuity samples. The argument here is that if there was indeed some year specific shocks to one or more subject areas, they would be faced by both the schools below as well as above the cutoff.

The results of this analysis are reported in table 11 and the corresponding graphs are presented in Figure 9. The first column reports results without controls while the second column includes controls. While there is a fall in the percentage of students scoring below the corresponding minimum criteria in each of the subject areas, this fall is the largest in writing. In addition, not only economically, but the falls in writing were statistically different from those in math and reading as well. In other words, the results in this table are consistent with those above and confirm the above findings. F schools in the discontinuity sample just below the cutoff tended to focus by far more in writing than in the other subject areas (relative to schools just above the cutoff), ruling out the fact that the focus on writing was due to year specific shocks.[21]

To sum, the F schools chose to concentrate on writing not because of any year specific shocks and irrespective of its distance from the cutoff, presumably because it was easiest to improve in. Case studies

[21] While the results reported here are for discontinuity samples 1 and 2, results are similar (and available on request) for the other discontinuity samples mentioned above (section 5.1.5) where both F and D schools fail the minimum criteria in reading and writing (math and writing) and F schools fail the minimum criteria in math (reading) also, unlike D schools.

reported in Goldhaber and Hannaway (2004) are very much consistent with this picture: *'Writing came first "because this is the easiest to pass"..."With writing there's a script; it's pretty much first we did this, then we did this, and finally we did that, and using that simple sequencing in your writing you would get a passing grade."'*

Telephone interviews conducted by me with school administrators in several F schools in different Florida districts also show a similar picture. They reveal widespread beliefs among school administrators that writing scores were much easier to improve in than reading and math scores. They say that they focused on writing in various ways after the program. They established a "team approach in writing" which introduced writing across the curriculum. This approach incorporated writing components in other subject areas also such as history, geography, etc. to increase the students' practice in writing. They also report that they introduced school wide projects in writing, longer time blocks in writing, and writing components in lower grades.

6 Conclusion

This paper analyzes the behavior of public schools facing a voucher system that embeds vouchers in an accountability regime. It focuses on the 1999 Florida program. Utilizing the institutional details of the program, it analyzes the incentives built into the system, and examines the behavior of public schools facing these incentives.

It focuses on two alternative ways in which the program incentives might have induced the threatened schools to behave. First, certain percentages of a school's students had to score above some pre-designated thresholds on the score scale to escape the second F grade. As a result, did the threatened schools tend to focus more on students below these cutoffs rather than equally on all students? Second, as per the program details, to avoid an F grade, schools needed to pass the minimum criteria in only one of the three subjects. Did this induce the threatened schools to focus more on one subject area rather than equally on all? If so, did they choose to focus on the subject area closest to the high stakes cutoffs? Alternatively, did they choose to focus on a specific subject that was perceived to be the easiest irrespective of the distances of the subject areas from the thresholds?

I find robust evidence that the threatened schools concentrated more on students expected to score just below the high stakes cutoffs and focused much more on writing compared to reading and math. The latter is consistent with the notion among Florida administrators that writing scores were considerably

easier to improve than scores in reading and math. Moreover, although the threatened schools focused more on students expected to score below the minimum criteria cutoffs, the improvement of the lower performing students does not seem to have come at the expense of the higher performing ones. Rather, there seems to have been a rightward shift of the entire score distribution in each of reading, math and writing with improvements more concentrated in score ranges just below the minimum criteria cutoffs.

These findings are informative from a policy point of view. They strongly suggest that the F schools responded to the incentives built into the system. This implies that policy can be appropriately targeted to affect public school behavior and to induce schools to behave in desirable ways. For example, if more attention on reading and math is warranted, it calls for a change in grading rules to give less weight to writing and more to reading and math. If more attention on comparatively higher performing students is desired, in addition to emphasis on low performing students, this calls for an inclusion of higher performing student scores in computation of F and D grades. Interestingly, two of the major elements of the grading criteria changes that went into effect in 2002 were to reduce the weight of writing and to increase those of reading and math; and extension of emphasis to scores of comparatively higher performing students also.

Effective policy making calls for an understanding of the responses of agents to specific rules of the policy, so that the lessons learnt can be used to create a more effective and stronger policy. This paper has contributed to this learning process and the findings promise to be valuable from the point of view of public school reform.

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 Number 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 volume 92, Number 5-6, 1371-1393.

Chakrabarti, Rajashri (2010),"Public School Incentives and the Test-Taking Population: Evidence from a Regression Discontinuity Analysis," mimeo, Federal Reserve Bank of New York.

Chay, Kenneth, Patrick McEwan and Miguel Urquiola (2005), "The central role of noise in evaluating interventions that use test scores to rank schools," American Economic Review, 95(4),

1310-1326.

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

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

Fan, Jianqing and Irene Gijbels (1996), "Local Polynomial Modeling and Its Applications", Chapman and Hall, London.

Figlio, David and Lawrence Getzler (2006), "Accountability, Ability and Disability: Gaming the System?", in T. Gronberg ed., Advances in Microeconomics, Elsevier.

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

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

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

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

Greene, Jay and Marcus Winters (2003), "When Schools Compete: The Effects of Vouchers on Florida Public School Achievement," Education Working Paper 2.

Greene, Jay (2001), "An Evaluation of the Florida A-Plus Accountability and School Choice Program," New York: Manhattan Institute for Policy Research.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression Discontinuity Design," Econometrica 69 (1): 201-209.

Holmstrom, B., and P. Milgrom (1991), "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," Journal of Law, Economics, and Organization, 7, 24-52.

Howell, William and Paul Peterson (2005), "The Education Gap: Vouchers and Urban Schools, Revised Edition", Washington D.C., Brookings Institution Press.

Hoxby, Caroline (2003a), "School Choice and School Productivity (Or, Could School Choice be the

tide that lifts all boats?)", in Caroline Hoxby (ed.) The Economics of School Choice, University of Chicago Press.

Hoxby, Caroline (2003b), "School Choice and School Competition: Evidence from the United States", Swedish Economic Policy Review 10, 11-67.

Imbens, Guido W., and Thomas Lemieux (2008), "Regression Discontinuity Designs: A guide to practice", Journal of Econometrics, 142(2), 615-635.

Jacob, Brian (2005), "Accountability, Incentives and Behavior: The Impacts of High-Stakes Testing in the Chicago Public Schools", Journal of Public Economics, 89, 761-796.

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

McCrary, Justin (2008),"Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," Journal of Econometrics, 142 (2): 698-714.

McMillan, Robert (2004),"Competition, Incentives, and Public School Productivity," Journal of Public Economics, 88, 1871-1892.

Neal, Derek and Diane W. Schanzenbach (2010),"Left Behind By Design: Proficiency Counts and Test-Based Accountability," The Review of Economics and Statistics, 92(2): 263-283.

Nechyba, Thomas (2003), "Introducing School Choice into Multi-District Public School Systems", in Caroline Hoxby (ed.), The Economics of School Choice, University of Chicago Press, Chicago.

Reback, Randall (2005), "Teaching to the Rating: School Accountability and Distribution of Student Achievement," Working Paper, Barnard College, Columbia University.

Rouse, Cecilia E. (1998), "Private School Vouchers and Student Achievement: Evidence from the Milwaukee Choice Program," Quarterly Journal of Economics 113(2), 553-602.

Silverman, Bernard W. (1998), "Density Estimation for Statistics and Data Analysis," New York: Chapman and Hall, 1986.

West, Martin and Paul Peterson (2006), "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments", The Economic Journal 116 (510), C46-C62.

Table 1: Effect of "Threatened Status" on percentage of students scoring in levels 1-5

(Sample of treated F and control D schools, Reading, Math and Writing)

	Reading		Math		Writing	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * level 1 * 1 year after program	-3.33***	-3.49***	-5.56***	-5.73***	-6.95***	-6.95***
	(1.25)	(1.16)	(1.44)	(1.45)	(1.27)	(1.21)
Treated * level 2 * 1 year after program	0.96	0.70	2.83***	2.69***	-10.07***	-10.21***
	(0.88)	(0.87)	(0.75)	(0.75)	(0.92)	(0.83)
Treated * level 3 * 1 year after program	2.32***	2.42***	2.17*	2.30*	11.04***	10.95***
	(0.73)	(0.61)	(1.23)	(1.20)	(1.26)	(1.35)
Treated * level 4 * 1 year after program	0.83	0.99	0.99**	1.13**	6.52**	6.70***
	(0.70)	(0.71)	(0.42)	(0.48)	(2.60)	(2.56)
Treated * level 5 * 1 year after program	-0.70***	-0.54***	-0.56***	-0.51**	0.39	0.37
	(0.14)	(0.17)	(0.22)	(0.23)	(0.46)	(0.47)
Treated * level 1 * 2 years after program	-4.13***	-2.78***	-7.52***	-7.20***	-6.78***	-6.59***
	(1.35)	(0.97)	(1.79)	(1.93)	(1.20)	(1.10)
Treated * level 2 * 2 years after program	2.21**	1.56*	3.73***	3.24***	-9.52***	-8.87***
	(0.87)	(0.86)	(0.96)	(1.05)	(1.17)	(1.20)
Treated * level 3 * 2 years after program	2.23**	1.78**	2.63***	2.72***	10.15***	9.89***
	(1.07)	(0.87)	(1.00)	(0.96)	(2.21)	(2.46)
Treated * level 4 * 2 years after program	0.58	0.30	1.24	1.39*	5.70**	5.33**
	(0.46)	(0.44)	(0.78)	(0.80)	(2.53)	(2.53)
Treated * level 5 * 2 years after program	-0.92***	-0.90***	-0.16	-0.22	1.26	1.10
	(0.31)	(0.30)	(0.30)	(0.34)	(0.95)	(0.98)
Treated * level 1 * 3 years after program	-5.47***	-4.73***	-5.96***	-5.10***	-7.05***	-7.06***
	(1.25)	(1.07)	(1.74)	(1.76)	(1.06)	(0.94)
Treated * level 2 * 3 years after program	2.86***	2.41***	4.48***	4.34***	-10.51***	-10.16***
	(0.74)	(0.60)	(0.86)	(0.83)	(1.34)	(1.31)
Treated * level 3 * 3 years after program	3.18***	2.89***	0.44	0.58	12.95***	12.94***
	(0.84)	(0.89)	(0.69)	(0.64)	(1.60)	(1.63)
Treated * level 4 * 3 years after program	0.45	0.35	0.10	0.17	5.60***	5.49***
	(0.60)	(0.63)	(0.80)	(0.90)	(1.72)	(1.80)
Treated * level 5 * 3 years after program	-0.99***	-0.91**	-0.13	-0.09	0.19	0.07
	(0.35)	(0.36)	(0.55)	(0.58)	(0.63)	(0.68)
Controls	Y	Y	Y	Y	Y	Y
Observations	10110	10110	10035	10035	10105	10105

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is percentage of students in school i scoring in level j in year t. The regression results are obtained from estimation of model 1 and its fixed effects counterpart. All regressions are weighted by the number of students tested. OLS regressions include level dummies and interactions of the level dummies with treated dummy and year dummies respectively. The FE columns include school by level fixed effects and interactions of level dummies with year dummies. Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of level dummies with each of these variables.

Table 2: Pre-program trend of F schools in levels 1-5, relative to D schools

	Reading		Math		Writing	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * level 1 * trend	0.65	0.60	0.76	1.09	0.39	0.57
	(1.60)	(1.41)	(1.58)	(1.59)	(0.67)	(0.78)
Treated * level 2 * trend	-0.06	0.01	1.48*	1.39	0.35	0.38
	(0.68)	(0.67)	(0.90)	(0.86)	(0.35)	(0.39)
Treated * level 3 * trend					-0.39	-0.37
					(0.56)	(0.58)
Treated * level 4 * trend					-0.16	-0.13
					(0.10)	(0.12)
Treated * level 5 * trend					-0.09	-0.08
					(0.05)	(0.05)
Controls	Y	Y	Y	Y	Y	Y
Observations	2030	2030	2020	2020	7150	7150

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is percentage of students in school i scoring in level j in year t. All regressions are weighted by the number of students tested. OLS regressions include level dummies and interactions of the level dummies with treated dummy and year dummies respectively. The FE columns include school by level fixed effects and interactions of level dummies with year dummies. This table reports results from estimation of model 2 and its fixed effects counterpart. Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of level dummies with each of these variables. Pre-program data are available only for levels 1 and 2 in reading and math.

Table 3: Mean reversion of the 98F schools, relative to 98D schools

dep. var. = % of students scoring in level i in school j in year t, i = {1,2}

	Reading			Math		
	OLS	OLS	FE	OLS	OLS	FE
98F * level 1 * trend	-1.70	-1.59	-1.56	0.64	0.26	0.14
	(1.52)	(1.51)	(1.49)	(1.88)	(1.98)	(1.79)
98F * level 2 * trend	0.50	0.41	0.42	2.21	2.09	2.07
	(0.89)	(0.90)	(0.91)	(1.35)	(1.38)	(1.29)
Controls	N	Y	Y	N	Y	Y
Observations	2728	2710	2710	2728	2710	2710

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is percentage of students in school i scoring in level j in year t. All regressions are weighted by the number of students tested. The table uses the sample of 98F and 98D schools. Pre-program data are available only for levels 1 and 2 in reading and math. OLS regressions include level dummies, interactions of level dummies with treated dummy and trend respectively. The FE columns include school by level fixed effects and interactions of level dummies with trend. Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of level dummies with each of these variables.

Table 4: The Issue of Sorting: Investigating demographic shifts
(Sample of F and D schools, 1994-2002)

	% white	% black	% hispanic	% asian	% american indian	% free/reduced price lunch eligible
	FE	FE	FE	FE	FE	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * program dummy	-1.64	-0.55	1.99**	-0.04	0.01	-0.16
	(1.12)	(1.11)	(0.95)	(0.18)	(0.10)	(1.27)
Treated * program * trend	0.84	-0.92	0.20	0.02	-0.01	-0.54
	(0.61)	(0.57)	(0.53)	(0.08)	(0.04)	(0.92)
Observations	4498	4498	4498	4498	4498	3076

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is the relevant demographic characteristic of school i in year t. This table reports results from the estimation of the fixed effects counterpart of model 3. All regressions include school fixed effects and also include trend, program dummy, interactions of trend with treated dummy and program dummy respectively.

Table 5: Is there a Stigma Effect of getting the Lowest Performing Grade? Effect of being Categorized in Group 1 on Performance in FCAT Writing
(Using Mean Scores and % of Students Scoring below 3, 1997-1998)

	Sample: Group 1, 2 Schools					
	dep. var. = mean score			dep. var. = % below 3 (i.e., % in levels 1 and 2)		
	OLS	FE	FE	OLS	FE	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Group 1 * trend	-0.01	-0.02	-0.02	-2.09	-2.06	-1.81
	(0.08)	(0.06)	(0.06)	(3.47)	(3.30)	(3.04)
Controls	N	N	Y	N	N	Y
Observations	314	314	314	314	314	314

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions are weighted by the number of students tested. Controls include racial composition of students, gender composition of students, percentage of students eligible for free or reduced-price lunches and real per pupil expenditure. OLS regressions include group 1 dummy and trend, FE regressions include school fixed effects and trend. The dependent variable in columns (1)-(3) range on a scale of 1-6; the dependent variable in columns (4)-(6) is a percentage.

Table 6: "Threat of Vouchers" Versus Stigma: Does Threatened School Response Change with Number of Private Schools in Near Vicinity?

	Reading		Math		Writing	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * level 1 *	-1.74***	-1.42**	-1.13**	-1.19**	-0.70	-0.70
program dummy * number	(0.67)	(0.66)	(0.57)	(0.55)	(0.63)	(0.62)
Treated * level 2 *	0.53*	0.48**	-0.42	-0.57	-1.81***	-1.96***
program dummy * number	(0.28)	(0.22)	(0.55)	(0.47)	(0.62)	(0.70)
Treated * level 3 *	0.36	0.17	0.20	0.30	-1.33	-1.27
program dummy * number	(0.37)	(0.39)	(0.43)	(0.26)	(1.20)	(1.15)
Treated * level 4 *	0.50*	0.42	0.55	0.64	3.31***	3.36***
program dummy * number	(0.30)	(0.32)	(0.36)	(0.46)	(0.90)	(0.95)
Treated * level 5 *	0.32**	0.30*	0.83***	0.85***	1.33***	1.35***
program dummy * number	(0.14)	(0.15)	(0.19)	(0.17)	(0.35)	(0.36)
Controls	Y	Y	Y	Y	Y	Y
Observations	9990	9990	9915	9915	9985	9985

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. The regression results are obtained from estimation of model 4 and its OLS counterpart. All regressions are weighted by the number of students tested. Controls include racial composition of students, gender composition of students, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of level dummies with each of these variables.

Table 7a: Testing Validity of Regression Discontinuity Analysis: Looking for Discontinuities in Pre-Program Characteristics at the Cutoff

Panel A	% White	% Black	% Hispanic	% Asian	% American Indian
	(1)	(2)	(3)	(4)	(5)
	2.92	-5.06	2.43	0.09	-0.16
	(7.24)	(11.39)	(6.73)	(0.28)	(0.06)
Panel B	% Multiracial	Male	% Free/Red. Pr. Lunch	Enrollment	Real PPE
	(6)	(7)	(8)	(9)	(10)
	-0.23	-1.21	-5.97	-14.45	-1.97
	(0.26)	(1.44)	(5.36)	(60.32)	(2.29)

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Standard errors are in parentheses and are clustered by the run variable (% of school's students at or above 3 in FCAT writing).

Table 7b: Testing for Discontinuity in the Density of the Run Variable

	1999
	(1)
Difference	-0.01
	(0.01)

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Standard errors are in parentheses and are clustered by the run variable (% of school's students at or above 3 in FCAT writing).

Table 8: Regression Discontinuity Analysis: Effect of "Threatened Status" on percentage of students scoring in levels 1-5, Reading, Math and Writing

Discontinuity Estimates	Reading		Math		Writing	
	(1)	(2)	(3)	(4)	(5)	(6)
1 year after program						
Level 1	-7.77**	-6.73***	-5.22*	-5.16**	-2.90***	-3.21***
	(3.42)	(2.28)	(2.88)	(2.12)	(0.99)	(0.75)
Level 2	0.21	0.29	-0.15	-0.35	-5.84*	-5.01*
	(1.32)	(1.24)	(1.28)	(1.88)	(3.07)	(3.01)
Level 3	3.24	2.61	2.52*	2.70**	1.29	0.90
	(2.55)	(1.83)	(1.35)	(1.18)	(3.05)	(2.89)
Level 4	4.22***	3.80***	3.12*	2.99***	4.14	4.24
	(1.29)	(0.42)	(1.66)	(1.00)	(2.62)	(2.85)
Level 5	0.05	0.05	-0.47	-0.42	1.71	1.69
	(0.47)	(0.42)	(0.38)	(0.42)	(1.35)	(1.47)
2 years after program						
Level 1	-12.25***	-11.15**	-13.16***	-11.07***	-3.85*	-3.88
	(4.06)	(5.08)	(3.04)	(3.30)	(2.26)	(2.37)
Level 2	4.11	4.73*	5.95***	5.77***	-1.92	-1.48
	(2.50)	(2.48)	(1.58)	(1.48)	(1.63)	(1.90)
Level 3	4.21*	3.52	3.42**	2.02	6.16*	6.33*
	(2.21)	(2.74)	(1.34)	(1.77)	(3.25)	(3.55)
Level 4	3.17**	2.50*	3.25*	2.84*	-0.59	-1.26
	(1.37)	(1.47)	(1.88)	(1.45)	(2.28)	(1.97)
Level 5	0.56	0.19	0.28	0.21	1.07	1.30
	(0.81)	(0.57)	(0.93)	(0.72)	(1.02)	(1.43)
3 years after program						
Level 1	-9.88**	-10.73***	-5.50	-5.41	-0.82	-1.33
	(4.07)	(3.72)	(8.77)	(9.32)	(1.91)	(1.83)
Level 2	1.22	1.55	0.21	1.22	-3.56***	-3.47**
	(2.06)	(1.92)	(1.63)	(1.57)	(1.31)	(1.43)
Level 3	3.92**	4.37*	1.80	1.58	-3.21	-1.30
	(1.69)	(2.24)	(3.19)	(3.38)	(4.71)	(4.10)
Level 4	4.73**	4.92***	2.86	2.29	2.22	2.02
	(1.97)	(1.47)	(4.17)	(4.49)	(2.61)	(2.85)
Level 5	0.06	-0.11	0.71	0.43	4.05***	3.70**
	(0.64)	(0.64)	(1.40)	(1.48)	(1.09)	(1.22)
Controls	N	Y	N	Y	N	Y

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Standard errors are in parentheses and are clustered by the run variable (% of school's students at or above 3 in FCAT writing). Controls include racial and gender composition of students, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures.

Table 9: Do Threatened Public Schools focus on the subject closest to cutoff?

	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)
Reading * 1 year after program	-0.32***	-0.33***	-0.34***	-0.27**
	(0.09)	(0.09)	(0.08)	(0.11)
Math * 1 year after program	-0.65***	-0.63***	-0.60***	-0.56***
	(0.12)	(0.12)	(0.11)	(0.10)
Writing * 1 year after program	-1.27***	-1.28***	-1.26***	-1.27***
	(0.20)	(0.21)	(0.24)	(0.16)
Low * 1 year after program	0.28	0.28	0.25	0.32*
	(0.19)	(0.20)	(0.21)	(0.18)
Mid * 1 year after program	0.20	0.18*	0.18	0.14
	(0.13)	(0.09)	(0.11)	(0.10)
Controls	N	N	Y	Y
Observations	390	390	378	378
p-values of differences:				
(Reading * 1 year after - Writing * 1 year after)	0.00	0.00	0.00	0.00
(Math * 1 year after - Writing * 1 year after)	0.00	0.00	0.00	0.00
(Low * 1 year after - Mid * 1 year after)	0.61	0.48	0.29	0.15

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. This table uses percentage of students below minimum criteria in reading, math and writing, each standardized by grade, subject and year to have a mean of zero and standard deviation of 1. The dependent variable is percentage of students below minimum criteria cutoff (standardized by grade, subject and year) in school i in subject s in year t. All regressions are weighted by the number of students tested. The regresion results are obtained from the estimation of model 5 and its fixed effects counterpart. The OLS columns include the three subject dummies, low and mid dummies. The FE columns include school by subject fixed effects. Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditure and interactions of the subject dummies with these variables.

Table 10a: Further delineating the behavior of public schools: Does subject rank matter?

	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)
Low * Reading * Year 2000	-0.85***	-0.82***	-0.22	-0.68**
	(0.15)	(0.17)	(0.20)	(0.30)
Low * Math * Year 2000	-0.26	-0.21	-0.01	-0.09
	(0.20)	(0.15)	(0.25)	(0.18)
Low * Writing * Year 2000	-1.15***	-1.19***	-1.18***	-1.17***
	(0.08)	(0.09)	(0.09)	(0.13)
Mid * Reading * Year 2000	0.01	0.01	-0.05	0.08
	(0.07)	(0.10)	(0.12)	(0.13)
Mid * Math * Year 2000	-0.39***	-0.43***	-0.45***	-0.44***
	(0.08)	(0.08)	(0.09)	(0.09)
Mid * Writing * Year 2000	-1.55***	-1.5***	-1.39***	-1.49***
	(0.26)	(0.25)	(0.29)	(0.17)
High * Reading * Year 2000	-0.25***	-0.31***	-0.35***	-0.24**
	(0.07)	(0.07)	(0.07)	(0.10)
High * Math * Year 2000	-0.83***	-0.72***	-0.61***	-0.59***
	(0.18)	(0.17)	(0.18)	(0.13)
High * Writing * Year 2000	-1.26***	-1.10***	-1.02***	-1.22***
	(0.27)	(0.38)	(0.31)	(0.41)
Controls	N	N	Y	Y
Observations	390	390	378	378

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. This table uses percentage of students below minimum criteria in reading, math and writing, each standardized by grade, subject and year to have a mean of zero and standard deviation of 1. The dependent variable is percentage of students below minimum criteria cutoff (standardized by grade, subject and year) in school i in subject s in year t. All regressions are weighted by the number of students tested. The OLS columns include the three subject dummies, low and mid dummies. The FE columns include school by subject fixed effects. Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditure and interaction of the subject dummies with these variables. The regression results are obtained from the estimation of model 6 and its fixed effects counterpart.

Table 10b: Further delineating the behavior of public schools: Does subject rank matter?

	(1)	(2)	(3)	(4)
p-values of differences:				
(Low * Writing * Year 2000) - (Low * Reading * Year 2000)	0.10	0.09	0.10	0.10
(Low * Writing * Year 2000) - (Low * Math * Year 2000)	0.00	0.00	0.00	0.00
(Low * Writing * Year 2000) - (Mid * Reading * Year 2000)	0.00	0.00	0.00	0.00
(Low * Writing * Year 2000) - (Mid * Math * Year 2000)	0.00	0.00	0.00	0.00
(Low * Writing * Year 2000) - (High * Reading * Year 2000)	0.00	0.00	0.00	0.00
(Low * Writing * Year 2000) - (High * Math * Year 2000)	0.09	0.00	0.01	0.00
(Mid * Writing * Year 2000) - (Mid * Reading * Year 2000)	0.00	0.00	0.00	0.00
(Mid * Writing * Year 2000) - (Mid * Math * Year 2000)	0.00	0.00	0.01	0.00
(Mid * Writing * Year 2000) - (Low * Reading * Year 2000)	0.06	0.02	0.00	0.01
(Mid * Writing * Year 2000) - (Low * Math * Year 2000)	0.00	0.00	0.00	0.00
(Mid * Writing * Year 2000) - (High * Reading * Year 2000)	0.00	0.00	0.00	0.00
(Mid * Writing * Year 2000) - (High * Math * Year 2000)	0.00	0.00	0.00	0.00
(High * Writing * Year 2000) - (High * Reading * Year 2000)	0.00	0.06	0.04	0.03
(High * Writing * Year 2000) - (High * Math * Year 2000)	0.21	0.38	0.31	0.17
(High * Writing * Year 2000) - (Low * Reading * Year 2000)	0.10	0.50	0.00	0.31
(High * Writing * Year 2000) - (Low * Math * Year 2000)	0.00	0.04	0.01	0.02
(High * Writing * Year 2000) - (Mid * Reading * Year 2000)	0.00	0.01	0.01	0.00
(High * Writing * Year 2000) - (Mid * Math * Year 2000)	0.00	0.10	0.06	0.08

Columns (1), (2), (3) and (4) respectively correspond to columns (1), (2), (3) and (4) of table 10a. P-values reported give the p-values of the F-tests that the differences of the corresponding coefficients in table 10a are zero.

Table 11: Regression Discontinuity Analysis: Did the Threatened Schools Focus more on Writing?

	Dependent Variable: % Below Minimum Criteria Cutoff	
	(1)	(2)
Reading	-0.46**	-0.40***
	(0.20)	(0.14)
Math	-0.33*	-0.33**
	(0.19)	(0.14)
Writing	-0.62**	-0.58**
	(0.28)	(0.29)
Controls	N	Y
p-values of differences:		
Writing - Reading	0.00	0.00
Writing - Math	0.00	0.00

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Standard errors are in parentheses and are clustered by the run variable (% of school's students at or above 3 in writing). Controls include racial and gender composition of students, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures.

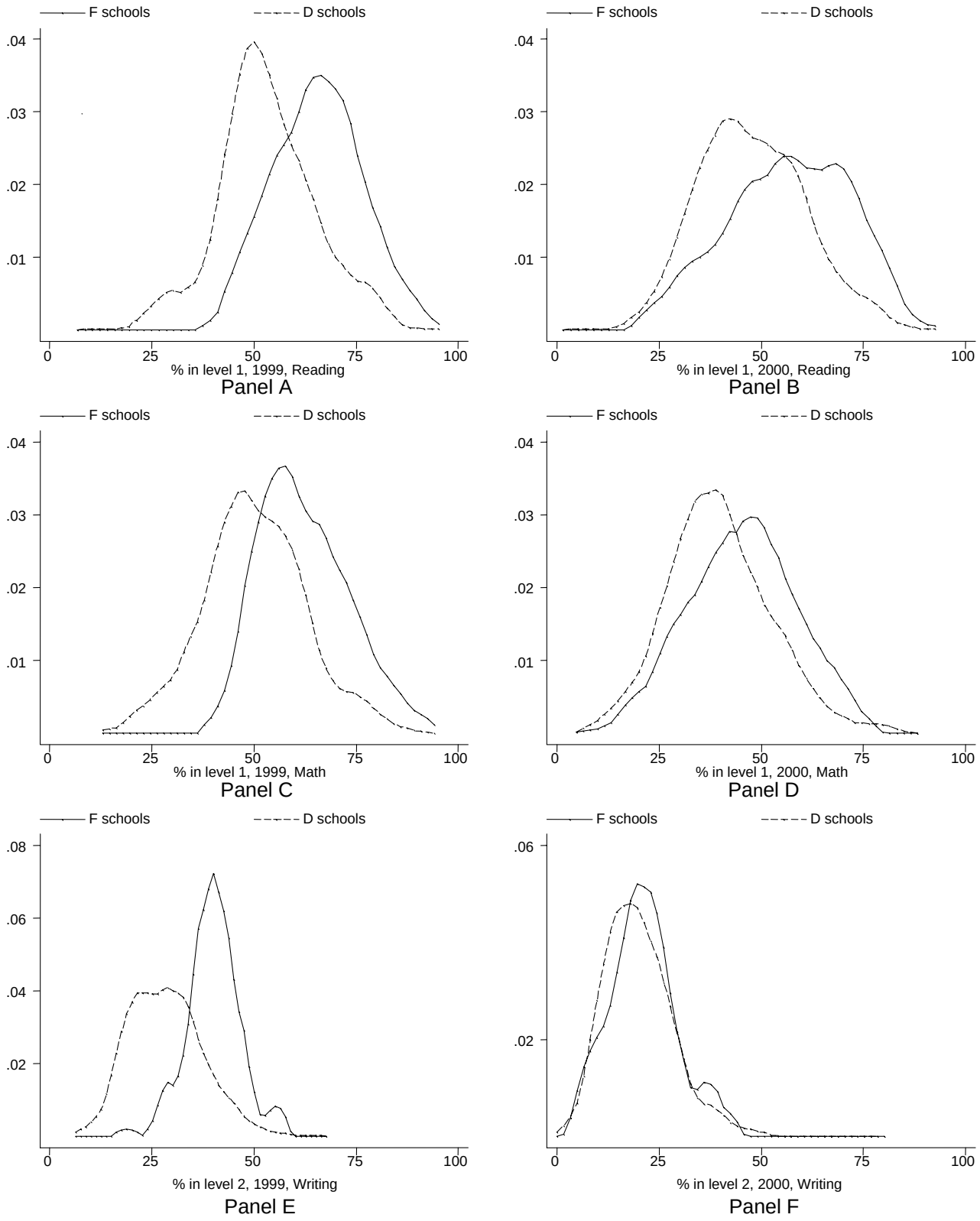
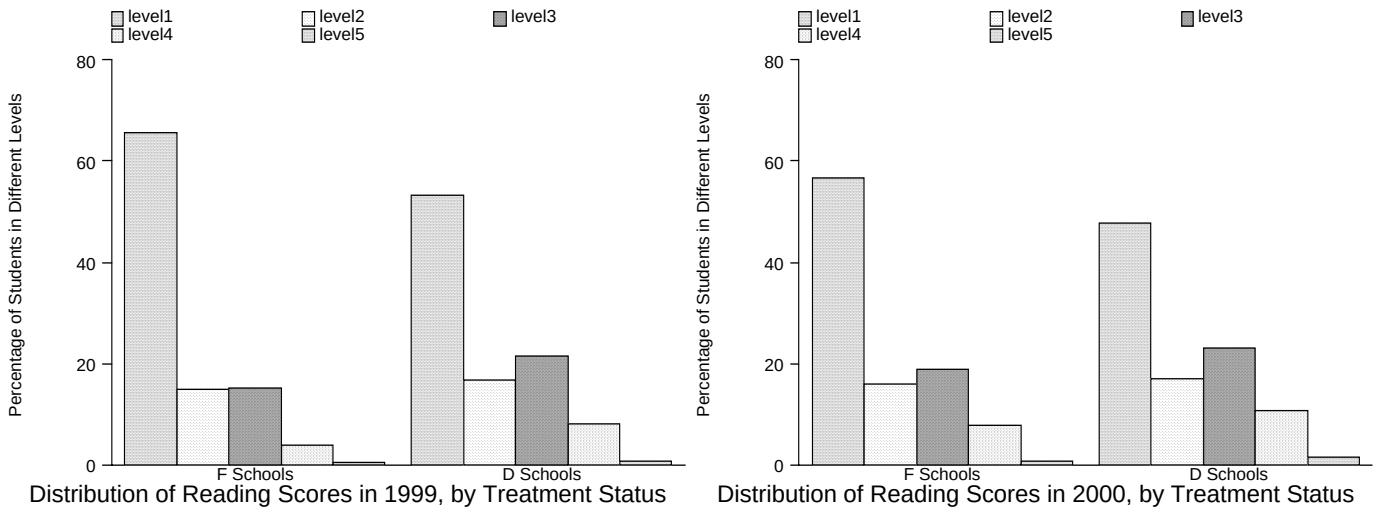
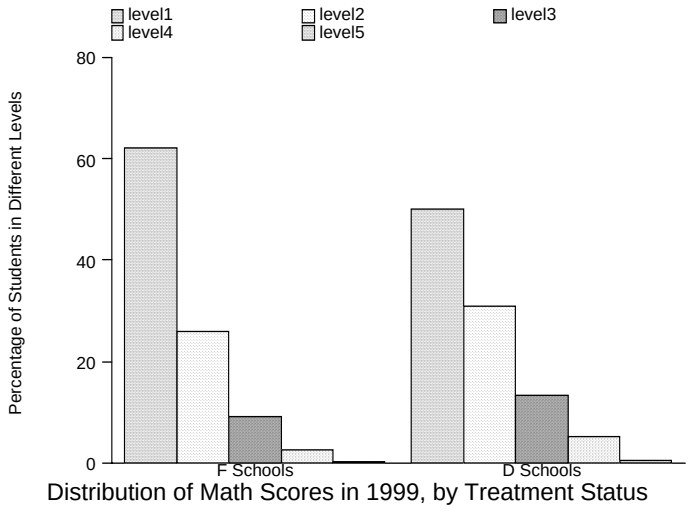
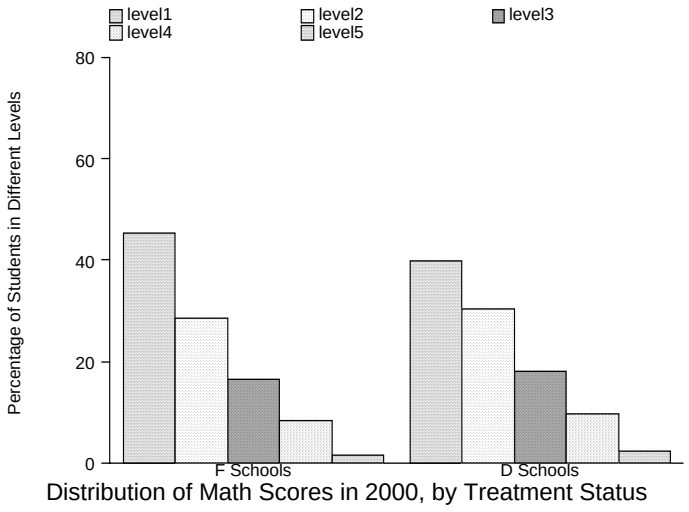
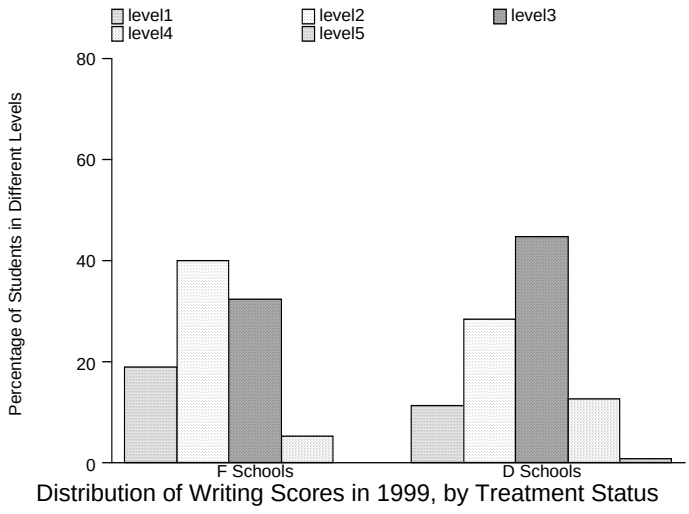


Figure 1. Distribution of percentage of students in level 1 Reading, level 1 Math and level 2 Writing, F and D Schools, 1999 and 2000









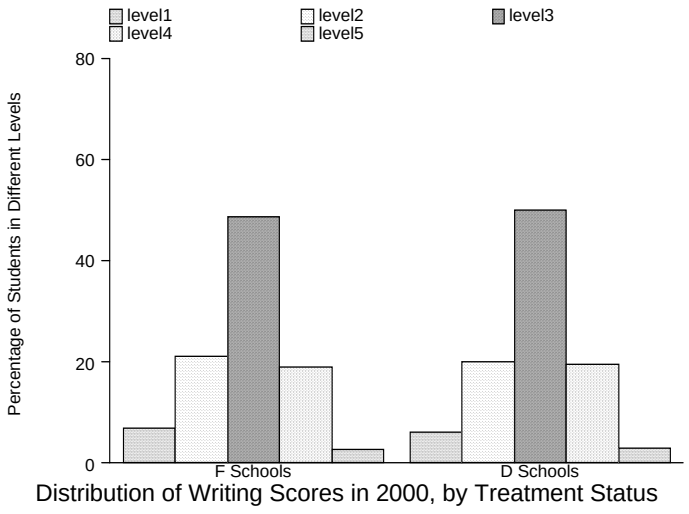


Figure 2. Distribution of Reading, Math and Writing Score for F and D schools (1999 and 2000)

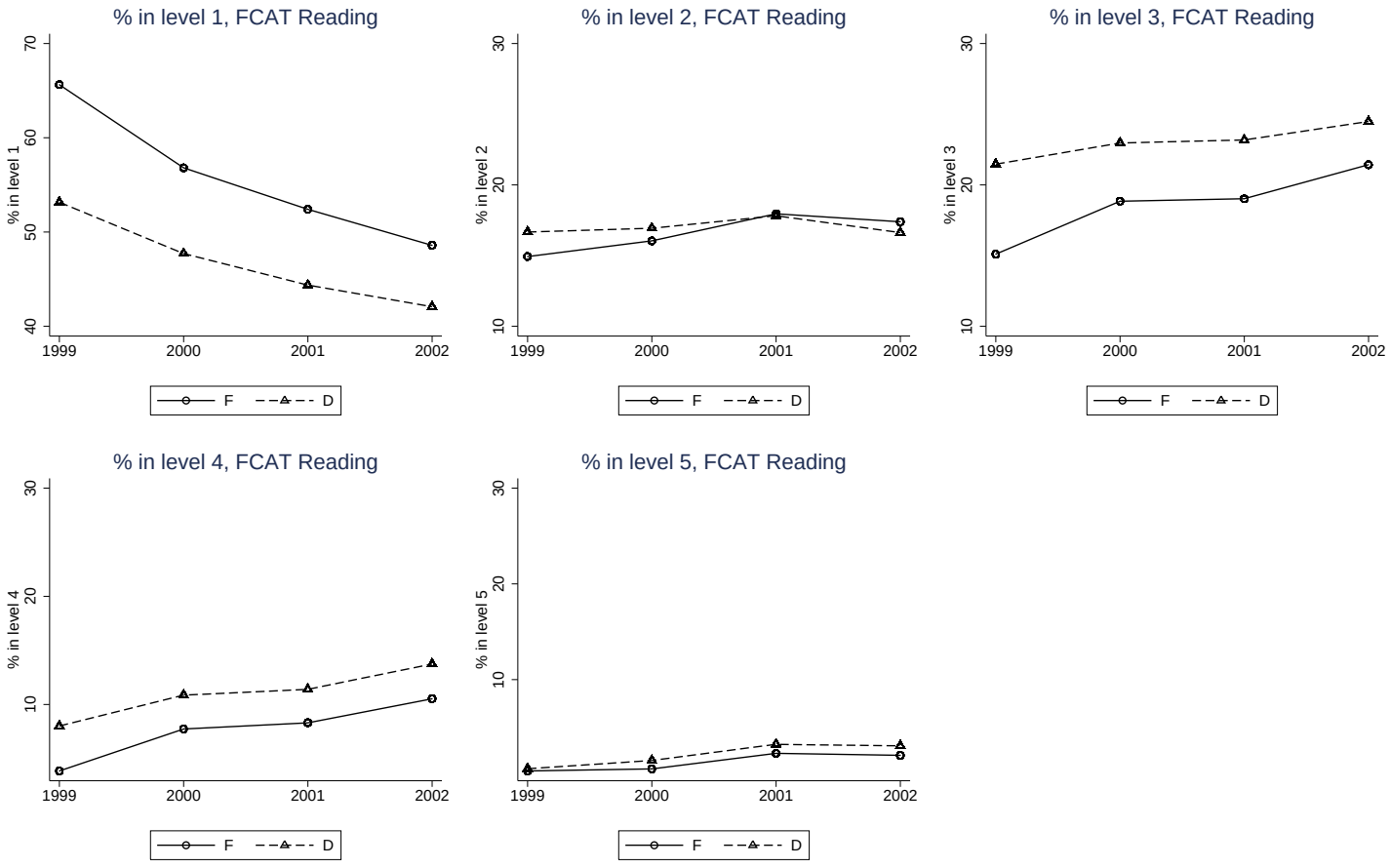


Figure 3. Percentage of Students in Levels 1−5, FCAT Reading, F and D Schools

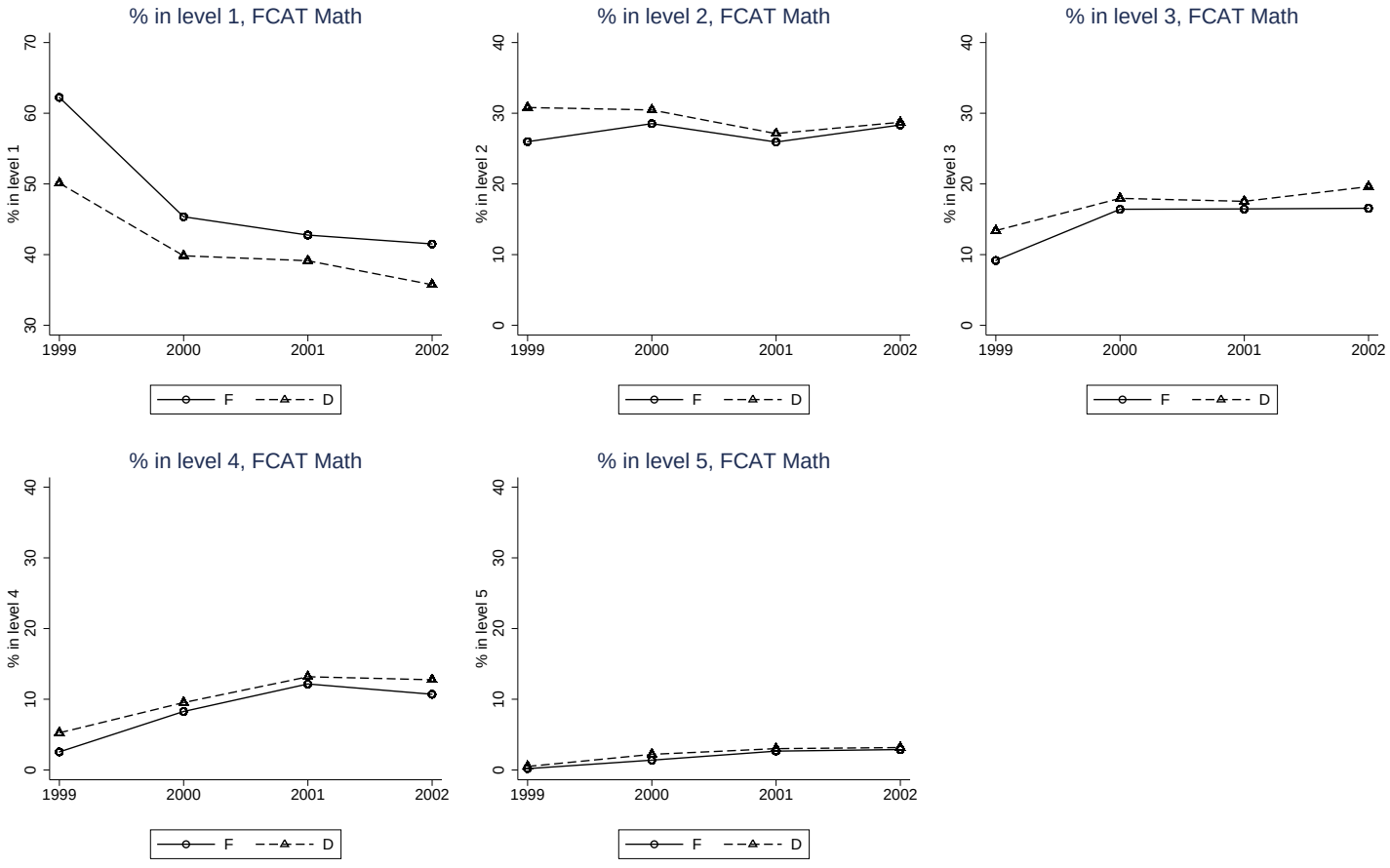


Figure 4. Percentage of Students in Levels 1−5, FCAT Math, F and D Schools

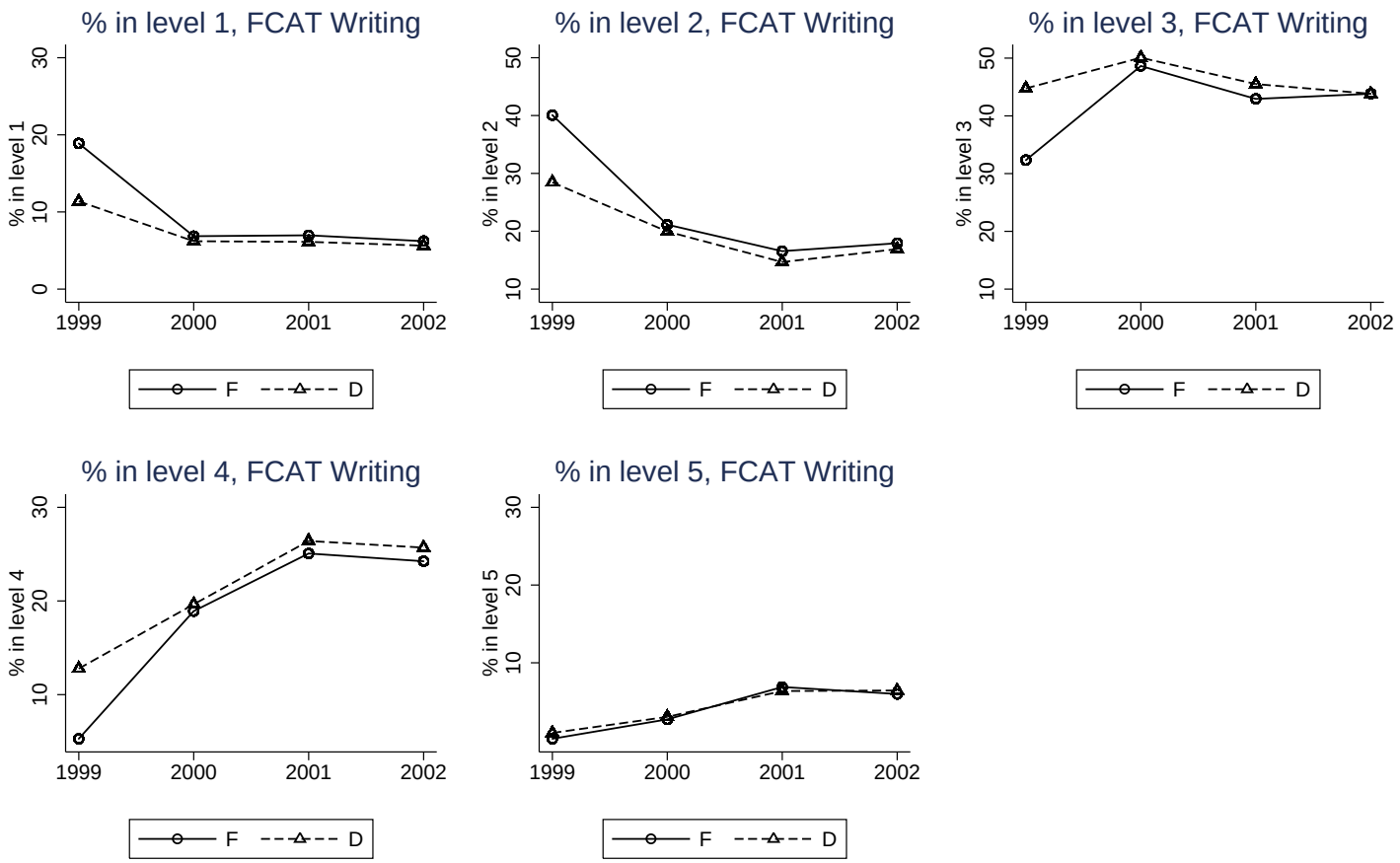


Figure 5. Percentage of Students in Levels 1−5, FCAT Writing, F and D Schools

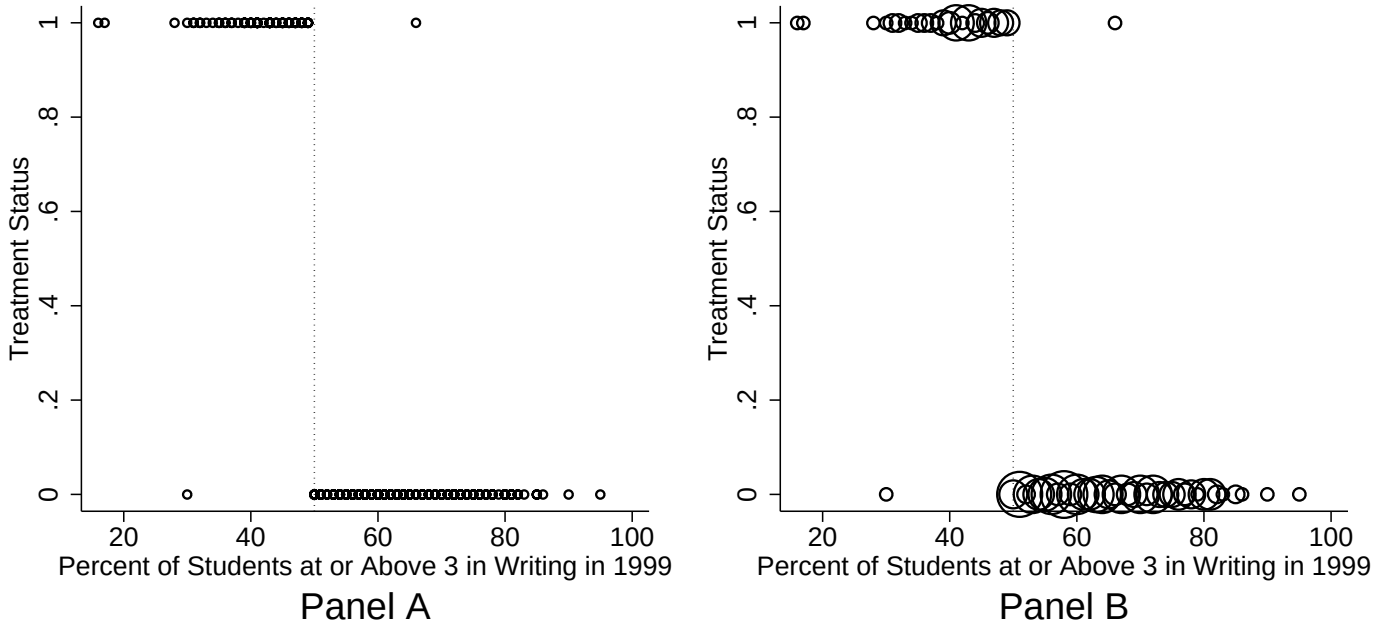


Figure 6. Regression Discontinuity Analysis:
Relationship Between % of Students at or Above 3 in Writing and Treatment Status

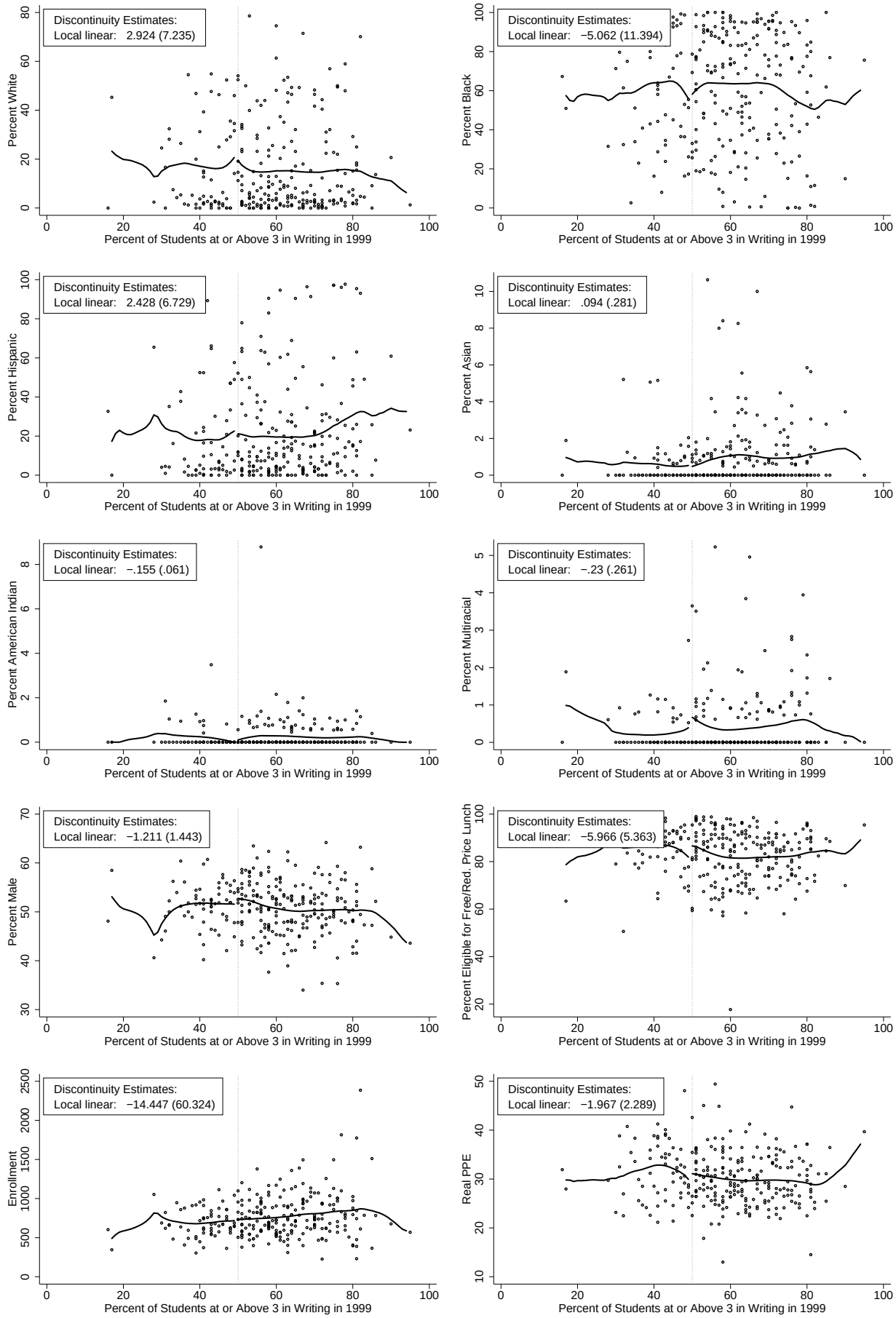


Figure 7. Testing Validity of Regression Discontinuity Design:
Pre−Program Characteristics Relative to the Cutoff

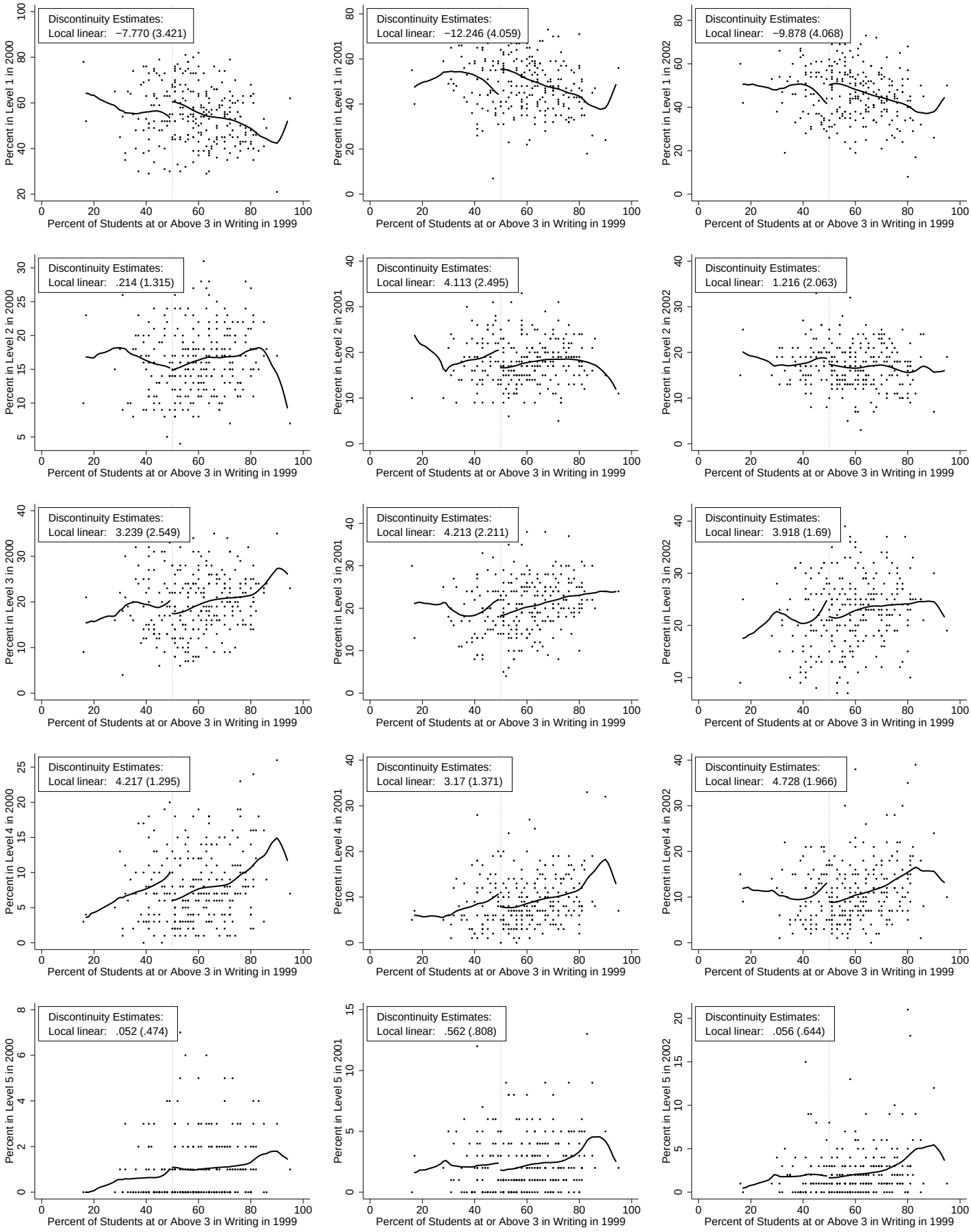


Figure 8a. Regression Discontinuity Analysis:
Effect of Treatment Status on FCAT Reading, Levels 1−5

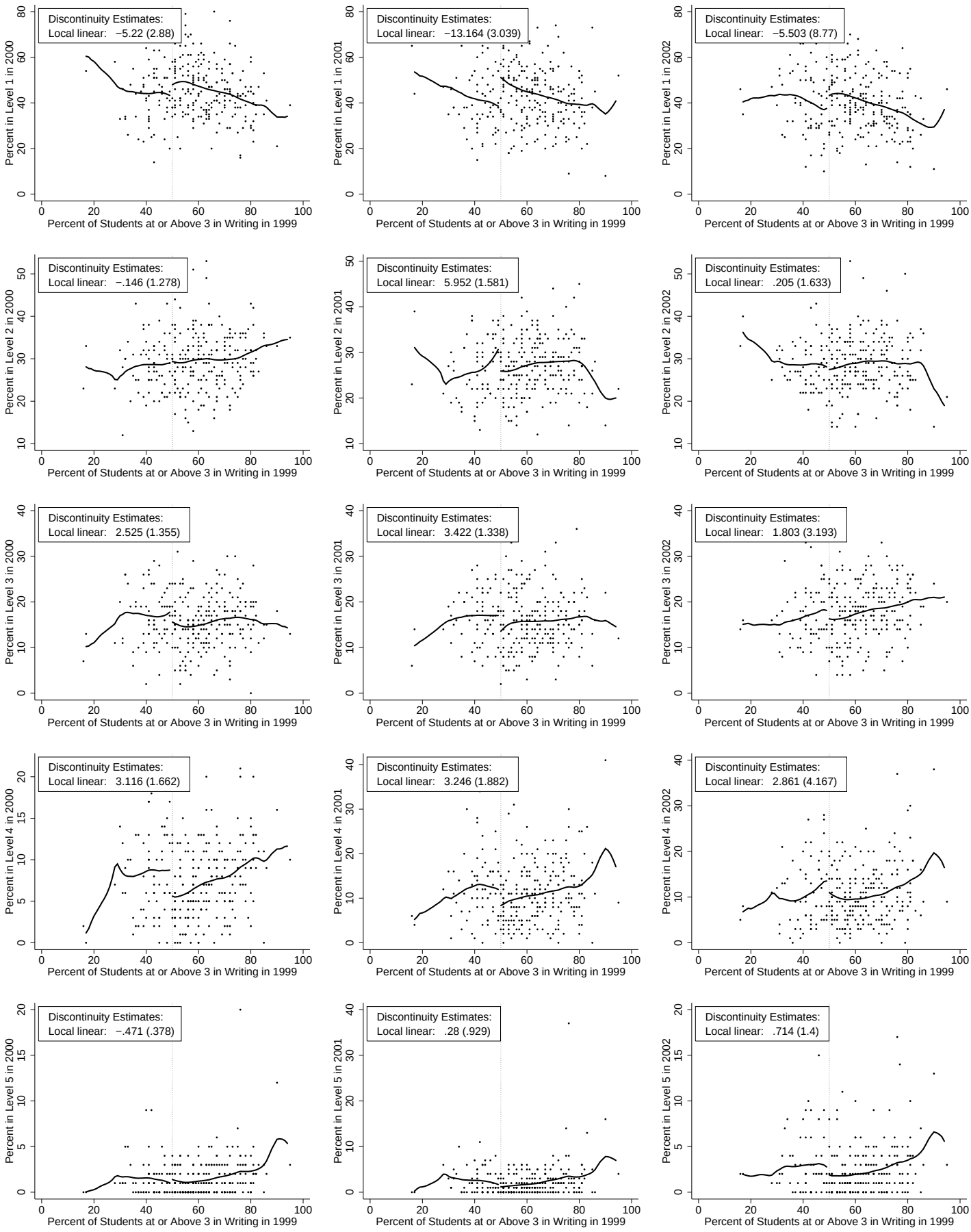


Figure 8b. Regression Discontinuity Analysis:
Effect of Treatment Status on FCAT Math, Levels 1−5

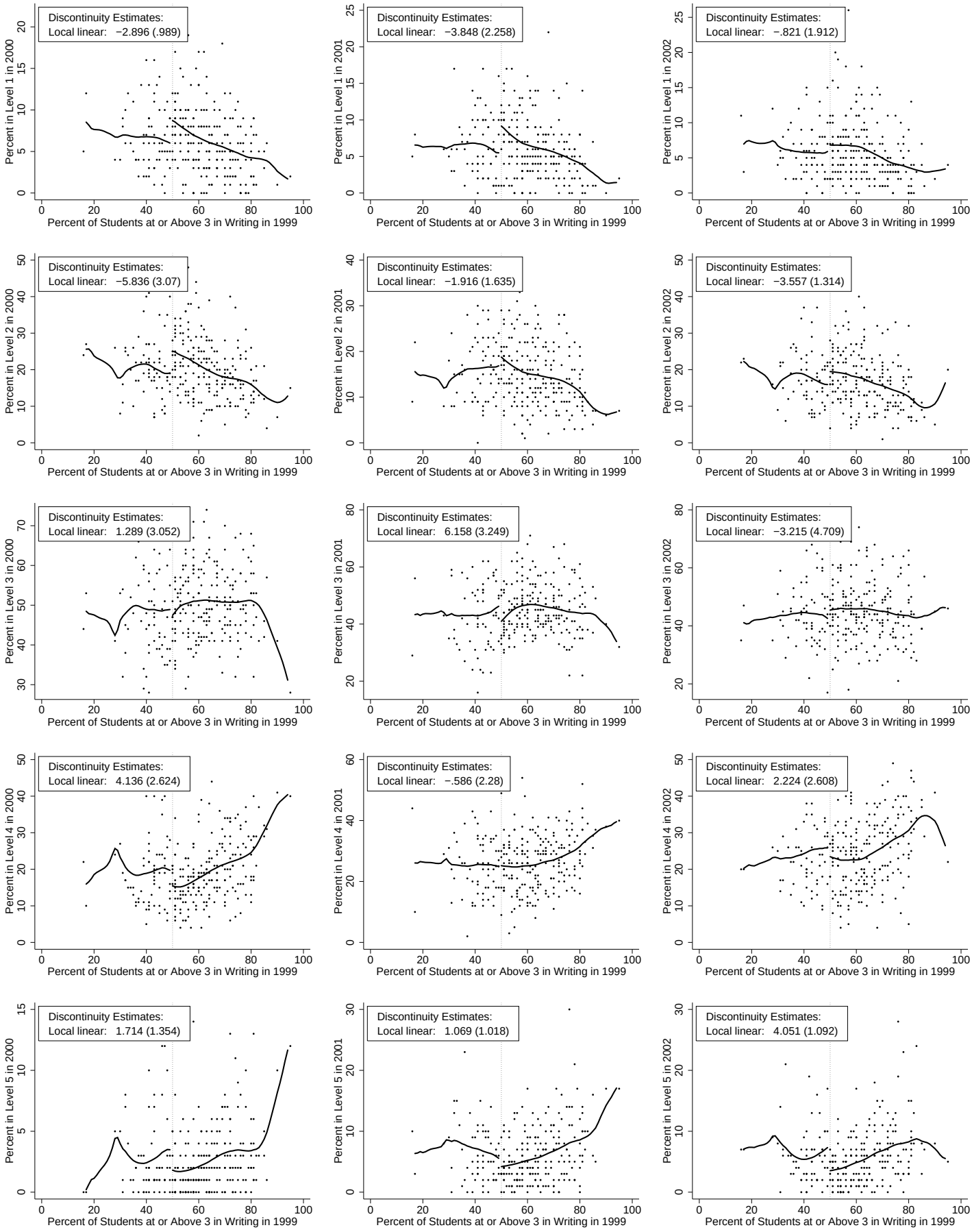
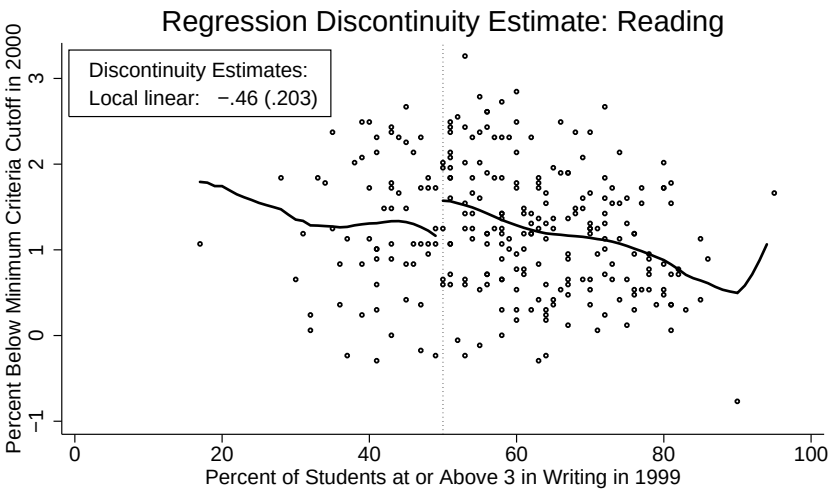
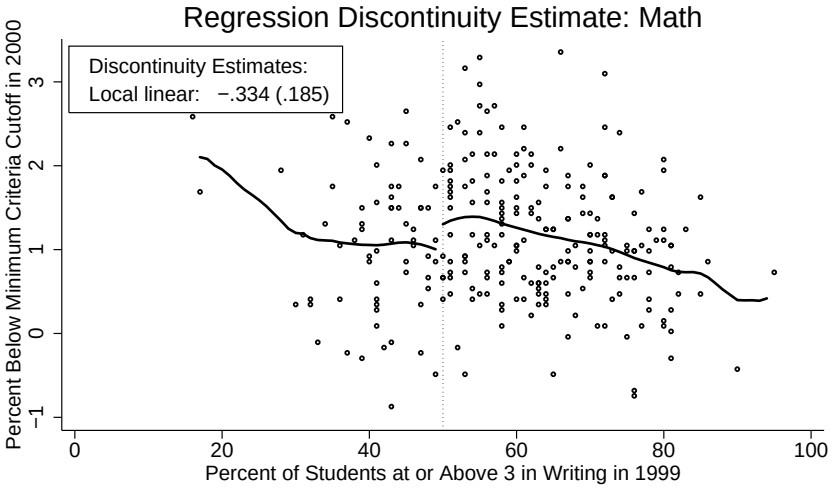


Figure 8c. Regression Discontinuity Analysis:
Effect of Treatment Status on FCAT Writing, Levels 1−5





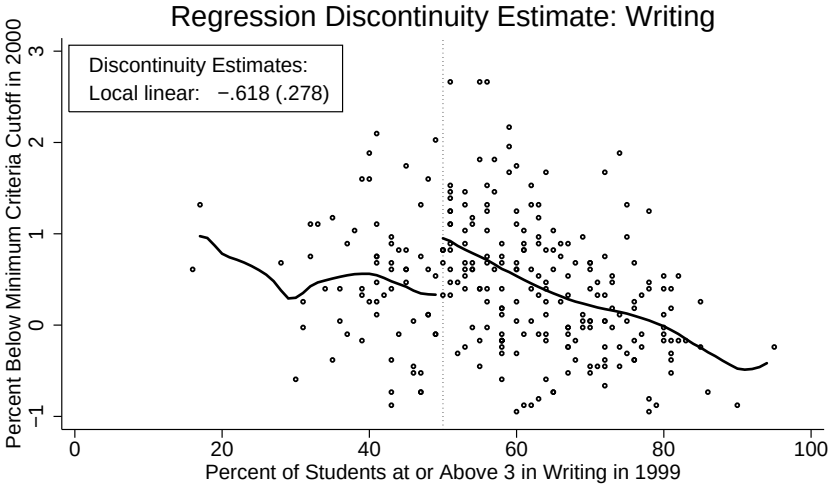


Figure 9. Regression Discontinuity Analysis: Did the Threatened Schools Focus More on Writing?

Note: The variable in the vertical axis is the percent of students below the minimum criteria cutoff in 2000 in the relevant subject area standardized by grade, subject, and year.