Vouchers, Public School Response and the Role of Incentives: Evidence from Florida*

Rajashri Chakrabarti†
Harvard University

Abstract

In this paper, I analyze the behavior of public schools facing vouchers. 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 threat of vouchers in Florida behaved. Florida schools getting an "F" grade are exposed to the threat of vouchers, while vouchers are implemented if they get another "F" grade in the next three years. Exploiting the institutional details of the 1999 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 as well as a regression discontinuity analysis, I find that the threatened schools tended to focus more on students below the minimum criteria cutoffs rather than equally on all, but interestingly, this improvement did not come at the expense of higher performing students. Second, consistent with incentives, they focused mostly on writing rather than reading and math. Finally, there is not much evidence of relative reclassification of low performing students into special education categories exempt from the calculation of grades. This is consistent with substantial costs associated with such reclassification during that period. These results are robust to controlling for differential pre-program trends, changes in demographic compositions, mean reversion and sorting. These findings have important policy implications and subsequent grading rule changes in Florida suggest that these changes have been a response to public school behavior.

Keywords: Vouchers, Incentives, Regression Discontinuity, Mean Reversion
JEL Classifications: H4, I21, I28

*I thank Steve Coate, Julie Cullen, Sue Dynarski, Ron Ehrenberg, David Figlio, Ed Glaeser, William Howell, Caroline Hoxby, Brian Jacob, Bridget Long, Paul Peterson, Shannon Seitz, Miguel Urquiola, Mike Waldman and seminar participants at Harvard University, Econometric Society conference, Association for Public Policy Analysis and Management conference, Society of Labor Economists conference for helpful discussions and the Florida Department of Education for providing the data used in this analysis. I would also like to thank the Program on Education Policy and Governance at Harvard University for its postdoctoral support. All errors are my own.

†J. F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. Email: rchakrab@fas.harvard.edu

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 vouchers 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, makes all students of a school eligible for vouchers if the school gets 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 are threatened by vouchers, but vouchers are implemented only if they get another "F" grade in the next three years. Vouchers are associated with a loss in revenue and also media publicity and visibility. Therefore, the threatened schools have a strong incentive to try to avoid the second "F", and thereby avoid vouchers. 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".[2] Therefore the threatened schools had an incentive to focus more on students expected to score just below these high stakes cutoffs

[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.

[2] The institutional details of the Florida program including the grading criteria are discussed in detail later. The Florida grading criteria underwent some important changes in 2002, as described later.

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.[3] 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. Third, according to Florida rules, scores of students in several special education categories (Exceptional Student Education (ESE) categories) were not included in the computation of grades. As a result, did the threatened schools tend to reclassify their low performing students into these "exempt" categories so as to remove them from the relevant test-taking pool and artificially boost scores?

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? The grading rules in Florida underwent some major changes in 2002. Did these 2002 policy changes bear any relationship to the public school response to the 1999 program?

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 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 are considerably

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

easier to improve than scores in reading or math. Finally, there is not much evidence that the threatened schools resorted to reclassification into "exempt" ESE categories after the program. As discussed in the paper, reclassification into ESE categories was associated with substantial costs during this period,[4] which might have induced the schools to focus more on other, less costly alternatives. These results are quite robust in that they withstand several sensitivity tests, and the results from the difference-in-differences analysis are qualitatively similar to those obtained from the regression discontinuity analysis. They are robust to controlling for pre-program trends, mean reversion, sorting, changes in demographic compositions and other observable characteristics of schools.

These findings strongly suggest that the threatened schools responded to incentives which in turn imply that policy can be targeted to shape public school behavior. Interestingly, the 2002 policy changes seem to have been largely a response to public school behavior. The new grading system gave less weight to writing scores and more to reading and math scores. Moreover, while under the 1999 grading system, the F grade and movement to a D depended solely on the percentages of students scoring below the minimum criteria cutoffs, the 2002 grading system also included the performance of comparatively higher performing students, while continuing to emphasize the performance of lower performing students. In contrast, the rules relating to the inclusion of the different special education categories in grade formation did not change. These policy changes suggest that there have been an interaction between public school response and policy in Florida.

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. Cullen and Reback (2002), Figlio and Getzler (2002) and Jacob (2005) show that schools facing such 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. Figlio (2003) 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

[4] Among other things, described later, the "McKay Scholarship Program for Students with Disabilities" in Florida acted as a major disincentive to such reclassification. Since this program makes every special education student in Florida public schools eligible for vouchers, reclassification in to ESE categories is associated with a threat of loss of the corresponding students.

lunches on testing days in an attempt to boost performance.

While these papers study the response of public schools facing accountability systems, the present paper studies public school response and behavior facing vouchers. 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 vouchers, an alternative form of public school reform. This study also uses a different estimation strategy than that used in the above literature. In addition to a difference-in-differences strategy, this paper also uses a regression discontinuity analysis unlike that in the above literature.

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 (2002) and Nechyba (2003). Modeling public school behavior, McMillan (2002) 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 (2004) 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 (2004), Chakrabarti (2005) and West and Peterson(2005).[5] Greene (2001, 2003) finds positive effects of the Florida program on the performance of treated schools. Analyzing the same program and using student level data from a subset of Florida districts, Figlio and Rouse (2004) find some evidence of improvement of the treated schools in the high stakes state tests, but these effects diminish in the low stakes, nationally norm-referenced test. Using student level data, West and Peterson (2005) 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

[5] 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).

focused on writing because it is the easiest to improve.[6] 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 (2005b) 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 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 (2004) and Figlio and Rouse (2004) 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 vouchers 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 become eligible for vouchers or "opportunity scholarships" if the school gets two "F" grades in a period of four years. A school getting an "F" grade for the first time is exposed to the threat of vouchers, but its students do not become eligible for vouchers unless and until it gets 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.[7] Following a field test in the school year 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. In the remainder of the paper, I refer to school years by the calendar year of the spring semester. 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

[6] 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.

[7] 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.

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.[8] 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.)[9] In the remainder of the paper, for 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,[10] 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. In the remainder of the paper, I will use the word "cutoff" and "minimum criteria cutoff" interchangeably to refer to these minimum criteria cutoffs (level 2 in reading and math and score 3 in writing).

While scores of all regular students were to be included in the computation of school grades, scores of students in only a few exceptional student education (ESE) and limited English proficient (LEP) categories were included in the calculation of grades. Specifically, ESE students belonging to the three categories of speech impaired, gifted and hospital/homebound and LEP students with more than two years in an ESOL (English for speakers of other languages) program were eligible to be included in school grade computation. Since 1998, Florida classified the special education students into twenty-one ESE categories,[11] therefore scores of students in eighteen ESE categories were not eligible to be included in the computation of grades.

[8] 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.

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

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

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

The 1999 grading system was replaced by a new system in 2002. Although the definitions of the achievement levels remained the same, the new system included learning gains of students in addition to their level scores in the computation of grades. School grades A-F under the new system corresponded to specific ranges on a point scale where higher points corresponded to higher grades. Under the 1999 grading system, the F grade and movement to a D depended solely on the percentages of students scoring below the minimum criteria cutoffs. Under the new system, improving scores of low performing students as well as students in other ranges of the score scale increased the total number of points of schools and contributed towards a higher grade. Moreover, the new system gave more weight to reading and math scores compared to writing scores. While higher scores of students in all the three subjects—reading, math and writing—added to the total number of points, learning gains of students in only reading and math added to the total number of points. The rules relating to the inclusion of various special education categories in grade formation, however, did not change.

3 Theoretical Discussion

This section and subsections 3.1-3.3 explore the 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 "threat of voucher" program designates a quality cutoff $\bar{q}$ such that 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 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 below the minimum criteria cutoffs.[12] 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 are not too high, the threatened schools should be expected to resort to such a strategy.

If schools 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 are 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 should be more prominent in level 2.

[12] Alternative 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.

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 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.[13] 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 on. 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.

3.1.3 Reclassifying Low Performing Students into Exempt ESE Categories

Since reclassifying low performing students in to excluded ESE categories serves to artificially lower the percentage of students below the minimum criteria cutoffs, marginal returns from such a reclassification are positive. However, there are costs associated with such a strategy. It has to be approved by the parents, a group of experts (such as physicians, psychologists, etc.) and increased classification is associated with increased special services. Moreover, too much classification may lead to investigations or audits by the Florida Department of Education.

The McKay Scholarship program acts as a farther disincentive to this sort of reclassification. Created in 1999 and fully implemented in the 2000-01 school year, this program makes every disabled Florida public school student eligible for vouchers to move to a private school (religious or non-religious) or to another public school. Thus reclassification of students in to special education categories is associated

[13] 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.

with a threat of loss of the student and the corresponding revenue. The threatened public schools will resort to reclassification into excluded ESE categories only if the returns from it justify the associated costs.

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, distribution of students in the various ESE categories, 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. The ESE data not only give information on total ESE membership, but data were also obtained on membership in each of the ESE categories in each Florida school for all years under consideration.

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 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5 and 6. 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.

During the period 1993-1997, Florida classified the exceptional students into 15 categories: educable mentally handicapped, trainable mentally handicapped, physically handicapped, physical/occupational therapy part-time, speech/language/hearing part-time, speech/language/hearing, visually handicapped part-time, visually handicapped, emotionally handicapped part-time, emotionally handicapped, specific learning disability part-time, specific learning disability, gifted part-time, hospital/homebound part-time and profoundly handicapped. The reported ESE categories changed in 1998. Starting from 1998 Florida classified students into 21 categories - educable mentally handicapped, trainable mentally handicapped, orthopedically handicapped, occupational therapy, physical therapy, speech impaired, language

impaired, deaf or hard of hearing, visually impaired, emotionally handicapped, specific learning disabled, gifted, hospital/homebound, profoundly mentally handicapped, dual-sensory impaired, autistic, severely emotionally disturbed, traumatic brain injured, developmentally delayed, established conditions and other health impaired. Detailed school level data on membership in each of the above ESE categories are available from 1993-2002.

School level data on grade distribution (K-12) of students are available from 1993-2002. Data on socio-economic characteristics include data on sex 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 Strategy

Under the Florida opportunity scholarship program, schools that received a grade of "F" in 1999 were directly threatened by the program in 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. Given the nature of the Florida program, the threat 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 Focusing on Students below the Minimum Criteria

As discussed above, if the treated schools tend to focus more on students they anticipate 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. 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.

To investigate whether the F schools resorted to such behavior, I look for shifts in the percentages of students scoring in the different levels (1-5) for the F schools relative to the D schools in the post-

program period. Using data from 1999 to 2002, I estimate the following model:

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

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 race, sex, 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 counterparts of this regression which includes school fixed effects (and hence has one less level dummy and one less interaction between level dummy and treated dummy). These regressions are run for each of the subject areas—reading, math and writing.

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 1994-1997. 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.[14] To investigate the issue of pre-existing trends, I estimate the following regression as well as its fixed effects counterpart using pre-program data:

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

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}.

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

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 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 calculated 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.[15] 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 for the 98F schools (relative to the 98D schools) during 1998-99.[16]

[15] 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.

[16] 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 problem (if any).

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

The issue of mean reversion arises here from the concern that the F schools may revert towards the mean to a greater extent than D schools just by virtue of their relatively low performance in 1999. Therefore regression discontinuity analysis, comparing schools in a narrow range on either side of the cutoff between F and D schools provides a useful way to get around the problem of mean reversion. The Florida 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 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, only F schools failed the minimum criteria in writing also, while D schools passed it. Therefore, in this sample the probability of treatment varies discontinuously as a function of the percentage of students scoring at or above 3 in 1999 FCAT writing. There exists a sharp cutoff at 50%—while schools below 50% faced a direct threat, those above 50% did 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 5 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 below 3 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 5, 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 corresponds to one school. These two panels show that in this sample, percentage of students scoring at or above 3 in writing uniquely predicts (except two schools) assignment to treatment and there is a discrete change in the probability of treatment at the 50% mark.

Ranking schools in terms of percentage of students scoring at and above 3 in FCAT writing, I first consider schools that lie within ± 7 percentage points of the 50% threshold and refer to it as discontinuity sample 1. Next, I further shrink the sample and pick schools that lie within a ± 5 percentage point range of the cutoff and call it discontinuity sample 2. Using each of these samples, I investigate whether the

F schools exhibit differential shifts in the percentage of students scoring in the various levels in the post program period. For this purpose, I estimate model (1) using each of these samples, except that the OLS regressions also include a smooth polynomial[17] in the selection variable, percentage of students scoring at or above 3 in FCAT writing.

I also consider two corresponding discontinuity samples where both F and D schools fail the minimum criteria in reading and writing (math and writing). F schools fail the minimum criteria in math (reading) also, unlike D schools. 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%.

5.1.4 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 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.[18] If there is evidence of a decline in percentage of students in lower levels in F schools relative to D schools, this would be driven by sorting only 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 this might happen.

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):

$$y_{it} = \alpha_0 + \alpha_1 F + \alpha_2 t + \alpha_3 (F * t) + \alpha_4 v + \alpha_5 (v * t) + \alpha_6 (F * v) + \alpha_7 (F * v * t) + \varepsilon_{it} \qquad (3)$$

[17] I experiment with three forms of the polynomial: a quadratic, a cubic and a quartic.

[18] 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.

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.

5.1.5 The Problem of Underestimation: Are D Schools Untreated?

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 above (both difference-in-differences and regression discontinuity estimates) would be underestimates. Note that this problem is likely to be more prominent in the regression discontinuity analysis.

To get around this problem, I rescale the effects obtained in the previous analyses by the difference in the probabilities of treatment of F and D schools, that is by calculating the corresponding Wald estimator.[19] I use pre-program data to calculate the probabilities that F and D schools respectively would fall into treatment the next year. These scaling factors are calculated both for the full sample of F and D schools and the discontinuity samples.

A problem here is that the system of assigning letter grades started in 1999. However, as described in section (5.1.2), I was able to assign F and D grades in 1998 using the state grading criteria and 1998 school scores. Using this sample of 98F and 98D schools and data on school grades in 1999, I calculate the above probabilities.[20] To calculate these probabilities for the discontinuity samples, I consider the set of schools that failed the minimum criteria in all three subject areas in 1998 (the 98F schools), and the set of 98D schools that failed the minimum criteria in reading and math in 1998, but passed the minimum criteria in writing. Ranking these schools in terms of their percentages of students scoring at and above 3 in 1998 FCAT writing, I consider the schools within a range of ± 5 (± 7) percentage points of the 50% cutoff for discontinuity sample 2 (discontinuity sample 1), and calculate the probabilities that these groups of 98F and 98D schools would fall into treatment the next year.

[19] I would like to thank Caroline Hoxby for suggesting this strategy.

[20] Note that 1998 is the first year that such grades can be calculated. This is because (after a field test in 1997) the FCAT reading and math tests were first administered in 1998.

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:

$$y_{ist} = \gamma_0 read + \gamma_1 math + \gamma_2 write + \gamma_3 low + \gamma_4 mid + \gamma_5(read * D00) + \gamma_6(math * D00) + \gamma_7(write * D00) + \gamma_8(low * D00) + \gamma_9(mid * D00) + \gamma_{10} X_{ist} + \varepsilon_{ist} \quad (4)$$

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 next explore these issues further by disaggregating the above effects. If the F schools choose to focus on the subject closest to the cutoff, then do they concentrate on the "low" subject irrespective of whether it is reading, math or writing or does the response in the "low" subject depends on the specific subject area? On the other hand, if they choose to focus on one subject area because of its relative ease, do they focus on it irrespective of its rank? To investigate these questions, I estimate the following

model (as well as the fixed effects counterpart of it). The coefficients of interest here are $\delta_5 - \delta_{13}$.

$$y_{ist} = \delta_0 read + \delta_1 math + \delta_2 write + \delta_3 low + \delta_4 mid + \delta_5(low * D00 * read) + \delta_6(low * D00 * math) + \delta_7(low * D00 * write) + \delta_8(mid * D00 * read) + \delta_9(mid * D00 * math) + \delta_{10}(mid * D00 * write) + \delta_{11}(high * D00 * read) + \delta_{12}(high * D00 * math) + \delta_{13}(high * D00 * write) + \delta_{14} X_{ist} + \varepsilon_{ist} \quad (5)$$

5.3 Reclassifying Low Performing Students into Exempt ESE Categories

This section describes the strategies I use to investigate whether F schools resorted to differential classification in to ESE categories after the program.

5.3.1 Looking for Shifts in Total ESE Classification

The dependent variable for this analysis is percentage ESE membership, i.e., total ESE membership as a percentage of total enrollment. After controlling for pre-program trends and post-program common shocks, I look for relative shifts in this variable in F schools after the program. Using data from 1998-2002, I begin with the following model:

$$pe_{it} = c + \phi_0 F + \phi_1 t + \phi_2(F * t) + \phi_3 v + \phi_4(v * t) + \phi_5(F * v) + \phi_6(F * v * t) + \phi_7 X_{it} + \varepsilon_{it} \quad (6)$$

where pe_{it} represents percentage ESE membership in school i in year t; v is the program dummy; $(F * t)$ allows for differential pre-program trend of F schools; v and $(v * t)$ control for post-program common intercept and trend shifts and X_{it} includes the set of school characteristics. The coefficients on the interaction terms $(F * v)$ and $(F * v * t)$ estimate the program effects - ϕ_5 gives the intercept shift and ϕ_6 the trend shift. Note that this specification constrains the post-program F school year effects to be homogeneous over time.

Next, I estimate an unrestricted model that no longer constrains the post-program year-to-year changes in F schools to be equal and allows for heterogeneous-in-time treatment effects.

$$pe_{it} = c + \sum_{j=1999}^{2002} \phi_j' D_j + \phi_0' F + \phi_1'(F * D_1) + \sum_{j=2000}^{2002} \phi_{2j}(F * D_j) + \phi_3 X_{it} + \varepsilon_{it} \quad (7)$$

where $D_1 = 1$ if year ≥ 1999, and 0 otherwise, and $F * D_1$ controls for any differential pre-program year effects of F schools. The coefficients ϕ_{2j} represent the effect of the program on the rate of ESE classification one, two and three years into the program. I also estimate the fixed effects counterpart of each of the above regressions.

5.3.2 Looking for Shifts in Classification in Excluded Categories relative to Included Categories

While trends in total ESE classification provide a summary picture, they are unlikely to provide a conclusive picture in terms of whether or not the F schools resorted to such reclassification of students. For example, absence of shifts in total ESE classification does not rule out the possibility that relative reclassification in excluded categories took place in the F schools. Similarly, positive shifts in total ESE classification does not necessarily imply an increase in classification in excluded categories. To have a closer look, I look for post-program shifts in classification in Excluded categories relative to Included categories in F schools, in comparison to D schools. Using data on percentage of total enrollment classified in Excluded and Included categories for F and D schools during 1998 through 2002, I estimate the following model:

$$z_{ijt} = constant + \theta_0 Exempt + \theta_1 t + \theta_2 F + \theta_3(Exempt * t) + \theta_4(Exempt * F) + \theta_5(F * t) + \theta_6(Exempt * F * t) + \theta_7 v + \theta_8(Exempt * v) + \theta_9(v * t) + \theta_{10}(F * v) + \theta_{11}(Exempt * v * t) + \theta_{12}(Exempt * F * v) + \theta_{13}(F * v * t) + \theta_{14}(Exempt * F * v * t) + \theta_{15} X_{ijt} + \varepsilon_{ijt} \quad (8)$$

where z_{ijt} indicates the total percentage of students in school i classified in category j in year t, j ={Excluded, Included}, "Exempt" is a dummy variable that takes the value of 1 for excluded categories and 0 otherwise, X_{ijt} includes the set of school characteristics and interactions of "Exempt" dummy with these variables. The coefficients of interest here are θ_{12} and θ_{14}—they capture any differential post-program intercept and trend shifts for F schools relative to D schools in excluded categories (relative to included categories), after controlling for pre-program trends and post-program common shocks. I also estimate a model that allows for heterogenous-in-time treatment effects:

$$z_{ijt} = constant + \theta_0' Exempt + \theta_1' F + \theta_2'(Exempt * F) + \sum_{k=1999}^{2002} \theta_{3k} D_k + \sum_{k=1999}^{2002} \theta_{4k}(Exempt * D_k) + \theta_5'(F * D_1) + \sum_{k=2000}^{2002} \theta_{6k}(F * D_k) + \theta_7'(Exempt * F * D_1) + \sum_{k=2000}^{2002} \theta_{8k}(F * Exempt * D_k) + \theta_9 X_{ijt} + \varepsilon_{ijt} \quad (9)$$

The coefficients of interest here are θ_{8k}, they capture the effect of the program on classification in excluded categories relative to included categories in F schools after one, two and three years after the program.

5.3.3 Classification in Mutable Excluded and Included Categories

The ESE categories vary in the extent of their severities. While some categories such as those with observable or severe disabilities or physical handicaps are comparatively non-mutable, others such as

learning disabilities and emotionally handicapped/disturbed are much more mild and comparatively mutable categories.[21]. Classification in these latter categories often has a large amount of subjective element to it and is hence easy to be manipulated. So if reclassification into excluded categories did take place, it is most likely to have taken place in these categories. Using data from 1998-2002 and models (8) and (9), I look for post-program shifts in classification in these mutable categories relative to included categories in the F schools (relative to D schools). "Exempt" is now a dummy variable that takes the value of 1 for learning disabled (or emotionally handicapped) and 0 otherwise.

5.3.4 Existence of Differential Pre-Program Trends in ESE Classification

The definition of some of the ESE categories changed in 1998, so that the categories in the pre-1998 period are not directly comparable to those in 1998-2002. Therefore, the above analysis includes two pre-program years: 1998 and 1999. In spite of changes in definitions, analysis of the pre-1998 data can provide us with some idea about differential pre-program trends and whether they are likely to be a major problem. Data on ESE classification in the pre-1998 period are available for 1993-97. Using these data I look for any differential trend in F schools in total ESE classification, relative classification into excluded categories as compared to included categories and relative classification in learning disabled (emotionally handicapped) categories compared to included categories.[22]

5.3.5 Using Regression Discontinuity Analysis to Investigate Relative Post-Program Shifts in ESE Classification in Florida

As outlined earlier, the design of the Florida program permits the use of this quasi-experimental design in analyzing the treatment effects in Florida. Regression discontinuity analysis serves to compare F and D schools within a narrow range of the cutoff between F and D. I use the same two discontinuity samples here as in section (5.1.3),—discontinuity sample 1 and discontinuity sample 2. The F and D schools in the discontinuity samples are very similar to each other (see table 4). Therefore regression discontinuity analysis promises to give more precise results on the relative classification, if any, in to ESE categories by the F schools and consistency of these results with those from the entire sample of F

[21] See Cullen (2003), Singer et. al. (1989) and Figlio and Getzler (2002)

[22] Learning disabled and emotionally handicapped categories were comparable in the two periods. However, while speech language and hearing impaired were reported as a single category in the earlier period, they were reported as separate categories in the latter period. Speech impaired is an included category under Florida law, while hearing and language impaired are not. Therefore the included categories cannot be completely separated from the excluded categories in the pre-1998 period. (The definition of the other included categories remained the same.) While interpreting the results from this analysis, this caveat should be kept in mind.

and D schools would attest to the robustness of the results.

Using the above two discontinuity samples, I look at the same three issues as in sections (5.3.1)-(5.3.3). Do the F schools in the discontinuity samples show a relative shift in total ESE classification after the program? Is there an increased classification in to excluded categories relative to included categories in F schools after the program? Is there an increased classification in the mutable excluded categories relative to included categories?

5.3.6 Using Change in Percentage of Students Tested to Investigate the Issue of Increased Classification into ESE Categories

Reclassification of low performing students into excluded ESE categories so as to remove them from the test-taking pool would lead to a decrease in the percentage of students tested in Florida.[23] Therefore, I analyze the patterns in percent of students tested in F and D schools before and after the program. More precisely, using data for reading, math and writing for 1998-2002 and after controlling for pre-program differences in trends and post-program common shocks, I look for any post-program shifts in percent of students tested in F schools (relative to D schools) in each of the subjects.

5.3.7 Reclassification into ESE categories: Are the above estimates underestimates?

As discussed in section (5.1.5), D schools are likely to be affected by the program as they are close to getting an "F". In such a case, the difference-in-differences estimates of classification into ESE categories, and especially the regression discontinuity estimates, are likely to be underestimates. To circumvent this problem, I use the strategy described in section (5.1.5), and scale the treatment effects by the corresponding differences in the probabilities of treatments of F and D schools.

6 Results

6.1 Did the Threatened Schools Focus on Students Expected to Score below the Cutoffs? Investigating Shifts in Percentages Scoring below the Cutoffs

This section looks for post-program shifts in the percentages of students scoring in the various levels for F schools in comparison to D schools. 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

[23] Increased classification accompanied by an increased percentage of regular students tested can in principle keep the percentage of students tested the same. However, it is not obvious that such a condition will hold in practice. See footnote 29.

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 distribution 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 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 distribition in 2000, with changes most concentrated in the crucial levels.

I next investigate whether these patterns continue to hold in a more sophisticated regression analysis also. Table 1 presents results on the effect of the program on percentage 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 race, sex, 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.[24] 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 a statistically significant decline, but the magnitudes never exceeded 1%. Moreover, the changes in

[24] Although the state still continued to grade the Florida schools on a scale of A though F, the grading criteria underwent some important changes in 2002, as described earlier. 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.

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. Levels 2, 3 and 4 percentages increase, most of which are statistically significant. 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 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, as would be dictated by the incentives created by the program. In writing, there is no evidence of a fall in the percentage of students scoring in level 5. 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 they expected to score below the minimum criteria cutoffs. More importantly, consistent with the incentives created by the program, while the declines in reading and math are concentrated in level 1, the decline in writing is most prominent in level 2, rather than level 1. 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 evidences 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 very small, always being less than 1% in magnitude. I later investigate whether this pattern continues to hold under a more precise regression discontinuity analysis.

Looking for Pre-Existing Trends

The results in Table 1 would be biased if there are pre-existing trends in the percentage of students scoring in the different levels in F schools relative to D schools. Table 2 investigates this issue. As discussed earlier, while pre-program data on the percentage of students in various levels are available for all levels in writing (1994-97), they are available for only levels 1 and 2 in reading and math (1998-99).[25] 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.

Mean Reversion

This section addresses the concern that the results reported in Table 1 may be biased by mean reversion. Using the strategy described in section 5.1.2, 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. As discussed earlier, the absence of relevant data in writing precludes the examination of mean reversion in writing using this method. However, the regression discontinuity analysis in the next section serves to get rid of this problem, if any.

Regression Discontinuity Analysis: Examining Focus on Students below Cutoffs

I use the two discontinuity samples as described in section 5.1.3, - discontinuity sample 1 (± 7) and discontinuity sample 2 (± 5). The summary characteristics of the F and D schools in these samples are reported in Table 4. All these numbers pertain to the pre-program year, 1999. The F and D schools are strikingly similar to each other in terms of various demographic characteristics (race, sex and percentage of students eligible for free or reduced-price lunches), ESE categorization (percentage of enrollment in all ESE categories together, percentage in excluded ESE categories, percentage in included ESE categories), real per pupil expenditure, mean scores (FCAT reading, math and writing)

[25] Data on percentage of students in all the five levels in reading and math are available in 1999, but not before that.

and number of students tested. Using discontinuity sample 2, Figure 6, panels C-E show the effects of the program on percentage of students scoring in levels 1, 3 and 5 in reading, one year after the program. Panels F-H show the corresponding effects in reading two years after program. In both years, the percentage of students scoring in level 1 dropped sharply around the 50% cutoff (panels C and F) implying that the program led to a decline in the percentage of students scoring in level 1. Breaks are also visible in level 3 for both years 2000 and 2001 (Panels D and G), but these are considerably smaller in magnitude than the level 1 breaks. There is no evidence of any break in the relationship in level 5 around the 50% cutoff.

Figure 7, Panels A-C show the effect of the program on percentage of students scoring in levels 1, 3 and 5 in math one year after program. Panels D-F show the corresponding effects in math two years after program. Panels G-I show the effect of the program on percentage of students scoring in levels 1, 2 and 3 in writing one year after the program (2000). (The graphs for 2002 in reading and math and for 2001 and 2002 in writing are similar to those in the other years for the corresponding subject and hence omitted.)[26]

In math, there is a sharp drop in the percentages of students scoring in level 1 close to the 50% cutoff in both the first and second years after program. Once again, breaks are visible in level 3 in both years in math, though they are much smaller than the corresponding level 1 breaks. While there is no evidence of any discontinuity in level 5 in 2000, there is evidence of a small break in relationship in 2001. (Note that the scales in the level 5 graphs are different from those in levels 1 and 2.) In writing, on the other hand, while drops are visible in both levels 1 and 2 around the 50% cutoff, the drop in level 2 is substantially larger than that in level 1.

I next examine whether these patterns continue to hold in a more rigorous analysis. Using discontinuity samples 1 and 2, Table 5 looks at the effect of the program on percentage of students scoring in different levels in F schools relative to D schools. The regressions report results from fixed effects regressions corresponding to model 1. The results from OLS are similar and hence not reported,—the OLS regressions also include a polynomial in the selection variable, the percentage of students scoring at or above level 3 in FCAT writing. I have experimented with three forms of the polynomial,- a quadratic, a cubic and a quartic - the results remain qualitatively similar for the different forms of the polynomial.

[26] The figures for the other levels are omitted for lack of space, they tally with the results reported in Table 5. All these graphs pertain to discontinuity sample 2, those for discontinuity sample 1 are qualitatively similar and correspond to the results in Table 5.

These results are available on request.

In each of reading and math, for each of the discontinuity samples and in each of the three years after the program, the F schools show a relative decline in the percentage of students scoring in level 1. These effects are statistically significant in all cases (except first year reading) and these effects always exceed the effects in the other levels. In addition, these effects are similar to the corresponding level 1 effects in the full sample of F and D schools. In writing, although the effects are somewhat smaller than that in the full sample, they are still economically large and statistically significant. The pattern also remains the same—the level 2 effects in each of the three years after the program and for each of the discontinuity samples exceed the corresponding level 1 effects. Moreover, the level 2 effects always exceed the changes in the other levels. 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 (less than 1%) decline in percentage of students scoring in level 5 reading and first year math, but there was no such evidence in writing or in the other years in math. But, as seen in table 5, the declines in reading and first year math no longer survive in the more precise regression discontinuity analysis. In fact, there is no evidence that the improvements of the low performing students came at the expense of the higher performing ones and there is even some evidence of a small increase in percentage of students in level 5 writing.

To sum up, there is strong evidence that the threatened schools concentrated more on low performing students (i.e., students expected to score below and close to the minimum criteria cutoffs) and the declines in the percentages of students just below the minimum criteria cutoffs exceeded the changes in percentages at all other ranges of the score scale. 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 lower performing 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.

Sorting

Another factor that might potentially bias the results is sorting. To investigate this issue, I analyze the

trends in the demographic composition of schools and investigate whether the program led to a shift in the demographic composition of the F schools relative to the D schools. Table 6 presents the estimation results for specification (6). 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 in the paper control for any change in demographic composition. To sum, it is unlikely that the patterns seen above are driven by sorting.

Addressing the Problem of Underestimation

A concern here is that the above estimates may be underestimates as the D schools are likely to be affected by the program to some extent. (Note, though, that the direction of the above effects are correct, but presumably the effects are even larger than the estimates presented in the above analysis.) Using the strategy described in section (5.1.5), the correct difference-in-differences and regression discontinuity estimates can be obtained by scaling up the corresponding estimates above by factors of 1.15 and 1.27 respectively.

6.2 Do Threatened Public Schools Focus on the Subject Closest to the Cutoff? Role of Differences in the Extent of Difficulties between Subjects

This section investigates whether the threatened schools facing the "threat of voucher" program chose to focus on a certain subject area. Table 7 presents the results from estimation of model 4. While columns (1)-(2) present the results without controls, columns (3)-(4) present those with controls.[27] The first column of each set reports the 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 7.

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

[27] Controls include race, sex, percentage of students eligible for free or reduced-price lunches, real per pupil expenditures and interactions of the subject dummies with these variables.

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.

Table 8a 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 5. 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 8b 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.

To sum, the F schools chose to concentrate on writing irrespective of its distance from the cutoff, presumably because it was easiest to improve in. Case studies 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 introduced school wide projects in writing, longer time blocks in writing, and

writing components in lower grades.

6.3 Do Threatened Schools tend to Reclassify Low Performing Students into Excluded ESE Categories

Total ESE classification

Table 9 looks for relative shifts in total ESE classification after the program. Columns (1)-(2) present results from model (6), while columns (3)-(4) present results from model (7). There is no evidence of any post-program shifts in ESE placement for the F schools relative to the D schools. Note that though the treatment effects are not significant, they are in most cases positive. Therefore, for each regression I conduct an F test to examine whether the treatment effects are jointly significant. As the p-values show, while the treatment effects are jointly significant for model (6), they are no longer so for model (7).

Thus, there is not much evidence in favor of a relative increase in total ESE classification in F schools after the program. However, absence of shifts in total ESE classification does not preclude the possibility of relative shifts in *excluded* categories in F schools. Therefore, to obtain a clearer picture, I next investigate whether there was an increased classification in excluded categories relative to included categories in the F schools in comparison to the D schools after program.

Classification in Excluded Categories relative to Included Categories

Table 10 looks at the effect of "threatened status" on classification in excluded categories relative to included categories. Columns (1) reports the results from estimation of model (9) and column (2) from it's fixed effects counterpart. The results from the model (8) are qualitatively similar and hence not reported, they are available on request.

There is no statistically significant evidence of any relative shifts in F schools, either in the included or in the excluded categories. Since the treatment effects on relative classification in excluded categories are positive, I also conduct an F-test to check whether they are jointly significant. As the p-values indicate, the treatment effects (relative classification in excluded categories) are also not jointly significant. I also estimate alternative forms of the above specification as well as model (8) that include district dummies and interactions of district dummies with year dummies. The results are very similar and hence are not reported here.[28] Figure 8 Panel A looks at the trend in total ESE classification in F and D schools

[28] Note that I also experiment with district dummies and interactions of district dummies with year dummies in the other analyses in this paper—when looking for effects on total ESE classification, investigating focus on specific subject areas and focus on students below minimum criteria. The results in all cases remain similar to those reported in the paper

while panel B graphs the trend in relative classification into excluded categories relative to included categories in F and D schools. Once again, consistent with the results in the previous and this section, there is no evidence of any relative reclassification in F schools relative to D schools.

Classification in Mutable Excluded Categories relative to Included Categories

The previous analysis does not find much evidence in favor of relative classification into excluded categories in F schools. However, this does not rule out the possibility that this kind of behavior took place in the F schools. This is because increased classification may have taken place in some specific categories which are more mutable and hence more amenable to manipulation, and consideration of all excluded categories together masks this kind of behavior. As argued earlier, if such reclassification did take place, it is most likely to have taken place in the mutable categories such as learning disabled and emotionally handicapped.

Columns (3)-(4) of table 10 investigate the effect of the program on classification into learning disabled categories relative to included categories in F schools. Although the treatment effects on relative classification into learning disabled categories in F schools are jointly significant, they are never statistically significant individually. Moreover, the magnitude of the effects in each of the three years after program are small. Columns (5)-(6) examine the effect on such classification in emotionally handicapped category. There is no evidence of any relative classification into this category in the threatened schools. To summarize, there is not much evidence in favor of a relative increase in classification in excluded categories in the treated schools. However, these effects may be contaminated by other factors such as the presence of pre-existing trends. I consider these issues below to check the sensitivity/robustness of the above findings.

Looking for pre-existing trends in ESE classification

Using data from 1993 through 1997, table 11 investigates the existence of pre-program trends in ESE classification in F schools relative to D schools. Columns (1)-(2) find no evidence of differential pre-program trends in F schools in total ESE classification. Columns (3)-(4) investigate the existence of differential pre-program trends in excluded categories relative to included categories in F schools (in comparison to D schools), columns (5)-(6) look for such differential trends in learning disabled and columns (7)-(8) in emotionally handicapped. There is no evidence that F schools tended to reclassify differentially in excluded categories (or mutable excluded categories) relative to the D schools before

and hence are omitted. They are available on request.

the program. To summarize, table 11 finds no evidence that the above post-program effects are biased by the existence of pre-program trends.

Regression Discontinuity Analysis: Examining the effect of Program on ESE Placement

Using discontinuity samples 1 and 2, table 12 investigates the effect of "threatened status" on total ESE placement. The results are qualitatively similar to those in the full sample. Although the fixed effects estimates show positive and significant treatment effects in the first year, the effects in the second and third years are no longer significant. Note that while the treatment effects are jointly significant in the OLS regressions, they jointly never reach standard levels of significance in the fixed effects regressions. Using the same discontinuity samples, table 13 investigates the effect of the program on relative classification in to excluded categories in F schools (relative to D schools). Columns (1)-(2) show that there is no statistically significant evidence that the F schools resorted to classification in to excluded categories, except in the first year after program for discontinuity sample 2. I also conduct an F test to check the joint significance of these effects. As the p-values reveal, these effects are never jointly significant.

Columns (5)-(8) investigate the effect of the "threatened status" on relative classification in the mutable excluded categories—learning disabled (columns (3)-(4)) and emotionally handicapped (columns (5)-(6)). Once again, there is no evidence that F schools resorted to such reclassification after the program.

Looking at the Effect of "Threatened Status" on Percentage of Students Tested

Another way to investigate whether such reclassification took place is to look for a shift in the percentage of students tested. If F schools tended to classify low performing students into excluded special education categories, this would lead to a decrease in the percentage of students tested. Table 14 investigates this issue. Using data on percentage of students tested in reading, math and writing during 1998-2002, it finds no evidence of a relative decline in percentage of students tested in either of the three years after program and in either of the subject areas tested. This further confirms the above finding.[29]

Thus there is not much evidence that the F schools tended to differentially classify low performing students in to excluded ESE categories. But this does not mean that the schools did not respond to

[29] Note that no change in percentage of students tested is consistent with an increased classification in to excluded ESE categories if the school increases the percentage of regular students tested. The school would resort to such a behavior only if the latter group of students are expected to score better than those reclassified in to excluded ESE categories. Both of these are quite stringent conditions. Moreover, it is not clear why the comparatively low performing students would choose to take the test while the comparatively better students would choose not to take it in the first place.

incentives. As discussed in the theoretical section, there are multiple ways in which the threatened schools can respond—they weigh the relative returns and costs in the different alternatives and choose the options that are least costly. The fact that they did not resort to such reclassification indicates that its costs did not justify its returns. Increased classification means increased provision of services which is often costly in spite of state financing of a large part of these services. Moreover, such classification has to be approved by the parents and a group of experts and too much reclassification might lead to audits by the Florida department of education. The existence of the McKay Scholarship program acts as a further deterrent to such classification. Since this program makes every disabled student in Florida public schools eligible for vouchers, reclassification is associated with a risk of loss of these students and a corresponding risk of loss of revenue.

Reclassification into ESE categories: Addressing the Problem of Underestimation

As discussed earlier, both the difference-in-differences and the regression discontinuity are likely to be underestimates because the D schools are likely to be affected by the program. This issue is more of a problem here because one might argue that the above absence of any robust evidence in favor of reclassification is driven by the fact that the D schools are not completely untreated. To circumvent this problem, I use the strategy described in sections (5.1.5) and (5.3.7). Using this strategy, the correct difference-in-differences estimates can be obtained by scaling up the above difference-in-differences estimates by a factor of 1.15, while the correct regression discontinuity estimates can be obtained by scaling up the corresponding estimates above by a factor of 1.27. Using the estimates in tables 9-10 and 12-13, it can be seen that the scaled up effects would still be very small and would always be less than 1.17%. This further serves to reinforce the above finding that there is not much evidence in favor of relative reclassification into special education categories by F schools.

7 Conclusion

This paper analyzes the behavior of public schools facing vouchers. It focuses on the 1999 Florida opportunity scholarship 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 three 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 and hence vouchers. 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 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 is perceived to be the easiest irrespective of the distances of the subject areas from the thresholds? Third, scores of students in certain special education categories were not eligible to be included in the calculation of grades. Did this rule induce the F schools to reclassify their low performing students in to these excluded categories so as to artificially inflate scores?

While there is not much evidence that the threatened schools resorted to such reclassification (there were substantial costs to this, as outlined in the paper), I find robust evidence that they 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 are 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 very 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 carve 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. In contrast, the rules relating to the inclusion of the different special education categories in grade formation did not change.

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 (2005),"Can Increasing Private School Participation in a Voucher Program Affect Public School Performance? Evidence from the Milwaukee Voucher Experiment," mimeo, Harvard University.

Chakrabarti, Rajashri (2004), "Impact of Voucher Design on Public School Performance: Evidence from the Florida and Milwaukee Voucher Programs", mimeo, Harvard University.

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 (2003), "The Impact of Fiscal Incentives on Student Disability Rates," Journal of Public Economics 87, 1557-1589.

Cullen, Julie and Randall Reback (2002), "Tinkering towards Accolades: School Gaming under a Performance Accountability System," mimeo, University of Michigan.

Figlio, David (2003), "Testing, Crime and Punishment", mimeo, University of Florida.

Figlio, David and Lawrence Getzler (2002), "Accountability, Ability and Disability: Gaming the System?", NBER Working Paper # 9307.

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 (2004), "Do Accountability and Voucher Threats Improve Low-Performing Schools?", mimeo, University of Florida and Princeton University.

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.

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.

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).

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

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.

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

Singer, Judith, Judith Palfrey, John Butler and Deborah Walker (1989), "Variation in Special Education Classification across School Districts: How does Where you Live Affect What You are Labeled?", American Educational Research Journal, 26 (2), 261-281.

West, Martin and Paul Peterson (2005), "The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments", Harvard University, Program on Education Policy and Governance, PEPG # 05-01.

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***	-3.32***	-5.56***	-5.57***	-6.95***	-6.97***
	(1.25)	(1.23)	(1.44)	(1.43)	(1.27)	(1.25)
Treated * level 2 * 1 year after program	0.96	0.96	2.83***	2.84**	-10.07***	-10.09***
	(0.88)	(0.87)	(0.75)	(0.74)	(0.92)	(0.93)
Treated * level 3 * 1 year after program	2.32***	2.31***	2.17*	2.18*	11.04***	11.02***
	(0.73)	(0.72)	(1.23)	(1.22)	(1.26)	(1.24)
Treated * level 4 * 1 year after program	0.83	0.82	0.99**	1.00	6.52**	6.51***
	(0.70)	(0.69)	(0.42)	(0.43)	(2.60)	(2.59)
Treated * level 5 * 1 year after program	-0.70***	-0.70***	-0.56***	-0.57**	0.39	0.37
	(0.14)	(0.12)	(0.22)	(0.21)	(0.46)	(0.45)
Treated * level 1 * 2 years after program	-4.13***	-4.12***	-7.52***	-7.50***	-6.78***	-6.77***
	(1.35)	(1.33)	(1.79)	(1.78)	(1.20)	(1.21)
Treated * level 2 * 2 years after program	2.21**	2.20**	3.73***	3.74***	-9.52***	-9.51***
	(0.87)	(0.88)	(0.96)	(0.95)	(1.17)	(1.18)
Treated * level 3 * 2 years after program	2.23**	2.22**	2.63***	2.61***	10.15***	10.16***
	(1.07)	(1.05)	(1.00)	(0.99)	(2.21)	(2.23)
Treated * level 4 * 2 years after program	0.58	0.57	1.24	1.25	5.70**	5.71***
	(0.46)	(0.47)	(0.78)	(0.80)	(2.53)	(2.50)
Treated * level 5 * 2 years after program	-0.92***	-0.91***	-0.16	-0.17	1.26	1.27*
	(0.31)	(0.30)	(0.30)	(0.29)	(0.95)	(0.93)
Treated * level 1 * 3 years after program	-5.47***	-5.46***	-5.96***	-5.98***	-7.05***	-7.02***
	(1.25)	(1.26)	(1.74)	(1.73)	(1.06)	(1.07)
Treated * level 2 * 3 years after program	2.86***	2.85***	4.48***	4.47***	-10.51***	-10.49***
	(0.74)	(0.73)	(0.86)	(0.85)	(1.34)	(1.35)
Treated * level 3 * 3 years after program	3.18***	3.17***	0.44	0.44	12.95***	12.97***
	(0.84)	(0.82)	(0.69)	(0.69)	(1.60)	(1.60)
Treated * level 4 * 3 years after program	0.45	0.45	0.10	0.10	5.60***	5.63***
	(0.60)	(0.60)	(0.80)	(0.79)	(1.72)	(1.71)
Treated * level 5 * 3 years after program	-0.99***	-0.98**	-0.13	-0.12	0.19	0.21
	(0.35)	(0.34)	(0.55)	(0.52)	(0.63)	(0.61)
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. All regressions include level dummies and interactions of the level dummies with treated dummy and year dummies respectively. The FE columns include school fixed effects. 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 and Writing)

	Reading		Math		Writing	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * level 1 * trend	0.65	0.66	0.76	0.88	0.39	0.44
	(1.60)	(1.43)	(1.58)	(1.74)	(0.67)	(1.07)
Treated * level 2 * trend	-0.06	-0.05	1.48*	1.61	0.35	0.40
	(0.68)	(1.32)	(0.90)	(1.64)	(0.35)	(0.77)
Treated * level 3 * trend					-0.39	-0.34
					(0.56)	(0.75)
Treated * level 4 * trend					-0.16	-0.11
					(0.10)	(0.24)
Treated * level 5 * trend					-0.09	-0.04
					(0.05)	(0.10)
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. All regressions include level dummies, interactions of level dummies with treated dummy and trend respectively. The FE columns include school fixed effects. 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 in relation to 98D schools

(Reading and Math, 1998-99)

	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	-1.72	0.64	0.26	0.25
	(1.52)	(1.51)	(1.25)	(1.88)	(1.98)	(1.50)
98F * level 2 * trend	0.50	0.41	0.27	2.21	2.09	2.06
	(0.89)	(0.90)	(1.07)	(1.35)	(1.38)	(1.38)
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. Regressions include level dummies, interactions of level dummies with treated dummy and trend respectively. The FE columns include school fixed effects. 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: Pre-program (1999) characteristics of F and D schools in Regression Discontinuity samples

	Discontinuity sample 1			Discontinuity sample 2		
	F (std. dev.)	D (std. dev.)	F-D [p-value]	F (std. dev.)	D (std. dev.)	F-D [p-value]
% black	64.68	63.45	1.23	66.39	62.40	3.99
	(28.39)	(26.47)	[0.84]	(27.96)	(27.18)	[0.55]
% hispanic	17.99	19.52	-1.53	16.52	20.40	-3.88
	(20.86)	(21.72)	[0.75]	(18.15)	(22.09)	[0.45]
% white	16.42	15.52	0.90	16.15	15.62	0.53
	(18.81)	(18.24)	[0.83]	(17.78)	(18.28)	[0.90]
% asian	0.47	0.69	-0.23	0.55	0.80	-0.25
	(0.70)	(1.04)	[0.47]	(0.75)	(1.88)	[0.54]
% american indian	0.19	0.31	-0.11	0.10	0.30	-0.20
	(0.70)	(1.24)	[0.64]	(0.38)	(1.21)	[0.43]
% male	51.22	52.18	-0.96	51.35	51.87	-0.51
	(4.00)	(5.07)	[0.37]	(4.43)	(5.12)	[0.69]
% free lunch eligible	86.3	85	1.29	85.48	84.60	0.87
	(8.34)	(11.38)	[0.58]	(8.56)	(11.46)	[0.74]
% ESE	16.20	15.39	0.84	15.51	15.29	0.22
	(5.40)	(5.70)	[0.51]	(5.35)	(5.62)	[0.87]
% in Excluded ESE categories	11.94	11.20	0.73	11.47	11.15	0.32
	(5.20)	(5.71)	[0.56]	(5.18)	(5.58)	[0.81]
% in Included ESE categories	4.27	4.16	0.10	4.05	4.14	-0.10
	(1.87)	(1.73)	[0.80]	(1.45)	(1.73)	[0.81]
% real per pupil expenditure	32.07	30.72	1.35	31.33	30.69	0.65
	(5.58)	(5.72)	[0.29]	(5.56)	(6.09)	[0.66]
FCAT reading score	246.97	246.43	0.53	248	246.46	1.54
	(16.28)	(16.11)	[0.88]	(14.29)	(15.73)	[0.68]
# tested in reading	99.97	104.08	-4.11	102.92	104	-1.08
	(27.05)	(34.67)	[0.57]	(27.45)	(33.91)	[0.89]
FCAT math score	266.19	268.19	-2.00	266.29	268.64	-2.35
	(14.15)	(15.21)	[0.55]	(14.90)	(15.06)	[0.52]
# tested in math	95.53	101.68	-6.15	99.17	102.09	-2.92
	(24.90)	(38.28)	[0.42]	(26.29)	(37.35)	[0.73]
FCAT writing score	2.44	2.53	-0.09	2.46	2.53	-0.08
	(0.13)	(0.12)	[0.00]	(0.13)	(0.12)	[0.01]
# tested in writing	100.75	104.40	-3.65	104.04	104.27	-0.23
	(26.54)	(34.91)	[0.61]	(27.97)	(34.30)	[0.98]

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

	Reading		Math		Writing	
	D. S. 1[1] FE (1)	D. S. 2 FE (2)	D. S. 1 FE (3)	D. S. 2 FE (4)	D. S. 1 FE (5)	D. S. 2 FE (6)
Treated * level 1 * 1 year after program	-2.32	-2.95	-5.80***	-6.49***	-3.92**	-3.80***
	(2.23)	(2.21)	(1.99)	(1.76)	(1.64)	(1.23)
Treated * level 2 * 1 year after program	-1.11	-1.53	-0.04	-0.16	-8.46***	-7.99***
	(0.98)	(1.31)	(1.79)	(2.04)	(1.72)	(2.04)
Treated * level 3 * 1 year after program	0.31	0.93	2.34	2.86**	3.05	2.10
	(1.69)	(1.55)	(1.26)	(1.26)	(1.99)	(2.14)
Treated * level 4 * 1 year after program	2.05*	2.27*	3.39***	3.39***	6.76***	7.01**
	(0.83)	(0.96)	(1.22)	(1.27)	(2.93)	(2.95)
Treated * level 5 * 1 year after program	-0.36	-0.40	0.26	0.47**	1.61*	1.64*
	(0.36)	(0.45)	(0.23)	(0.23)	(0.85)	(0.96)
Treated * level 1 * 2 years after program	-4.72**	-6.33***	-8.48***	-10.04***	-3.71**	-3.84**
	(2.38)	(2.43)	(2.30)	(1.93)	(1.76)	(1.74)
Treated * level 2 * 2 years after program	1.64	2.00	2.65	3.24	-3.50*	-4.34**
	(1.39)	(1.37)	(1.97)	(2.23)	(2.00)	(1.85)
Treated * level 3 * 2 years after program	0.91	1.49	2.00	2.27**	4.99	3.80
	(1.82)	(1.79)	(0.83)	(0.94)	(3.04)	(2.43)
Treated * level 4 * 2 years after program	1.73**	2.47***	2.96*	3.40*	0.61	2.18
	(0.83)	(0.86)	(1.70)	(1.89)	(3.58)	(3.27)
Treated * level 5 * 2 years after program	0.28	0.24	0.91*	1.10***	1.51	2.07
	(0.59)	(0.67)	(0.36)	(0.33)	(1.19)	(1.32)
Treated * level 1 * 3 years after program	-4.12*	-6.10*	-6.16**	-6.81**	-3.21**	-3.31**
	(2.48)	(3.31)	(3.08)	(3.12)	(1.81)	(1.65)
Treated * level 2 * 3 years after program	0.67	1.20	2.07	0.42	-6.38***	-7.92***
	(1.72)	(1.98)	(1.63)	(1.57)	(2.28)	(1.99)
Treated * level 3 * 3 years after program	-0.02	0.87	1.08	1.92	2.20	2.11
	(2.33)	(2.53)	(1.49)	(1.61)	(4.19)	(4.01)
Treated * level 4 * 3 years after program	2.27**	3.77***	2.25	3.36*	4.46	5.83**
	(1.08)	(1.18)	(1.99)	(1.98)	(3.40)	(3.58)
Treated * level 5 * 3 years after program	0.13	0.16	1.04	1.36	2.56***	3.07***
	(0.47)	(0.57)	(1.07)	(1.19)	(0.65)	(0.76)
Controls	Y	Y	Y	Y	Y	Y
Observations	1645	1565	1645	1565	1645	1565

*, **, ***: 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. [1] D.S. stands for discontinuity sample. All regressions include school fixed effects, level dummies and interactions of the level dummies with treated dummies and year dummies respectively. 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. The results from OLS are similar and hence not reported. These OLS regressions also include a polynomial in the selection variable, the percentage of students scoring at or above level 3 in writing.

Table 6: 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 7: 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***	-0.34***
	(0.09)	(0.09)	(0.08)	(0.08)
Math * 1 year after program	-0.65***	-0.65***	-0.60***	-0.59***
	(0.12)	(0.11)	(0.11)	(0.12)
Writing * 1 year after program	-1.27***	-1.28***	-1.26***	-1.27***
	(0.20)	(0.21)	(0.24)	(0.23)
Low * 1 year after program	0.28	0.28	0.25	0.23
	(0.19)	(0.20)	(0.21)	(0.20)
Mid * 1 year after program	0.20	0.21	0.18	0.19
	(0.13)	(0.09)	(0.11)	(0.11)
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	0.62	0.29	0.30

*, **, ***: 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 4 and its fixed effects counterpart. The OLS columns include the three subject dummies, low and mid dummies. The FE columns include school fixed effects, two subject dummies and low and mid dummies. 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 8a: 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***	-0.47**	-0.22	-0.21
	(0.15)	(0.23)	(0.20)	(0.31)
Low * Math * Year 2000	-0.26	-0.02	0.01	-0.08
	(0.20)	(0.15)	(0.25)	(0.21)
Low * Writing * Year 2000	-1.15***	-1.25***	-1.18***	-1.18***
	(0.08)	(0.09)	(0.09)	(0.11)
Mid * Reading * Year 2000	0.01	-0.10	-0.05	-0.09
	(0.07)	(0.10)	(0.12)	(0.11)
Mid * Math * Year 2000	-0.39***	-0.43***	-0.45***	-0.39***
	(0.08)	(0.07)	(0.09)	(0.07)
Mid * Writing * Year 2000	-1.55***	-1.36***	-1.39***	-1.40***
	(0.26)	(0.25)	(0.29)	(0.24)
High * Reading * Year 2000	-0.25***	-0.30***	-0.35***	-0.33***
	(0.07)	(0.08)	(0.07)	(0.09)
High * Math * Year 2000	-0.83***	-0.79***	-0.61***	-0.64***
	(0.18)	(0.15)	(0.18)	(0.15)
High * Writing * Year 2000	-1.26***	-0.94***	-1.02***	-1.05**
	(0.27)	(0.30)	(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 fixed effects, two subject dummies and low and mid dummies. 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 5 and its fixed effects counterpart.

Table 8b: 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	0.00	0.10	0.01
(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	0.01	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	0.00	0.00	0.00
(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	0.03	0.00	0.01
(High * Writing * Year 2000) - (High * Reading * Year 2000)	0.00	0.05	0.04	0.10
(High * Writing * Year 2000) - (High * Math * Year 2000)	0.21	0.39	0.31	0.37
(High * Writing * Year 2000) - (Low * Reading * Year 2000)	0.10	0.10	0.00	0.10
(High * Writing * Year 2000) - (Low * Math * Year 2000)	0.00	0.00	0.01	0.05
(High * Writing * Year 2000) - (Mid * Reading * Year 2000)	0.00	0.01	0.01	0.03
(High * Writing * Year 2000) - (Mid * Math * Year 2000)	0.00	0.10	0.06	0.10

Columns (1), (2), (3) and (4) respectively correspond to columns (1), (2), (3) and (4) of table 8a. P-values reported give the p-values of the F-tests that the differences of the corresponding coefficients in table 8a are zero.

Table 9: Effect of "Threatened Status" on total ESE/Special Education placement
(Sample of treated F and control D schools, 1998-2002)

	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)
Program dummy	0.11	0.09		
	(0.23)	(0.25)		
Program dummy * trend	-0.29	-0.23		
	(0.28)	(0.27)		
Treated * Program dummy	0.53	0.33		
	(0.43)	(0.41)		
Treated * Program dummy * trend	0.51	0.56		
	(0.51)	(0.49)		
Treated * 1 year after program			0.59	0.43
			(0.39)	(0.37)
Treated * 2 years after program			0.79	0.73
			(0.49)	(0.52)
Treated * 3 years after program			0.78	0.73
			(0.59)	(0.65)
Controls	Y	Y	Y	Y
Year dummies	N	N	Y	Y
Observations	2553	2553	2553	2553
p-value[1]	0.05	0.08	0.46	0.57

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. [1] p-value of F-test of joint significance of treatment effects on treated schools. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is percentage ESE membership in school i in year t. Columns (1)-(2) report results from estimation of model (6), columns (3)-(4) from model 7. OLS columns include treated dummy, FE columns include school fixed effects. Columns (1)-(2) also include trend, program dummy, interaction of trend with treated dummy and interaction of program dummy with trend. Columns (3)-(4) include year dummies and interaction of D_1 dummy ($D_1 = 1$ if year $\geq$ 1999) with treated dummy. Controls include race, sex, percentage of students eligible for free or reduced-price lunches and grade distribution of students.

Table 10: Effect of "Threatened Status" on classification in excluded relative to included categories
(Sample of treated F and control D schools, 1998-2002)

	All Excluded Versus Included		Mutable Excluded Versus Included			
			Learning Disabled Versus Included		Emotionally Handicapped Versus Included	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * 1 year after program	0.16	0.08	0.08	0.08	0.05	0.05
	(0.15)	(0.13)	(0.12)	(0.13)	(0.12)	(0.13)
Treated * 2 years after program	0.11	0.09	0.04	0.05	0.07	0.06
	(0.20)	(0.22)	(0.19)	(0.20)	(0.21)	(0.21)
Treated * 3 years after program	0.36	0.34	0.25	0.27	0.30	0.28
	(0.18)	(0.23)	(0.21)	(0.23)	(0.22)	(0.24)
Exempt * treated * 1 year after program	0.27	0.27	0.01	0.02	-0.01	-0.01
	(0.32)	(0.34)	(0.23)	(0.24)	(0.14)	(0.14)
Exempt * treated * 2 years after program	0.55	0.56	0.47	0.47	-0.04	-0.04
	(0.53)	(0.56)	(0.39)	(0.41)	(0.22)	(0.24)
Exempt * treated * 3 years after program	0.05	0.06	0.13	0.13	-0.25	-0.25
	(0.70)	(0.71)	(0.49)	(0.52)	(0.22)	(0.23)
Observations	5106	5106	5106	5106	5106	5106
p-value[1]	0.12	0.15	0.01	0.02	0.29	0.33

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. [1]p-value of F-test of the program effect on relative classification in excluded categories in treated schools. Robust standard errors adjusted for clustering by school district are in parentheses. The dependent variable is percentage of students in school i classified in category j in year t. LD stands for learning disabled category and EH for emotionally handicapped category. All columns report results from the unrestricted model (9). Results from model (8) are similar and hence not reported. The FE columns include school fixed effects while the OLS columns include a treated dummy. Controls include race, sex, percentage of students eligible for free or reduced-price lunches and grade-distribution of students.

Table 11: Pre-program Trend of D and F Schools, 1993-97

(Total Special Education Classification, Excluded relative to Included categories and Mutable Excluded relative to Included categories)

	Total ESE Classification		All Excl. categories vs. Included		Mutable excluded categories vs Included			
					LD vs. Included		EH vs. Included	
	OLS	FE	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treated * trend	-0.04	0.00	0.00	0.05	0.05	0.06	0.06	0.06
	(0.13)	(0.90)	(0.10)	(0.16)	(0.10)	(0.11)	(0.10)	(0.11)
Exempt * treated * trend			-0.09	-0.09	-0.14	-0.14	-0.09	-0.09
			(0.12)	(0.23)	(0.10)	(0.15)	(0.11)	(0.14)
Controls	Y	Y	Y	Y	Y	Y	Y	Y
Observations	2450	2450	4900	4900	4900	4900	4900	4900

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. OLS columns include an F dummy, FE columns include school fixed effects. The dependent variable in columns (1)-(2) is percentage ESE membership in school i in year t, the dependent variable in columns (3)-(8) is percentage of students in school i classified in category j in year t. All columns include trend, columns (3)-(8) also include an Exempt dummy and interactions of Exempt dummy with trend and treated dummies respectively. LD and EH stand for learning disabled and emotionally handicapped respectively. Controls include race, sex, percentage of students eligible for free or reduced-price lunches and grade distribution of students.

Table 12: Regression Discontinuity Analysis:
Effect of "Threatened Status" on total ESE placement

	Discontinuity Sample 1		Discontinuity Sample 2	
	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)
Treated * 1 year after program	0.52	0.69**	0.78	0.92**
	(0.48)	(0.35)	(0.51)	(0.40)
Treated * 2 years after program	0.00	0.31	0.14	0.70
	(0.87)	(0.81)	(0.81)	(0.57)
Treated * 3 years after program	0.38	0.52	0.43	1.01
	(0.99)	(0.90)	(1.00)	(0.87)
Year dummies	Y	Y	Y	Y
Observations	415	415	395	395
p-value[1]	0.04	0.12	0.00	0.19

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. [1] p-value of F-test of joint significance of treatment effects on treated schools. Robust standard errors adjusted for clustering by school district are in parentheses. The OLS columns include a treated dummy, the FE columns include school fixed effects. The OLS columns include a cubic in the selection variable, the percentage of students scoring at or above level 3 in writing. All columns include year dummies and interaction of D_1 dummy ($D_1 = 1$ if year ≥ 1999) with treated dummy. Controls include race, sex, percentage of students eligible for free or reduced-price lunches and grade distribution of students.

Table 13: Regression Discontinuity Analysis: Effect of "Threatened Status" on classification in excluded relative to included categories

	All Excl. categories vs. Incl.		Mutable Excluded categories vs Included			
	D. S. 1	D. S. 2	LD vs. Included		EH vs. Included	
			D. S. 1	D. S. 2	D. S. 1	D. S. 2
	(1)	(2)	(3)	(4)	(5)	(6)
Treated * 1 year after program	0.07	0.03	-0.06	-0.09	-0.10	-0.12
	(0.17)	(0.18)	(0.16)	(0.14)	(0.15)	(0.13)
Treated * 2 years after program	0.33	0.20	0.13	0.11	0.03	0.00
	(0.41)	(0.32)	(0.30)	(0.24)	(0.31)	(0.26)
Treated * 3 years after program	0.50	0.52	0.10	0.25	0.06	0.19
	(0.77)	(0.75)	(0.67)	(0.63)	(0.63)	(0.59)
Exempt * treated * 1 year after program	0.55	0.87*	0.20	0.14	0.23	0.29
	(0.42)	(0.46)	(0.20)	(0.25)	(0.20)	(0.21)
Exempt * treated * 2 years after program	-0.35	-0.30	-0.03	0.08	-0.11	0.07
	(0.86)	(0.63)	(0.47)	(0.41)	(0.40)	(0.41)
Exempt * treated * 3 years after program	-0.47	-0.02	0.06	0.30	-0.12	-0.06
	(1.33)	(1.21)	(0.74)	(0.81)	(0.66)	(0.70)
Observations	830	790	830	790	830	790
p-value[1]	0.17	0.18	0.99	0.88	0.53	0.46

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. [1] p-value of F-test of joint significance of treatment effects on treated schools. Robust standard errors adjusted for clustering by school district are in parentheses. LD and EH stand for learning disabled and emotionally handicapped respectively. All columns include school fixed effects and report results from fixed effects counterpart of model (9). Controls include race, sex, percentage of students eligible for free or reduced-price lunches and grade-distribution of students. The results from OLS are similar and hence not reported. These OLS regressions also include a polynomial in the selection variable, the percentage of students scoring at or above level 3 in writing.

Table 14: Effect of "Threatened Status" on percentage of students tested
(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 * 1 year after program	-0.60	-0.69	-1.15	-1.12	0.15	0.12
	(0.94)	(1.04)	(0.93)	(1.06)	(0.89)	(0.93)
Treated * 2 years after program	0.74	0.11	0.75	0.53	0.62	0.64
	(1.24)	(1.11)	(0.99)	(1.00)	(1.31)	(0.94)
Treated * 3 years after program	0.34	-0.07	0.17	0.10	0.68	0.46
	(0.93)	(1.10)	(1.32)	(1.11)	(0.82)	(0.90)
Controls	Y	Y	Y	Y	Y	Y
Observations	2525	2525	2511	2511	4491	4491
p-value	0.29	0.90	0.03	0.44	0.62	0.89

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions include year dummies and interactions of year dummies with treated dummy. The OLS columns include a treated dummy, the FE columns include school fixed effects. Controls include race, sex, percentage of students eligible for free or reduced-price lunches and real per pupil expenditure.

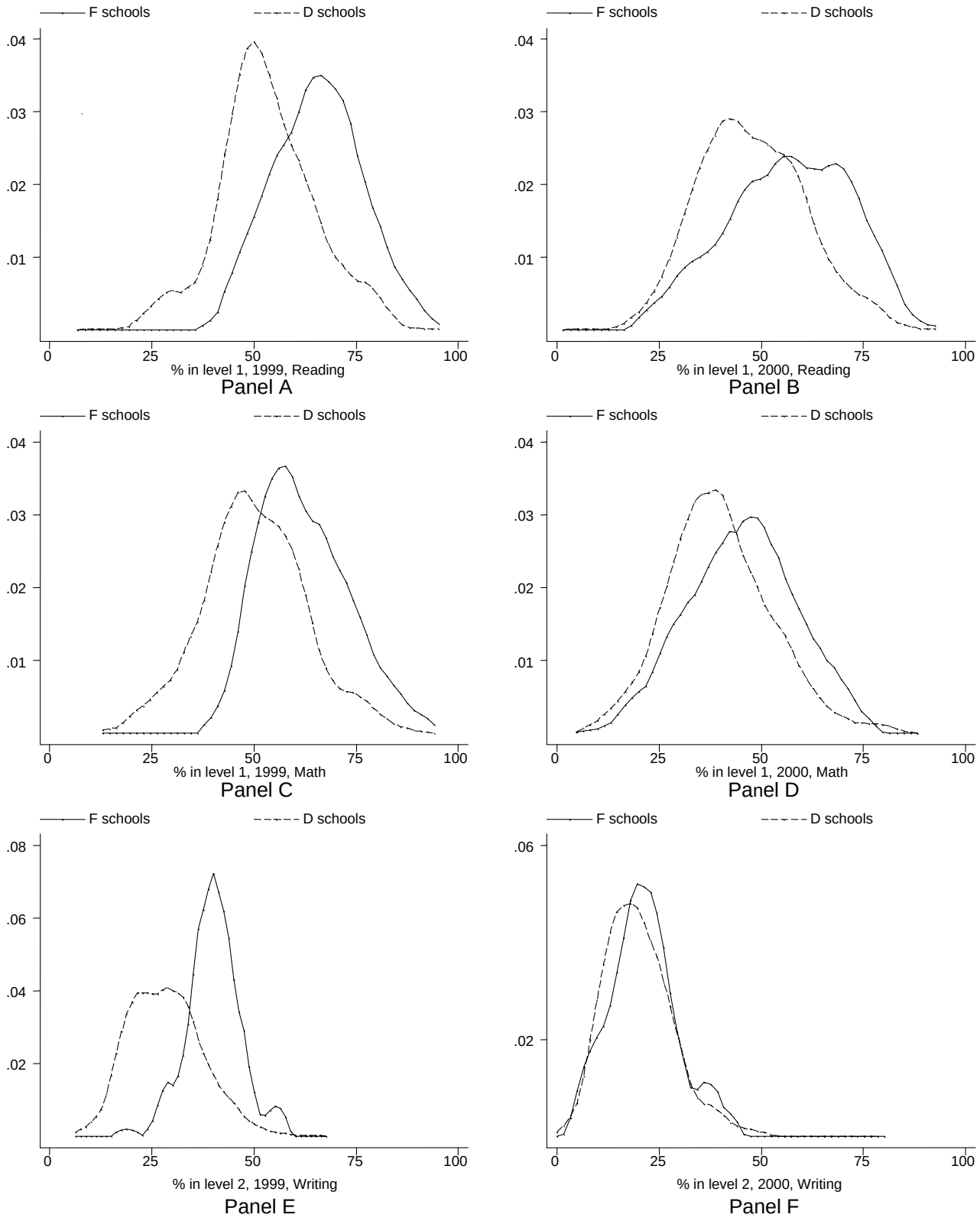


Figure 1. Distribution of percentage of students in level 1 Reading, level 1 Math and level 2 Writing, F and D Schools, 1999 and 2000

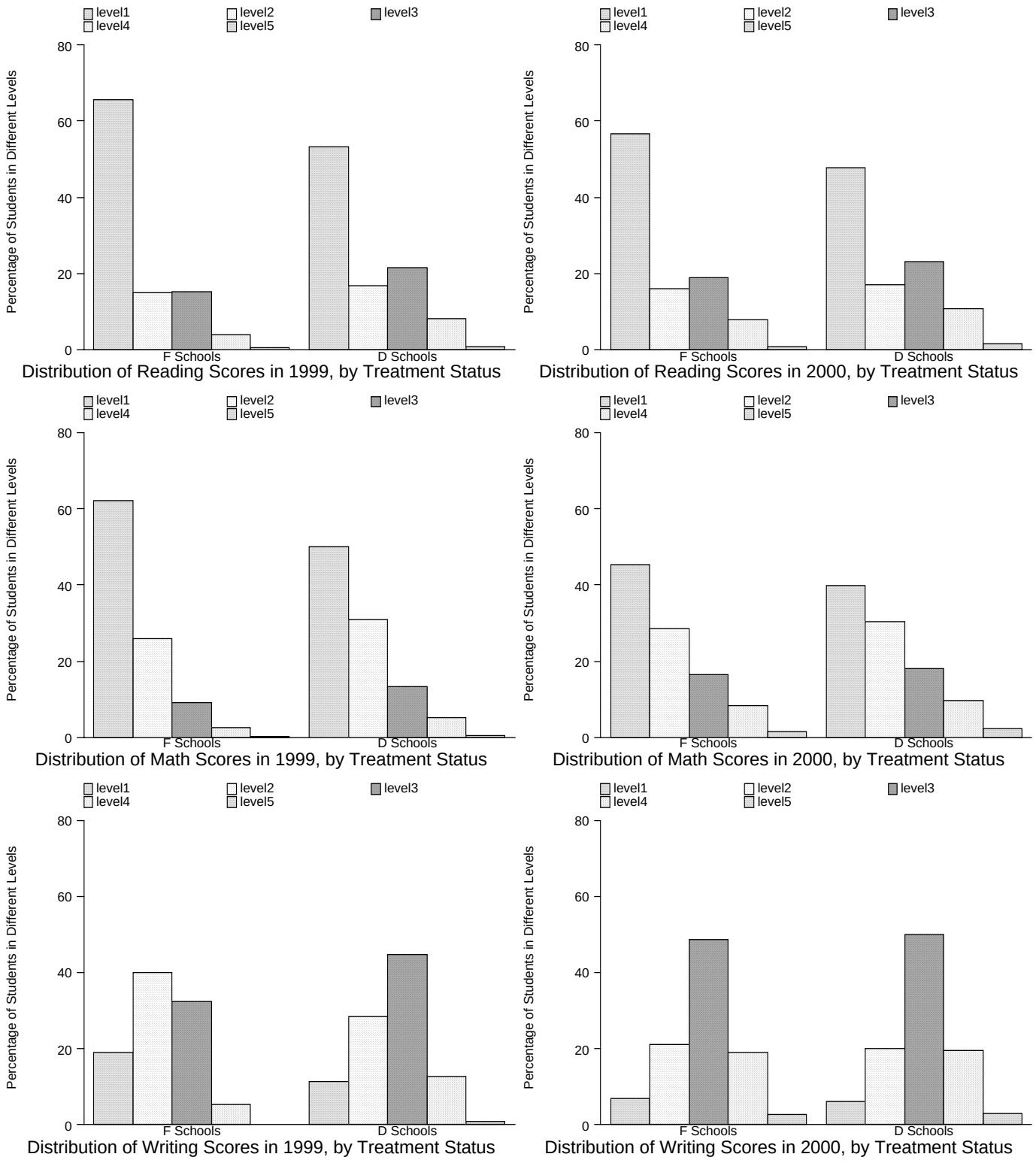


Figure 2. Distribution of Reading, Math and Writing Score for F and D schools (1999 and 2000)

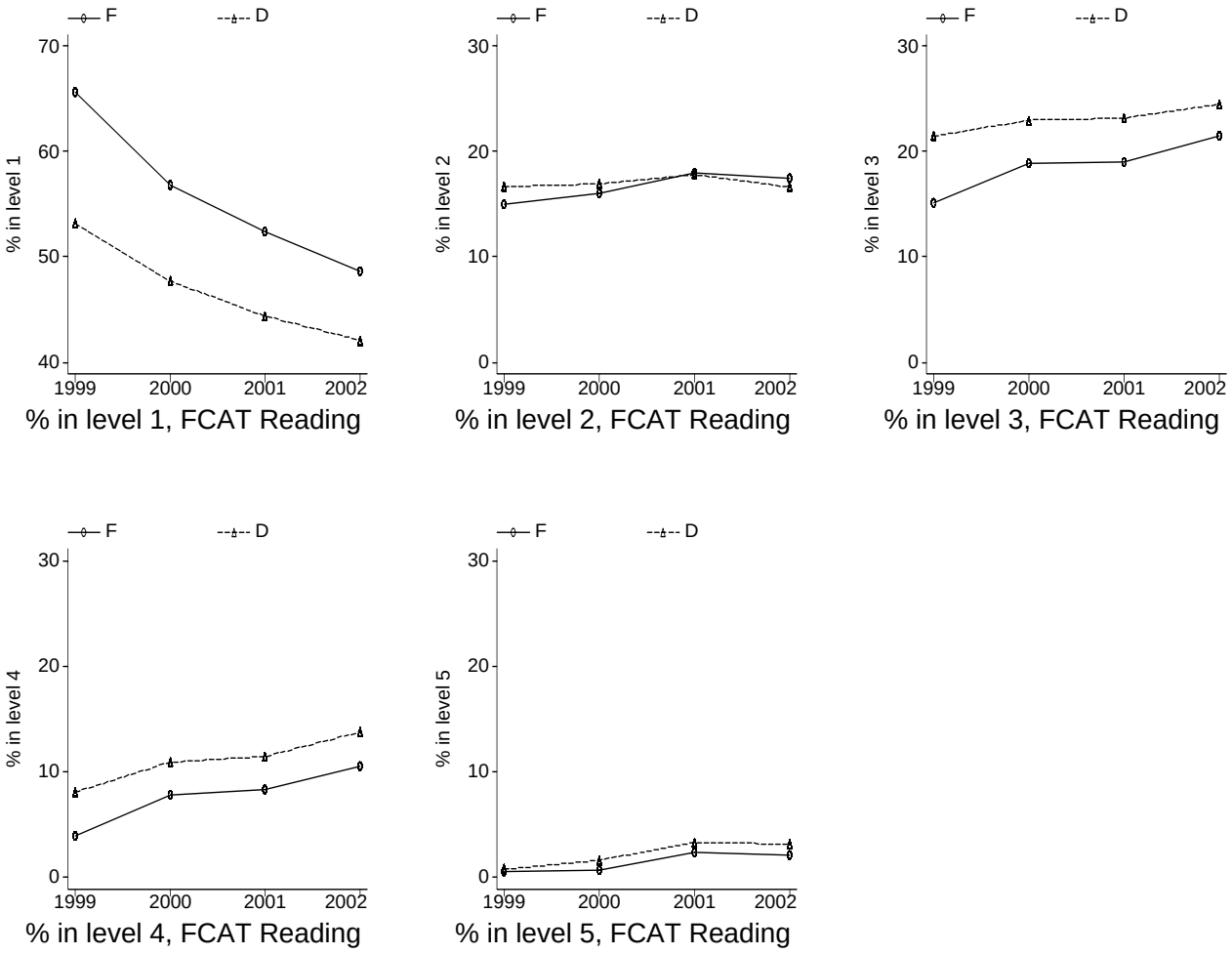


Figure 3. Percentage of students in levels 1-5, FCAT Reading, F and D schools

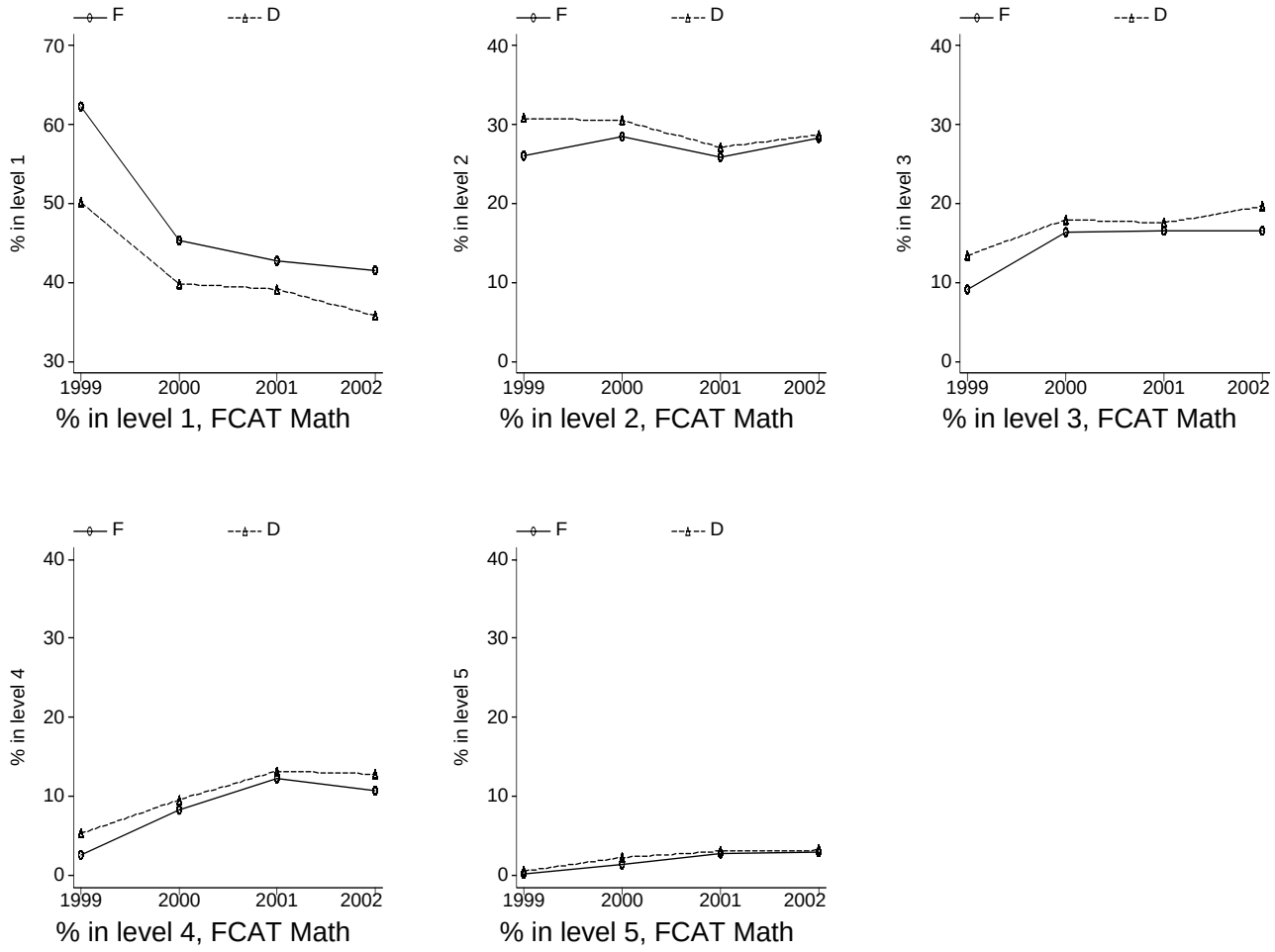


Figure 4. Percentage of students in levels 1-5, FCAT Math, F and D schools

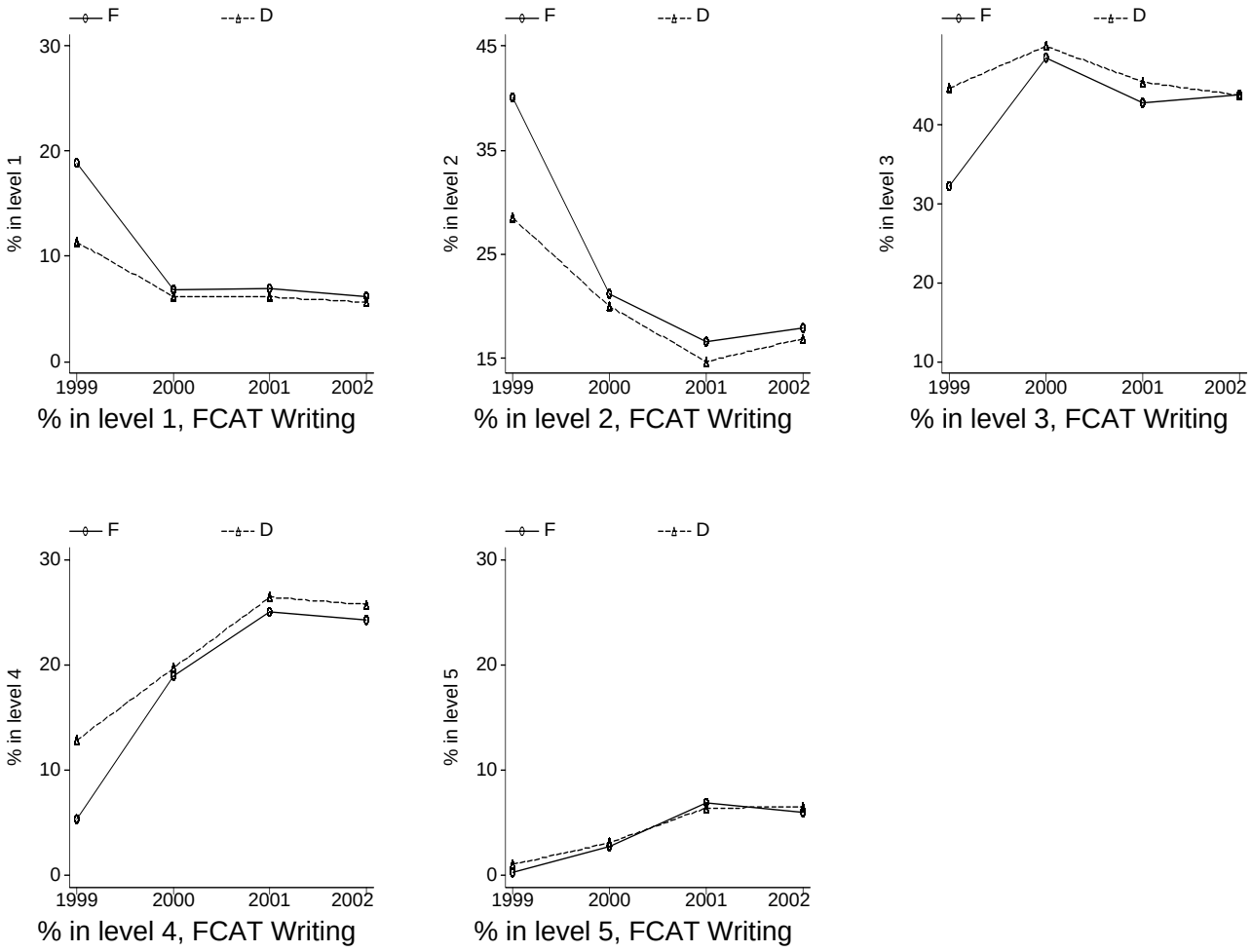


Figure 5. Percentage of students in levels 1-5, FCAT Writing, F and D schools

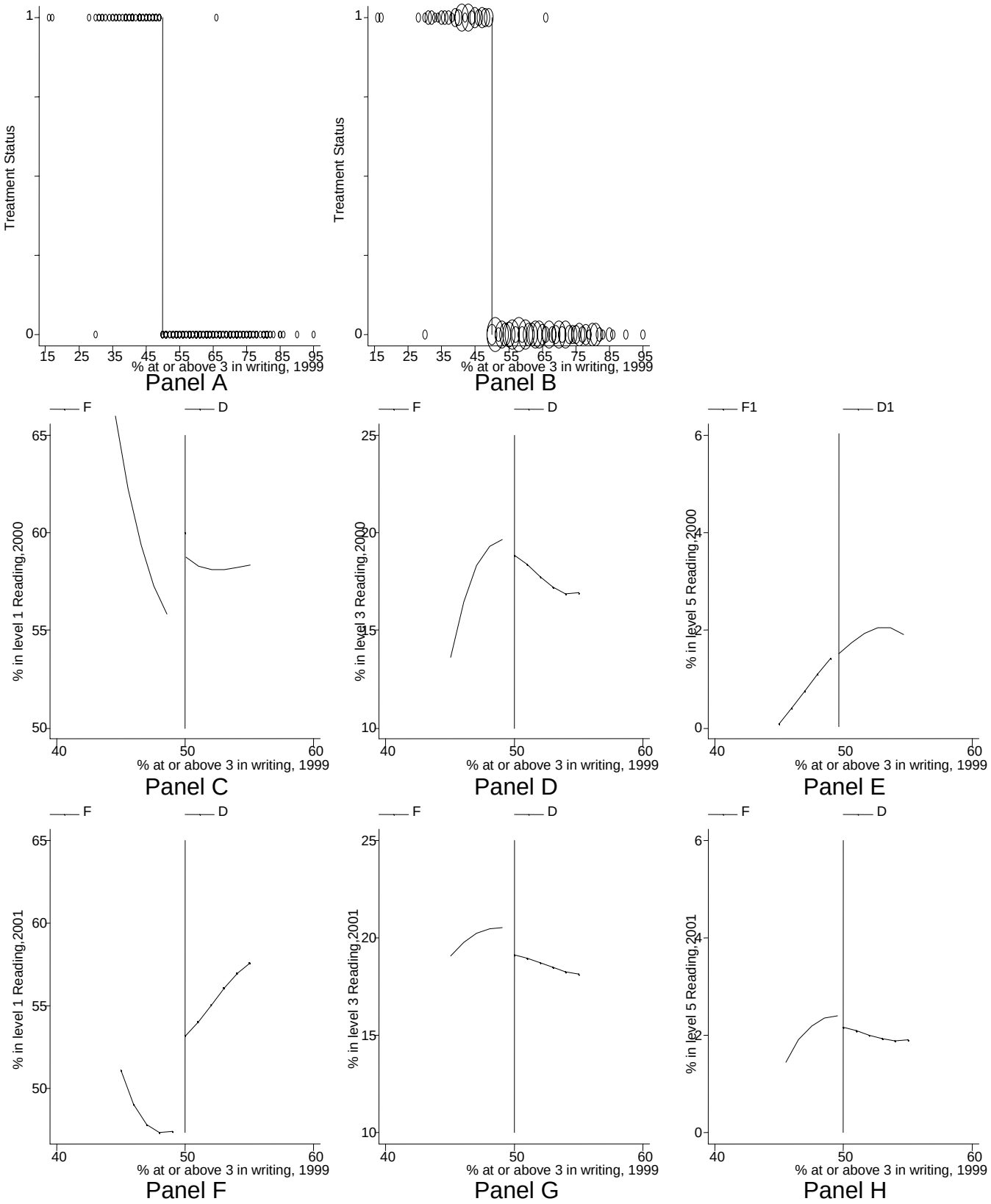


Figure 6. Regression Discontinuity Analysis:
Relationship between % of students at or above 3 in writing and Treatment Status (Panels A-B) and Effect of Treatment Status on FCAT Reading (Panels C-H)

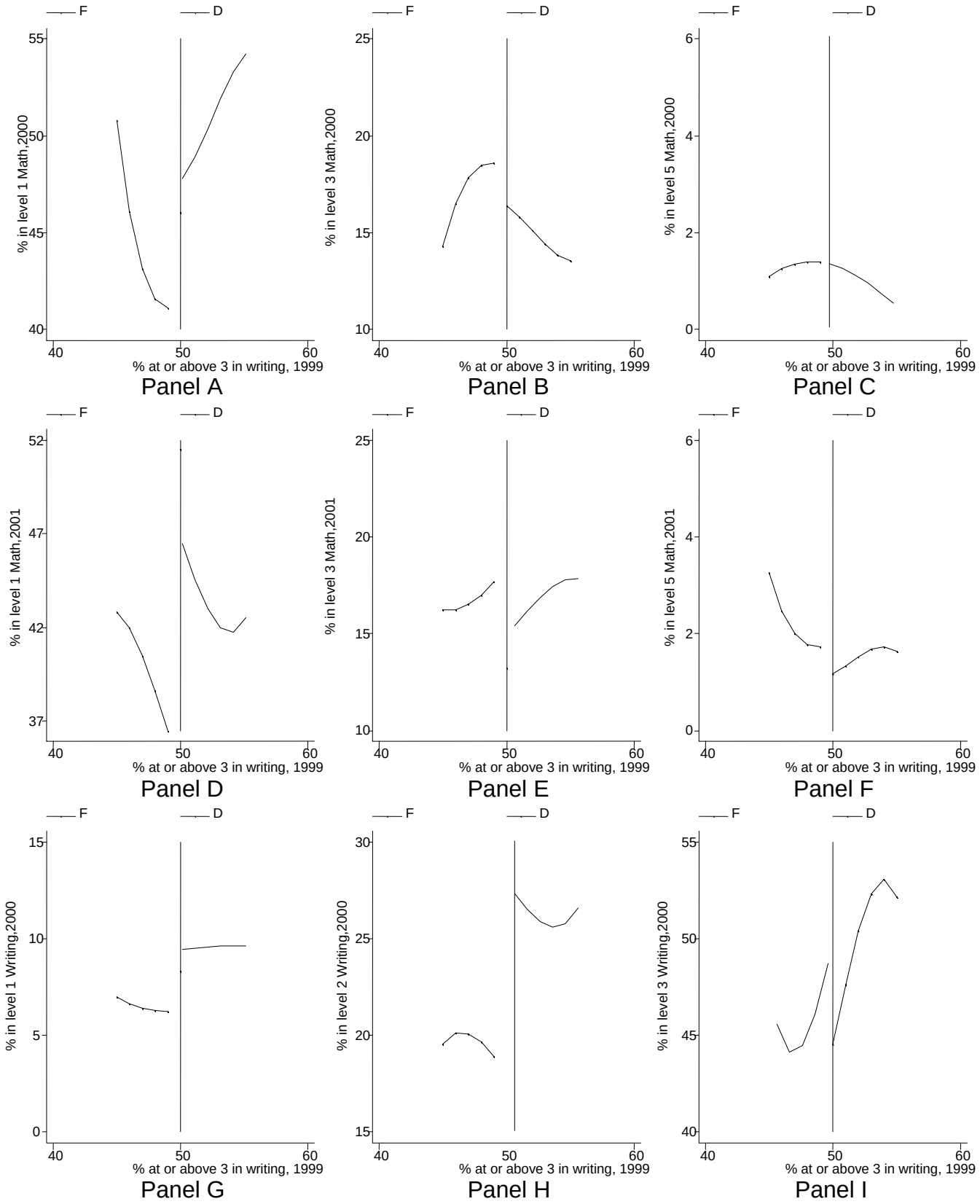
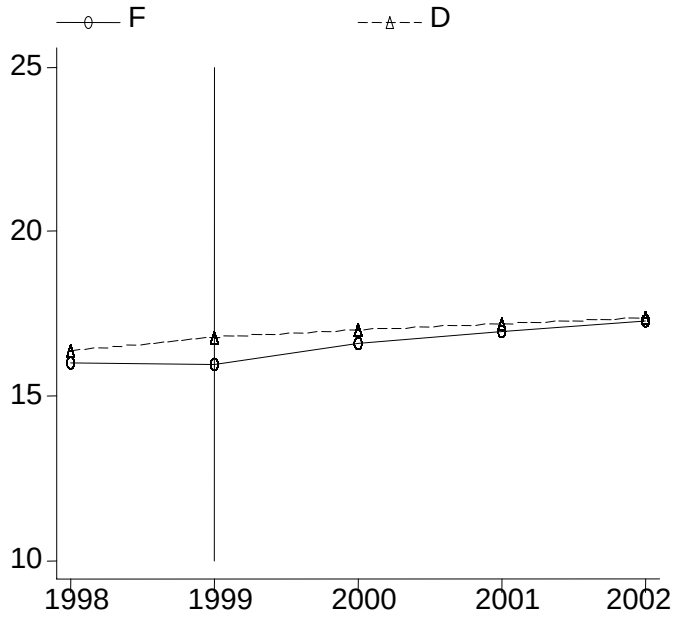
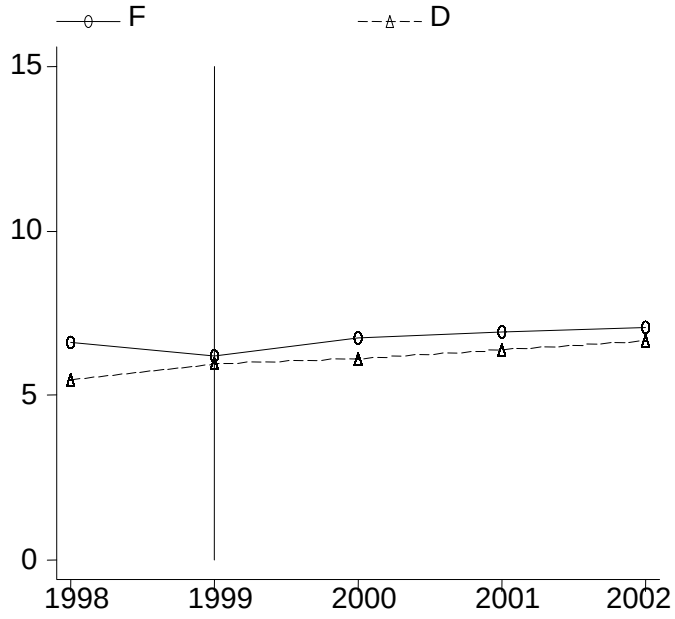


Figure 7. Regression Discontinuity Analysis:
Effect of Treatment Status on FCAT Reading (Panels A-F) and Writing (Panels G-I)



Panel A. Total ESE Classification



Panel B. Classification into Excluded Relative to Included ESE Categories

Figure 8. Classification into ESE Categories, F and D schools