Federal Reserve Bank of New York
Staff Reports

Vouchers, Responses, and the Test-Taking Population: Regression Discontinuity Evidence from Florida

Rajashri Chakrabarti

Staff Report no. 486
March 2011

Vouchers, Responses, and the Test-Taking Population: Regression Discontinuity Evidence from Florida

Rajashri Chakrabarti

Federal Reserve Bank of New York Staff Reports, no. 486
March 2011
JEL classification: H4, I21, I28

Abstract

While there is a rich literature that investigates whether accountability regimes induce schools to manipulate their test-taking population by strategically excluding weaker students, no study thus far investigates whether voucher programs induce schools to engage in similar strategic behavior. This paper analyzes a Florida program that embedded vouchers in an accountability regime. Specifically, it investigates whether the threat of vouchers and the stigma associated with the Florida program induced schools to strategically manipulate their test-taking population. Under Florida rules, scores of students in several special-education and limited-English-proficient (LEP) categories were not included in the computation of school grades. Did this rule induce the threatened schools to reclassify some of their weaker students into these "excluded" categories so as to remove them from the effective test-taking pool? Using a regression discontinuity strategy, I find evidence in favor of strategic reclassification into the excluded LEP category in high-stakes grade 4 and entry-grade 3. In contrast, I find no evidence that the program led to reclassification into excluded special-education categories, which is consistent with the substantial costs of classifying into special-education categories during this period. These findings have important policy implications.

Key words: vouchers, incentives, regression discontinuity

Chakrabarti: Federal Reserve Bank of New York (e-mail: rajashri.chakrabarti@ny.frb.org). For helpful discussions, the author thanks David Figlio, Caroline Hoxby, Brian Jacob, Sarah Turner, and seminar participants at Duke University, the University of Florida, Harvard University, the University of Maryland, and the Massachusetts Institute of Technology, as well as participants at conferences sponsored by Northwestern University, the American Economic Association, the American Education Finance Association, the Econometric Society, the Association for Public Policy Analysis and Management, and the Society of Labor Economists. She also thanks the Florida Department of Education for the data used in this analysis. Noah Schwartz provided excellent research assistance. The views expressed in this paper are those of the author and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

1 Introduction

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

The Florida voucher program, known as the "opportunity scholarship" program, embeds a voucher program within a school accountability system. Written into law in June 1999, the Florida voucher program made all students of a school eligible for vouchers if the school received two "F" grades in a period of four years. Thus, it can be looked upon as a "threat of voucher" program—schools getting an "F" grade for the first time were directly threatened by vouchers, but vouchers were implemented only if they got another "F" grade in the next three years. Vouchers were associated with a loss in revenue (equivalent to state aid per pupil for each student) and also negative media publicity and visibility. Moreover, the "F" grade, being the lowest performing grade, was likely associated with shame and stigma. Therefore, the threatened schools had a strong incentive to try to avoid the second "F". This paper studies some alternative ways in which the threatened schools might have responded to the incentives built into the program.[1]

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

Under Florida rules, scores of limited English proficient (LEP) students who were in an ESOL (English for speakers of other languages) program for less than two years were not included in the calculation of school grades. Did this induce the threatened schools to classify some of their weaker students into this "excluded" LEP category so as to remove them from the relevant test taking pool that counted towards grade formation? Similarly, 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 classify their low-performing students into these "excluded" ESE categories, so as to remove them from school grade calculations and artificially boost scores?

Using data obtained from the Florida Department of Education and a regression discontinuity estimation strategy that exploits the institutional details of the Florida program, I find that the program led to increased classification into the excluded LEP category in high stakes grade 4 and the entry grade to high stakes grade (grade 3)[2] in the first year after program. Specifically, the threatened schools classified an additional 0.31% of their total students in the excluded LEP category in grade 4 and an additional 0.36% of their total students in this category in grade 3 in the first year after program. In contrast, I do not find any evidence that the threatened schools resorted to increased classification into "excluded" ESE categories in any of the three years after program. As discussed later in the paper, classification into ESE categories was associated with substantial costs during this period[3], which might have acted as a disincentive to this form of reclassification.

behavior of the 1999 threatened schools during these three years.

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

[3] Among other things, described later, the "McKay Scholarship program for Students with Disabilities" in Florida acted as a major disincentive to such classification. Since this program made every special education student in Florida public schools eligible for vouchers, reclassification into ESE categories was associated with a threat of losing the corresponding students (and corresponding revenue).

This study is related to two strands of literature. The first strand investigates whether schools facing accountability systems and testing regimes respond by gaming the system in various ways. This relates to the moral hazard problems associated with multidimensional tasks under incomplete observability, as pointed out by Holmstrom and Milgrom (1991). Cullen and Reback (2006), Figlio and Getzler (2006) and Jacob (2005) find evidence of classification of low-performing students into excluded disabled categories; Jacob (2005) finds evidence of teaching to the test, preemptive retention of students and substitution away from low-stakes subjects; Jacob and Levitt (2003) find evidence of teacher cheating; Reback (2008), Ladd and Lauen (2010) and Neal and Schanzenbach (2010) find evidence in favor of differential focus on marginal students; Figlio (2006) finds that low-performing students were given longer suspensions during the testing period than higher performing students for similar crimes; Figlio and Winicki (2005) find that schools faced with accountability systems increased the caloric content of school lunches on testing days in an attempt to boost performance.

The second strand of literature analyzes the effect of voucher programs on public school performance and behavior. Hoxby (2003a, 2003b) and Chakrabarti (2008b) find that the expansion of the Milwaukee voucher program in the late 1990's led to an improvement of the treated schools facing the program. Figlio and Hart (2010) find evidence in favor of improvement of threatened schools facing voucher-threats via the Florida tax credit scholarship program. A wide volume of literature finds that "threat of vouchers" and stigma associated with the Florida program have led to an improvement of the corresponding treated schools (Greene (2001, 2003), Chakrabarti (2008a), Figlio and Rouse (2006), West and Peterson (2006)). Rouse et al. (2007) and Chiang (2009) find evidence in favor of persistence of achievement gains in the medium-run of students who attended voucher-threatened schools in Florida. Both studies also find evidence in favor of

behavioral changes of these schools,—such as more focus on instruction and teacher development. Chakrabarti (2010) finds that threatened schools facing the same program in Florida tended to focus more on students expected to score just below the minimum criteria cutoffs. Goldhaber and Hannaway (2004) and Chakrabarti (2010) also find evidence that the threatened schools tended to overwhelmingly focus on writing, rather than reading and math (passing in one subject was sufficient to escape an "F").

Thus, while there are studies that investigate whether accountability regimes lead affected schools to re-classify their low-performing students into excluded categories, there is no study thus far that investigates whether public schools facing voucher systems resort to this form of strategic behavior as well. This paper fills this important gap. In spite of the evidence on accountability systems, it would be instructive to know whether public schools resort to similar strategic classification facing a voucher system tied to accountability,—an alternative form of public school reform. In general, evidence on gaming or strategic behavior of schools facing voucher programs is sparse. To the best of my knowledge, there are only three studies thus far (all mentioned above) that investigate three other forms of strategic behavior in the context of voucher programs,—teaching to the test (Figlio and Rouse (2006)), strategic focus on marginal students (Chakrabarti (2007)) and differential focus on subject areas (Goldhaber and Hannaway (2004) and Chakrabarti (2007)). For a fuller understanding of the behavior of public schools facing such voucher systems tied to accountability, it is imperative to understand whether they resort to manipulation of their test-taking population in an effort to artifically inflate their grades/scores. This study contributes in that direction.

2 Institutional Details

The Florida Opportunity Scholarship Program was signed into law in June 1999. Under this program, all students of a public school became eligible for vouchers or "opportunity scholarships" if the school received two "F" grades in a period of four years. A school receiving an "F" grade for the first time was exposed to the threat of vouchers and stigma, but its students did not become eligible for vouchers unless and until it got a second "F" within the next three years.

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

The system of assigning letter grades to schools started in the year 1999,[4] and they were based on the FCAT reading, math and writing tests. The state designated a school an "F" if it failed to attain the minimum criteria in all three FCAT subjects (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 or above.[5]

Scores of all regular students were included in the computation of school grades. However, scores of students in only some 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 as well as LEP stu-

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

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

dents with more than two years in an ESOL program were included in school grade computations. In contrast, scores of LEP students who were in an ESOL program for less than two years were not included in the computation of grades, nor were scores of ESE students in eighteen ESE categories. Florida classified ESE students into 21 ESE categories in total,—educable mentally handicapped, trainable mentally handicapped, orthopedically handicapped, occupational therapy, physical therapy, speech impaired, language impaired, deaf or hard of hearing, visually impaired, emotionally handicapped, specific learning disabled, gifted, hospital/homebound, profoundly mentally handicapped, dual-sensory impaired, autistic, severely emotionally disturbed, traumatic brain injured, developmentally delayed, established conditions and other health impaired. From now on, I will refer to the "less than two years in an ESOL program" category as the "excluded" LEP category and "2 years or more in an ESOL program" category as the "included" LEP category. Similarly I will refer to the speech impaired, gifted, and hospital/homebound categories as "included" ESE categories, and to the other ESE categories as "excluded" ESE categories.

3 Data

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

School-level data were also obtained on the distribution of students in the various ESE categories. In addition to information on total ESE enrollment, these data also report enrollment in each of the ESE categories in each Florida school for the years 1999 through 2002.

School-level data on the grade distribution (K-12) of students are available from 1999-2002. Data on socio-economic characteristics include data on gender composition, race composition and percent of students eligible for free or reduced-price lunches. School finance data consist of several measures of school-level and district-level per pupil expenditures and are available for the years under consideration.

4 Empirical Strategy

Under the Florida opportunity scholarship program, schools that received a grade of "F" in 1999 were directly threatened by "threat of vouchers" and stigma,–the former in the sense that all their students would be eligible for vouchers if the school received another "F" grade in the next three years. I will refer to these schools as "F schools" from now on.[6] The schools that received a "D" in 1999 were closest to the F schools in terms of grade, but were not directly threatened by the program. I will refer to them as "D schools" in the rest of the paper. Given the nature of the Florida program, the threat of vouchers faced by the 1999 F schools would be applicable for the next three years only. Therefore, I study the behavior of the F schools (relative to the D schools) during the first three years of the program (that is, upto 2002).

I use a regression discontinuity analysis to analyze the effect of the program. The analysis essentially entails comparing the response of schools that barely missed D and received an F with schools that barely got a D. The institutional structure of the Florida program allows me to follow this strategy. The program created a highly non-linear and discontinuous relationship between the percentage of students scoring above a pre-designated threshold and the probability that the

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

school's students would become eligible for vouchers in the near future, which enables the use of such a strategy.

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

Using the sample of F and D schools that failed the minimum criteria in both reading and math in 1999, Figure 1, Panel A illustrates the relationship between assignment to treatment (i.e. facing the threat of vouchers) and the schools' percentages of students scoring at or above 3 in FCAT writing. The figure shows that all but one of the schools in this sample that had less than 50% of their students scoring at or above 3 actually received an F grade. Similarly, all schools (except one) in this sample that had 50% or a larger percentage of their students scoring at or above 3 were assigned a D grade. Note that many of the dots correspond to more than one school; Figure 1, Panel B illustrates the same relationship where the sizes of the dots are proportional to the number of schools at that point. The smallest dot in this figure corresponds to one school. These two panels show that in this sample, the percentage of students scoring at or above 3 in writing indeed uniquely predicts (except two schools) assignment to treatment and there is a discrete change in the probability of treatment at the 50% mark.

I also consider two corresponding samples where both F and D schools fail the minimum criteria in reading and writing (math and writing). According to the Florida rules, F schools would fail

the minimum criteria in math (reading) also, unlike D schools. I find that indeed in these samples, the probability of treatment changes discontinuously as a function of the percentage of students scoring at or above level 2 in math (reading) and there is a sharp cutoff at 60%. However, the sizes of these samples are considerably smaller than above and the samples just around the cutoff are considerably less dense. So I focus on the first sample above, where the D schools passed the writing cutoff and the F schools missed it, and both groups of schools missed the cutoffs in the other two subject areas. The results reported in this paper are from this sample. Note, though, that the results from the other two samples are qualitatively similar.

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

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

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

Hahn, Todd and van der Klaauw (2001) show that γ_1 is identified by the difference in average outcomes of schools that just missed the cutoff and those that just made the cutoff, provided the conditional expectations of the other determinants of Y are smooth through the cutoff. Here, γ_1

identifies the local average treatment effect (LATE) at the cutoff.

The estimation can be done in multiple ways. In this paper, I use local linear regressions with a triangular kernel and a rule of thumb bandwidth suggested by Silverman (1986). I also allow for flexibility on both sides of the cutoff by including an interaction term between the run variable and a dummy indicating whether or not the school falls below the cutoff. I estimate alternate specifications that do not include controls as well as those that use controls.[7] Assuming the covariates are balanced on both sides of the cutoff (I later test this restriction), the purpose of including covariates is variance reduction. They are not required for the consistency of γ_1.

To test the robustness of the results, I also experiment with alternative bandwidths. The results remain qualitatively similar and are available on request. In addition, I also do a parametric estimation where I include a third order polynomial in the percentage of students scoring at or above 3 in writing and interactions of the polynomial with a dummy indicating whether or not the school falls below the cutoff. I also estimate alternative functional forms that include fifth order polynomial instead of a third order polynomial and the corresponding interactions.[8] The results remain very similar in each case and are available on request.

4.1 Testing Validity of the Regression Discontinuity Analysis

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

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

[8] I use odd order polynomials because they have better efficiency (Fan and Gijbels (1996)) and are not subject to boundary bias problems, unlike even order polynomials.

Next, I examine whether the use of a regression discontinuity strategy is valid here. As discussed above, identification of γ_1 requires that the conditional expectations of various pre-program characteristics be smooth through the cutoff. Using the strategy outlined above, I test if that was indeed the case. Note though that there is not much reason to expect manipulation or selection in this particular situation. The program was announced in June 1999 while the tests were given a few months before, in January and February of 1999. Also, any form of strategic response with the objective of precise manipulation of test scores likely takes quite some time. It is unlikely that the schools had the time or information to manipulate the percentage of students above certain cutoffs before the tests.

Nevertheless, I check for continuity of pre-determined characteristics at the cutoff, using the strategy outlined above. The corresponding graphs are presented in Figures 2A and 2B and the discontinuity estimates in Table 1. Figure 2A considers pre-program (1999) demographic and socio-economic characteristics, while Figure 2B considers classification in excluded and included LEP and ESE categories in the pre-program (1999) period. The discontinuity estimates are never statistically distinguishable from zero. Visually examining the graphs, it seems that unlike in the cases of the other pre-determined characteristics, there is a small discontinuity in the variable, "percentage of school's students eligible for free or reduced price lunches". But the discontinuity is small and not statistically significant (with a p-value of 0.28). Also, note that even if it were statistically significant, with a large number of comparisons, one might expect a few to be statistically different from zero by sheer random variation. So, from the above discussion, it seems reasonable to say that this case passes the test of smoothness of predetermined characteristics through the cutoff.

Following McCrary (2008), I also test whether there is unusual bunching at the cutoff. Using

the density of the run variable (percentage of students at or above 3 in writing in 1999) and the strategy above, I test for a discontinuity in the density of the run variable at the cutoff. As can be seen from Table 2, there is no evidence of a statistically significant discontinuity in the density function at the cutoff in 1999.

5 Results

Having established that the use of a regression discontinuity strategy in this setting is valid, I next look at the effect of the program on the behavior of threatened schools. For reference, let's first look at the behavior of these same schools in the pre-program period. Figure 2B and Table 1 (Panels C-E) look at the LEP and ESE classification in excluded and included categories in 1999, the year just before program. There is no evidence that the schools that would be threatened the next year behaved any differently than the non-threatened schools in excluded or included LEP classification in any of the high stakes or low stakes grades. Nor is there any evidence of any differential classification in excluded or included ESE categorization in 1999. The picture in the post-program period is very different, as seen below.

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

First, consider the excluded category. In the first year after program, the table finds that the program led to a statistically significant increase in the percentage of students classified in excluded LEP categories in the high-stakes grade 4 and the entry grade 3. In contrast, there is no evidence of an increase in the low-stakes grade 2 or the high-stakes grade 5. The estimates suggest that in the first year after program, schools facing the threat of vouchers classified an additional

0.31% of their total students in the excluded LEP category in grade 4 and an additional 0.36% of their students in grade 3. Since it might have been difficult to do the classification all at once, the administrators might have chosen to phase out the process to the entry grade 3. In the second year after program (column 2), there is evidence of positive and statistically significant shifts in the excluded LEP category in grades 4 and 5 in the threatened schools. Compared with the effects in the first year, it seems that the increase in grade 5 (grade 4) in the second year was generated by the increased classification in grade 4 (grade 3) in the first year after program. There does not seem to have been any new classification in the second year after program. Similarly, there is no evidence of any new classification in the third year after the program(column 3).

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

Figures 3A and 3B display the effects of the program on classification in excluded and included LEP categories graphically. While the estimates presented in the table includes controls, the graphs display results of estimations without controls. As might be expected from the results of the continuity tests above, the patterns are very similar and do not depend on inclusion of controls.

The above results can be summarized as follows. In the pre-program period, there is no evidence that the would-be threatened schools behaved any differently than the would-be non-threatened schools in terms of categorization of students in excluded or included LEP categories in any of the high-stakes or low-stakes grades or in terms of classification in excluded or included ESE categories. In contrast, the program led to increased classification of students into the excluded LEP category in the high-stakes grade 4 and the entry grade 3 in the first year after program.

There is no evidence of any new classification in this category either in the second or third years after program. Nor is there any evidence of differential classification in the included category in any of the three years after program. Students classified into the excluded LEP category in grade 4 in the first year after program would not count in school grades either in the current year or in the following year (that is, in both high stakes grades 4 and 5). Students classified into the excluded LEP category in grade 3 would not count the following year when they would be in the high stakes grade 4. So the findings above suggest that the threatened schools attempted to remove certain students from the effective test-taking pool, both in the current year and in the following year, by classifying them into the excluded LEP category.

Tables 4, 5 and 6 look at the effect of the program on ESE classification. Table 4 looks at the effect on total ESE classification. The dependent variable for this analysis is percentage ESE enrollment, i.e., total ESE enrollment as a percentage of total enrollment. The estimates show that there is no evidence in favor of any differential classification in the threatened schools at the cutoff.

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

To have a closer look, Table 5 looks at the effect of the program on classification in excluded (Panel A) and included (Panel B) ESE categories. The dependent variable here is percentage of total enrollment classified in excluded (Panel A) and included (Panel B) categories. The estimates show no evidence that the threatened schools resorted to relative classification into excluded categories in any of the three years after program. Nor is there any evidence of differential

classification in the included categories.

The ESE categories vary in the extents 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 disabled and emotionally handicapped are much more mild and comparatively mutable categories.[9] Classification in these latter categories often has a large amount of subjective element to it and hence could be easily manipulated. The above 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; 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. If such classification did take place, it is most likely to have taken place in such mutable categories.

Table 6 investigates the effect of the program on relative classification in mutable excluded categories,—learning disabled (Panel A) and emotionally handicapped (Panel B). There is no evidence that the threatened schools tended to differentially classify students into either learning disabled or emotionally handicapped categories. Figure 4 Panels A, B, C and D look at the effect of the program on classification in total excluded, included, emotionally handicapped, and learning disabled categories, respectively. As earlier, the graphs display results from regression discontinuity estimations that do not include controls while those in the tables include controls. The graphical patterns in Figure 4 mirror closely the results obtained in Table 6. The discontinuities are either small or indistinguishable from zero and they are never statistically significant.

To summarize, I find no evidence that the treated schools resorted to strategic classification into excluded ESE categories. But this does not mean that the schools did not respond to

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

incentives. There are multiple ways in which the threatened schools might have responded—they likely weighed the relative returns and costs in the different alternatives and chose the options that were least costly. The fact that they did not resort to classification into excluded ESE categories indicates that perhaps its costs did not justify its returns. There were considerable costs associated with a strategy of classification into ESE categories. It had to be approved by the parents and a group of experts (such as physicians, psychologists, etc.). Increased classification meant increased provision of services which was often costly in spite of state financing of a large part of these services.

The McKay Scholarship program acted as a further disincentive to this sort of reclassification. Created in 1999 and fully implemented in the 2000-01 school year, this program made 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 was associated with a risk of loss of the student and the corresponding revenue.

5.1 The Problem of Underestimation: Are D Schools Untreated?

While D schools did not directly face the threat of vouchers, they were close to getting an "F" and hence were likely to face an indirect threat. If this was indeed the case, the program effects shown above could be underestimates.

To get around this problem, I rescale the effects obtained above by the difference in the probabilities of treatment of F and D schools, that is by calculating the corresponding Wald estimator.[10] 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 sample.

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

A problem here is that 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. Using this sample of 98F and 98D schools and data on school grades in 1999, I calculate the above probabilities.[11] To calculate these probabilities for the discontinuity sample, 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 or above 3 in 1998 FCAT writing, I consider the schools within the same Silverman bandwidth as above, and calculate the probabilities that these groups of 98F and 98D schools would fall into treatment the next year. The corresponding scaling factors, thus calculated, are 1.15 and 1.20 for the full sample and the discontinuity sample, respectively. This implies that the underestimation-corrected results are similar to above and underestimation is not a major problem here.

5.2 Compositional Changes of Schools and Sorting

If there is differential student sorting or compositional changes in the treated schools, then the above effects can be in part be driven by those changes. None of the threatened schools received a second "F" grade in 2000 or 2001, and therefore none of their students became eligible for vouchers. Thus, the concern about vouchers leading to sorting is not applicable here. However, the F and D grades can lead to a differential sorting of students in these two types of schools.[12]

To investigate this issue further, I examine whether the demographic composition of the treated

[11] 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.

[12] 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.

schools saw a relative shift after the program. I use the same regression discontinuity strategy outlined above, but the dependent variables are now demographic variables (% white, % black, % hispanic, % asian, % american indian, % multiracial, % male, % free/reduced price lunch and enrollment).

The results of this analysis are presented in Table 7. As can be seen, there is no evidence of any differential shift in the treated schools in any of the characteristics in any of the three years after program, except for % asian in the second year after program. As discussed above, with a large number of comparisons, it is reasonable to expect a few to be statistically different from zero due to sheer random variation. So from the above analysis, it seems safe to conclude that the results obtained above are not being driven by differential changes in composition of schools or student sorting.

"Threat of Vouchers" versus Stigma

As noted earlier, the Florida accountability-tied voucher system subjected "F" schools to both the "threat of vouchers" and stigma. So a question may be whether the effects above are driven by "threat of vouchers" or "stigma". Note that the objective of this paper is not to assess the effects of the individual components of the system, but rather to identify the effects of the whole voucher system that involves both "threat of vouchers" and stigma. The effects obtained above captures the effects of the whole program. A voucher system tied to accountability (as in Florida) involves both "threat of vouchers" and stigma,—it is instructive to know how such a voucher system as a whole affects incentives and behavior, and that is what this paper aimed to do.

6 Conclusion

This paper analyzes the behavior of public schools facing a voucher system that embedded vouchers in an accountability regime. It focuses on the 1999 Florida program. Utilizing the institutional

details of the program, it analyzes the incentives built into the system, and examines the behavior of public schools facing these incentives.

It focuses on two alternative ways in which the program incentives might have induced the threatened schools to behave. First, scores of LEP students who were in an ESOL program for less than two years were not counted in the calculation of grades. As a result, threatened schools might have had an incentive to reclassify their low-performing students into this excluded LEP category so as to remove them from school grade calculations. Did this actually happen in practice? Second, scores of students in several special education categories were not eligible to be included in the computation of grades. Did this rule induce the threatened schools to reclassify their low-performing students into these excluded categories so as to artifically inflate scores?

I address these questions in this paper. Using data obtained from Florida Department of Education and a regression discontinuity technique, I find evidence of increased classification by threatened schools in the excluded LEP category in high-stakes grade 4 and entry grade 3 in the first year after program. There is no evidence of any differential classification in the included LEP category in any of the grades following the program. For reference, there was no difference in excluded or included LEP classification between the threatened and non-threatened schools before the program. To summarize, the findings suggest that in the first year after program the threatened schools tended to differentially reclassify students into the excluded LEP category in an effort to remove them from the effective test taking pool both in the current year and in the following year.

The patterns in ESE classification present a different picture. There is no evidence that the program led to a differential classification into excluded (or included) ESE categories in the threatened group of schools in any of the years after program, when the threat was in effect.

Classification into ESE categories was associated with substantial costs which might have acted as a disincentive toward this form of classification. While there were other costs (as outlined in the paper), the main cost was posed by Florida's McKay scholarship program that made any ESE student in Florida's public schools eligible for the McKay vouchers. These vouchers entitled them to move to private schools and were funded by public school revenue. In other words, classification into special education categories bore with it the threat of losing the student and the corresponding revenue. There were multiple ways in which threatened schools could have responded to the program. It is likely that the schools weighed the costs and benefits of various possible avenues and chose to respond in the least costly ways. So, it is not surprising that I do not find any evidence in favor of strategic classification into excluded ESE categories.

These findings have important policy implications. They suggest that schools facing vouchers tied to accountability regimes might chose to behave strategically to classify their low-performing students into excluded categories in an effort to remove them from the effective test taking pool. It follows that when designing policies that incorporate vouchers (or sanctions) tied to accountability regimes, policymakers should be wary of creating exemptions for certain groups of students as they might create adverse incentives to game the system. Understanding the responses of public schools to various school reform policies is indispensable for an effective policy design and this study has contributed in this direction.

References

Chakrabarti, Rajashri (2008a),"Impact of Voucher Design on Public School Performance: Evidence from Florida and Milwaukee Voucher Programs," Federal Reserve Bank of New York Staff Paper Number 315.

Chakrabarti, Rajashri (2008b),"Can Increasing Private School Participation and Monetary

Loss in a Voucher Program Affect Public School Performance? Evidence from Milwaukee," Journal of Public Economics volume 92, Number 5-6, 1371-1393.

Chakrabarti, Rajashri (2010),"Vouchers, Public School Response and the Role of Incentives: Evidence from Florida," Federal Reserve Bank of New York Staff Paper Number 306.

Chiang, Hanley (2009),"How Accountability Pressures on Failing Schools Affects Student Achievement," Journal of Public Economics volume 93, 1045-1057.

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

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

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

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

Figlio, David and Cassandra Hart (2010),"Competitive Effects of Means-Tested Vouchers," National Bureau of Economic Research Working Paper Number 16056.

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

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

Figlio, David and Joshua Winicki (2005), "Food for Thought? The Effects of School

Accountability Plans on School Nutrition", Journal of Public Economics, 89, 381-394.

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

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

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

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

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

Hoxby, Caroline (2003a), "School Choice and School Productivity (Or, Could School Choice be the tide that lifts all boats?)", in Caroline Hoxby (ed.) The Economics of School Choice, University of Chicago Press.

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

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

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

Jacob, Brian and Steven Levitt (2003), "Rotten Apples: An Investigation of the Prevalence

and Predictors of Teacher Cheating", Quarterly Journal of Economics, 118 (3).

Ladd, Helen F. and Douglas L. Lauen (2010), "Status Versus Growth: The Distributional Effects of School Accountability Policies", Journal of Policy Analysis and Management 29(3), 426-450.

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

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

Reback, Randall (2008), "Teaching to the Rating: School Accountability and Distribution of Student Achievement," Journal of Public Economics 92, June 2008, 1394-1415.

Rouse, Cecilia E., Jane Hannaway, David Figlio and Dan Goldhaber (2007), "Feeling the Florida Heat: How Low Performing Schools Respond to Voucher and Accountability Pressure," CALDER (National Center for Analysis of Longitudinal Data in Education Research) Working Paper 13.

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

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

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

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

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses.

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

	1999
Difference	-0.01
	(0.01)

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

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

	% of Students in Excluded Category			% of Students in Included Category		
	1 Year After	2 Years After	3 Years After	1 Year After	2 Years After	3 Years After
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 2	0.292	0.170	0.024	0.115	0.477	-0.594
	(0.228)	(0.268)	(0.232)	(0.300)	(0.685)	(0.534)
Observations	123	124	120	123	124	120
R^2	0.532	0.573	0.537	0.664	0.652	0.735
Grade 3	0.362*	0.295	0.261	-0.422	-0.212	0.611
	(0.181)	(0.278)	(0.264)	(0.484)	(0.436)	(0.835)
Observations	121	122	124	121	122	124
R^2	0.540	0.483	0.579	0.568	0.603	0.605
Grade 4	0.314**	0.365*	0.144	0.039	-0.220	0.532
	(0.118)	(0.218)	(0.295)	(0.310)	(0.360)	(0.399)
Observations	119	124	121	119	124	121
R^2	0.403	0.523	0.398	0.531	0.435	0.546
Grade 5	0.270	0.317***	0.449	0.011	0.065	0.276
	(0.253)	(0.097)	(0.380)	(0.391)	(0.455)	(0.259)
Observations	116	117	122	116	117	122
R^2	0.430	0.544	0.444	0.325	0.365	0.397

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, and real per pupil expenditure.

Table 4: Effect of the Program on Total ESE Classification: A Regression Discontinuity Analysis

	% of Students in ESE Categories		
	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
	0.437	-1.142	-0.648
	(0.403)	(0.830)	(0.873)
Observations	130	132	132
R^2	0.920	0.838	0.733

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, pre-program (1999) percentage of students in ESE categories, and real per pupil expenditure.

Table 5: Effect of the Program on Classification into Excluded and Included ESE Categories: A Regression Discontinuity Analysis

Panel A	% in Excluded ESE Categories		
	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
	0.699	-0.433	-0.086
	(0.565)	(1.062)	(1.060)
Observations	130	132	132
R^2	0.920	0.852	0.740
Panel B	**% in Included ESE Categories**		
	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
	-0.236	-0.633	-0.493
	(0.286)	(0.413)	(0.379)
Observations	130	132	132
R^2	0.837	0.706	0.570

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, pre-program (1999) percentage of students in Excluded (Panel A) or Included (Panel B) ESE categories, and real per pupil expenditure.

Table 6: Effect of the Program on Classification into Learning Disabled and Emotionally Handicapped Categories: A Regression Discontinuity Analysis

Panel A	% in Learning Disabled Category		
	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
	-0.178	-0.445	0.347
	(0.261)	(0.478)	(0.407)
Observations	130	132	132
R^2	0.801	0.727	0.643
Panel B	% in Emotionally Handicapped Category		
	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
	0.083	-0.138	0.038
	(0.158)	(0.179)	(0.235)
Observations	130	132	132
R^2	0.925	0.886	0.788

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses. All regressions control for racial composition, gender composition, percent of students eligible for free or reduced price lunch, pre-program (1999) percentage of students in Learning Disabled (Panel A) or Emotionally Handicapped (Panel B) category, and real per pupil expenditure.

Table 7: Are Compositional Changes or Sorting Driving Results? Investigating Demographic Shifts using a Regression Discontinuity Analysis

	1 Year After Program	2 Years After Program	3 Years After Program
	(1)	(2)	(3)
% White	1.048	0.725	-0.746
	(5.770)	(5.112)	(4.400)
Observations	134	134	133
R^2	0.024	0.019	0.020
% Black	-6.796	-9.315	-9.438
	(11.089)	(10.382)	(10.020)
Observations	134	134	133
R^2	0.020	0.020	0.021
% Hispanic	5.287	8.111	9.867
	(7.539)	(7.569)	(7.993)
Observations	134	134	133
R^2	0.006	0.010	0.013
% Asian	0.540	0.547*	0.423
	(0.333)	(0.287)	(0.332)
Observations	134	134	133
R^2	0.039	0.038	0.040
% American Indian	-0.079	-0.067	-0.106
	(0.050)	(0.085)	(0.073)
Observations	134	134	133
R^2	0.009	0.011	0.017
% Male	-0.137	0.683	0.699
	(0.591)	(0.728)	(0.839)
Observations	134	134	133
R^2	0.011	0.015	0.011
% Free/Red. Price Lunch	-1.422	-2.295	0.462
	(4.842)	(3.796)	(4.081)
Observations	134	133	133
R^2	0.017	0.026	0.019
Total Enrollment	-11.809	-12.593	15.901
	(52.498)	(43.411)	(42.119)
Observations	134	134	133
R^2	0.014	0.030	0.011

*, **, ***: significant at the 10, 5, and 1 percent level, respectively. Robust standard errors adjusted for clustering by school district are in parentheses.

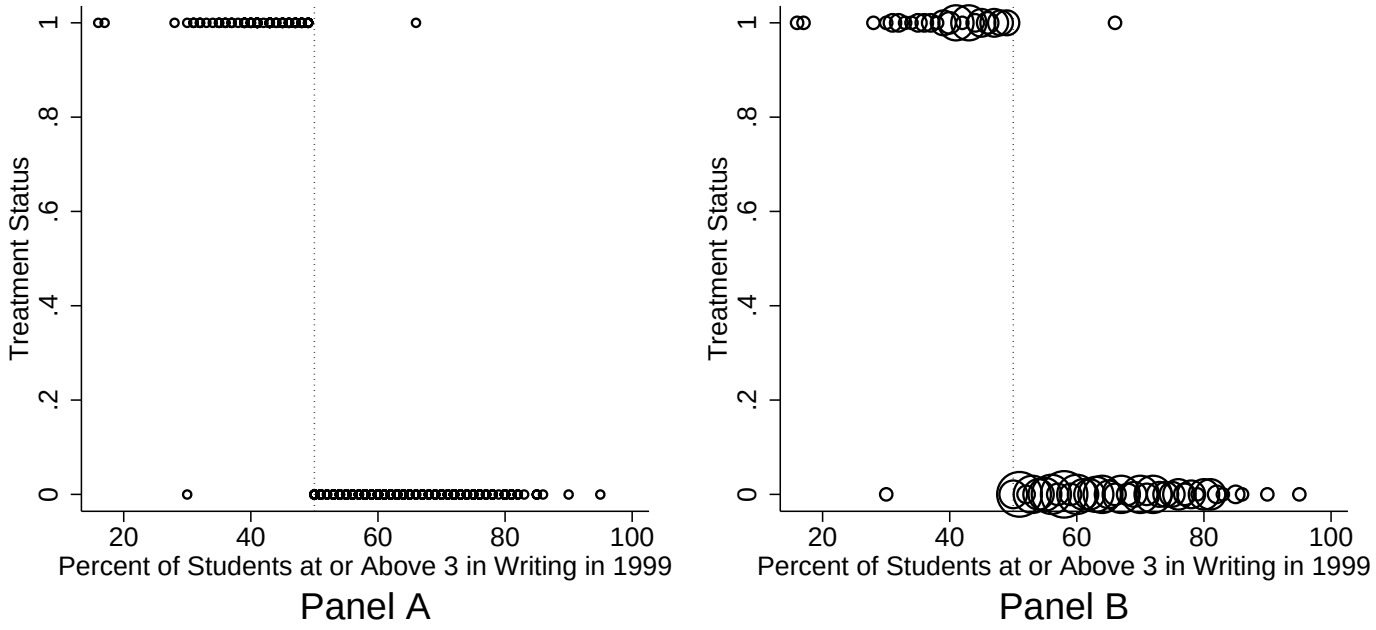


Figure 1. Regression Discontinuity Analysis:
Relationship Between % of Students at or Above 3 in Writing and Treatment Status

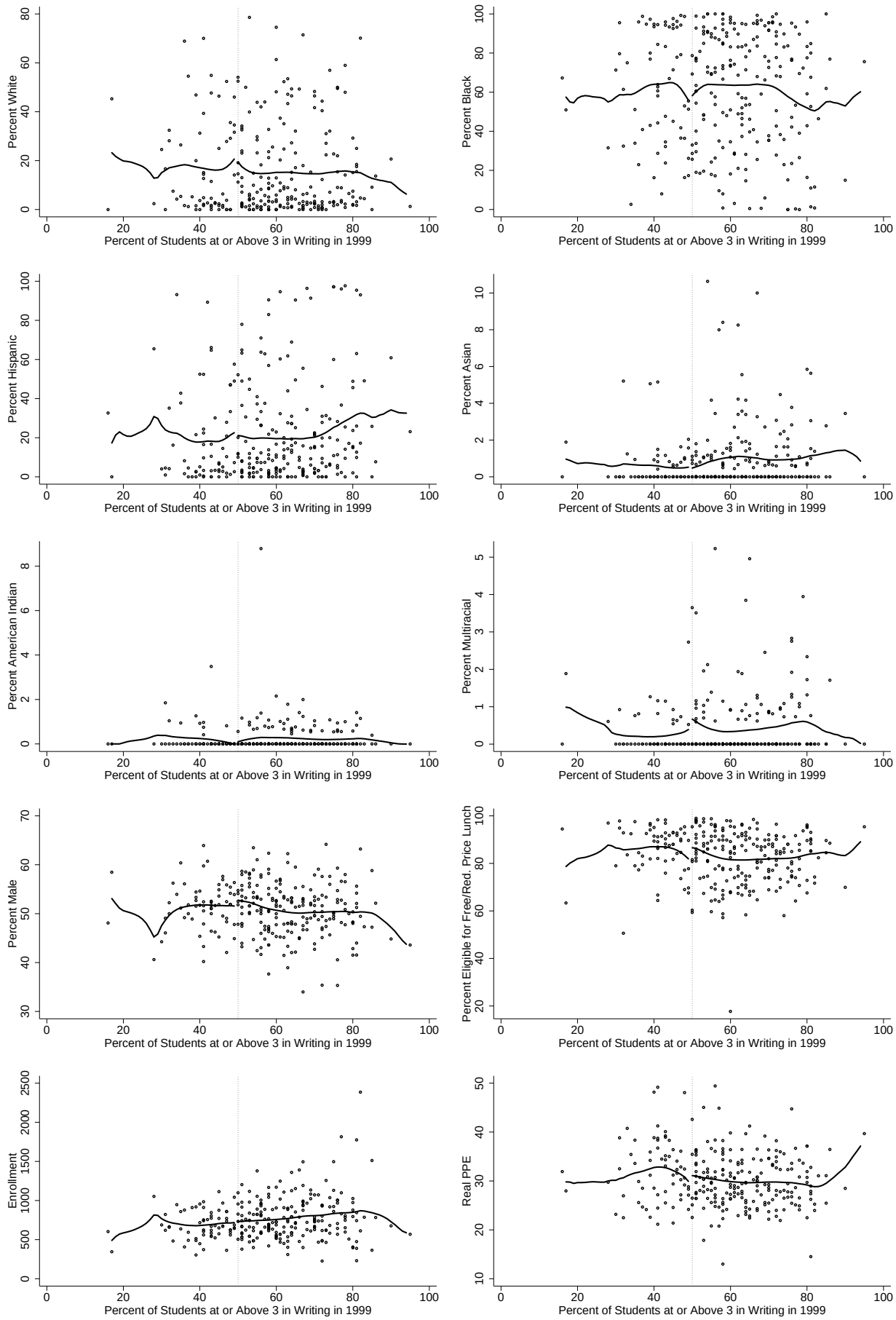


Figure 2A. Testing Validity of Regression Discontinuity Design:
Pre−Program Characteristics Relative to the Cutoff

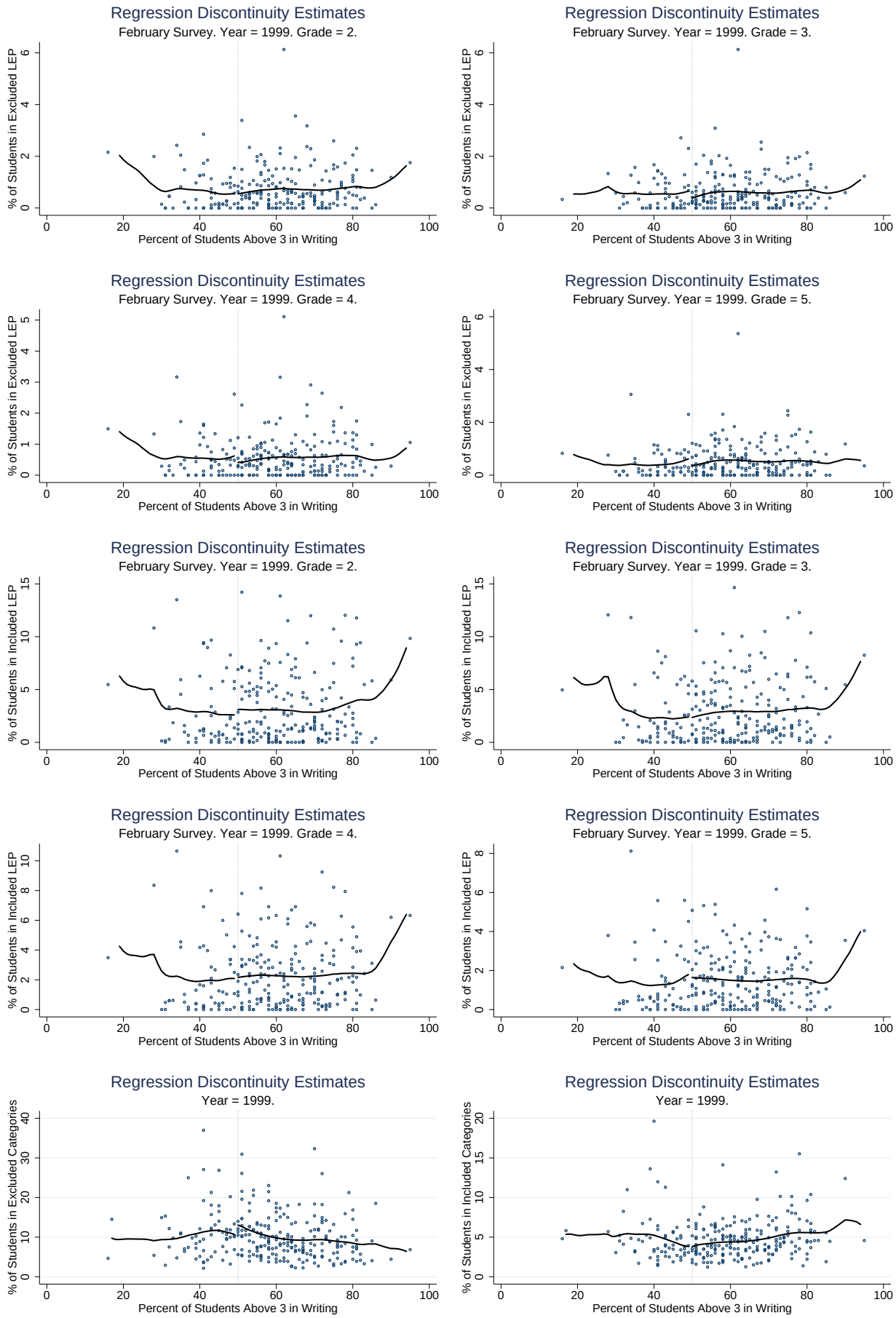


Figure 2B. Testing Validity of Regression Discontinuity Design: Examining Classification in Excluded and Included LEP and ESE Categories Relative to Cutoff in Pre−Program Period

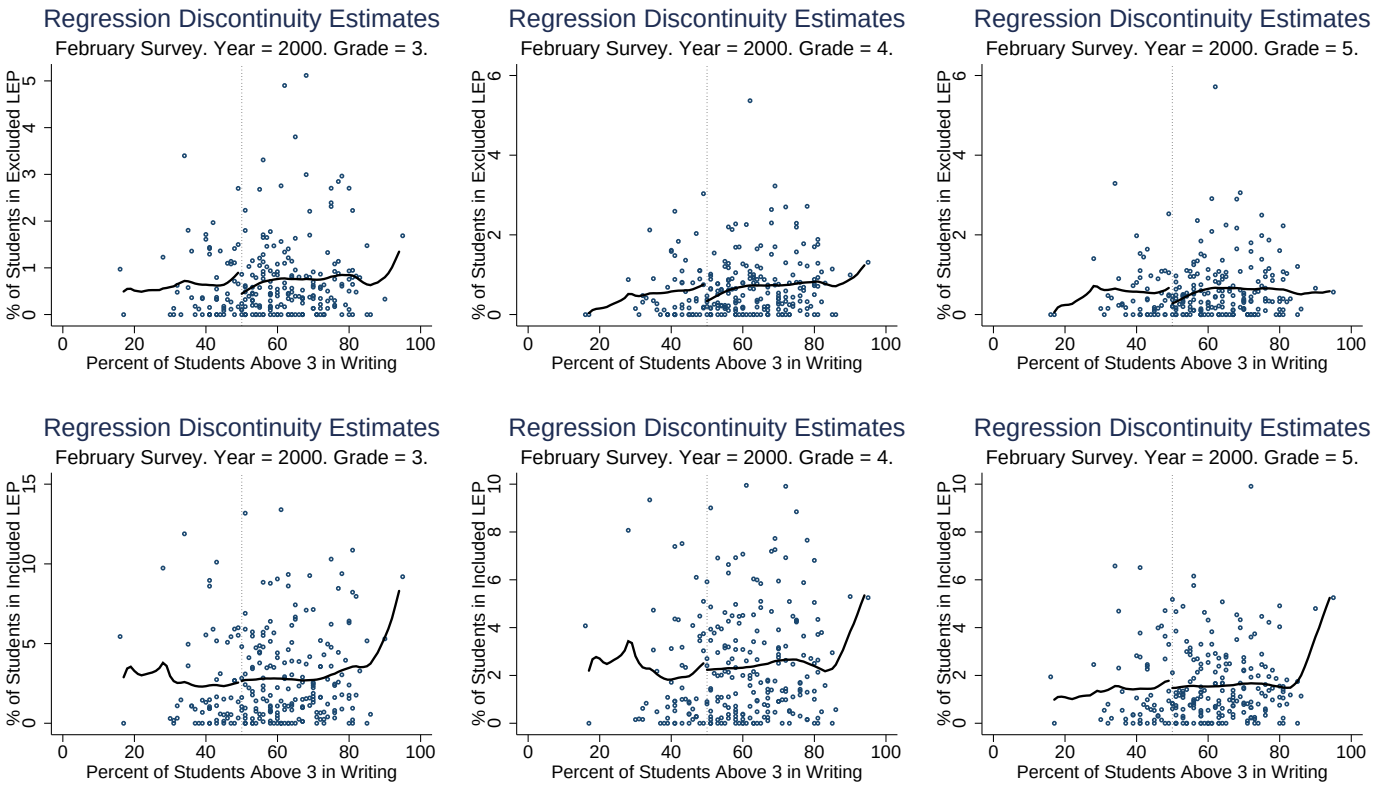


Figure 3A. Examining the Effect of the Program on Classification in Excluded and Included LEP Categories, 2000

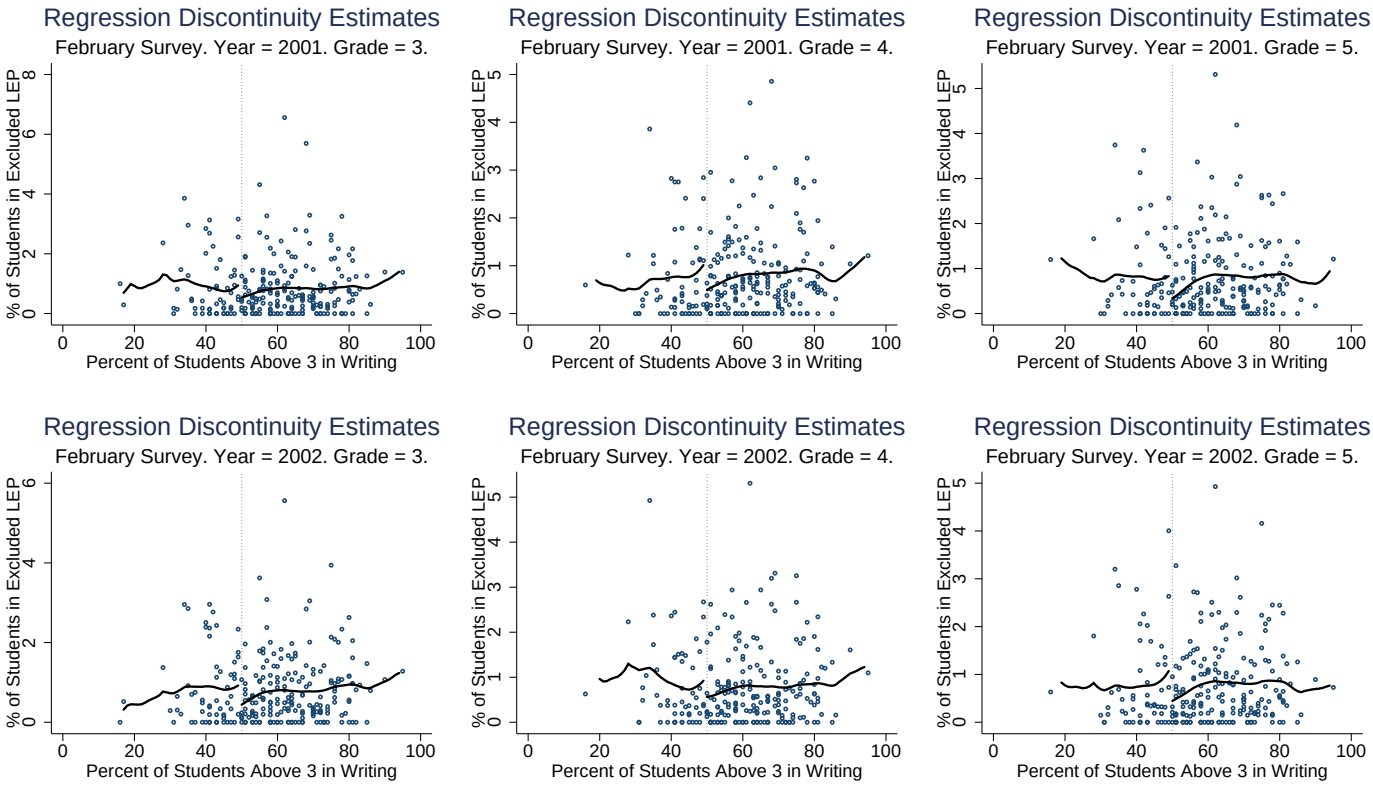
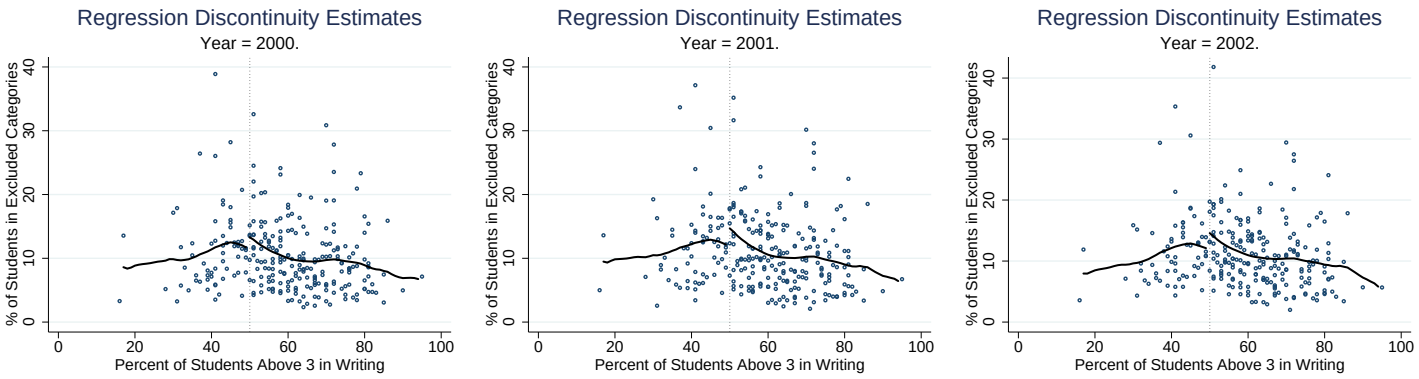
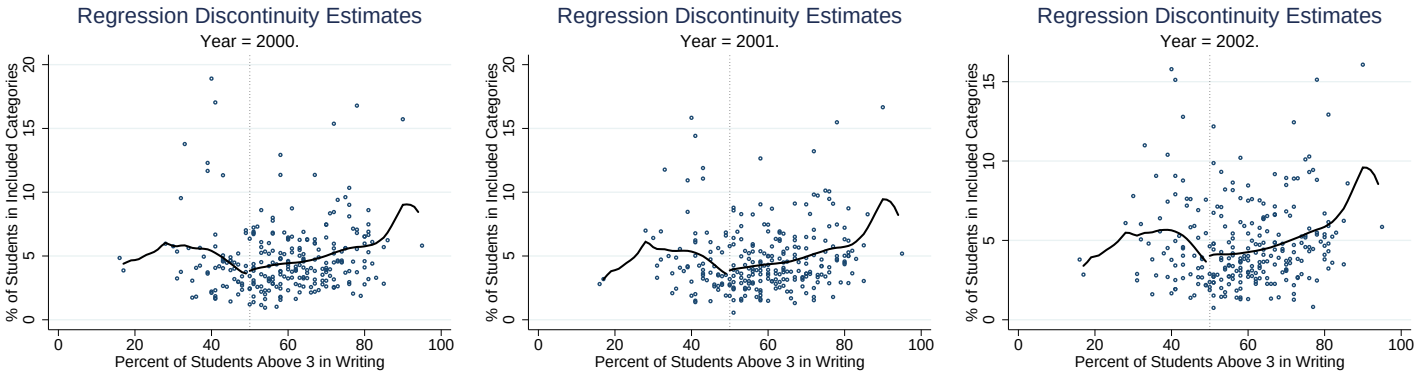


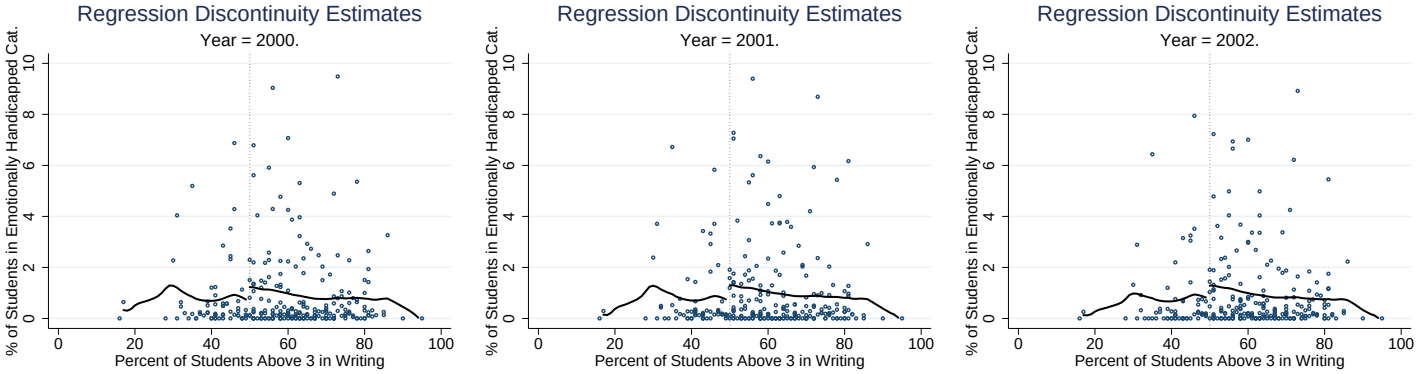
Figure 3B. Examining the Effect of the Program on Classification in Excluded and Included LEP Categories, 2001 and 2002



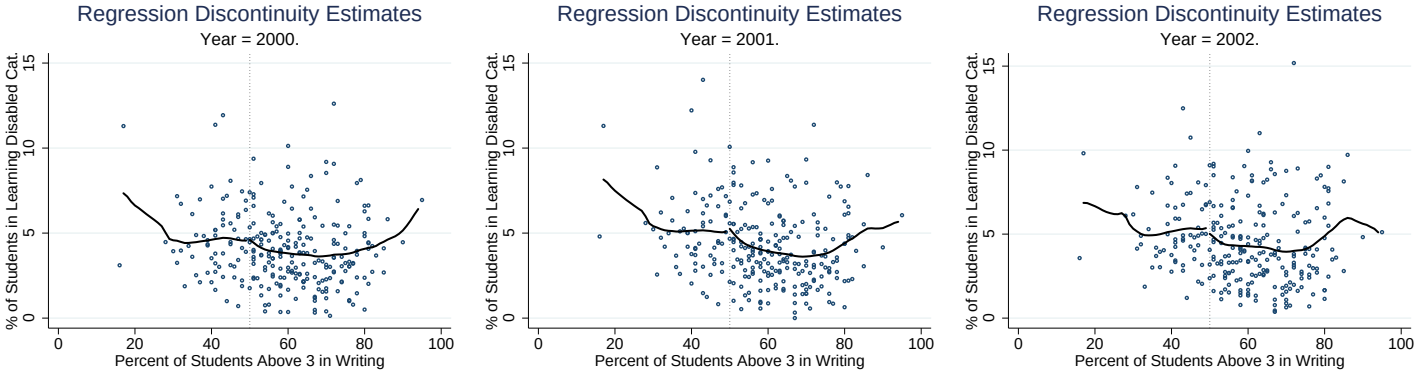
Panel A. Effect on Classification in Excluded ESE Categories



Panel B. Effect on Classification in Included ESE Categories



Panel C. Effect on Classification in Emotionally Handicapped Category



Panel D. Effect on Classification in Learning Disabled Category

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