Rajashri Chakrabarti

Program Design, Incentives, and Response: Evidence from Educational Interventions

- In an effort to reform K-12 education, policymakers have introduced vouchers in some U.S. school districts, enabling students to transfer from public to private schools.
- The different designs of two school voucher programs—the Milwaukee and Florida programs—have had different effects on public school incentives and performance.
- In Milwaukee, vouchers were imposed from the outset; in Florida, schools were first threatened with vouchers and thus had an incentive to avoid them.
- The Florida public schools' efforts to avoid vouchers resulted in performance effects that far exceeded those of Milwaukee's program.
- Program design is critical: Policies that present failing public schools with functional and credible sanctions are best suited to provide the results intended by policymakers.

1. Introduction

Concerns that U.S. students are not performing as well as their counterparts in other developed countries on international math and science tests have led to widespread demands for the reform of K-12 education in the United States. Of the various reforms under consideration, school voucher reform is at the forefront.

Vouchers are scholarships that make students eligible to transfer from public to private schools. A basic feature of all publicly funded voucher programs in the United States is the funding of vouchers by public school revenue, so that money always "follows" students. In other words, schools that lose students lose their corresponding funding. Schools therefore recognize the financial implications of vouchers and have an incentive to avoid being subject to voucher programs.

This article investigates the role of program design in the context of two such educational interventions in the United States—the Milwaukee and Florida school voucher programs—and analyzes the effects of design on public school incentives and performance.[1] We demonstrate that variations in program design have markedly different outcomes for public schools affected by vouchers.

The Milwaukee program, introduced in 1990, was the first voucher program in the country. Implemented in 1999, the

Rajashri Chakrabarti is an economist at the Federal Reserve Bank of New York.
rajashri.chakrabarti@ny.frb.org

The author thanks Andrew Haughwout for helpful suggestions. Noah Schwartz provided excellent research assistance. All errors are the author's. The views expressed 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.

Florida program was the nation's third, following Cleveland's. The Milwaukee and Florida voucher programs share the basic feature of funding by public school revenue. But there are crucial differences. Milwaukee's is a means-tested program targeting low-income students while Florida's embeds a voucher program in a full-fledged accountability system.

Using test-score data from Milwaukee and Florida and implementing a difference-in-differences estimation strategy, our study estimates the impact of each program by comparing the post-program results of the affected schools with a comparable set of control schools. Controlling for potentially confounding pre-program time trends and post-program common shocks, we find that the performance effects of the Florida program far exceed those of Milwaukee's program. These results are quite robust in that they hold after controlling for other confounding factors, such as mean reversion and a possible stigma effect; they also withstand several sensitivity tests.

Our findings have important policy implications, which we consider in the context of New York State's federal, state, and city accountability programs. These programs include New York City's accountability policy, known as the "Progress Report" policy, and the federal No Child Left Behind (NCLB) law, as implemented by New York State.

Our study is organized as follows. Section 2 describes the Milwaukee and Florida voucher programs. In Section 3, we discuss the incentives created by the programs and the corresponding responses that might be expected from the affected public schools. Our data and empirical strategy are reviewed in Sections 4 and 5, respectively. Section 6 presents our results, and Section 7 considers policy implications.

2. Institutional Details

The first publicly funded school voucher program in the United States was established in Milwaukee, Wisconsin, in 1990. The Milwaukee Parental Choice Program made the city's entire low-income public school population eligible for vouchers. Specifically, starting in the 1990-91 school year, the program made all Milwaukee public school students with family income at or below 175 percent of the poverty line eligible for vouchers to attend nonsectarian private schools.

In contrast, the Florida Opportunity Scholarship Program, introduced in 1999, can be looked upon as a "threat-of-voucher" program. Here, failing public schools were threatened with the imposition of vouchers, with vouchers implemented *only if* schools failed to meet a government-designated cutoff quality level. The institutional details of the Milwaukee and Florida programs are summarized in Table 1.

The Florida Department of Education classified schools according to five grades: A, B, C, D, or F. The state assigned school grades based on Florida Comprehensive Assessment Test (FCAT) reading, math, and writing scores. For FCAT reading and math, it categorized students into five achievement levels—1 lowest, 5 highest—that correspond to specific ranges on the raw-score scale. Using current-year data, the Department of Education assigned an "F" grade to a school if it was below the minimum criteria in reading, math, and writing; a "D" if it was below the minimum criteria in one or two of the three subject areas; and a "C" if it was above the minimum criteria in all three subjects, but below the higher performing criteria in all three. In reading and math, at least 60 percent (50 percent) of students had to score level 2 (3) and above; in writing, at least 50 percent (67 percent) had to score 3 and above to meet the minimum (high-performing) criteria in that respective subject.[2]

The major difference in program design between the Milwaukee and Florida programs is that in Milwaukee vouchers were imposed at the outset, whereas in Florida failing schools were first threatened with vouchers, with vouchers introduced only if the schools failed to show adequate improvement in performance.

Under the Florida Opportunity Scholarship Program, all public school students became eligible for vouchers, or

[1] Our study focuses on the impact of alternative voucher designs on public school performance. A growing body of literature analyzes the many issues associated with school vouchers. Nechyba (1996, 1999, 2000, 2003) analyzes distributional effects of alternative voucher policies in a general equilibrium framework; Epple and Romano (1998, 2002) and Chakrabarti (2009) investigate sorting attributable to vouchers; Manski (1992) considers the impact of vouchers on public school expenditure and social mobility; and McMillan (2004) and Chakrabarti (2008b) model the quality of public schools facing vouchers. Hoxby (2003a, b) and Chakrabarti (2008a) study the effects of the Milwaukee voucher program, while Greene (2001), Greene and Winters (2003), Figlio and Rouse (2006), West and Peterson (2005), and Chakrabarti (2007, 2008a) study the effects of the Florida program.

[2] In 1999, seventy-eight schools received an "F" grade. Students in two of those schools became eligible for vouchers. In 2000, four elementary schools received an "F," although none became eligible for vouchers. In 2001, no schools received an "F" grade. In 2002, sixty-four schools received an "F." Students in ten of those schools became eligible for vouchers. In 2003, students in nine schools became eligible for vouchers; in 2004, the figure was twenty-one schools.

Table 1

Comparison of Milwaukee and Florida Voucher Programs

Milwaukee Program	Florida Program
• First U.S. voucher program	• Third U.S. voucher program
• Started in 1990-91 school year	• Started in 1998-99 school year
• Public school students with family income at or below 175 percent of the poverty line eligible for vouchers to attend nonsectarian private schools	• Vouchers contingent on school performance
• Private schools were not permitted, by law, to discriminate against students who apply with vouchers: – Had to accept all students unless oversubscribed – If oversubscribed, had to choose students randomly	• Schools classified according to five grades: A, B, C, D, F (A-highest, F-lowest) – Grades based on the Florida Comprehensive Assessment Test (FCAT) reading, math, and writing scores – F, if below the minimum criteria in reading, math, and writing – D, if below the criteria in one or two of the three subjects – C, if above the minimum criteria in all three subjects, but below the higher performing criteria in all three
• Average voucher amount equaled the state aid per pupil, and vouchers were financed by an equivalent reduction of state aid to the school district	• Students categorized into five achievement levels in FCAT reading and math (1-lowest, 5-highest)
• 1990-91 and 1996-97: – Average voucher amounts were $3,346 – Vouchers as a percentage of total revenue per pupil were 45.23 percent	• Minimum criteria: – Reading and math: at least 60 percent must score level 2 and above – Writing: at least 50 percent must score level 3 and above
	• High-performing criteria: – Reading and math: at least 50 percent must score level 3 and above – Writing: at least 67 percent must score level 3 and above
	• All students of a public school became eligible for vouchers if the school received two "F" grades in a period of four years
	• Private schools were not permitted, by law, to discriminate against students who apply with vouchers – Had to accept all students unless oversubscribed – If oversubscribed, had to choose students randomly
	• Average voucher amount equaled the state aid per pupil, and vouchers were financed by an equivalent reduction of state aid to the school district
	• 1999-2000 and 2001-02: – Average voucher amounts were $3,330 – Vouchers as a percentage of total revenue per pupil were 41.55 percent

Source: Information and data provided in various Florida Department of Education and Milwaukee Department of Public Instruction reports.

"opportunity scholarships," if the school received two "F" grades in a period of four years. Therefore, a school that received an "F" for the first time was exposed to the threat of vouchers, but did not face them unless and until it got a second "F" within the next three years. Thus, the major difference in program design between the Milwaukee and Florida programs is that in Milwaukee vouchers were imposed at the outset, whereas in Florida failing schools were first threatened with vouchers, with vouchers introduced *only* if the schools failed to show adequate improvement in performance.

Apart from the above differences, the design of the two programs was strikingly similar. In both programs, private schools could not, by law, discriminate against students who applied with vouchers—the schools had to accept all students unless they were oversubscribed, in which case they had to choose students randomly. Indeed, the application form did not ask questions about the student's race, sex, parents' education, past scores, or prior records (for example, truancy, violence). The questions were specifically worded only to ascertain whether the student was eligible for the program.[3] The system of funding for the Milwaukee and Florida voucher programs was also very similar. Under each program, the average voucher amount was equal to the state aid per pupil, and vouchers were financed by an equivalent reduction of state aid to the school district. Thus, state funding was directly tied to student enrollment, and enrollment losses due to vouchers were reflected in a revenue loss for the public school.[4] The average voucher amounts under the Milwaukee (1990-91 through 1996-97) and Florida (1999-2000 through 2001-02) programs were $3,346 and $3,330, respectively. During these periods, vouchers as a percentage of total revenue per pupil were 45.23 percent in Milwaukee and 41.55 percent in Florida.

[3] While the schools could not employ any selection criteria for the voucher students, this was not the case for nonvoucher students in the same school. Also note that the private schools had the choice of whether to participate in the program. However, if they decided to participate, they were required by law to accept all students or to choose students randomly, if oversubscribed.

3. Discussion: Effects of the Programs on Public School Incentives and Response

What incentives would be created by the aforementioned program rules, and how would one expect the affected public schools to respond? Consider a public school subject to the Florida program, a school that has just received its first "F" grade ("F-school" hereafter). The school realizes that if it can avoid another "F" grade in the next three years, it can escape vouchers and the monetary loss and embarrassment associated with them.[5] Therefore, it would have an incentive to improve its scores so as not to receive a second "F" grade. In contrast, if the same school were subject to a Milwaukee-type voucher program—in which vouchers have already been introduced—the school could not avoid vouchers (and the revenue loss)

> *In a Florida-type [voucher] program, the threatened public schools . . . have more of an incentive to respond in order to improve their scores and escape vouchers.*

altogether. In this case, improvements would serve to retain or attract a few students, but the effect would be marginal compared with that of a Florida-type program. In a Florida-type program, the threatened public schools (schools that have received their first "F" grade) have more of an incentive to respond in order to improve their scores and escape vouchers.[6] Thus, the key difference between the two programs is that in the Milwaukee program, vouchers have already been implemented, whereas the Florida program first threatens the schools and gives them a window to respond, and an adequate response can preclude sanctions. Sanctions (vouchers) are implemented only if the schools fail to attain the predesignated standard.

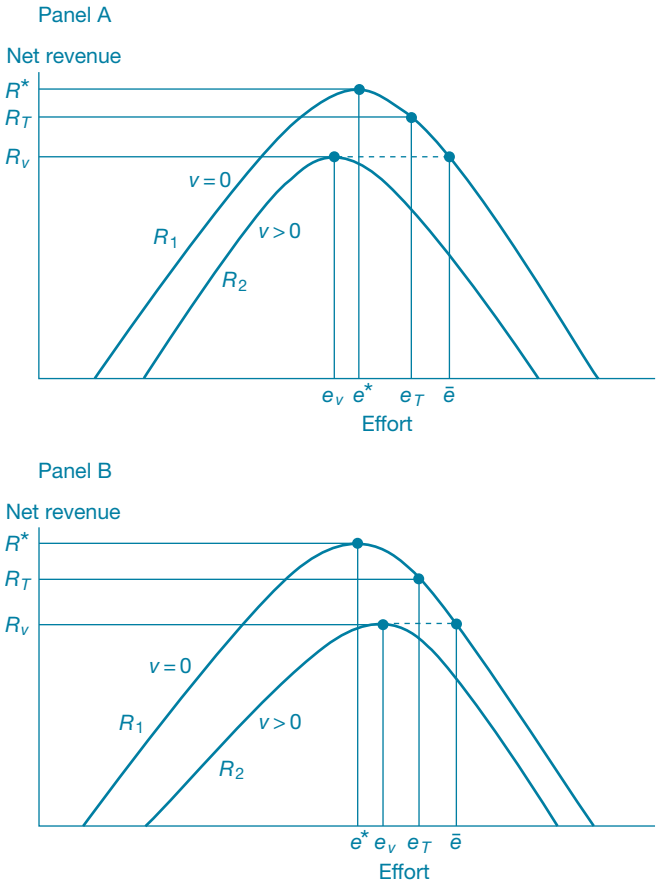
[4] We focus on the Milwaukee program up to 1996-97. The reason is that following a 1998 Wisconsin Supreme Court ruling, there was a major shift in the program when religious private schools were allowed to participate. Moreover, the financing of the Milwaukee program underwent some crucial changes, so that the voucher amounts and the revenue loss per student due to vouchers were not comparable between the Florida and second-phase Milwaukee programs. See Chakrabarti (2008a) for an analysis of how the shift in the Milwaukee voucher program affected public school incentives and performance as well as for a comparison of public school responses in the two phases of the program. We focus on the Florida program up to 2001-02. This period is chosen because schools that received an "F" grade in 1999 would face the threat of vouchers only through 2002.

[5] The loss of students due to vouchers leads to a decrease in both revenue and costs for the school. But for a school operating at full capacity, the cost savings due to the loss of students are marginal, while the loss in revenue is not. This effect is a major reason why public schools do not support vouchers.

[6] For a formal proof, see Chakrabarti (2008b).

Chart 1

Analyzing the Effect of "Voucher Threat" versus Vouchers



The intuition above is shown in Chart 1. Let R_1 illustrate the initial net revenue function of the public school. The public school chooses the effort to maximize net revenue. Let this equilibrium effort be denoted by e^* and the corresponding net revenue by R^*. Now assume that Milwaukee-type vouchers are introduced. This leads to a downward shift of the net revenue function—the new net revenue function is denoted by R_2 and the corresponding optimum effort and net revenue by e_V and R_V, respectively.[7] Panel A of the chart illustrates the case in which $e_V < e^*$, and panel B the case in which $e_V > e^*$. The chart implies that any target effort in the range $(e_V, \bar{e}]$ under a threat-of-voucher regime will induce an effort strictly greater than e_V. For example, assume that the policymaker implements a target effort, e_T. Satisfying this target would lead to a net revenue of R_T while failing to satisfy it would lead to the introduction of vouchers and corresponding revenue of $R_V (< R_T)$. Therefore, the school has an incentive to implement an effort of $e_T (> e_V)$.

[7] For formal proofs, see Chakrabarti (2008b).

4. Data

The Florida data consist of school-level data on test scores, grades, socioeconomic characteristics of schools, and school finances; they are obtained from the Florida Department of Education. School-level data on test scores are obtained from the Florida Comprehensive Assessment Test. Mean scale scores (on a scale of 100-500) on grade 4 reading and grade 5 math are available for 1998-2002. Mean scale scores (on a scale of 1-6) on the Florida grade 4 writing test are available for 1994-2002.

Data on socioeconomic characteristics include sex composition (1994-2002), percentage of students eligible for free or reduced-price lunch (1997-2002), and race composition (1994-2002), and are obtained from the school indicators database of the Florida Department of Education. This study refers to school years by the calendar year of the spring semester. School finance data consist of several measures of school-level and district-level per-pupil expenditures, and are obtained from the school indicators database and the Office of Funding and Financial Reporting of the Florida Department of Education.

The Wisconsin data consist of school-level data on test scores, socioeconomic characteristics of schools, and per-pupil expenditures (both at the school and district levels). The data are obtained from the Wisconsin Department of Public Instruction, the Milwaukee Public Schools, and the Common Core of Data of the National Center for Education Statistics. School-level data on test scores are obtained for 1) the Third Grade Reading Test (renamed the Wisconsin Reading Comprehension Test, or WRCT, in 1996) and 2) the grade 5 Iowa Test of Basic Skills (ITBS). School scores for the WRCT, which was first administered in 1989, are reported in three "performance standard categories": percentage of students below, percentage of students at, and percentage of students above the standard.[8] Data for these three categories are available for 1989-97. School-level ITBS reading data (mean scores) are available for 1987-93; ITBS math data (mean scores) are available for 1987-97.

5. Empirical Strategy

5.1 Florida

In Florida, the schools that received an "F" grade in 1999 were directly exposed to the threat of vouchers because all their students would be eligible for vouchers if the school received another "F" in the next three years. These F-schools constitute the group of treated schools. Schools that received a "D" grade in 1999 were closest to the F-schools in terms of grade, but were not directly treated by the program. These "D-schools" constitute the group of control schools. The treatment and control groups consist of 65 and 457 elementary schools, respectively.[9] Because the program was announced in June 1999 and the grades were based on tests held in February 1999, we classify schools into treatment and control groups on the basis of their pre-program scores and grades.

If the F-schools and D-schools have similar trends in scores in the pre-program period, any shift of the F-schools compared with the D-schools in the post-program period can be attributed to the program.

The identifying assumption here is that if the F-schools and D-schools have similar trends in scores in the pre-program period, any shift of the F-schools compared with the D-schools in the post-program period can be attributed to the program. To test this assumption, we first run the following fixed-effects regression (and its ordinary least squares [OLS] counterpart) using only pre-program data:

$$(1) \qquad s_{it} = f_i + \alpha_0 t + \alpha_1 (F * t) + \alpha_2 X_{it} + \varepsilon_{it},$$

where s_{it} is the mean score of school i in year t, f_i are school-fixed effects, t denotes a time trend, F is a dummy variable taking a value of 1 for F-schools and 0 for D-schools, $F*t$ is an interaction between the F dummy and trend, X_{it} denotes the set of school characteristics, and ε_{it} is a stochastic error term. Scores considered in the Florida part of the analysis include mean school scores in FCAT reading, FCAT math, and FCAT writing. The pre-program difference in trend of the F-schools is captured in α_1.

If F-schools and D-schools have similar pre-program trends, we investigate whether the F-schools demonstrate a higher improvement in test scores in the post-program era using specification 2 below. If the treated F-schools demonstrate a differential pre-program trend, then in addition to estimating this specification, we estimate a modified version in which we control for the pre-program differences in trends.

We estimate a completely unrestricted and nonlinear model that includes year dummies to control for common year effects and interactions of post-program year dummies with

[8] The method of reporting ITBS math and WRCT reading scores changed in 1998. Therefore, we use pre-1998 scores.

[9] We restrict our analysis to elementary schools because there were too few middle and high schools that received an "F" grade in 1999 (seven and five, respectively) to justify analysis.

the F-school dummy to capture individual post-program year effects:

$$(2) \quad s_{it} = f_i + \sum_{j=1999}^{2002} \beta_{0j} D_j + \sum_{j=1999}^{2002} \beta_{1j}(F * D_j) + \beta_2 X_{it} + \varepsilon_{it},$$

where $D_j, j = \{1999, 2000, 2001, 2002\}$ are year dummies for 1999, 2000, 2001, and 2002, respectively. While the above specification includes school-fixed effects, we also estimate an OLS counterpart to it. OLS regressions corresponding to both specifications 1 and 2 include a dummy for the treatment group *F*. Note that this is absorbed in the fixed-effects regressions because it is a time-invariant school effect.

Specification 2 does not constrain the post-program year-to-year gains of the F-schools to be equal and allows the program effect to vary across years. The coefficients $\beta_{1t}, i = 2000, 2001, 2002$ represent the effect of one, two, and three years into the program, respectively, for the F-schools. Given the nature of the Florida program, the 1999 threatened schools (that is, the schools that received an "F" grade in 1999) would be exposed to the threat of vouchers for the next three years only. Therefore, we track the performance of the threatened schools (relative to the control schools) for three years after the program—2000, 2001, and 2002—when the threat of vouchers would be in effect.

The above specifications assume that the D-schools were not affected by the program. Although the D-schools did not face any direct threat from the program, they might have faced an indirect threat because they were close to receiving an "F" grade.[10] Therefore, we next allow the F-schools and D-schools to be different treated groups (with varying intensities of treatment) and compare their post-program improvements, if any, with 1999 "C-schools," which are the next grade up in the scale using the above specifications after adjusting for another treatment group. It should be noted that since D-schools and C-schools may face the threat to some extent, our estimates may be underestimates (lower bounds), but not overestimates.

5.2 Milwaukee

Our strategy is based on and is similar to that of Hoxby (2003b). Since students in the Milwaukee Public Schools eligible for free or reduced-price lunch were also eligible for vouchers, the extent of treatment of the Milwaukee schools depended on the percentage of students eligible for free or reduced-price lunch.[11] Using this information, Hoxby classifies the Milwaukee schools into two treatment groups based on the percentages of students eligible for free or reduced-price lunch—"most treated" (at least two-thirds of students eligible in the pre-program period) and "somewhat treated" (less than two-thirds of students eligible in the pre-program period).

We classify the schools into three treatment groups (in contrast to Hoxby's two) based on their pre-program (1989-90 school year) percentage of students eligible for free or reduced-price lunch. Thus, our treatment groups are more homogenous as well as starker from each other. Additionally, to test the robustness of our results, we consider alternative samples obtained by varying the cutoffs that separate the different treatment groups, departing from the Hoxby approach. The 60-47 (66-47) sample classifies schools that have at least 60 percent (66 percent) of students eligible for free or reduced-price lunch as "more treated," schools with such population between 60 percent (66 percent) and 47 percent as "somewhat treated," and schools with such population less than 47 percent as "less treated." We also consider alternative classifications, such as "66" and "60" samples, where there are two treatment groups—schools that have at least 66 percent (60 percent) of students eligible for free or reduced-price lunch are designated as more treated schools, and schools with such population below 66 percent (60 percent) as somewhat treated schools. Since there were very few middle and high schools in the Milwaukee Public Schools and student participation in the Milwaukee Parental Choice Program was mostly in the elementary grades, we restrict our analysis to elementary schools.

Since students in the Milwaukee Public Schools eligible for free or reduced-price lunch were also eligible for vouchers, the extent of treatment of the Milwaukee schools depended on the percentage of students eligible for free or reduced-price lunch.

[10] In fact, there is some anecdotal evidence that D-schools may have responded to the program. The superintendent of Hillsborough County, which had no F-schools in 1999, announced that he would take a 5 percent pay cut if any of his thirty-seven D-schools received an "F" grade on the next school report card. For more information, see Innerst (2000).

[11] Under the Milwaukee program, all households at or below 175 percent of the poverty line are eligible to apply for vouchers. Households at or below 185 percent of the poverty line are eligible for free or reduced-price lunch. However, the cutoff of 175 percent is not strictly enforced (Hoxby 2003a), and households within this 10 percent margin are often permitted to apply. In addition, there were very few students who fell in the 175 percent-185 percent range, while in fact 90 percent of students eligible for free or reduced-price lunch qualified for free lunch (Witte 2000). Students below 135 percent of the poverty line qualified for free lunch.

Our control group criteria are also based on Hoxby (2003b). Since all schools in Milwaukee were potentially affected by the program, Hoxby constructs a control group that consists of Wisconsin schools outside Milwaukee that satisfy the following criteria in the pre-program period that: 1) had at least 25 percent of their population eligible for free or reduced-price lunch, 2) had black students who make up at least 15 percent of the population, and 3) were urban. Her control group consists of twelve schools.

For our control schools, we designate schools that are located outside Milwaukee but within Wisconsin, satisfy Hoxby's first two criteria, and have locales as similar as possible to the Milwaukee schools. Note that all of these characteristics pertain to the pre-program school year 1989-90.[12]

Using each sample, we investigate how the different treatment groups in Milwaukee responded to the "voucher shock" program. Using specification 3 below, we first test whether the pre-program trends of the untreated and the different treated groups were the same. We then estimate OLS and fixed-effects versions of specification 4 below. If we observe differences in pre-existing trends between the different treated groups of schools, then in addition to estimating specification 4, we estimate modified versions of the specification that control for pre-existing differences in trends:

$$(3) \qquad s_{it} = f_i + \gamma_0 t + \sum_k \gamma_{1k} (I_k * t) + \gamma_2 X_{it} + \varepsilon_{it}$$

$$(4) \qquad s_{it} = f_i + \sum_{j=1989}^{2007} \delta_{0j} D_j + \sum_{j=1989}^{2007} \delta_{1kj} (I_k * D_j) + \gamma_2 X_{it} + \varepsilon_{it},$$

where s_{it} denotes scores of school i in period t; D_j, $j = \{1989, \ldots, 2007\}$ are year dummies for 1989 through 2007, respectively; $k \in \{MT, ST, LT\}$ for the WRCT and $k \in \{MT, ST\}$ for the ITBS, where MT denotes "more treated," ST denotes "somewhat treated," and LT denotes "less treated." The scores considered are mean scores in ITBS reading and ITBS math as well as percentages of students above the standard in WRCT reading.

6. Results

Table 2 presents baseline characteristics of treated and control groups in Florida and Wisconsin. It shows that the more treated schools in Florida were indeed similar to the more treated schools in Wisconsin and, except in one case, the differences between them were not statistically significant. Similarly, the control schools in Florida were similar to the control schools in Wisconsin, and the differences between them were not statistically significant.

Table 2

Pre-Program Demographic Characteristics of Florida and Wisconsin More Treated and Control Schools

Percent

Panel A: More Treated Schools

		Wisconsin		Florida–Wisconsin	
	Florida	66-47	60-47	66-47	60-47
Black	62.79	66.55	62.90	-3.76	-0.10
	(28.23)	(32.22)	(29.58)	[0.56]	[0.99]
Hispanic	18.95	18.07	14.81	0.88	4.14
	(23.40)	(24.54)	(21.86)	[0.87]	[0.36]
White	17.18	10.21	17.38	6.97	-0.20
	(19.54)	(10.68)	(16.55)	[0.07]	[0.96]
Male	51.38	52.25	52.33	-0.87	-0.95
	(4.84)	(2.60)	(2.58)	[0.34]	[0.22]
Free or reduced-price lunch	85.80 (9.95)	84.5 (6.48)	82.9 (9.04)	1.3 [0.50]	2.9 [0.12]

Panel B: Control Schools

	Florida	Wisconsin	Florida–Wisconsin
Black	18.12	22.37	-4.25
	(14.17)	(12.93)	[0.10]
Hispanic	15.49	14.84	0.17
	(21.23)	(6.02)	[0.86]
White	63.59	60.85	2.73
	(22.33)	(12.80)	[0.49]
Male	51.38	50.63	0.76
	(4.84)	(2.29)	[0.43]
Free or reduced-price lunch	50.14 (17.51)	44.95 (11.66)	5.19 [0.10]

Source: Author's calculations.

Notes: The group of Florida more treated and control schools is composed of F-schools and C-schools, respectively. Samples 66-47 and 60-47 are described in Section 5.2 of the article. Standard deviations are in parentheses; *p*-values are in brackets.

However, the treated schools were somewhat different from the control schools within each state. The reason is that Wisconsin schools outside Milwaukee were considerably more advantaged than schools in Milwaukee. We arrived at this control group despite using the strategy (following Hoxby [2003a, b]) of selecting control schools as similar as possible to Milwaukee's more treated schools in terms of pre-program characteristics.

It is important that both the more treated schools and the control groups be similar across the two programs in terms of pre-program characteristics as well as across the two locations. As a result, for purposes of comparing effects across the two

[12] The more treated and control group characteristics are presented in Table 2. In the 66-47 sample, the somewhat treated (less treated) group had an average of 55.4 percent (37.17 percent) of students eligible for free or reduced-price lunch, 50.99 percent (45.37 percent) who were black, and 4.09 percent (3.83 percent) who were Hispanic.

programs, we use the C-schools in Florida as the control group. Noticeably, the control group in Wisconsin was very similar to the C-schools in Florida and was not statistically different from them in terms of any characteristics (Table 2). Still another reason for selecting the C-schools as the control group in Florida was that while the D-schools were more similar to the more treated F-schools in terms of grade and demographics, they were very close to receiving an "F" grade; hence, to some extent they perceived an indirect threat and to some extent were treated by the program.

Because of differences between the treated and control schools, one might argue that in the absence of the program, the control group would have evolved differently from the more treated group. However, multiple years of pre-program data allow us to check (and control) for any differences in pre-program trends of these groups. In this way, we can dispose of any level differences between the treated and control groups as well as control for differences in pre-program trends, if any. It seems likely that once we control for differences in trends as well as in levels, any remaining differences between the treated and control groups will be minimal. In other words, our identifying assumption is that if the treated schools followed the same trends as the control schools in the immediate pre-program period, they would have evolved similarly in the immediate post-program period in the absence of the program. We also control for time-varying observable characteristics. School-fixed effects remove any time-invariant unobservable characteristics. Note that while time-varying unobserved characteristics cannot be directly controlled for, they did not drive the results as long as the F-schools did not experience a differential shock in unobserved characteristics that coincided with the timing of the program.

6.1 Florida

Considerable anecdotal evidence suggests that F-schools have responded to the voucher program. Just after the program's inception, Escambia County implemented a 210-day extended school year in its F-schools (the typical duration was 180 days), introduced an extended school day at least twice a week, and added small-group tutoring on afternoons and Saturdays and longer time blocks for writing and math instruction. To curb absenteeism, the county started an automated phone system to contact parents when a child is absent. Miami-Dade County hired 210 additional teachers for its twenty-six F-schools, switched to phonics instruction, and encouraged parents (many of whom were dropouts) to go back to school for a high-school-equivalency diploma. Broward County reduced its class size to eighteen to twenty students in its low-performing schools and increased services for children whose primary language is not English. Palm Beach County targeted its fourth-grade teachers for coaching and began more frequent and closer observation of teachers in its F-schools (Innerst 2000). Carmen Varela-Russo, Associate Superintendent of Technology, Strategic Planning,

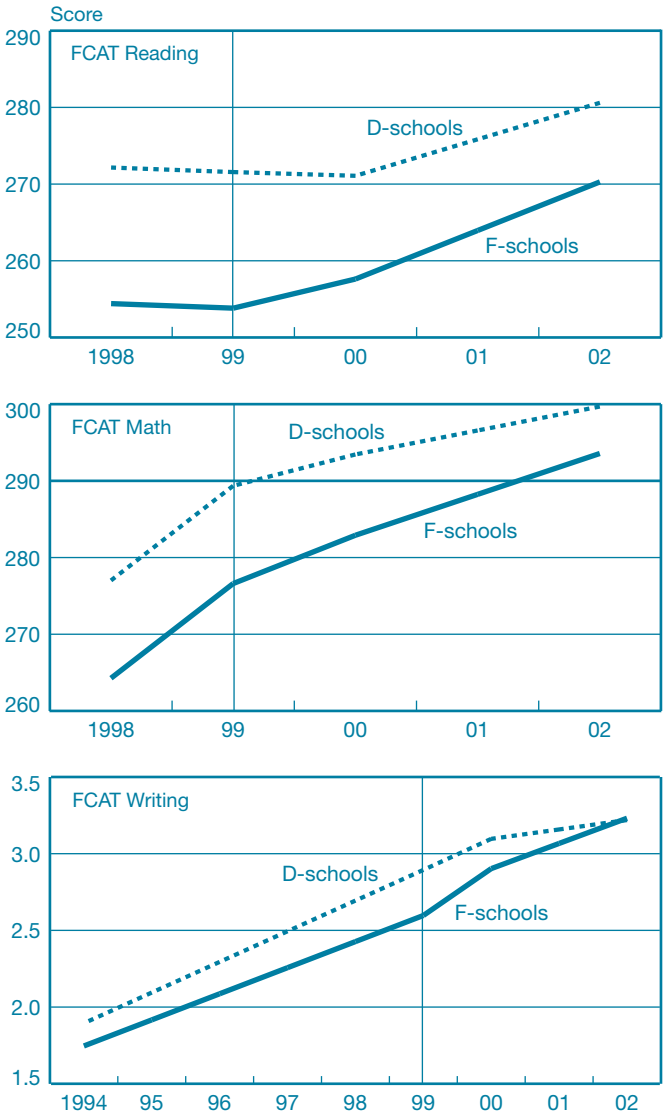
Considerable anecdotal evidence suggests that F-schools have responded to [Florida's] voucher program.

and Accountability at Broward County Public Schools, described the situation this way: "People get lulled into complacency . . . the possibility of losing children to private schools or other districts was a strong message to the whole community" (Innerst 2000). The analysis below investigates whether the data in Florida support this behavior.

Chart 2, which depicts trends in reading, math, and writing scores in F-schools and D-schools, shows that 1999 was the watershed year. In both reading and math, the F-schools had similar trends before the program. However, the F-schools showed improvement relative to the D-schools after the program, and the gap between F- and D-schools narrowed. In writing, while the F-schools were deteriorating relative to the D-schools before the program, this pattern changed after it. The F-schools showed improvement relative to the D-schools to the extent that they successfully closed the "F" to "D" gap after the program.

We now turn to our estimation results. All regressions control for ethnicity (the percentage of students in different racial categories in a school), the percentage of male students, the percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditures. Table 3 presents pre-program trends in reading, math, and writing. It reveals that F-schools have no significant differences in trend compared with D-schools in reading and math, although they exhibit a small, negative differential trend in writing. Compared with C-schools, F-schools exhibit a negative differential trend in reading and writing, but no significant differential trend in math. D-schools exhibit a negative trend in reading and a positive trend in math and writing compared with C-schools. Whenever there is a difference in pre-program trends, our reported regressions control for these differences by including interactions between trend and the respective treated dummies.

Chart 2
Florida "Threat-of-Voucher" Program



Source: Author's calculations.

Note: FCAT is the Florida Comprehensive Assessment Test.

Table 4, columns 1-3 present the effects of the Florida threat-of-voucher program on F-school reading, math, and writing scores compared with those for D-schools. All models reported include school-fixed effects. The results from our OLS estimation are similar to the fixed-effects estimates and hence are not reported. The regressions for writing include interactions of the "F" dummy with trend to control for differences in pre-program trends seen above.[13,14] The table shows economically large, positive, and statistically significant effects in each subject area and year.

D-schools are considered as an additional treatment group in Table 4, columns 4-6. Here, we see how the program affects F-schools (more treated) and D-schools (less treated) compared with C-schools. All columns control for differences in pre-existing trends between groups. The results show positive, significant year effects in reading, math, and writing for F-schools in each of the years after the program's implementation. Although many of the D-school effects are also positive and significant, the F-school shifts are statistically larger in each year.[15] The F-school effects are economically meaningful as well. In reading, relative to the base year, F-schools showed a 3.6 percent improvement in the first year after the program, a 4.2 percent improvement after the second year, and a 6.3 percent improvement after the third year. In math, F-schools showed a 3.4 percent, 4.2 percent, and 4.5 percent improvement in the first, second, and third years, respectively, after implementation of the program. In writing, the percentage improvement was around 15 percent. At the end of 2002 (three years after program implementation), the pre-program gap between F-schools and C-schools was closed by 37.08 percent in reading, 30.31 percent in math, and around 75 percent in writing.

Our results show considerable improvement in the F-schools after the program compared with the control schools.

In summary, based on different samples, different subjects, and both OLS and fixed-effects estimates, our results show considerable improvement in the F-schools after the program compared with the control schools. Although D-schools show non-negligible improvement (at least in reading and writing), their improvement is considerably less than and statistically different from that of the F-schools.

[13] Note that the table reports only the coefficients that reflect program effects; therefore, the coefficient corresponding to this interaction term (which captures the differential pre-existing trend) is not reported. Pre-existing trends are reported in Table 3.

[14] The regressions for reading and math (columns 1 and 2) do not include this interaction term because there is no evidence of differential pre-program trends in reading and math for F-schools and D-schools (Table 3). Note that the results with inclusion of this term remain very similar.

[15] Here, we test whether the F-school effects are statistically different from the D-school effects against the null hypothesis that they are equal.

Table 3

Pre-Program Trend of F-, D-, and C-Schools in Florida

	Sample of F- and D-Schools						Sample of F-, D-, and C-Schools					
	FCAT Reading		FCAT Math		FCAT Writing		FCAT Reading		FCAT Math		FCAT Writing	
	OLS (1)	FE (2)	OLS (3)	FE (4)	OLS (5)	FE (6)	OLS (7)	FE (8)	OLS (9)	FE (10)	OLS (11)	FE (12)
Trend	0.41	-0.05	13.20***	13.02**	0.20**	0.21**	2.66**	2.70	9.79***	10.20***	0.18***	0.19***
	(0.56)	(0.47)	(0.55)	(0.61)	(0.008)	(0.003)	(0.57)	(0.36)	(0.53)	(0.38)	(0.01)	(0.002)
F * trend	-1.78	-2.01	-0.98	-0.72	-0.05***	-0.04***	-3.80	-4.77***	2.46	1.96	-0.03***	-0.03***
	(2.47)	(1.46)	(1.44)	(1.48)	(0.011)	(0.007)	(2.29)	(1.41)	(1.51)	(1.43)	(0.01)	(0.01)
D * trend							-2.29***	-2.69***	3.46***	2.79***	0.02**	0.02***
							(0.66)	(0.57)	(0.60)	(0.67)	(0.007)	(0.003)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Observations	1,013	1,013	1,006	1,006	2,948	2,948	2,386	2,386	2,377	2,377	6,982	6,982
R^2	0.58	0.93	0.59	0.90	0.64	0.80	0.76	0.95	0.74	0.93	0.65	0.82

Source: Author's calculations.

Notes: FCAT is the Florida Comprehensive Assessment Test; OLS is ordinary least squares regression; FE is fixed-effects regression. Controls include race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure. Huber-White standard errors are in parentheses. All regressions are weighted by the number of students tested.

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

Table 4

Effect of "Threatened Status" on FCAT Reading, Math, and Writing Scores

Sample of Treated F- and Control D-Schools in Florida

	Reading FE (1)	Math FE (2)	Writing FE (3)	Reading FE (4)	Math FE (5)	Writing FE (6)
Treated * one year after program	4.85***	6.78***	0.35***			
	(1.68)	(1.63)	(0.04)			
Treated * two years after program	3.30*	7.25***	0.37***			
	(1.71)	(1.82)	(0.04)			
Treated * three years after program	7.08***	5.35***	0.43			
	(1.78)	(2.00)	(0.05)			
Less treated * one year after program				3.53***	0.97	0.05**
				(0.76)	(0.85)	(0.02)
Less treated * two years after program				5.52***	2.54***	0.00
				(0.80)	(0.94)	(0.02)
Less treated * three years after program				7.94***	3.47***	-0.03
				(0.87)	(0.92)	(0.02)
More treated * one year after program				9.32[b]***	8.96[b]***	0.39[b]***
				(1.87)	(1.59)	(0.04)
More treated * two years after program				10.75[a]***	11.00[b]***	0.37[a]***
				(1.87)	(1.77)	(0.04)
More treated * three years after program				16.03[b]***	11.94[b]***	0.39[a]***
				(1.91)	(1.95)	(0.05)
School-fixed effects	Y	Y	Y	Y	Y	Y
Year dummies	Y	Y	Y	Y	Y	Y
Controls	Y	Y	Y	Y	Y	Y
Observations	2,550	2,524	4,476	5,933	5,909	10,587
R^2	0.77	0.76	0.85	0.86	0.83	0.86
p-value[c]	0.00	0.00	0.00	0.00	0.00	0.00

Source: Author's calculations.

Notes: FCAT is the Florida Comprehensive Assessment Test. FCAT scores for reading and math are for the period 1998-2000; FCAT scores for writing are for the period 1994-2002. FE is fixed-effects regression. Huber-White standard errors are in parentheses. Controls include race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure. All regressions are weighted by the number of students tested.

[a] More treated significantly different from less treated at 5 percent level.
[b] More treated significantly different from less treated at 1 percent level.
[c] p-value of F-test of the program effect on treated schools.

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

Table 5

Pre-Program Trend of More Treated, Somewhat Treated, and Less Treated Schools in Milwaukee

	WRCT (Percentage above)		ITBS Reading		ITBS Math	
	OLS	FE	OLS	FE	OLS	FE
	(1)	(2)	(3)	(4)	(5)	(6)
Trend	-3.84	-4.34**	-4.09	-3.45	-3.04*	2.52**
	(2.33)	(2.16)	(4.11)	(3.42)	(1.66)	(0.98)
More treated * trend	-3.08	-2.03	4.01	-1.88	0.56	0.32
	(3.41)	(3.35)	(3.69)	(2.73)	(1.97)	(1.40)
Somewhat treated * trend	-4.41	-3.61	3.14	2.12	0.73	0.31
	(3.01)	(2.67)	(4.05)	(3.17)	(1.83)	(1.21)
Less treated * trend	-2.33	-3.23				
	(3.61)	(3.10)				
Observations	242	242	411	411	410	410
R^2	0.50	0.87	0.30	0.56	0.30	0.71

Source: Author's calculations.

Notes: WRCT is the Wisconsin Reading Comprehension Test; ITBS is the Iowa Test of Basic Skills; OLS is ordinary least squares regression; FE is fixed-effects regression. Controls include race, sex, and percentage of students eligible for free or reduced-price lunch.

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

6.2 Milwaukee

The Milwaukee analysis uses the 66-47 sample. Estimation results for pre-program trends are presented in Table 5. The results show no statistical difference in trends between the various treated and control groups in any subject area.

Table 6 examines the effect of the Milwaukee "voucher shock" program on the WRCT (the percentage above), ITBS reading, and ITBS math scores of different treated groups. Except for the positive and statistically significant effect in WRCT reading in the test's second year, there is no statistically significant evidence of any effect of the program. Although the second year's somewhat treated effect in ITBS math is statistically significant, it is more than the corresponding more treated effect.[16]

Thus, the results in Milwaukee are mixed. The program seems to have had a positive and significant effect in the second year after the program's implementation, at least in the WRCT.

[16] Since the ITBS was administered in Milwaukee as a district assessment program, we do not have data on non–Milwaukee, Wisconsin, schools for this test. As a result, our comparison group is the less treated group of schools. Since the comparison group is also treated to some extent, we expect our estimates for the ITBS to be lower bounds.

These results seem to be robust in that they are replicated in the analysis with other samples.[17] Chart 3 presents the trends in ITBS scores for the various groups. As expected, there is no evidence of any program effect.

The results show no statistical difference in trends between the various treated and control groups in any subject area. . . . Except for the positive and statistically significant effect in [Wisconsin Reading Comprehension Test] reading in the test's second year, there is no statistically significant evidence of any effect of the program. . . . Thus, the results in Milwaukee are mixed.

[17] These results are not reported here, but are available from the author.

Table 6

Effect of the Milwaukee "Voucher Shock" Program

	WRCT (1)	ITBS Reading (2)	ITBS Math (3)
Somewhat treated * one year after program	2.03	4.15	-1.35
	(2.81)	(4.49)	(2.94)
Somewhat treated * two years after program	5.38**	7.83	6.14*
	(2.43)	(5.17)	(3.38)
Somewhat treated * three years after program	5.01	6.78	2.47
	(3.03)	(5.31)	(3.31)
More treated * one year after program	-0.92	1.12	-4.02
	(3.33)	(3.86)	(3.26)
More treated * two years after program	6.06*	6.59	4.36
	(3.14)	(5.15)	(3.83)
More treated * three years after program	5.69	2.85	-2.22
	(3.98)	(5.18)	(3.54)
School-fixed effects	Y	Y	Y
Year dummies	Y	Y	Y
Controls	Y	Y	Y
Observations	1,195	717	1,127
R^2	0.58	0.55	0.60
p-value[a]	0.11	0.62	0.27

Source: Author's calculations.

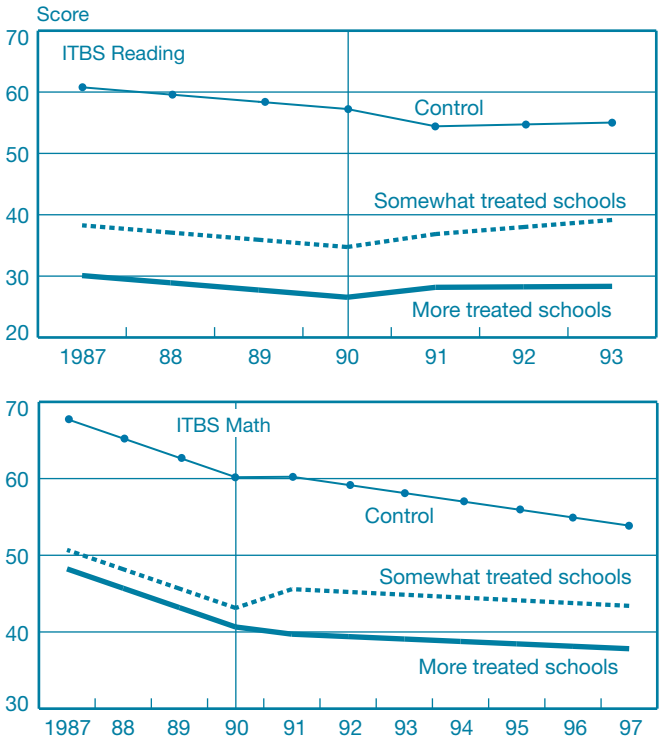
Notes: WRCT is the Wisconsin Reading Comprehension Test; ITBS is the Iowa Test of Basic Skills. Huber-White standard errors are in parentheses. All regressions include school-fixed effects and control for race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure.

[a] p-value of the F-test of joint significance of more treated shift coefficients.

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

Chart 3

Milwaukee "Voucher-Shock" Program



Source: Author's calculations.

Note: ITBS is the Iowa Test of Basic Skills.

7. Robustness Checks

7.1 Mean Reversion

Several factors might bias the results; we consider each factor and its potential solutions. First is the issue of mean reversion. Mean reversion is the statistical tendency whereby high- or low-scoring schools tend to score closer to the mean subsequently. Because the F-schools scored low in 1999, a natural question would be whether the improvement in Florida is driven by mean reversion rather than the voucher program. Since we conduct a difference-in-differences analysis, our estimates will be tainted by mean reversion only if F-schools mean-revert to a greater extent than do the D-schools or the C-schools, or both.

To investigate mean reversion, we examine whether and by how much schools that received an "F" grade in 1998 improved during the 1998-99 academic year compared with those that received a "D" (or "C") grade in 1998. Since these years fall within the pre-program period, the gain can be taken to approximate the mean-reversion effect and can be subtracted from the post-program gain of F-schools compared with D-schools (or C-schools) to get at the mean-reversion-corrected program effect.

The accountability system of assigning letter grades to schools began in 1999. The pre-1999 accountability system classified schools into four groups, designated 1 (low) to 4 (high). However, using the state grading criteria and data on the percentage of students in different achievement levels in each FCAT reading, math, and writing, we assigned letter grades to schools in 1998 and implemented the above strategy. Schools receiving "F," "D," and "C" grades in 1998 using this procedure are referred to as "98F-schools," "98D-schools," and "98C-schools," respectively.

Using Florida data for 1998 and 1999, we demonstrate in Table 7, panel A, that when compared with the 98D-schools, the 98F-schools show no evidence of mean reversion either in reading or math, although there is mean reversion in writing. Compared with the 98C-schools (panel B), there is no evidence of mean reversion in reading; both 98D-schools and 98F-schools show comparable amounts of mean reversion in math; only 98F-schools show mean reversion in writing.

Table 7

Mean Reversion of 98F-Schools Compared with 98D- and 98C-Schools, 1998-99

Panel A: 98F- and 98D-Schools

	Dependent Variable: FCAT Score, 1998-99		
	Reading FE (1)	Math FE (2)	Writing FE (3)
Trend	2.01***	14.02***	0.04***
	(0.43)	(0.49)	(0.01)
98F * trend	-0.65	1.17	0.14***
	(1.14)	(1.19)	(0.02)
Observations	1,353	1,354	1,355
R^2	0.93	0.91	0.85

Panel B: 98F-, 98D-, and 98C-Schools

	Dependent Variable: FCAT Score, 1998-99		
	Reading FE (1)	Math FE (2)	Writing FE (3)
Trend	1.76***	9.71***	0.03***
	(0.35)	(0.36)	(0.01)
98F * trend	-0.55	4.63***	0.14***
	(1.12)	(1.16)	(0.02)
98D * trend	0.16	4.22***	0.01
	(0.54)	(0.58)	(0.01)
Observations	2,605	2,608	2,608
R^2	0.96	0.94	0.87

Source: Author's calculations.

Notes: FCAT is the Florida Comprehensive Assessment Test; FE is fixed-effects regression. All regressions control for race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure. The ordinary least squares regressions include 98F- and 98D-school dummies. In the sample of 98F- and 98D-schools, the standard deviations of FCAT reading, math, and writing are 18.9, 18.05, and 0.30, respectively. In the sample of 98F-, 98D-, and 98C-schools, the standard deviations of FCAT reading, math, and writing are 21.16, 21.56, and 0.31, respectively.

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

Table 8

Is There a Stigma Effect of Getting the Lowest Performing Grade? Effect of Being Categorized in Group 1 on FCAT Writing Scores

	Using FCAT Writing Scores, 1997-98					
	Sample: Group 1, 2 Schools			Sample: Group 1, 2, 3 Schools		
	OLS (1)	FE (2)	FE (3)	OLS (4)	FE (5)	FE (6)
Trend	0.52***	0.52***	0.48***	0.48***	0.48***	0.46***
	(0.04)	(0.03)	(0.04)	(0.02)	(0.01)	(0.02)
Group 1 * trend	-0.01	-0.02	-0.02	0.03	0.01	0.02
	(0.08)	(0.06)	(0.06)	(0.07)	(0.05)	(0.05)
Group 2 * trend				0.03	0.04	0.04
				(0.04)	(0.03)	(0.03)
Controls	N	N	Y	N	N	Y
Observations	314	314	314	1,361	1,361	1,358
R^2	0.49	0.84	0.85	0.52	0.87	0.87

Source: Author's calculations.

Notes: FCAT is the Florida Comprehensive Assessment Test; OLS is ordinary least squares regression; FE is fixed-effects regression. Huber-White standard errors are in parentheses. All regressions are weighted by the number of students tested; controls include race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure. The OLS regressions include group 1 and group 2 dummies.

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

7.2 Stigma Effect of Getting the Lowest Performing Grade

A second concern in Florida is the potential stigma effect of receiving a performance grade of "F." If there is such a stigma, the F-schools will try to improve only to avoid this stigma rather than in response to the program. We use several alternative strategies to investigate this possibility. First, although the system of assigning letter grades to schools started in 1999, Florida had an accountability system in the pre-1999 period when schools were categorized into four groups, designated 1 (low) to 4 (high), based on FCAT writing and reading and math norm-referenced test scores. Using FCAT writing data for two years (1997 and 1998), we investigate whether the schools, which were categorized in group 1 in 1997, improved in relation to the 1997 group 2 and group 3 schools in 1997-98.[18] Our rationale is that if a stigma effect is associated with getting the lowest performing grade, the group 1 schools should improve relative to the group 2 and 3 schools, even in the absence of the threat-of-voucher program.

If there is [a low-performance] stigma, the F-schools will try to improve only to avoid this stigma, rather than in response to the program.

Table 8, using pre-program FCAT writing scores, shows that no such stigma effect exists—group 1 schools display no improvement relative to the group 2 or group 3 schools.

Second, all the schools that received an "F" grade in 1999 received higher grades in 2000, 2001, and 2002. Therefore, although the stigma effect on F-schools may be operative in 2000, this is not likely to be the case in 2001 or 2002 since none of the F-schools received an "F" grade in the preceding year (2000 or 2001, respectively). However, the F-schools would face the threat of vouchers until 2002, so any improvement in

[18] We do not use the pre-1999 reading and math norm-referenced test (NRT) scores because different districts used different NRTs during this period, which varied in content and norms. Also, districts often chose different NRTs in different years. Thus, these NRTs were not comparable across districts and across time. Moreover, since districts could choose the specific NRT to administer each year, the choice was likely related to time-varying (and also time-invariant) district-unobservable characteristics that also affected test scores.

2001 and 2002 would provide evidence in favor of the threat-of-voucher effect and against the stigma effect. F-schools showed strong gains in both 2001 and 2002—a result that provides further support for the threat-of-voucher effect and against the stigma effect.

7.3 Sorting

Another factor relates to sorting in the context of the Milwaukee voucher program. Vouchers affect public school quality not only through direct public school response but also through changes in student composition and peer quality brought about by sorting. These three factors are then reflected in the public school scores.[19] This issue is important in Milwaukee because over the years students have left the city's public schools with vouchers. In contrast, no Florida school became eligible for vouchers in 2000 or 2001. Therefore, the program effects (for each of the years 2000, 2001, and 2002) are not likely to be tainted by this factor.[20] Moreover, as we discuss shortly, the demographic compositions of the different groups of schools remained very similar across the years under consideration.

We also examine whether the demographic composition of the different Milwaukee treated groups changed over the years (Table 9). No such evidence is found. Only a few of the coefficients are statistically significant, and they are always very small in magnitude. They imply changes of less than 1 percent, more precisely, ranging between 0.22 percent and 0.65 percent. This result suggests that sorting was not an important factor. Note that we conducted the same exercise for Florida as well and found no evidence of any relative shift of the demographic composition of the F-schools compared with the D-schools or C-schools.

Vouchers affect public school quality not only through direct public school response but also through changes in student composition and peer quality brought about by sorting.

[19] See Hsieh and Urquiola (2006) for a discussion.

[20] This does not mean that the Florida program was not credible. Ten schools received a second "F" grade in 2002, nine schools in 2003, and twenty-one in 2004; all of these students became eligible for vouchers.

Table 9
Effect of Milwaukee Program on Demographic Composition of Schools
Percent

	Black (1)	Hispanic (2)	Asian (3)
Less treated * program	0.90	0.40	0.04
	(1.59)	(0.83)	(0.37)
Somewhat treated * program	-0.25	1.06	0.53
	(1.35)	(0.63)	(0.37)
More treated * program	-1.0	1.57	0.65*
	(1.34)	(0.81)	(0.37)
Less treated * program * trend	0.22	0.16	0.24***
	(0.32)	(0.15)	(0.07)
Somewhat treated * program * trend	0.70	-0.12	0.29***
	(0.25)	(0.13)	(0.07)
More treated * program * trend	0.08	-0.39***	-0.22***
	(0.23)	(0.14)	(0.07)
Observations	1,228	1,226	1,216
R^2	0.95	0.97	0.91

Source: Author's calculations.

Notes: Huber-White standard errors are in parentheses. All regressions are weighted by the number of students tested. All columns include a time trend, a program dummy that takes a value of 1 after the program, and an interaction between program dummy and trend.

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

A Comparison of Program Effects in Florida and Milwaukee

Since Florida and Milwaukee are in different regions, we argue that our comparison of the effects of the two programs is fair and reasonable. First, as mentioned earlier, apart from the crucial design differences between the two programs, the other features of the programs were very similar. In both programs, private schools could not discriminate against voucher applicants. Also, the method of funding for the two programs, the average voucher amounts, and the per-pupil revenue losses from vouchers were very similar. Second, state and local revenues constituted very similar proportions of total revenue during the relevant periods—the percentages of revenue from state and local sources were 51 percent and 41 percent, respectively, in Florida, and 55 percent and 36 percent, respectively, in Milwaukee. Third, the demographic characteristics of the more treated and control schools in Florida were very similar, both economically and statistically, to those of the more treated and control schools in Milwaukee

Table 10

Comparison of Results from Florida "Threat-of-Voucher" and Milwaukee "Voucher-Shock" Programs Using Standardized Reading and Math Scores

					Corrected for Mean Reversion			
	Reading		Math		Reading		Math	
	Wisconsin WRCT (1)	Florida FCAT (2)	Wisconsin ITBS (3)	Florida FCAT (4)	Wisconsin WRCT (5)	Florida FCAT (6)	Wisconsin ITBS (7)	Florida FCAT (8)
More treated * one year after program	-0.06	0.47***	-0.24	0.45***	-0.06	0.47***	-0.24	0.24***
More treated * two years after program	0.38*	0.50***	0.26	0.55***	0.38*	0.50***	0.26	0.34***
More treated * three years after program	0.35	0.80***	-0.13	0.60***	0.36	0.80***	-0.13	0.39***

Source: Author's calculations.

Notes: Reading test scores are from the Wisconsin Reading Comprehensive Test (WRCT), 1989-97, and the Florida Comprehensive Assessment Test (FCAT) Reading, 1998-2002. Math test scores are from the Iowa Test of Basic Skills (ITBS) Math, 1986-97, and the FCAT Math, 1998-2002. All figures are respective sample standard deviations. All figures are obtained from regressions that contain school-fixed effects, year dummies, interactions of year dummies with the respective treatment dummies, race, sex, percentage of students eligible for free or reduced-price lunch, and real per-pupil expenditure. Standard deviation of FCAT reading scores = 20; standard deviation of FCAT math scores = 20; standard deviation of WRCT (percentage above) reading scores = 16; standard deviation of ITBS reading scores = 18.45; standard deviation of ITBS math scores = 16.71. For standard deviations corresponding to the mean reversion sample, see the notes to Table 4.

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

(Table 2). Fourth, we repeat our analysis by comparing the improvement in Milwaukee with that of a large urban district in Florida: Miami-Dade County (the state's largest school district). The results are very similar and hence are not reported here. Finally, and perhaps most importantly, since we follow a difference-in-differences strategy in trends, any level or even trend differences between the two regions (that are common to schools in that region) are differenced out. It is unlikely that any remaining difference, which differentially affects the *trends* in the two regions *only* in the post-program period, will be large.

Table 10 compares the effects of the Florida and Milwaukee programs on their respective more treated schools both before and after correcting for mean reversion. Figures are based on data in Tables 4 and 6, and all numbers are expressed in terms of their respective sample standard deviations. Columns 1-4 present results before correcting for mean reversion; columns 5-8 present results corrected for mean reversion. Pre-correction results show positive and significant effect sizes in each of the years and subject areas in Florida, which always exceed the corresponding Milwaukee effect sizes (which are not significant, except in second-year reading). Mean-reversion-corrected effect sizes are obtained by subtracting the effect size attributed to mean reversion (obtained from expressing the relevant coefficients in Table 7, panel B, in terms of respective standard deviations) from the F-school effect sizes (obtained from expressing the more treated coefficients in Table 4, columns 4-6, in terms of respective sample standard deviations) in each of the three years after the program. The estimates in reading are the same as those described earlier. In math, although the effect sizes fall in Florida, they are still positive and considerably larger than those in Milwaukee. In reading (math), relative to the control schools, the F-schools show an improvement of 0.47 (0.24) standard deviations in the first year after the program, 0.5 (0.34) standard deviations after the second year, and 0.8 (0.39) standard deviations after the third year. Mean-reversion-corrected effect sizes in writing are 0.29, 0.25, and 0.29 in the first, second, and third years, respectively, after the program. Note that since none of the F-schools received an "F" grade in either 2000 or 2001, the mean-reversion-corrected effect sizes attributed to the Florida program in the second and third years may be underestimates.

8. Lessons for New York City

Our analysis of school voucher programs implies that policies that threaten underperforming public schools (or other agents) with functional and credible sanctions can induce them to respond in a way intended or desired by the policymaker. This finding has important implications for some educational policies in New York City. These include New York City's own accountability policy, also known as the "Progress Report" policy, and the federal No Child Left Behind law, as implemented by New York State.

The Progress Report policy was introduced in New York City in 2007. It rates schools on a scale of A to F, with grades based on three components: school environment, student performance, and student progress. A school's environment score is based on attendance rates and responses from surveys given to teachers, parents, and students. The other two scores are based on student performance in state math and English Language Arts (ELA) examinations. While student performance measures rely on level scores, student progress measures rely on growth or changes in student scores over years. The program attaches consequences to the letter grades. Higher grade (A) schools are eligible for increases in per-pupil funding and bonuses for principals. Schools receiving "F" or "D" grades are required to implement "school improvement measures and target setting." Low-performing (F- and D-schools) are also threatened with changes in their principal, and possible restructuring and closure if they continue to receive poor grades. The program also makes students in F-schools eligible to transfer to better performing schools.

As in Florida's voucher program, public schools in New York face valid sanctions if they fail to perform. Therefore, incentives faced by New York's low-performing schools are similar to those faced by the F-schools in Florida, and one would expect a similar response from them.

Although the Progress Report program does not have a voucher element, it is in many ways similar to the Florida voucher program; indeed, its design was based on the Florida program. Like the Florida program, it embeds sanctions in an accountability framework with consequences/sanctions imposed on low-performing schools if they fail to improve. Additionally, the criteria of the New York City program that make students in low-performing schools eligible to transfer to other higher performing schools are similar to those of Florida's program. The only distinction is that in New York, students can transfer to public schools only—not to private schools, as in the Florida program. The threat of removal of the principal and the possibility of restructuring are sanctions imposed over and above the transfer option. These sanctions are credible and pose a valid threat to administrators. For example, as reported in Rockoff (2008), "Seven schools receiving an F and two schools receiving a D were told in December of 2007 that they would be closed immediately or phased out after the school year 2007-08. . . . Additionally, 17 percent of the remaining F-school principals (and 12 percent of the D-school principals) did not return in the school year 2008-09, relative to 9 percent of principals receiving a C, B, or A grade."

Thus, as in Florida's voucher program, public schools in New York face valid sanctions if they fail to perform. Therefore, incentives faced by New York's low-performing schools are similar to those faced by the F-schools in Florida, and one would expect a similar response from them. Accordingly, the above analysis would indicate that low-performing schools under the Progress Report program would have an incentive to improve. In fact, there is some evidence in favor of such improvement. In 2009, 82 percent of students passed in math and 69 percent in English, up from 42 percent and 38 percent, respectively, in 2002. Earlier, all five boroughs of New York City ranked toward the bottom in the state; now Queens and Staten Island rank toward the top in elementary-school math scores. The racial achievement gap in passing rates has been closed by half in some tests. (Statistics are from Elissa Gootman and Robert Gebeloff, *New York Times*, August 4, 2009.) Gootman and Gebeloff also report:

> At Public School 398 in Brownsville, Brooklyn, 77 percent of students passed the math tests this year and 60 percent passed English, up from 56 and 43 percent last year. Gene McCarthy, a fifth-grade teacher, attributed the improvement to "a tremendous amount of test prep," but said that with a little creativity on his part, "ultimately I think it's learning." The principal, Diane Danay-Caban, said at P.S. 398, which had struggled for years with low scores and discipline problems, she has come to feel that the push to raise scores has brought genuine gains in knowledge.

Rockoff and Turner (2008) find that schools labeled "F" improved their performance in both ELA and math, with larger effects in math. Winters (2008), analyzing the same program, finds improvement of F-schools in math, although he finds no such effect in ELA.

NCLB, a major reform of the Elementary and Secondary Education Act, was signed into law on January 8, 2002. The states, including New York, implemented it soon thereafter. In compliance with the law, New York established Adequate Yearly Progress (AYP) targets, and all schools were graded on the basis of the targets. AYP is determined based on each school's progress toward meeting the state proficiency level for all students in English language arts, mathematics, science, as well as the high-school graduation rate. Schools are held accountable for the achievement of students of different races and ethnic groups, students with disabilities, students with limited English proficiency, and students of low-income families. Schools must also have an average over two years of 95 percent of their students participating in state tests. If a school does not meet requirements in any one of these categories, it is said to miss AYP. Schools that receive Title I money are subject to NCLB sanctions if they miss AYP in two consecutive years. A school missing AYP for two consecutive years is required to provide public school choice to students. A school missing AYP for three consecutive years is required to provide supplemental educational services (such as tutoring) in addition to the above sanctions. Missing AYP for four consecutive years leads to corrective action in addition to the above sanctions; for five consecutive years, it results in restructuring in addition to the above sanctions. Thus, sanctions start with two years of missed AYP and escalate from there.

Only a fraction of eligible students took advantage of the transfer option in New York as well as in the nation as a whole. This result is attributable mainly to two factors: the absence of an adequate number of spaces in nearby schools and the lack of adequate information.

While NCLB does not have any voucher component, the accountability-sanctions component is similar in spirit to that of Florida's voucher program. In fact, the design of NCLB was based on that program. As in the Florida program, NCLB first threatens failing schools with sanctions, and sanctions are introduced only if the schools fail to meet the predesignated targets in the following years.[21] Therefore, one would expect similar incentives to be created by NCLB and threatened schools to respond in a way similar to the F-schools under the Florida program. In other words, one would expect schools threatened by the NCLB sanctions to improve their performance in an effort to make AYP. However, it should be emphasized that these incentives and responses would be applicable only if the sanctions are credible and pose a valid threat to the affected schools. Under NCLB, though, implementation of the sanctions has been largely limited. For example, only a fraction of eligible students took advantage of the transfer option in New York as well as in the nation as a whole. This result is attributable mainly to two factors: the absence of an adequate number of spaces in nearby schools and the lack of adequate information. For example, as reported in the New York *Daily News*, "Some parents of kids in failing schools told the *Daily News* they weren't even aware they could transfer out, and some were turned away from better schools that are already overcrowded" (February 3, 2008).

The challenge to policymakers in [accountability] programs is to establish—and enforce—credible sanctions that function as valid threats to the agents (here, public schools).

In summary, both New York City's Progress Report program and NCLB have the potential to induce improvement from threatened schools, but the incentives and response ultimately depend on how functional and credible the threats under consideration are. The challenge to policymakers in such programs is to establish—and enforce—credible sanctions that function as valid threats to the agents (here, public schools). Only in such cases would the agents have an incentive to respond in the direction intended or deemed appropriate by the policymakers.

9. Conclusion

This article examines the role of program design in the context of two educational interventions in the United States—the Florida and Milwaukee school voucher programs. Even though both programs involve vouchers, their designs are quite different: the Milwaukee program makes low-income Milwaukee public school students eligible for vouchers, while the Florida system ties vouchers to low school performance. Specifically, Florida students become eligible for vouchers if

[21] Note, though, that while under NCLB all low-performing schools face stigma (embarrassment) due to public reporting of scores and grades, only Title I schools (schools that receive Title I money) are subject to sanctions.

and only if their school receives two "F" grades in a period of four years. This study shows that program design matters; indeed, the design differences have had very different incentive and performance effects on schools subject to the two programs. Specifically, the Florida program led to considerably larger improvements from the threatened schools compared with corresponding schools under the Milwaukee program. These findings are robust to several sensitivity checks.

The lessons drawn from our analysis are applicable to some of New York City's educational policies. These policies include the No Child Left Behind Act, as implemented by the state, and New York City's "Progress Report" policy. While neither of these programs has voucher components, both are accountability programs that have consequences for schools that fail to perform. In that sense, one would expect the incentives and responses generated by these programs to be similar to those created by the Florida program. Hence, the threatened schools could be expected to improve in an effort to avoid the sanctions. In fact, there is some evidence of such improvement in the affected schools, especially in schools treated by New York City's Progress Report program. However, the extent of the responses and the performance effects ultimately depends on the credibility of the sanctions and the validity of the threat posed to the affected schools.

References

Chakrabarti, R. 2007. "Vouchers, Public School Response, and the Role of Incentives: Evidence from Florida." Federal Reserve Bank of New York Staff Reports, no. 306, October.

———. 2008a. "Can Increasing Private School Participation and Monetary Loss in a Voucher Program Affect Public School Performance? Evidence from Milwaukee." Journal of Public Economics 92, no. 5-6 (June): 1371-93.

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

———. 2009. "Do Vouchers Lead to Sorting under Random Private-School Selection? Evidence from the Milwaukee Voucher Program." Federal Reserve Bank of New York Staff Reports, no. 379, July.

Epple, D., and R. E. Romano. 1998. "Competition between Private and Public Schools, Vouchers, and Peer Group Effects." American Economic Review 88, no. 1 (March): 33-62.

———. 2002. "Educational Vouchers and Cream Skimming." International Economics Review 49, no. 4 (November): 1395-1435.

Figlio, D. N., and C. E. Rouse. 2006. "Do Accountability and Voucher Threats Improve Low-Performing Schools?" Journal of Public Economics 90, no. 1-2 (January): 239-55.

Greene, J. P. 2001. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." Manhattan Institute for Policy Research Civic Report (with Florida State University and the Program on Education Policy and Governance at Harvard University), February.

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

Hoxby, C. M. 2003a. "School Choice and School Competition: Evidence from the United States." Swedish Economic Policy Review 10, no. 2: 9-65.

———. 2003b. "School Choice and School Productivity: Could School Choice Be a Tide that Lifts All Boats?" In C. M. Hoxby, ed., The Economics of School Choice, 287-342. National Bureau of Economic Research Conference Report. Chicago: University of Chicago Press.

Hsieh, C., and M. Urquiola. 2006. "The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program." Journal of Public Economics 90, no. 8-9 (September): 1477-1503.

Innerst, C. 2000. "Competing to Win: How Florida's A+ Plan Has Triggered Public School Reform." Available at http://www.edreform.com/published_pdf/Competing_To_Win_How_Floridas_A_Plus_Plan_Has_Triggered_Public_School_Reform.pdf.

Manski, C. F. 1992. "Educational Choice (Vouchers) and Social Mobility." Economics of Education Review 11, no. 4 (December): 351-69.

McMillan, R. 2004. "Competition, Incentives, and Public School Productivity." Journal of Public Economics 88, no. 9-10 (August): 1871-92.

Nechyba, T. J. 1996. "Public School Finance in a General Equilibrium Tiebout World: Equalization Programs, Peer Effects, and Private School Vouchers." National Bureau of Economic Research Working Paper no. 5642, June.

———. 1999. "School-Finance-Induced Migration and Stratification Patterns: The Impact of Private School Vouchers." Journal of Public Economic Theory 1, no. 1 (January): 5-50.

———. 2000. "Mobility, Targeting, and Private-School Vouchers." American Economic Review 90, no. 1 (March): 130-46.

———. 2003. "Introducing School Choice into Multidistrict Public School Systems." In C. M. Hoxby, ed., The Economics of School Choice, 145-94. National Bureau of Economic Research Conference Report. Chicago: University of Chicago Press.

Rockoff, J. E., and L. J. Turner. 2008. "Short-Run Impacts of Accountability on School Quality." National Bureau of Economic Research Working Paper no. 14564, December.

References (Continued)

West, M. R., and P. E. 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 no. 05-01.

Winters, M. A. 2008. "Grading New York: An Evaluation of New York City's Progress Report Program." Manhattan Institute for Policy Research Civic Report no. 55, November.

Witte, J. F. 2000. The Market Approach to Education: An Analysis of America's First Voucher Program. Princeton, N.J.: Princeton University Press.