SCHOOL VOUCHER STUDIES

<TARGET><BILL></BILL><SUBJECT>SCHOOL VOUCHER STUDIES</SUBJECT><COMM>SEDC27</COMM></TARGET>

SCHOOL VOUCHERS: EXAMINING THE EVIDENCE

Other books from the Economic Policy Institute

The State of Working America 2000-2001

The Class Size Policy Debate

Risky Business:
Private Management of Public Schools

Where's the Money Gone?
Changes in the Level and Composition of Education Spending

School Choice:
Examining the Evidence

SCHOOL VOUCHERS: EXAMINING THE EVIDENCE

By Martin Carnoy

ECONOMIC POLICY INSTITUTE

Washington, D.C.

Martin Carnoy, a professor of education and economics at Stanford University, has written extensively about education, labor markets, and the changing international economy. Some of his recent books include *The New Global Economy in the Information Age* (with M. Castells, S. Cohen, and F. H. Cardoso), *Decentralization and School Improvement* (with Jane Hannaway), *Faded Dreams: the Economics and Politics of Race in America,* and, most recently, *Sustainable Flexibility: Work, Family, and Community in the Information Age*, published by Harvard University Press and the Russell Sage Foundation. He is also the editor of the *International Encyclopedia of the Economics of Education*. Prof. Carnoy came to the School of Education at Stanford in 1969, where he helped build the International and Comparative Education Program.

Acknowledgments: I would like to thank Dominic Brewer, Brian Gill, Doug Harris, Larry Mishel, Bella Rosenberg, and Jennifer King Rice for their helpful comments and critiques of earlier drafts, and Tom Kane and Susanna Loeb for providing feedback at critical junctures. Obviously, the responsibility for the final product is mine.

ECONOMIC POLICY INSTITUTE
1660 L Street, NW, Suite 1200
Washington, D.C. 20036

http://www.epinet.org

ISBN: 0-944826-94-6

Table of contents

Executive summary

School vouchers have been in the limelight for a decade. The basic argument is that giving parents public funds to send their children to private schools will stimulate innovation and competition among schools. Although vouchers lack broad public support, parents in low-income inner cities are more likely to favor alternatives to traditional public education, and this interest has stimulated small pilot programs in a few urban school districts. Such programs have the potential to inform public debate about vouchers' strengths and weaknesses, but they have been evaluated mainly by researchers who openly and actively support vouchers. Yet the media tend to report results from these analyses without necessary caveats and alternative views. Now that the push for vouchers has reached the federal government through President Bush's education initiative, the urgency for a balanced perspective has become more important than ever.

Do school vouchers improve student performance? A review of the evidence finds that vouchers' effects on student achievement are almost certainly smaller than claimed by pro-voucher researchers. Although programs in many cities were designed to be like randomized-trial medical experiments—with high validity and reliability—common problems in implementation may have compromised validity and produced misleading results. Moreover, the results are marked by broad inconsistencies across grades, academic subjects, and racial groups.

Recent highly publicized research involving Florida schools also highlights the difficulty in attributing test score gains to vouchers, since many of these programs involve not only vouchers but also school grading systems and others variables at the same time. The same researchers who found large effects from earlier voucher programs also found large voucher effects in Florida. But a closer look reveals that most of the gains could have been caused by the school grading system, not vouchers. In three states with school grading systems—Texas, North Carolina, and Florida before vouchers—low-performing schools (sometimes referred to as "F" schools) produced gains quite similar to those of the Florida voucher program. Thus, the "scarlet letter" effect from identifying low-performing schools is as plausible an explanation for the test score gains as is the voucher threat.

Identifying the effects of programs is a challenging task, especially for vouchers. As the evidence slowly comes in, a balanced analysis suggests that voucher effects may exist, but they are significantly smaller than voucher proponents would have the public and the media believe.

Introduction

School vouchers have been in the limelight for almost a decade, mainly at the state and local level. But with George Bush's candidacy and his election to the presidency, they have now become a national issue. At the same time, voucher advocates have produced new reports claiming that students using vouchers improve their academic performance and that the threat of the availability of vouchers leads to improved student performance in public schools. The results of these reports have been widely—and largely uncritically—circulated in the press. They give the impression that vouchers are the solution to the educational woes of minority students in "failing" public schools. The problem is real, but do these studies support their claims? The empirical findings on the educational effects of vouchers deserve a closer look.

The idea of public funding of private schools is not new, nor does it belong exclusively to conservative free market reformers. In the 1960s and early 1970s, academics on the left, such as Christopher Jencks (1966), argued that vast differences between the quality of public schooling for inner-city blacks and suburban whites could not be resolved within the structure of a residentially segregated public education system. Jencks argued for a policy concept introduced by Milton Friedman (1955) more than a decade earlier. Friedman proposed to offer public funds to families that could be used only for education but in any educational institution, public or private. Such "vouchers" would serve to give families increased choice of the kind of education their children received. Friedman saw vouchers as a way to break the "monopoly" of the public sector over education and increase consumer choice, hence economic welfare. Jencks saw vouchers as a way of improving educational opportunities for a historically discriminated-against group within American society. Both shared a distrust of the state—Friedman of the bureau-centric state interfering with "democratic" markets, Jencks of the class/race-centric state reproducing inequality through public education. But conditions may have changed in the last 40 years. While there is still a glaring gap between achievement of black and white students, the gap has

been considerably narrowed. In the last decade the progress seems to have stopped, but it is unclear what the causes of the continued gap might be.[1] The voucher issue therefore has two different political origins. One is a conservative, free market ideology that prefers private to public provision of any services, and the second is the practical demand of low-income parents for better schooling, public or private. Even if private schools were no more effective than public schools, market reformers would insist that vouchers make parents and children better off because of choice and competition, and that private school choice should be made available to all parents, regardless of income. But the demand in inner cities for better schooling is based not on free market ideology but on academic results.

Whatever the origin of their support for vouchers, advocates have been attempting to support two claims: first, that private schools supported by public funds actually can do a better job than public schools of educating the children most at-risk of school failure, whether because vouchers are a route to smaller classes and better teachers, or because private schools are superior in other respects; and second, that vouchers increase incentives for public schools to improve by threatening low-performing public schools with the loss of students to competing private schools.

In the last few years, the leading proponent of the idea that private schools are demonstrably more effective at educating low-income African American students and an effective mechanism for improving public education has been Harvard Professor Paul Peterson. The research support for these claims is controversial, in large part because the Peterson group's statistical analysis seems always tilted to favor a positive result for vouchers. The history of such tilting is no longer just support generated for the alleged greater effectiveness of private education; it has also carried over into the claims regarding vouchers as a stimulus to better public schooling. In February 2001, Jay Greene, now a researcher at the Manhattan Institute, published a short paper assessing the impact of the Florida voucher plan on "failing" schools. All of these studies bear extremely close scrutiny.

This study reviews the recent empirical research in these two areas: (1) the effect of vouchers on student achievement, particularly for low-income minorities enabled to go to private schools; and (2) the effect of the threat of vouchers on low-performing public schools.

Among its findings:

- Research on the effect of vouchers in Milwaukee and Cleveland showed anywhere from no effects to small effects of vouchers for mainly African American students. Studies in Cleveland suggest that the achievement gains after two years in existing religious schools for voucher students

were higher in one subject, science. Voucher students in for-profit private schools did significantly worse than non-voucher students in one study, but did better and then worse according to another. The much larger size of the voucher in Milwaukee (about $5,500 currently) than in Cleveland (maximum $2,500) also suggests that, whether test scores in private schools are higher or not, a larger voucher induces many more families to transfer their children to private schools and induces more private schools to offer educational services to low-income students.

- Research in Dayton, New York, and Washington (conducted and evaluated by voucher proponent Paul Peterson and his colleagues) show no significant test score gains for Hispanic and white voucher recipients, but it did find gains for African Americans that were statistically significant overall in New York and Washington and marginally significant in Dayton (in reading only). But several methodological issues make these comparisons of achievement gains problematic. These issues include the inability to ensure that participants are available for follow-up evaluation; the inability to explain differences in outcome by grade/age and ethnic cohort; inconsistent inclusion and exclusion of data on students who experience either large gains or large drops in test scores.

- Findings that the threat of vouchers for students in "failing" (F) public schools caused math and writing gains among Florida's lowest-performing schools to increase significantly more than the gains of higher-performing schools are plagued by methodological problems. The research tends to overestimate the effect of being designated an F school, and it offers no evidence that the higher estimated test-score gain by an F school should be attributed to the threat of vouchers.

Chapter 1

Do vouchers for private education raise student achievement?

In the most recent salvo in the voucher debate, Paul Peterson and his colleagues (William Howell of the University of Wisconsin, Patrick Wolf of Georgetown University, and David Campbell, also of Harvard (Howell et al. 2000)) announced in August 2000 that their voucher experiments in New York, Washington, D.C., and Dayton, Ohio showed that at least some pupils—African Americans—achieve better in private than in public schools. The finding was widely hailed by voucher supporters across the political spectrum as showing that private schools could solve a problem public schools apparently could not—the lagging achievement of low-income inner-city black children.

As Robert Reich wrote in the *Wall Street Journal* (Reich 2000), "[e]vidence mounts that vouchers do work for kids who use them. A new study of students in New York, Washington, and Dayton, Ohio—conducted by researchers at Harvard, Georgetown, and the University of Wisconsin—found that after two years, the average performance of black students who switched to private schools was 6% higher than that of students who stayed behind in public schools. So why not simply 'voucherize' all education funding and let students and their parents select where they can get the best education?" And, as William Safire commented in the *New York Times* (Safire 2000), "This hard evidence is not what teacher unionists want to hear....The Harvard study shows Bush is on the right side of this. He should embrace the successful voucher students and joyfully join the controversy...."

But soon after the results were presented, another member of the Peterson team, David Myers, contractor for the New York City part of the research, challenged Peterson's interpretation, arguing that the New York results—even for African American students—were not convincing enough to support the Peterson group's policy conclusions. Earlier voucher studies in Milwaukee and Cleveland seemed to support this more carefully worded view.

Who is right? Even if we thought that voucher proponents were willing to limit vouchers programs to low-income, inner-city families, how

sanguine should we be that such inner-city (read African American) pupils would gain by switching to private schools?

The short answer is that the three-city study is not nearly as reliable as its authors claim. As a basis for educational policy, it should be interpreted cautiously. It is possible that a more structured private school environment with smaller classes and higher-achieving peers could help African Americans make greater gains than if they stayed in public schools. It is also possible that improvements to public schools would yield comparable improvements. But that said, the Peterson results may misrepresent gains that typical low-income African American students can make by switching to private schools. Using statistical techniques not easily understood by the media or the public, the studies' methodology is laced with potential biases. In the context of an intense ideological push for privatizing education, the question to ask is not *whether* these latest Peterson-group reports overestimate private school effects, but *by how much*.

In four cities—Dayton, New York, Washington, and Charlotte, N.C. (where data were released more recently)—the Peterson team built evaluations into the voucher plans themselves. Evaluating these evaluations is not easy, because, with the exception of New York, the researchers have not publicly released their data (the New York data were the most transparent because they were reported by grade). Earlier, though, in Milwaukee and Cleveland, the Peterson evaluations were constructed after the fact, and they included responses to research originally carried out by those not politically committed to vouchers. The Peterson estimates in those studies have a different character. For one, "experimental" controls were weaker or nonexistent. More important, the data were available to others and so were subject to re-analysis.

The Milwaukee voucher experiment

The longest-running voucher initiative in the U.S. is Milwaukee's. It began in 1991 on the initiative of Polly Williams, an African American Wisconsin legislator. The $2,500 vouchers were awarded by lottery to low-income families, 75% African Americans, to be used only in secular private schools. Schools had to accept the voucher as full payment (parents could not top it up). Initially, seven private schools agreed to take voucher students. Although the legislature set a maximum of 1,500 vouchers to be awarded, this number was never attained during the five years of the program. Enrollment increased steadily but slowly, from 341 in 1990-91 to 830 in 1994-95. The number of schools participating also increased, from seven in 1990-91 and six in 1991-92 to 11 in 1992-93 and 12 in 1994-5 and 1995-96.

The legislature commissioned University of Wisconsin professor John Witte and his colleagues to study the students who received vouchers and compare their achievement with similar students in public schools. Witte et al. found high levels of satisfaction among families receiving vouchers (Witte, Sterr, and Thorn 1995). Yet, when they analyzed achievement differences between those Milwaukee pupils who used vouchers to attend "choice schools" and Milwaukee public school pupils of similar socioeconomic background, race, and ethnicity, Witte et al. found that, generally, voucher students did no better in either math and reading. The one exception was a statistically significant *negative* effect of attending choice schools on reading scores in the second year of the program (1991-92). According to Witte et al., many of the poorest choice students left the program at the end of that second year. The authors also estimated the achievement effect controlling for the number of years the choice students had been in a private school. Again, private school voucher students did no better than public school students in either math or reading. The only effect that approached statistical significance was a negative reading score for those who had been in private schools for two years.

Witte et al. admitted that such an analysis has its limits, since many new students were being added to the private school sample every year, and a large fraction (about 30%) left the sample. The proportion leaving the sample was about the same for public school pupils. So the sample of private and public school pupils differed from year to year.

Other factors also changed in Milwaukee from year to year. The initial voucher was about one-half of Milwaukee's public school per-pupil spending. The voucher rose quickly, with private schools demanding and getting a higher voucher, until it was close to the primary school public cost per pupil when special education costs were accounted for. This is a major reason that more private schools were attracted into the program and more students could be accommodated in later years. Even so, over the course of the experiment, several of the participating private schools closed, including some due to bankruptcy.

In 1996 Peterson and his colleagues obtained the Milwaukee data and published their own study, using a "quasi-experimental" design that compared achievement of those who got vouchers in the lottery with those who did not. Peterson claimed that Witte had misspecified his model by comparing private school pupils with those who remained in public schools but had not necessarily applied for vouchers (Greene, Peterson, and Du 1996). In contrast, Greene et al. assumed that selection into the voucher program by lottery had resulted in a random sample of *applicants* being chosen to attend private schools. Thus, applicants should be the relevant pool from

which to draw comparison groups. The results of their comparison showed pupils attending private schools making significant gains in both math and reading over the students who applied for vouchers but ended up attending public school. The gains were found in the third and fourth years of the voucher program.

Witte (1997) countered that students who applied but did not get vouchers included students who had gotten vouchers but were rejected by the private schools. In addition, many of those pupils who were in the "control group" (those who had applied for vouchers but not gotten them) could not be located to measure their later test scores as public school students. Some others who had originally applied for vouchers and did not get them attended private religious schools assisted by a parallel, privately funded choice program, Partners for Advancing Values in Education (PAVE), so were not included in the control group. Since these students were more likely to have more educated and motivated parents than those who stayed in public schools, the control group was not necessarily a random sample of those who did not get vouchers.[2]

A third party, Princeton economist Cecilia Rouse, then took the same data, reworked them and found that students in private schools made faster gains in math (after the second year), but none in reading (Rouse 1998a). The gains in math were statistically significant but relatively small. Rouse compared the choice students (those who had been selected to get a voucher, whether or not they had actually used it) with both Greene et al.'s comparison group and a sample of Milwaukee public school students, similar to that used by Witte et al. She corrected all three samples for an implicit set of student characteristics that are invariant over time (including but not limited to, native ability, race, ethnicity, socioeconomic background, as well as other, "unobservable," student attributes) but may be correlated with students' families applying to get vouchers. These are called student "fixed effects." She argues that her results disagree with Witte et al.'s because the latter restricted their samples to students for whom prior test scores were available (her fixed effects variable does not depend on measuring prior test scores) and disagree with Greene et al.'s because Greene et al.'s reading results disappear when student fixed effects are included. A second Rouse paper found that gains for low-income Milwaukee public school students in smaller classes were higher than the gains of voucher students in private schools, which also were characterized by much smaller class sizes than in Milwaukee public schools (Rouse 1998b).

In 1997 the Wisconsin legislature expanded the voucher program to 15,000 low-income students, and included religious schools. The legislation was upheld by the Wisconsin Supreme Court. Initially, about 8,000

students took up the vouchers, which continued to be worth about the cost of Milwaukee's per pupil spending on primary education ($5,500 in 1997). In the first year, about one-third of voucher takers under this expanded program were already in private schools but qualified because of their low family incomes. By the school year 2001-02, about 10,000 children will use vouchers at over 100 mostly religious private schools (Williams 2000).[3] This is a significant fraction of Milwaukee's 100,000 public school students. Even if only 7,500 of the voucher students were not already in private schools and transferred from public schools, the voucher program has shifted almost 8% of Milwaukee's public school students to private schools. This suggests that, given a large enough voucher, many low-income families will take advantage of it, and at least some new schools will come into the market. However, no one knows whether voucher students are performing better in this expanded program because, unlike public school students, they are not required by the legislature to take state tests, and no evaluation program is written into the legislation. We also know little about how many students who took up vouchers returned to public schools after a year or two in a private school.

The Cleveland voucher program

Cleveland's voucher program was approved by the Ohio legislature in June 1995 and began in the 1996-97 school year with a maximum voucher of $2,500. Voucher recipients were chosen by lottery and received a fixed percentage of tuition charged by private schools, the percentage depending on the family's income level. Students whose family income was at or above 200% of the poverty line received 75% of the school's tuition up to $2,500, and those below the poverty line received 90%, up to $2,500.

The Cleveland program differed from the Milwaukee experiment in several important aspects. In Cleveland, more than twice as many vouchers were offered as in Milwaukee (3,700 versus 1,500). Unlike Milwaukee families, Cleveland families had to add to the voucher to attend private schools, both because the voucher covered only part of tuition and because private schools could charge tuition higher than the voucher. As in Milwaukee, the program got off to a slow start, with only about 1,500 students taking advantage of the vouchers.[4] A fraction (about 25%) of Cleveland's vouchers were offered to families with children already in private schools, and vouchers in Cleveland could be used in religious schools, as they later could in the expanded Milwaukee program. About 80% of families in the Cleveland program sent their children to Catholic and other parochial schools.

Nearly all of the others went to Hope Schools—two private for-profit

schools created by David Brennan, a wealthy entrepreneur and major contributor to Ohio's Republican Party, to take advantage of voucher availability. Brennan had been instrumental in getting the voucher program through the Ohio legislature, but was later unable to raise the value of the voucher once he realized that his schools were losing money at the $2,500 level. He subsequently converted the Hope Schools into charters to take advantage of higher levels of financing. This left almost all voucher students attending religious schools. On December 11, 2000, the 6th Circuit Federal Court of Appeals upheld a lower court ruling that these vouchers gave unconstitutional aid to religious schools. The program's future will be decided by the U.S. Supreme Court.

Evaluations of Cleveland vouchers are in even greater disagreement than the Milwaukee analyses. In Cleveland, almost as soon as the voucher program got under way, evaluations were conducted on different sets of students by two different groups of researchers. The Ohio legislature contracted researchers from the University of Indiana, headed by Kim Metcalf, to conduct the state's evaluation. Metcalf et al. could get "baseline" spring 1996 second-grade state test scores for almost all public school pupils. So the researchers focused on the almost 200 third-grade pupils who had received a voucher and switched from public to private schools in fall of 1996 (Metcalf et. al. 1998). They used as their outcome measure the scores on a test they gave to entire third-grade classrooms that contained their sample students.

Independently, the Peterson group (Greene, Howell, and Peterson) was also evaluating the Cleveland voucher program, but it focused on the two Hope Schools and looked at students in all grades, not just the third. The Peterson group did not collect data on previous tests taken by voucher students, but rather estimated increases in scores by comparing scores on tests *they* gave students in fall 1996 with scores on the same test in spring 1997. Greene et al. found higher levels of parent satisfaction in the Hope Schools and significant test score gains from fall 1996 to spring 1997 by students who started in one of grades K-3 in 1996 (Greene, Howell, and Peterson 1998). These findings were almost immediately criticized for using as baseline a test given right after the summer, when students have "lost" skills learned in the previous year (AFT 1997).

Because the Peterson group had already given tests twice in the Hope Schools (fall and spring), the schools' administration declined to test the children again using the Metcalf group's instrument. Metcalf et al. had to do a separate analysis of the Hope students. Their results of the second-grade test scores suggest that those students who used vouchers to get into non-Hope and Hope schools were similar socioeconomically to the sample

taken of students remaining in Cleveland public schools, but they had higher initial test scores than public school students (these differences were not large). Metcalf et al.'s results for the 94 non-Hope School third-grade voucher students for whom they had second-grade scores showed no significant differences in the gains posted by voucher students and students who stayed in public schools, once socioeconomic background differences were accounted for.

The Peterson group was able to obtain Metcalf's data and almost immediately criticized the researchers' results on three main grounds: they limited their analysis to third graders, they used a second-grade test taken in public school the previous year that is not an accurate measure of baseline test scores, and they left the Hope Schools out of the analysis. They also claimed that Metcalf et al. used a statistical technique that underestimated treatment effects. But when the Peterson group reanalyzed Metcalf's data not taking into account second-grade scores, it found significant gains only for private school students in language and science. Controlling for second-grade test scores, even these gains were statistically insignificant at the 5% significance level (Peterson, Greene, and Howell 1998).

Yet another study by the Peterson group confirms that reported gains for students in the Hope Schools depend heavily on the low scores on the initial test in fall 1996. Any later scores compared with those initial test scores result in substantial reported gains, but using spring 1997 as a base results in consistent declines in math and reading scores for students in the Hope Schools (Peterson, Howell, and Greene 1999).

The Metcalf group continued to evaluate the voucher plan, following the third-grade cohort into fourth grade and testing the large cohort of voucher students in first grade. As in Milwaukee, the attrition rate from private schools was substantial from the first to the second year of the program.[5] Second-year results showed that fourth-grade voucher students in the established, non-Hope private schools scored significantly higher than did public school students in language and science but not in other subjects. Students in the Hope Schools, however, scored significantly lower in all subjects than did either public school students or voucher students in non-Hope Schools (Metcalf et al. 1999). The differences in scores between voucher students in non-Hope Schools and public school students were lower when socioeconomic differences were taken into account, but Hope students still had significantly lower scores. As was the case in Milwaukee, Metcalf et al. found that private school classes were marked by fewer students per teacher than were classes in public schools.

To summarize, the Milwaukee and Cleveland research on the effect of vouchers showed anywhere from no effects to small effects of vouchers for

mainly African American student groups. Studies in Cleveland suggest that the achievement gains after two years in existing (religious) schools for voucher students were higher in one subject (science). Voucher students in for-profit private schools (the Hope Schools) did significantly worse than non-voucher students in one study (Metcalf et al.), but did better and then worse according to another study (Greene et al.).

Significantly, in both cities, different researchers, taking somewhat different approaches to the data, came up with different results. And in both cities, the Peterson group's evaluation of the voucher program reported the largest gains for students in private schools.

The new voucher research

The Peterson group's new round of research concerns efforts by well-financed voucher advocates to fund scholarships (vouchers) for low-income children to attend private schools. The programs establish lotteries for parents who apply, give applicants a baseline test, award scholarships to applicants at random, then later test children who did and did not receive the scholarships. Some families who get vouchers do not actually send their children to private school, either because they cannot come up with the extra tuition or because they cannot find convenient private schools to accept their children.[6]

Results for Dayton, New York, and Washington show no significant test score gains for Hispanic and white voucher recipients. Gains in the Howell et al. study are reported as National Percentile Ranking (NPR) points, which run from 0 to 100, with a national median of 50. This measurement allows the gains to be compared with test score gains reported in other studies, such as the Tennessee class size experiment. Gains for African Americans are found to be statistically significant overall in New York and Washington and marginally significant in reading in Dayton.[7] Reported gains are largest in Washington, D.C. (**Table 1**). As shown in **Figure A**, the aggregate math gain for African Americans across the three cities is about 5.5 percentile points in year 1. But math scores fail to increase significantly in year 2. The reading gain is negligible in year 1 and is about 6 percentile points in year 2. As shown in Table 1, gains for other ethnic groups are not significantly different from zero in either years 1 or 2.

Gains can also be expressed relative to the statistical variance of test scores in the sample of African Americans taking the test. This is called the "effect size." Gains are expressed in the proportion of a standard deviation in test score represented by the percentile point gains. One standard deviation is the difference in score between the average score (50th percentile)

TABLE 1 Impact on test score performance in three cities of switching to a private school (gain measured in National Percentile Ranking points)

	African American pupils				All other ethnic groups			
	Year 1		Year 2		Year 1		Year 2	
City/ grade/test	Gain	Sample size	Gain	Sample size	Gain	Sample size	Gain	Sample size
New York								
Grade 2/5								
Math	7.0***	623	4.1*	497	-2.1	817	-3.2	497
Reading	4.6**	623	4.5**	497	-1.3	817	0.2	497
D.C.								
Overall								
Math	7.3**	891	9.9***	700	8.5	39	5.8	44
Reading	-9.0**	891	8.1**	700	6.3	39	-5.6	44
Grade 2/5								
Math	9.8***	620	10.0***	490				
Reading	-5.1	620	8.6**	490				
Grade 6/8								
Math	1.5	270	12.8*	210				
Reading	-19.0***	270	7.8	210				
Dayton								
Grade 2/8								
Math	0.4	296	5.3	273	-0.8	108	0.0	96
Reading	6.1	296	7.6*	273	2.8	108	-0.4	96

* statistically significant at 10% level
** significant at 5% level
*** significant at 1% level
Note: Grade signifies grade entered in first year of study.

Source: Howell, Wolf, Peterson, and Campbell (2000), Tables 2A, 2B, and 2C.

and the 84th percentile on the up side or the 16th percentile on the down side. If the gain represents one-fourth of a standard deviation, it means, approximately, a move from the 50th to the 60th percentile. The percentile point gains for African Americans in Figure A translate into 0.3 standard deviations for math in year 1, with little if no gain in year 2, and, for reading, little gain in year 1 but about a 0.25 standard deviation effect size in year 2.

As shown in **Figure B**, if the major increase in math gains by D.C. middle school students between years 1 and 2 (from negligible to high positive gains) is excluded from the aggregate, the math gains for blacks decline between year 1 and year 2. Furthermore, if the huge turnaround in reading scores among D.C. elementary and middle school students is excluded from the aggregate, Dayton plus New York reading scores show a negligible aggregate gain after year 1.[8]

FIGURE A Test score gains for African Americans switching to a private school, New York, Washington, and Dayton

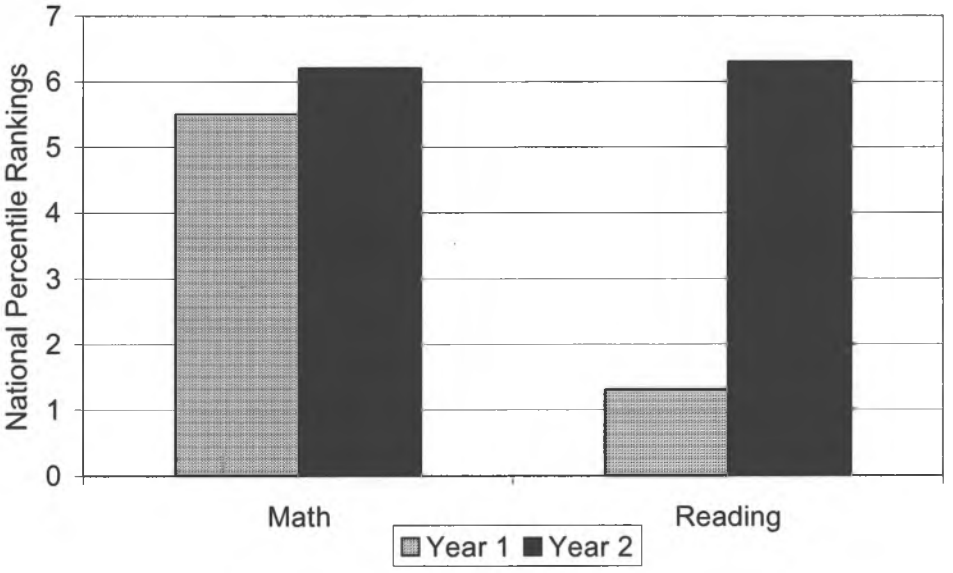
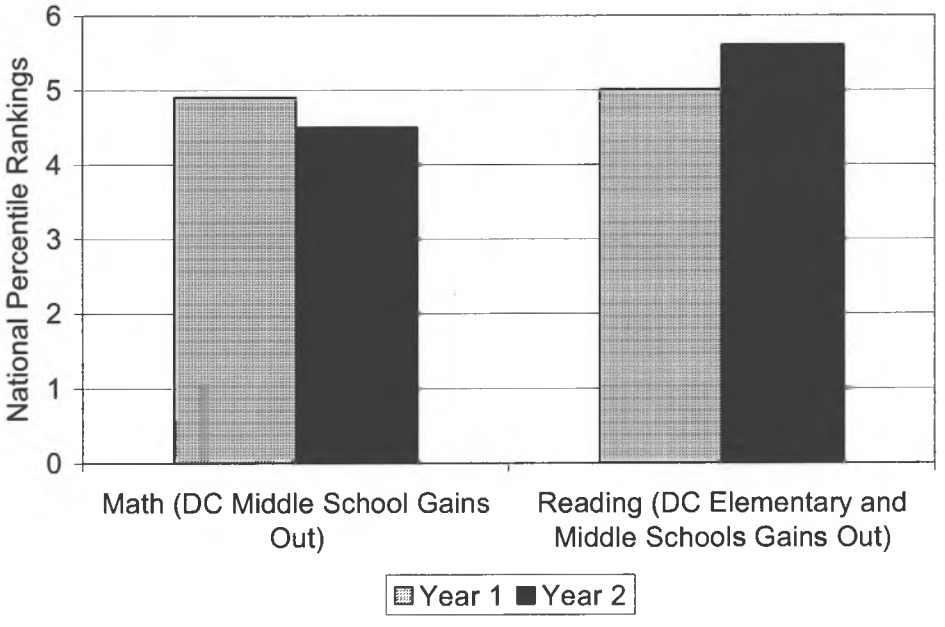


FIGURE B Test score gains for African Americans switching to a private school, Washington gains omitted



Source for figures A and B: see note to Table 1.

Thus, the combined results from two cities, Dayton and New York, suggest that for African American students the main effect of switching to a private school occurs in the year after the switch, and, as shown in Table 1, for other ethnic group (mainly Hispanic) students there is no effect at all. Using the same methodology, the Charlotte gains (not shown in table) are found to be about 6 percentile points in both reading and math. These are not broken down by ethnic group, but 80% of the sample is African American (Greene 2000).

Closer scrutiny

Several methodological issues make these comparisons of achievement gains problematic.

The 'disappointment' and Hawthorne effects problem. Comparing students already in public and private schools has a major disadvantage: private school students may come from more motivated families and have survived selection processes. Solving this problem requires an "experiment:" randomly assigning students to private and public schools.

The Peterson group draws on the statewide Tennessee class size experiment (Mosteller 1995) for much of the experimental method it uses to test the effects of vouchers. In Tennessee, students and teachers were randomly assigned to "normal" size classes (about 24 students) or classes reduced to about 15 students. All the students were followed and tested over the next 12 years. Researchers were able to distinguish those students who had stayed for several years in small classes, the students who had spent only one year in a small class and then switched, and the control group—those who stayed in normal size classes from kindergarten through the third grade. The Tennessee experiment, while randomly choosing the "treatment" (small class) and "control" groups (usual number of pupils in the classroom), was not a "blind" trial, as many medical experiments are. In a truly blind trial controls are given a placebo and do not know whether they are receiving the treatment. Education experiments can never fulfill this condition—families know their child's class size or whether they get a voucher. This makes education experiments subject to a "Hawthorne effect," where the fact that participants know they are involved in a treatment to produce a positive impact can cause them to try harder. The motivation of families who were rejected for the treatment (i.e., the controls) can also be affected by the experiment itself.

The voucher experiments in various cities have families apply for a voucher, give baseline tests to all applicants, then randomly select some to

get vouchers to attend private schools.[9] But the students in these experiments are not necessarily representative of low-income urban students. Families applying for vouchers whose children attend public schools are more motivated to switch their children and more dissatisfied with public schools than are average low-income parents, most of whom do not apply. Not receiving a voucher for parents already dissatisfied with their child's schooling could have an adverse "disappointment" effect on the child's performance. [10]

The differential gains recorded in these experiments may therefore be due partly to lower gains by discouraged voucher rejectees rather than greater gains by recipients. Stanford psychologist Claude Steele has done research showing that test scores are significantly affected by the self-perception of test-takers (Steele and Aronson 1998). Although pupils who did not get a voucher were selected randomly, they and their parents may still feel "unlucky" and less efficacious. For a better comparison, voucher experiments would need also to draw a random sample of pupils from urban public schools whose low-income parents do not apply for vouchers, and give them the initial and follow-up tests. These pupils would come from families who are probably more satisfied with their current situation.

Peterson and Howell (2001) have responded to this criticism by claiming that the level of satisfaction with their children's public schooling by those parents who did not receive a voucher did not decline significantly in the first year following the voucher lottery, but it did in the second. Neither did rejected parents' participation rate in school activities decline. Since the chance of getting a voucher was only one in 20, it is not likely, the authors argue, that getting rejected produces any "sore loser" effect. But this is not very convincing. Peterson and Howell fail to mention that parents who applied for the voucher had rather low satisfaction with their children's schooling to start with. Continued low levels of satisfaction (or even declines after two years) may be enough to affect their children's test scores negatively.

Peterson and Howell also claim that, although there are some signs of a Hawthorne effect, the gains continue to increase in year 2 for African Americans, a trend which suggests that voucher gains persist and are not the result of a Hawthorne effect. The claim that the test score gains persist are based mainly on a large D.C. turnaround in second-year middle school math scores and second-year elementary and middle school reading scores. When these are excluded, there is no second-year increase (see Figure B).

Non-returnees at follow-up. Another problem is self-selection for the follow-up evaluation. Voucher researchers measure academic gains by convincing families to bring children in on a weekend to take follow-up math

and reading tests. As in medical trials, high participation rates may require inducements. For those families who received vouchers, the New York inducement was that children would have to take the test to continue getting a voucher. Researchers used only moral suasion in other cities. For those who received but did not use the voucher and for those in the control group (who did not get a voucher), the inducement was typically $20 plus eligibility for a voucher in the future. Participation rates varied, with the highest rates in New York (about 66% in the second-year follow-up) and about 50% in D.C. and Dayton. The participation rate in Charlotte was particularly low, at 40%. All these are considerably lower than in medical trials.

The Peterson group deals with participation problems by estimating the probability that a student with a certain initial test score and set of family characteristics and attitudes would participate in each follow-up test and then weighting actual scores according to this probability. Probability functions were estimated separately for the control group and for those who received vouchers, using data researchers had gathered on the original questionnaires and the original test scores for all the students who had applied to be in the program. Using these probability functions, the students actually participating in the follow-up tests received a weight that was the inverse of the probability that a student with those characteristics would come back and take a follow-up test. For example, if a student had a set of parent characteristics that made it likely that he/she would participate in the follow-up, he/she was given a lower weight in the calculation of the estimated follow-up test. Thus, students who came back to take follow-up tests but had characteristics that made them unlikely follow-up participants got a bigger weight, so test scores would be more "representative" of the original group of students.

The researchers could not do much more than this to correct for no-shows. But the procedure is hardly free of potential bias. It assumes that follow-up test scores for the many who didn't take the tests would be the same as scores for those who did show up and had similar initial scores and similar parent characteristics. But we really don't know how follow-up scores of no-shows might be related to their not showing up to take the follow-up test. For example, those who did not show up to take the first-year or second-year follow-ups may have had indications from their performance during the school year that they might score low, even though they did reasonably well on the baseline test. This might have been especially true for private school students, or, alternatively, for public school students. Thus, the large non-participation rates could easily have reflected behavior that systematically biased the relative gains of voucher recipients and non-recipients.

Bias issues in taking up the voucher offer. Yet another problem is bias in who takes up the voucher offer. The vouchers, which range from $1,200 to $1,700, depending on the city, are not large enough to cover tuition at most of the private schools available. Many families receiving vouchers were unable to use them. In New York, 62% of families whose children started out in public school used the scholarships for two years; in Dayton and Washington, 53% used the voucher in the first year, with an unreported drop-off in the second year.

Voucher takers in each city, as would be expected, have higher income than non-takers. Critics have argued that this income difference biases results. But the researchers have made a valid attempt to deal with the problem by comparing the controls with all students who were offered the voucher, not only those who actually used it. Here is how it works. Those who receive a voucher and use it may be self-selecting when they choose to go to private school—they may be the better students from higher-educated families, with greater chances to make test score gains. But voucher recipients are randomly chosen, so a good "instrument" in this case is whether the student received a voucher—if there is a major difference between voucher recipients and users, the "instrument" should pick that up. To test whether the measured or unmeasured characteristics of voucher users would produce such a result, the researchers estimated the private school effect conditioned on the probability that someone who got a voucher actually used it.[11]

The single cohort problem. In New York, the only students who made significant gains were African Americans who switched to private schools when they were entering the fifth grade and whose gains were large enough to produce a significant average gain for the entire New York sample of African Americans.

Results for African American students in Dayton also have a strange inconsistency. Certain cohorts—those who entered second, fourth, and sixth grades in the first year of the experiment—had large National Percentile Ranking percentage-point gains in combined math and reading test scores for the two years in private schools, while those in the other grades did not. According to David Howell (who kindly provided unpublished data by grade), African American voucher recipients finishing private school third grades at the end of the study's second year made large two-year gains (in combined math and reading). Those finishing fourth grade performed slightly worse than pupils who did not get vouchers. Those finishing private school fifth grades made large gains, those finishing private school sixth grades made small losses, those finishing seventh grade made large gains, and

TABLE 2 Two-year test score gains for African American students from switching to private schools in Dayton and Washington, math and reading combined, by grade (National Percentile Ranking points)

Grade (in second year of trial)	Dayton point estimate	D.C. point estimate
3	19.7	10.2
4	-1.9	8.1
5	14.5	3.5
6	-1.8	9.7
7	17.0	11.4
8	-5.7	4.0
9	-14.9	-3.6

Source: Data provided to author by William G. Howell.

those finishing eighth and ninth grades made large losses compared to non-voucher students. Only in Washington, D.C. are achievement gains of voucher recipients attending private schools relatively consistent across grades (**Table 2**).With gains so variable by cohort, it is fair to ask, as did Mathematica's David Myers *(New York Times,* September 2000) concerning his New York study for the Peterson group, whether one can claim that students in private schools do better than those in public schools. Shouldn't we, instead, wonder what conditions produced such large gains for some cohorts but not for others?

Tankers and leapers. First-year results were reported in 1999. In Dayton and Washington, D.C., the first-year estimates in those earlier reports excluded "tankers" (test takers whose scores fell more than 1.5 standard deviations) and "leapers" (test takers whose scores rose more than 2 standard deviations). In the more recent second-year report, tankers and leapers are not excluded, changing the first-year results considerably. For example, the 7% gain in math scores reported in year 1 for black students in Dayton is reduced to zero. The Washington math score gain in grades 2-5 rises from 7% in the first-year report to 10% in the latest report, and the reading score in grades 6-8 drops from –8% to –19%. The rationale for not excluding leapers and tankers from the second round of estimates is that, if they stayed high or low, it indicated a more "permanent" effect. But the sample size in both Dayton and Washington dropped considerably in the second year, particularly among African Americans in Washington and highest of all in the highly volatile Washington grades 6-8. How was this drop in sample size related to tankers and leapers? Did more "divers" among voucher

students not show up for the second round of testing? Did more tankers and other low-scoring students leave private schools at the end of the first year, as occurred in Cleveland (Metcalf et al. 1999)? If so, this could have affected the results as to gains in scores. One simple way to test for this would be to present second-year results with tankers and leapers excluded, giving readers an insight into the robustness of the results.

Erratic results. A final problem is erratic results. Big differences between first- and second-year gains in Washington, D.C. may relate to which students failed to show up for testing in the second year. Students might have failed to participate in the second-year testing either because of negative first-year experiences in private schools, or because of disappointment with the first-year testing result. For such students, the probability is higher that they would do badly again than that they would do well. If they leave the sample, that alone could drive up the second-year result.

Chapter 2

Do vouchers improve failing public schools?

In the latest round of voucher advocacy research, Jay Greene, another member of the Peterson group, recently announced that the threat of vouchers in Florida for students in "failing" public schools caused math and writing gains among Florida's lowest-performing schools to increase significantly more than the gains of higher-performing schools (Greene 2001a). The finding was widely publicized as "proving" that vouchers were an effective policy tool for improving education.

In 1999, Florida adopted the A+ accountability system, which included a provision that awarded vouchers to students in schools that "failed" repeatedly. Florida grades schools as A, B, C, D, or F, based on the average scores students achieve on the Florida Comprehensive Assessment Test (FCAT). If a school receives Fs two out of four years, it becomes eligible for some form of corrective action, including but not limited to the offer of vouchers to its students to attend other schools, public or private. In the 1999-2000 school year, two Pensacola schools met the failing criteria and lost 53 children to private schools and 85 children to other public schools.

Greene used results on reading, math, and writing tests by school for the years 1998-99 and 1999-2000 to test the notion that "performance of students on academic tests improves when public schools are faced with the prospect that their students will receive vouchers" (Greene 2001a, 2). He finds that all 78 schools that received an F grade in 1999 (66 primary schools, seven middle schools, and four high schools) received a higher grade in 2000. The gains by F schools were also much higher than those for schools ranked A-D. To get the "voucher effect," Greene compares schools that "were probably very much alike in many respects" (Greene 2001a, 7), namely higher-scoring F schools and lower-scoring D schools. [12] The only thing that differentiates these two types of schools, according to Greene, is that the F's have the threat of vouchers hanging over them, and the D's do not. He concludes from this comparison that the higher-scoring F schools did significantly better on the math and writing tests, with "effect sizes" (the difference in high F and low D scores compared to the standard devia-

tion of the scores in the sample of schools being compared) of 0.12 for reading, 0.30 for math, and 0.41 for writing. This difference, he claims, is the effect that can be assigned to the voucher threat.

Part of this difference, Greene recognizes, may be due to an effect known as "regression to the mean." We would expect that individuals or groups of individuals scoring particularly low in one year would score higher in the next year, not because of any action taken but because of simple variation in performance. Similarly, high scorers in any year are likely to score lower in the following year. Baseball batting averages are a good example of this phenomenon. Players who have been in the majors for several years and had bad years in 1999 will, on average, have higher batting averages in 2000, not necessarily because of the threat of being sent to the minors (even though that threat exists), but because the normal variation associated with batting over a whole season makes it likely that hitters doing badly in a given year will do better rather than worse. The opposite is true for players who had particularly good years. Greene checks for this phenomenon by comparing gains of higher-scoring F schools with lower-scoring F schools. He finds that in reading and math, the higher-scoring F schools have higher gains than lower-scoring F schools. On these grounds, he dismisses the regression to the mean effect.

The Greene analysis has major defects that fall into two categories. First, his statistical analysis tends to overestimate the effect of being designated an F school. His interpretation of the size effects is also probably too large. Second, whatever the correct test score gain caused by a school getting an F grade, Greene presents no evidence that this should be attributed to the threat of vouchers. Florida's school-grading program is relatively new, but was in effect before 1999, the year the voucher threat was first used. How much larger was the effect in that year compared to previous years, when an F designation carried stigma but no voucher threat? Greene does not answer this question, but the data are available to do so. Other states, such as Texas and North Carolina, also have "scarlet letter" designations that trigger sanctions but not vouchers. Do F-graded schools in those states make larger gains than D schools?

Mis-estimating the 'scarlet letter' effect. Gregory Camilli and Katrina Bulkley (2001), professors at Rutgers University, re-estimated the differential gains of F-designated schools, using the same database available to Greene. They chose to compare *all* students who took the test rather than just the "standard" student population (which excludes "special" categories of students) used by Greene. There is no reason to believe that this changes the estimated gains, although Greene argues that F schools are

likely to have more special students when all students are included, dampening their estimated gains compared to gains when only standard students are included. However, a deeper problem exists: FCAT may have been given to two different populations of students in 1998-99 and 1999-2000.[13] In the earlier year, *all* students in school at the time of the test (in early spring) were given the test. In 1999-2000, only those students who were enrolled in that particular school in October took the FCAT. This would tend to increase test score gains because students who have been in school all year are likely to do better on the test than are students who changed schools during the year. This is particularly true for low-income students who are more likely to attend low-scoring schools (Rumberger 1996). This selection bias alone would cast doubt on Greene's results.

Yet, Camilli and Bulkley raise three other objections to Greene's statistical treatment of the data. They argue that Greene inadequately corrects for regression to the mean; that he aggregates data across primary, middle, and high schools; and that he overestimates size effects by inappropriately using the standard deviation of school mean test scores as the reference variable instead of the much larger variation of student test scores. In Greene's answer to Camilli and Bulkley, he rejects all three criticisms. Who is right?

Greene interprets the regression-to-the-mean problem as a floor effect (when scores are very low, they can only go up). However, as Camilli and Bulkley correctly point out, regression to the mean is not mainly the result of a floor effect but of "noise"—unexplainable variation that tends to raise low scores and reduce high scores in any give year toward the mean the next year. Camilli and Bulkley correct for this effect by (1) estimating the predicted school test score in year 2 as a function of school test score in year 1, (2) calculating the gain between second-year score and predicted second-year score, and (3) estimating the relation between this gain and school designations (in A, B, C, D, and F categories). They argue that this gives a "truer" estimate of the effect of school designation, corrected for gains associated with regression to the mean. Greene claims that their form of estimation "overcorrects" for the potential voucher effect because it "takes away" part of the gain of F-designated schools that could have come as a response to vouchers. He would be right if there were little noise in school test scores in a given year, so that all or almost all of the variation were due to policy effects such as school designation. But given what we know about school test score variation from year to year, this is unlikely.

Since Greene sees the regression-to-the-mean problem as a floor effect, he tries to test for it by comparing the gains of low-scoring F-designated schools with high-scoring F-designated schools. He finds no signifi-

cant differences in reading and math score gains between those schools, but significantly larger gains in writing for low-scoring F schools. He concludes that there is no regression-to-the-mean effect and that the F designation itself (the voucher threat) is the main explanation for the larger gain of low-scoring schools. But Haggai Kupermintz, a statistician at the University of Colorado, shows that low-scoring schools within all groups (A, B, C, D, E, and F) make larger gains than high-scoring schools in reading and math, and, that in math and especially writing, low-scoring F schools make much larger gains than high-scoring F schools (Kupermintz 2001, Fig. 2). Yet, even correcting for regression to the mean, Camilli and Bulkley and Kupermintz find a significant effect on math and writing achievement gain associated with the F designation. The estimated effect is not much different from the one Greene estimates when he compares the gains of high-scoring F-designated schools and low-scoring D-designated schools—what he calls the "hard" voucher effect.

The question of whether or not it is proper to divide the analysis into levels of schooling depends on what one wants to know. Camilli and Bulkley's analysis tells us that the big gains in both reading and math are in the seven middle schools that received an F designation. Middle schools made smaller relative gains in writing. High schools that received F's actually went down relative to other schools in math and reading and made about the same gains in writing, and primary schools with F designations made large relative gains in writing, smaller relative gains in math, and no relative gains in reading. Thus, the underlying results by school level and academic area suggest that an F designation has, at best, an uneven impact across school levels and subject areas. Greene's analysis aggregates these results, gaining some statistical significance but losing interesting and important information.

Should the size effects be measured in terms of the variance of scores among schools or among individual students? This depends on the kinds of comparisons being made. If we want to compare effect sizes of gains on different tests within the sample of Florida schools, Greene's use of test score variance among schools is a valid reference for effect size. But that is not the comparison Greene makes. He argues that educational researchers consider "effect sizes of 0.1 to 0.2 standard deviations to be small, effects of 0.3 to 0.4 standard deviations as moderate, and gains of 0.5 or more standard deviations are thought of as large" (Greene 2001a, 8). These are effect sizes based on gains compared to the variation of *individual* student achievement scores, which are much larger than the variation of average school scores. If the gains associated with receiving an F designation are to be compared with, for example, the effect size of the Tennessee class size

reduction experiment (as Greene does), he is claiming an effect on *students* in F schools of a net six-point gain in math scores (Camilli and Bulkley estimate a smaller gain). This gain, to be comparable to the effect sizes of educational research on individual student gains, should be compared to the standard deviation of individual student scores in the FCAT, not of school scores. This makes the effect size of the math gains in F-designated schools quite small (less than 0.1 standard deviations) and the writing gains moderate (about 0.23), as noted in the Camilli and Bulkley critique.

Incorrectly attributing gains at F schools to the threat of vouchers. Throughout Greene's analysis, he claims that the higher gains of F schools (corrected for regression to the mean) must be the result of voucher threat. However, F schools may also tend to raise their test scores more than other schools in other situations where there is no voucher threat. Being branded an F school may itself carry sufficient stigma to cause F schools to raise their test scores, whether or not there is a voucher threat. If that is the case, we (and Greene) have no way of knowing whether vouchers were the cause of higher scores in Florida in 1999-2000.

One way to test that hypothesis is to estimate the net effect of the F designation in Florida in the years before the voucher threat. Another is to make similar estimates for other states that rate schools as failing. Although Greene did not correct properly in his analysis for regression to the mean, it is worth comparing the relative performance of F schools to non-failing schools in these other situations using Greene's flawed methodology. This simulates what Greene might have found were he to do his (flawed) statistical analysis in Florida before the A+ plan was implemented, or in Texas or North Carolina, which have no statewide voucher program.

Data are available on school performance in Florida beginning in 1996-97, a year after the state implemented a testing program and categorized schools by the proportion of students passing the state tests. Doug Harris has analyzed these data and compared them to Greene's results for the first year of the A+ program (see Appendix A for the complete Harris study). Harris also compares his and Greene's results using Camilli and Bulkley's corrections for potential regression to the mean.

Table 3 presents Greene's and Harris' results using the regression to the mean using Greene's method of comparing the gains of high-scoring F schools to the gains of low-scoring D schools. Harris uses 1997-98 gains (pre-voucher), and Greene uses 1999-2000 gains (post-voucher). The gains in math at F schools before vouchers were introduced were larger than in the post-voucher years, but gains in reading and writing are larger post-voucher. In **Table 4** Harris adjusts both his and Greene's estimates for re-

TABLE 3 Adjusting for regression to the mean, Greene's approach (effect size)

	Harris: Florida ratings (1997-98)			Greene: Florida A+ (1999-2000)		
	Reading	Math	Writing	Reading	Math	Writing
Low 2 (D)	0.097	0.142	0.817	0.184	0.259	0.694
High 1 (F)	0.053	0.374	0.856	0.222	0.346	0.882
Difference	-0.044	+0.132	+0.039	+0.038	+0.087	+0.186

Source: Harris, Appendix A.

TABLE 4 Adjusting for regression to the mean, Camilli and Bulkley approach

	Harris: Florida grading			Greene: Florida A+		
	Reading	Math	Writing	Reading	Math	Writing
4 (A+B)	+0.012	+0.023	+0.041	+0.029	+0.005	NA
3 (C)	- 0.009	- 0.026	- 0.041	-0.001	- 0.000	NA
2 (D)	- 0.031	- 0.082	- 0.151	-0.001	- 0.025	NA
1 (F)	- 0.040	+0.199	- 0.091	+0.021	+0.062	NA

Source: Harris, Appendix A.

TABLE 5 Isolating the effect of a low-performance rating, Texas

	Gains in reading	Number	Gains in math	Number
Lower-scoring 'acceptable'	1.4212	1,338	2.1548	560
Higher-scoring 'low performing'	1.5417	45	3.8719	62
Low-performance effect	0.1205		1.1717	
Low-performance effect measured in standard deviations	0.018		0.184	
Low-performance effect in Florida measured in standard deviations	0.12		0.30	

Note: Gains are measured in terms of the Texas Learning Index, set at 70 for the minimum-expected learning level in each grade.

Source: Brownson; see Appendix B.

gression to the mean using the Camilli and Bulkley approach. Again, the math gains were much larger for F schools before the voucher threat, and reading gains were slightly larger relative to other schools once vouchers were introduced. If vouchers were the reason that F-designated schools did so much better in 1999-2000, as Greene claims, it is difficult to understand why F schools made larger (and significant) relative gains in math without any voucher sanction and why reading skills made insignificant gains even with a voucher sanction.

Amanda Brownson of the University of Texas' Dana Center duplicated the Greene analysis for Texas schools (see Appendix B). She compared the gains on the Texas Assessment of Academic Skills (TAAS) between the academic years 1996-97 and 1997-98 and the years 1998-99 and 1999-2000 for "low-performing" designated schools (this corresponds to the F designation in Florida) with "exemplary," "recognized," and "acceptable" schools' gains. Brownson argues that, because the Texas assessment is older than Florida's, schools in Texas have been making gains over a longer period than in Florida. She shows that TAAS scores increased steadily throughout the 1990s, but tailed off by 1999-2000. For both sets of comparisons of gains in scores, she found similar results for Texas as Greene found in Florida, even though Texas has no voucher threat.[14]

When Brownson compares the gains in lower-scoring "acceptable" schools with higher-scoring low-performance schools in 1998-99/1999-2000, she finds a very small net effect for low-performing schools in reading gains but a larger, statistically significant effect in math score gains. In both tests, the effect sizes are smaller than in Florida (here it is appropriate to compare relative gains to the standard deviation in mean scores among schools in order to correspond to the Greene results). These results are shown in **Table 5**. (Note that Brownson divides the test score gains by the standard deviation based on Greene's approach. Therefore, Tables 5 and 6 below should not be compared with Tables 3 and 4 above.)

Table 6 shows the same comparison for the early years. The difference in higher-scoring low-performance schools and lower-performing acceptable schools is statistically significant for both reading and math scores, and the effect sizes are close to the effect sizes estimated by Greene for Florida. Brownson's argument that gains were likely to be larger for low-performing schools in earlier years is borne out. More important, her estimates show that, using Greene's methodology, the level of relative gains made by failing schools in Florida is also made by failing schools in Texas, *although Texas uses no voucher threat.* Brownson also corrected her estimates for regression to the mean, using the Camilli-Bulkley method. This correction strengthens her conclusion that failing schools in Texas made

TABLE 6 Isolating the effect of a low-performance rating, 1996-97

	Gains in reading	Number	Gains in math	Number
Lower-scoring 'acceptable'	2.5565	1,714	3.3672	1,750
Higher-scoring 'low performing'	3.1990	55	4.6690	55
Low-performance effect	0.6425		1.3018	
The Texas low-performance effect measured in standard deviations	0.115		0.243	
The Florida voucher effect measured in standard deviations	0.12		0.30	

Note: Gains are measured in terms of the Texas Learning Index, set at 70 for the minimum-expected learning level in each grade.

Source: Brownson; see Appendix B.

relative gains as large or larger than the gains Greene attributes to Florida's voucher plan.[15] Like Florida and Texas, North Carolina has a "strong" accountability system that sanctions schools for continued "low performance." In North Carolina, the four main categories of schools are "exemplary," "meets expectations," "no recognition," and "low performing." Sanctions in North Carolina also do *not* include the threat of vouchers. Duke University's Helen Ladd, a well-known public policy analyst who has written about accountability (Ladd 1996) and about choice in New Zealand (Fiske and Ladd 2000), duplicated Greene's analysis for North Carolina (see Appendix C). She grouped schools into the four categories used by the state to characterize school performance over time[16] and examined two different measures of changes in student performance by school from 1997 to 1998: the change in the percent of students at or above grade level in each year of school based on reading, math, and writing scores (performance composite), and the gain in test scores minus the expected gain in test scores in each year (growth composite). These are the two standards by which schools in North Carolina are judged.

Ladd found that, for both measures of change, "low-performing" schools had a significantly greater positive change than any other school type (**Tables 7A** and **7B**). She also duplicated Greene's Florida comparison of high-scoring "low-performing" schools with low-scoring "no-recognition" schools. This comparison in North Carolina shows that high-scoring low-performing schools had a higher gain on both measures of gain than did no-recognition schools, and this difference was statistically significant (**Tables 8A** and **8B**). Ladd concludes that in North Carolina low-perform-

TABLE 7A Changes in performance composite, North Carolina, 1997-98

1997 school evaluation	Avg. difference in composite	Number of schools
Exemplary	2.37	514
Expected growth	3.43	380
No recognition	4.94	551
Low performing	10.68	114

TABLE 7B Changes in growth composite, North Carolina, 1997-98

1997 school evaluation	Avg. difference in composite	Number of schools
Exemplary	2.11	514
Expected growth	4.66	380
No recognition	7.48	551
Low performing	10.98	114

The change for low-performing schools compared to schools with higher evaluations is statistically significant at $p<.0001$.

Source: Ladd and Glennie, Appendix C.

TABLE 8A Changes in performance composite, North Carolina, 1997-98, comparing 'most similar' schools

1997 school evaluation	Avg. difference in composite	Number of schools
Lower-scoring no recognition	5.91	272
Higher-scoring low performing	9.22	57

TABLE 8B Changes in growth composite, North Carolina, 1997-98, comparing 'most similar' schools

1997 school evaluation	Avg. difference in composite	Number of schools
Lower-scoring no recognition	7.24	272
Higher-scoring low performing	9.46	57

The difference between the gains for the two school types is statistically significant at $p<.0001$.

Source: Ladd and Glennie, Appendix C.

ing schools made significantly larger achievement gains than other school categories, *with no voucher threat.*

Thus, Greene's claim that vouchers caused the observed gains in Florida may or may not be true, but the evidence he presents is not sufficient to support his case. We observe that schools designated as failing in Florida before the A+ voucher plan was implemented showed even larger gains in math scores. We also observe similarly large test score gains for schools designated as failing in states without vouchers. This suggests that using larger gains to Florida's F schools in 1999-2000 as evidence that a voucher threat improves low-performing public schools is at best a stretch.

In sum, Greene's statistical estimates have problems, but, more important, he should have been much more careful in attributing the larger gains he found for schools designated as failing to the voucher threat component of the Florida A+ plan. In studies that duplicate Greene's study for Florida in years before the A+ voucher plan was implemented, and in studies of Texas and North Carolina—states that publicize school "failure" but do not use a voucher threat—F-designated schools consistently make larger gains. Given these results, we have no reason to believe that the larger rise in student performance in low-performing schools in Florida in 1999-2000 was due to the threat of vouchers.

Parenthetically, we also cannot argue that the larger test score gains in failing schools in Texas and North Carolina were the result of those states' accountability systems. To prove any of these cases, we would have to show that individual schools' test scores behaved in a particular way over time and then changed significantly when the voucher threat or (in states without vouchers) other policies designed to change their behavior appeared.

Chapter 3

What have we learned?

Voucher evaluations in the U.S. have now gone through several phases. They all analyze voucher experiments aimed at low-income children, mostly African Americans, in medium and large cities, and now, with the Florida study, in states. The first experiments were publicly financed and offered vouchers by lottery. But vouchers were generally *under*-subscribed, and those applicants who did not receive vouchers were not necessarily randomly rejected nor carefully followed up. Thus, evaluations in the first phase generally chose as a "control" group students in public schools who resembled voucher students socioeconomically. These first-phase evaluations were also carried out using similar data by various researchers, some voucher advocates and others not.

The second plans were privately financed, and evaluations were designed to conform to medical experiments, where treatment and control groups are selected randomly. Vouchers were offered to a large group of low-income applicants by lottery. Vouchers were fully subscribed, and both voucher recipients and non-recipients were followed up with tests at the end of the first and second years of the experiment. In comparing pupils who receive vouchers to those who do not in this fashion, bias from differential motivation and socioeconomic background is allegedly eliminated.

The third round of studies have moved into the analysis of the effects of a voucher *threat* on low-performing public schools.

The results of the first round of studies in Milwaukee and Cleveland suggest that parents who receive vouchers and use them (actually send their children to private schools) are more satisfied with their schools than are parents of similar socioeconomic background whose children attend public schools. There is general agreement on that point. There is also general agreement that the vouchers enabled low-income parents who otherwise would not have been able to do so to send their children to private schools. Yet, the sample of low-income, urban parents seeking vouchers does not represent the average low-income urban parent with children in public school. Parents who file for vouchers are, for one, more dissatisfied than other parents. To determine whether private schools are more satisfying to low-income parents than public schools would require taking a random sample of all low-income parents in a particular city with children in public

and private schools and randomly re-assigning students to public and private schools. Parents whose children were assigned to private schools might still be more satisfied than before, but the differences would probably be much smaller than when only the children of dissatisfied parents are switched.

The results are less clear on the achievement effects of vouchers in Milwaukee and Cleveland. Results varied according to which researchers did the studies. The Peterson group produced the most favorable results for vouchers in each of the two cities. When all the results are compared, it appears that voucher-using (choice) students in Milwaukee probably made greater gains by their third and fourth years in private schools—at least in math—than did students in public schools. But the achievement effect was not large, and only a fraction of voucher students stayed in private schools for this long even though the voucher fully covered tuition. In Cleveland, the most reliable results suggest that, after two years, choice students who used their vouchers to attend existing (religious) private schools made greater gains in science than did non-choice students but not in other subjects. Students who used their vouchers in the commercial schools created to take advantage of the voucher plan did significantly worse compared to other students, both students in public schools and voucher users in religious private schools. As in Milwaukee, attrition rates of voucher users from private schools were large over the first and second years of the program.

The Milwaukee and Cleveland cases also indicate that small vouchers of $2,500 or less, such as in the early years of the Milwaukee experiment and in Cleveland, limit the number of low-income families that will actually use the voucher, either because the families cannot supply the extra tuition or because the number of private schools made available at that level of voucher funding is too limited. The eventually much larger Milwaukee voucher increased the number of private schools entering the market and also apparently made the voucher more attractive to low-income families. Nevertheless, we have no information on how the greater number of Milwaukee voucher students are performing compared to their counterparts in Milwaukee's public schools.

The results from the second round of voucher studies show similar satisfaction gains as in Milwaukee and Cleveland, but much larger achievement gains from using a voucher in private schools, at least for African American students. Studies were available only to one set of researchers. All students attended existing private schools, many of them religious, which makes the favorable results not inconsistent with results in Cleveland, although they were still much larger in Dayton and particularly Washington, D.C.

The authors of the second round of studies claim that their results are

better than those done in Milwaukee and Cleveland because of the truly experimental design of the evaluation. However, this strategy does not speak to other issues. The new studies suffer from uncorrected potential biases, including "disappointment effects" of families that did not receive vouchers, low participation rates in follow-up tests, the concentration of gains in particular small cohorts in the sample that the researchers do not attempt to explain, and possibly non-random declines in participation between baseline testing and the first and second years of follow-up testing.

The Peterson group cites Carolyn Hoxby's description of randomized field trials as the "gold standard" of social science research.[17] Hoxby's characterization certainly has merit. But to the extent that the Peterson group's results depend upon instrumental variables and upon weighting to correct for non-participation, they are no longer reporting results of a randomized field trial, but rather are reporting an empirical study with many of the dangers of assuming the relevance of characteristics that the "gold standard" attempts to avoid.

The Peterson group model has yet another problem. Low-income urban pupils attending private schools may do better because private schools are able to select their students. To the degree that a private school can construct a *peer* environment that is conducive to learning in ways that public schools cannot because they must take all comers, the influence of peers on student achievement may be more positive than in a public school. The ability to select students is not a feature of private education that voucher advocates care to stress, because, if this is the source of a positive private school effect, it can imply no condemnation of public schools that are unable to select students. Further, peer effects can run out quickly as private education expands in inner cities.[18] The Cleveland results showing small positive gains to voucher students who entered existing private religious schools and significant relative *declines* in test scores for students who used their vouchers in the commercial Hope Schools could be partly the result of peer effects, not just the relative quality of teaching in different types of schools.

The Peterson group could deal with this problem if it tested a random sample of students already in the private schools attended by voucher recipients, identified them by school, and estimated the peer effect (average test score of non-voucher students in each school) on the score of voucher recipients attending the school. It may not be easy to get the private schools to allow such testing, simply because they would then be subjecting themselves to evaluation. But without such information, it is difficult to understand the source of private school advantage, if such advantage even exists.

Peterson and his colleagues should have an interest in knowing whether it is peer effect or school characteristics that are producing their positive

results for voucher students attending private schools. They claim, contrary to Cecilia Rouse's results in Milwaukee, that smaller class size is not an important factor in explaining the higher gains in private schools of African American students. Indeed, in a recent paper, they were unable to explain why, in New York City, African American students realized a positive effect on test scores from attending a private school while Latino students did not. Private school characteristics as reported by parents, including class size, did not explain test score differences.[19] If peer effects are important in explaining whether students using vouchers make significant gains over their public school counterparts, then we might infer that a voucher plan is likely to benefit relatively few low-income students, i.e., mainly those who can get into existing schools with already better-performing students. Correspondingly, it should be kept in mind that small vouchers of $1,700 (New York City and Washington, D.C.) or $1,500 (President Bush's plan) would be unlikely to induce the creation of many new private schools for low-income voucher students.

Objective evaluations of the currently much larger Milwaukee voucher plan would provide additional information on the extent of voucher benefits. Since relatively large numbers of low-income students have apparently shifted from public to private schools (and some back), and voucher students attend approximately 100 different schools, it would be possible to assess differences in effects by type of school and whether gains are due to student selection or school quality.

In the study of competition effects, the voucher studies, as exemplified by Greene's Florida analysis, attempt to measure the "macro" effect of vouchers on public school improvement. These studies are more indirect than the studies of individual effects in experiments, so they require even more care in estimating voucher effects and interpreting them. Just the opposite has occurred. Jay Greene's study draws attention to Florida's voucher program, but tends to overestimate the degree to which failing schools actually do better than other types of schools and tells us little about whether the threat of vouchers actually makes low-performing public schools do better. Indeed, as the studies of relative gains made in an earlier year in Florida and by failing schools in Texas and North Carolina show, even without a voucher threat failing schools made relative gains as large or larger than in Florida in 1999-2000, when vouchers were introduced. If "failing" schools consistently make larger test score gains than higher-scoring schools in such a wide variety of situations, the fact that they did so in Florida in 1999-2000, when Florida's A+ voucher plan went into effect, is not evidence that the voucher threat was responsible for the larger gains. Much more rigorous empirical tests are required to make that case.

APPENDIX A

What caused the effects of the Florida A+ program: ratings or vouchers?

by Doug Harris, Economic Policy Institute

Criticisms of Jay Greene's analysis of the Florida A+ program fall into two categories: problems in estimating the effect size, and problems in attributing the effect size to the A+ program. This appendix focuses on the second problem. How much of the gain identified by Greene was caused by vouchers and how much was caused by the embarrassment of being labeled an "F" school?

It is often difficult to determine the effects of any individual policy. Governments often change multiple policies at the same time, making it difficult to understand which change had what impact. This same problem arises in the case of the A+ program. The state of Florida assigned labels to schools (A-F) *and* funded vouchers for schools that fell in the F category. This means that A+ is really two programs, which could have been implemented separately. Vouchers could have been funded regardless of school ratings, and the ratings could have been made without vouchers.

In November 1995, the Florida Department of Education (FDOE) released ratings to the public for each school in the state based on test scores from 1993-94 and 1994-95. Level 1 was called "critically low" and included 158 schools. The rules included *non-voucher* sanctions for schools that remained on this list for three consecutive years. However, these sanctions would only occur after three years and only after a series of hearings with the state Board of Education and appeals by the school district. Even if the district were found to be negligent through this process, the state board was not required to take action. Regardless, no sanctions were ever imposed.

Despite the apparent weakness of the sanction threat, anecdotal evidence suggests the schools worked hard to improve their scores to avoid further public embarrassment.[20] Many schools subsequently increased their ratings through test score improvement, decreasing the number of "critically low" schools from 158 to 71 in 1995-96 and to 30 in 1996-97.

In 1999, the state added a provision that students in schools designated as "failing" for two consecutive years would be offered a voucher that could be used in any other school, private or public. This was called the A+ program. At the beginning of the 1999-2000 school year, students in two schools were offered vouchers. They were chosen based on 1998 ratings and a "long history of failure," even though the two-consecutive-year provision of the A+ program

was not yet in effect. In 2000-01, no schools qualified for vouchers; all the 1999-2000 F schools managed to rise to an acceptable rating.

The analysis here attempts to compare the effects of the pre-voucher ratings with the A+ program, which includes both ratings and vouchers. If the effects of the earlier ratings were similar to the ratings-plus-vouchers included in A+, then this would provide evidence that vouchers did not add much value.

Analysis

The general method used here is simple. First, the test score gains achieved by "critically low" schools due to the ratings released in 1995 are estimated. Second, these gains are compared to those of the F schools studied by Greene and others, which also had the threat of vouchers. The methodology follows closely the analysis by Greene (2001), Camilli and Bulkley (2001), Brownson (Appendix B), and Ladd and Glennie (Appendix C).

While the general method is simple, the implementation is somewhat more complicated. One potential problem is that different schools used different tests in both reading and math during 1995-97, just before the FCAT was introduced. In these subject areas, schools were allowed to select from a menu of nationally norm-referenced tests. This is a common circumstance in education research, and it is common practice to make the results comparable by using either percentiles or effect sizes (gains adjusted based on the standard deviation). In this case, the FDOE reports the percentage of students in the school who reach the national 50th percentile on whichever test was taken. The writing scores, in contrast to reading and math, are based on a single statewide test taken by all students. In this case, the FDOE reports the percentage of students who receive a "3" or higher. Therefore, in the case of writing, adjustments are not necessary to compare scores across schools.

One possible effect of the new tests is that schools may achieve gains by "teaching to the test," i.e., teaching test-taking techniques and focusing their teaching on the kinds of material covered in the state exams. This is much more likely after 1999, since major changes were made in Florida testing at about that time. Specifically, Florida created its own test, FCAT, which replaced the nationally normed tests used previously for reading and math. Based on discussions with FDOE officials and other experts, it does not appear that any such changes occurred in the years immediately before the original rating system when into effect. Therefore, the real gains in achievement from A+ may be smaller than they appear.

A third and related issue is that the more recent data provided by the FDOE for the A+ program evaluations is based on raw scores, whereas the older data used here is necessarily reported in terms of "percentages above a set standard." Unfortunately, raw scores and percentages have different statistical properties. This assumption could prove problematic, especially if the "shape of the distribution" of student ability is different across schools. In other words, the test score average across schools can differ, but the variation around that

average cannot. This study tests for possible problems using the A+ data, which includes both raw scores and percentages. The use of these two measures does not significantly affect the results, as discussed later.

A fourth possible issue is that this analysis uses data for the years 1996-97 and 1997-98. The A+ ratings were released in 1999 based on data from the spring of that year. Greene then compares this with the following year's scores. Programs tend to have their biggest impacts in the early years when they receive the most attention. Therefore, it would be beneficial to analyze Florida's pre-voucher rating system using data from 1994-95 and 1995-96 to coincide with that program's first year. However, data for earlier years were not available. It is likely that the effects from pre-voucher ratings would be larger if the earlier data were used in the analysis.

A final issue is that the school rating system changed with the A+ program. Schools were still placed in categories, but the A+ ratings were based on higher standards than previously. This change could make the voucher effect look smaller, since the difficulty of further gains increases as students reach higher levels. For example, moving from the 20th percentile to the 30th percentile is likely to be easier than moving from the 30th to the 40th. The changes in tests and standards shifted 76 schools into the "F" (critically low) category that otherwise would have been in the "D" (level 2) category had the standards remained unchanged. These schools had already made large gains due to the pre-voucher ratings, discussed below.[21]

Results

Table A-1 indicates the gains by school rating and subject area for both the pre-voucher ratings and the A+ program. All of the analysis presented here is based on elementary school scores. The main reason for this restriction is that only one school was rated "critically low" at either the middle or high school level.

Schools not rated critically low were put into three higher categories, 2, 3, and 4. This rating system includes one fewer category than the A+ system. Therefore, to simplify the analysis, the A and B schools were combined into one group. This grouping will not affect the conclusions, since the focus of the analysis is on the low-scoring schools, not A and B schools. All numbers in this table, and throughout this appendix, are gain scores divided by the standard deviation, producing "effect sizes."[22]

Based on Table A-1 alone, it would appear that the A+ program is more effective than the pre-voucher grading system. However, this is quite misleading. Gains in test scores can be separated into three categories: (1) statistical "noise," e.g., a construction project going on near the testing site, which distracts students; (2) the average gain made by everyone, i.e., "trends," which may occur due to idiosyncrasies in the test and other factors; and (3) the unique gains of individual schools due to their own efforts, which are perhaps influenced by incentives created by state policy. The goal of the analysis is to separate the unique gains from statistical noise and trends. One way to account for

TABLE A-1 A simple comparison between the pre-voucher ratings and the A+ program (effect size)

	Harris: Florida ratings (1997-98)			Greene: Florida A+ (1999-2000)		
	Reading	Math	Writing	Reading	Math	Writing
All	-0.001 (1,486)	-0.007 (1,486)	-0.008 (1,486)	0.083 (2,486)	0.178 (2,486)	0.538 (2,486)
4 (A+B)	-0.013 (793)	-0.021 (793)	-0.085 (793)	0.052 (510)	0.143 (510)	0.473 (510)
3 (C)	0.007 (535)	0.006 (535)	0.080 (535)	0.066 (1,223)	0.169 (1,223)	0.529 (1,223)
2 (D)	0.035 (132)	0.019 (132)	0.097 (132)	0.143 (583)	0.229 (583)	0.612 (583)
1 (F)	0.045 (28)	0.374 (28)	0.198 (28)	0.251 (76)	0.367 (76)	1.024 (76)

Source: Author's analysis.

TABLE A-2 Adjusting gains for trends: the grand means (effect size)

	Harris: Florida ratings (1997-98)			Greene: Florida A+ (1999-2000)		
	Reading	Math	Writing	Reading	Math	Writing
4 (A+B)	-0.012	-0.014	-0.077	-0.031	-0.035	-0.065
3 (C)	0.006	0.013	0.088	-0.017	-0.009	-0.009
2 (D)	0.034	0.026	0.105	0.060	0.051	0.074
1 (F)	0.044	0.381	0.206	0.158	0.189	0.486

Source: Author's analysis.

trends is to subtract the sample mean (shown in the first row of Table A-1) from the four individual gains, shown by rating. This yields the "grand means" in **Table A-2**.

The results in Table A-2 imply a quite different conclusion. The gains in math, for instance, are now nearly twice as large in the pre-voucher ratings for the F schools (0.381 versus 0.189). However, more adjustments are necessary. The second source of variation, statistical noise, causes regression to the mean. Greene attempts to deal with this by comparing the low-scoring D schools with the high-scoring F schools. **Table A-3** includes this comparison.[23]

If this were the correct adjustment, the results would be ambiguous. The gains in math are considerably larger in the pre-voucher ratings compared with

TABLE A-3 Adjusting for regression to the mean, Greene's approach (effect size)

	Harris: Florida ratings (1997-98)			Greene: Florida A+ (1999-2000)		
	Reading	Math	Writing	Reading	Math	Writing
Low 2 (D)	0.097	0.142	0.817	0.184	0.259	0.694
High 1 (F)	0.053	0.374	0.856	0.222	0.346	0.882
Difference	-0.044	+0.132	+0.039	+0.038	+0.087	+0.186

Source: Author's analysis.

TABLE A-4 Adjusting for regression to the mean, Camilli and Bulkley approach

	Harris: Florida grading			Greene: Florida A+		
	Reading	Math	Writing	Reading	Math	Writing
4 (A+B)	+0.012	+0.023	+0.041	+0.029	+0.005	NA
3 (C)	- 0.009	- 0.026	- 0.041	-0.001	- 0.000	NA
2 (D)	- 0.031	- 0.082	- 0.151	-0.001	- 0.025	NA
1 (F)	- 0.040	+0.199	- 0.091	+0.021	+0.062	NA

Source: Author's analysis.

the A+ ratings. For writing and reading, the gains are larger for A+. However, the ongoing discussion among researchers since Greene's first paper seems to have produced agreement, even by Greene, that this is not the right approach. The standard approach in education statistics is used by Camilli and Bulkley, in which the gain scores are obtained by subtracting the second year score from the expected second year score. The expected score comes from a regression of the second year score on the first year score. Greene does not report these results in either his original paper or his response. Camilli and Bulkey do report these results, which are presented on the right hand side of **Table A-4**.

Greene is critical of this approach because *larger* real gains due to the program will, paradoxically, make the effect look *smaller*. Therefore, he suggests excluding the F schools from the regression. In theory, he is correct. However, it is highly unlikely that dropping 76 observations out of a total of 2,400 would have any meaningful impact on the results. This is confirmed by performing the same analysis as above but dropping the critically low schools from the regression. Only one of the effect sizes changes by more than 10% in any direction. Most of them changed by less than 5%, as expected.

All of the new analysis above for the pre-voucher ratings is based on the percentage of students above the 50th percentile in math and reading and the

percentage achieving a 3 in writing. This method is different than Greene's analysis based on raw scores. To see whether this difference has any impact on the results, the percentages were converted into raw scores using data from 1999-2000, which include both percentages and raw scores. Specifically, it is assumed that the statistical relationship between the raw scores and the percentages remained roughly the same over the five-year period, which is likely.[24] After creating the estimated raw scores for the earlier years, the results in Table A-4 were re-estimated for reading. The results were similar, implying that the use of percentages does not significantly affect the results.

Conclusion

Based on all of the results above, it does not appear that vouchers have a significant positive impact on public school performance. Using Greene's own approach,[25] the average effect size across subjects (excluding writing) for the pre-voucher ratings is 0.08 standard deviations, compared with 0.10 for the A+ program. This means that *the real effect size of vouchers is only 0.02 standard deviations, which is quite small even by Greene's own standards*. There is still some uncertainty about these numbers, due to the factors mentioned earlier. Two of these factors indicate that the real effect is even smaller than 0.02; only one factor indicates that it is larger. Given this, it is fairly clear that the effects are small, and considerably lower than Greene's original estimates.

These results should not be surprising to those familiar with how schools work. Outside incentives may be effective, but there is no reason to believe that vouchers, based on a theory of market competition, will work better than school grades or other approaches. Regardless, tens of thousands of students remain at a major disadvantage because they attend "critically low" schools. Finding incentives that improve these schools remains an important task.

APPENDIX B

A replication of Jay Greene's voucher effect study using Texas performance data

by Amanda Brownson, Charles A. Dana Center, The University of Texas at Austin

In an article published in February 2001 by the Manhattan Institute, Jay Greene (2001) examined growth in student performance in Florida schools using achievement test scores from 1999 to 2000. The purpose of the investigation was to determine whether the accountability system, coupled with sanctions in the form of vouchers enabling students to leave schools that received an F-rating for two consecutive years, is having an effect on student performance. Specifically, Greene investigated the effect of vouchers on student performance in schools that received a state-assigned rating of F, which is their lowest of five accountability ratings, to discover whether the potential for losing students who would take their vouchers to other schools resulted in improved test scores.[26]

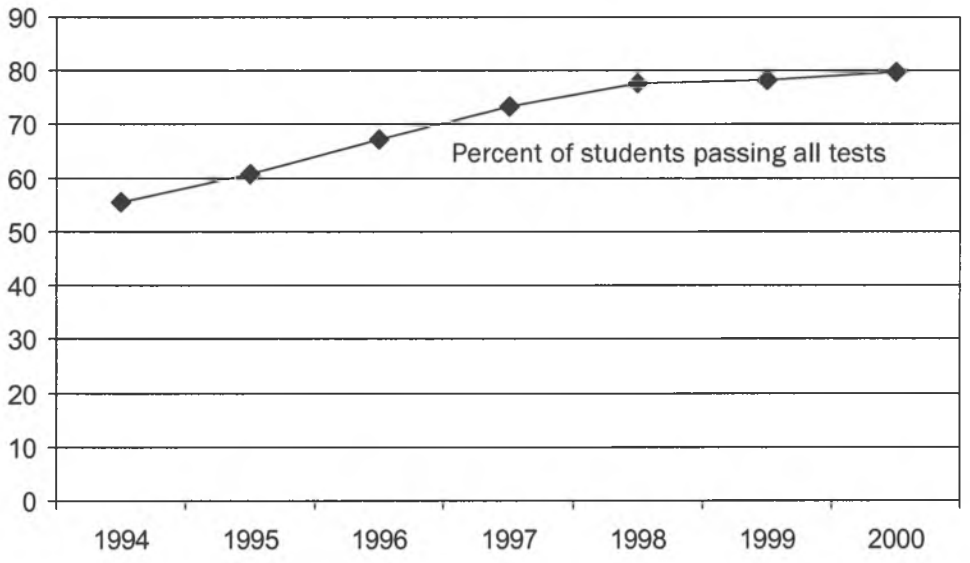
Greene found that campuses with an accountability rating of F improved at a faster rate than the higher-rated campuses in the state. From this analysis, he hypothesized that the threat of vouchers, which he says affect F-rated schools more than others, is the cause for this faster improvement. However, there a number of other possible explanations for this improvement unrelated to the threat of vouchers. Regression to the mean, which is the idea that extreme scores are likely to move to the average on subsequent measurements, has been discussed by Camilli and Bulkley as a possible explanation for this improvement.[27]

Purpose

A replication of Greene's analysis using data from Texas schools can shed light on this question. Texas, like Florida, administers statewide assessments and bases state-assigned accountability ratings on those assessments. In Texas, ratings range from "exemplary" to "low performing." These ratings have been assigned since 1993 and are based largely on student performance on the state's criterion-referenced test, the Texas Assessment of Academic Skills (TAAS).[28] However, Texas has not adopted a voucher policy.[29] Therefore, similar improvement in the "low-performing" schools in Texas cannot be attributed to a voucher threat.

This analysis uses a measure of student performance on TAAS called the Texas Learning Index (TLI).[30] The TLI scores were aggregated across grade levels for each campus in the state of Texas.[31] Data came from the Texas Education Agency's Academic Excellence Indicator System (AEIS) database and included all students who took the test.

FIGURE B-1 Percentage of Texas students passing all sections of the TAAS, 1994-2000



Source: Author's analysis.

This analysis replicates Greene's using data from two different two-year time periods. The first analysis uses data on student performance growth from the two-year period from 1999 to 2000. Growth from 1996 to 1997 is examined as well, because the accountability system in Texas has been in place longer than in Florida—the current version of Florida's system began in 1998. A quick look at overall Texas achievement scores in **Figure B-1** indicates that test scores were improving at a faster rate in the earlier years of TAAS administration than they are currently. The leveling off of the growth rate in Texas may indicate that it is now more difficult to demonstrate large gains. As a result, the 1996-97 data may more accurately reflect the current situation in Florida.

Finally, this analysis uses an approach described by Camilli and Bulkley in which second-year scores are regressed on first in order to compute an expected score, and then campus ratings are regressed on the difference between the expected scores and the actual second year scores.

TAAS improvement by accountability ratings

An examination of the Texas data reveals many of the same patterns that Greene found in Florida, despite the absence of a voucher threat in Texas. **Table B-1** indicates gains in reading and math test performance between the 1998-99 school year and the 1999-2000 school year aggregated at the campus level and measured by the TLI. These results are presented in standard deviations of the 1998-99 TLI scores.[32] A statistical test of mean differences was conducted to deter-

TABLE B-1 Mean differences in the Texas Learning Index between 1999 and 2000

	Mean difference (2000-1999)		
Performance category	Reading effect	Math effect	Number
Exemplary	0.011	0.066	1,094
Recognized	0.040	0.099	1,781
Acceptable	0.138	0.218	2,982
Low performing	0.456	0.773	78

Source: Author's analysis.

TABLE B-2 Mean differences in the Texas Learning Index between 1996 and 1997

	Mean difference (1997-1996)		
Performance category	Reading effect	Math effect	Number
Exemplary	0.062	0.085	385
Recognized	0.163	0.245	1,268
Acceptable	0.356	0.490	3,931
Low performing	0.771	1.2	99

Source: Author's analysis.

mine if the differences on the change variable were statistically significant between campuses within each rating category.

As was true in Florida, schools with lower ratings had larger gains between 1999 and 2000 than did schools with higher ratings. In Texas, low-performing schools demonstrated gains in both reading and math that were more than three times as high as the gains for schools with a rating of "acceptable." Tests for statistical significance on the reading exam revealed that there were statistically significant differences ($p<.05$) in the average change in TLI between all groups, except between "exemplary" and "recognized" campuses. There were statistically significant differences ($p<.05$) between all campus performance groups in math.

The 1996 to 1997 data (**Table B-2**) reveal similar patterns, though with larger gains for all groups.[33] Again, the low-performing campuses gained almost three times as much as the acceptable campuses during the same period, and this time there were statistically significant differences ($p<.01$) between all performance groups on both tests.

TABLE B-3 Isolating the effect of a low-performance rating using average TLI gains between 1999 and 2000

	Gains in reading	Number	Gains math	Number
Lower half of acceptable	0.213	1338	0.338	560
Upper half of low performing	0.231	45	0.608	62
Low-performance effect	0.018		0.274	
The Florida voucher effect	0.12		0.30	

Source: Author's analysis.

TABLE B-4 Isolating the effect of a low-performance rating using average TLI gains between 1996 and 1997

	Gains in reading	Number	Gains math	Number
Lower half of acceptable	0.456	1714	0.628	1750
Upper half of low performing	0.570	55	0.871	55
Low-performance effect	0.115		0.243	
The Florida voucher effect measured in standard deviations	0.120		0.30	

Source: Author's analysis.

Isolating a 'low-performance' effect

Greene attempted to isolate a "voucher effect" by comparing the higher-performing F schools to the lower-performing D schools. He defined higher-performing F schools as those that had average achievement that was above the mean for F schools, and lower-performing D schools as those that had average achievement below the mean for D schools. In an effort to replicate that analysis, higher-performing campuses rated low performing were compared to lower-performing campuses rated acceptable. This analysis was repeated on Texas data to see if low-performing campuses in Texas produced comparable gains without the threat of vouchers. Results reveal that the upper half of the distribution of campuses rated low performing grew at a faster rate than did the acceptable campuses in the bottom half of the distribution. **Table B-3** shows the growth of both lower-scoring acceptable campuses and higher-scoring low-performing campuses in Texas. Again, all growth is reported in standard deviations of 1999 scores for both the Texas data and the Florida data that Greene reports.

For these years, the Florida effects appear larger, with F-rated schools showing a growth of 0.12 standard deviations in math compared to 0.018 for

Texas, and growth of 0.3 in reading compared to 0.184 for Texas. In both Florida and Texas, the difference was significant for the math test, and not for the reading test.

A replication of this process using gains from 1996 to 1997 shows similar differences, and the Texas low-performance effect now appears to be comparable to the effect that Greene found in Florida. **Table B-4** compares the growth of low-scoring acceptable campuses and higher-scoring low-performing campuses for the 1996 to 1997 data.

For data from 1996 to 1997, the differences between the higher-performing low-performing campuses and the lower-performing acceptable campuses was statistically significant ($p<.05$) for both reading and math gains.

Greene used a second approach to account for possible regression to the mean. He regressed change scores on higher-scoring F and lower-scoring F schools, holding constant prior achievement scores. He argued that if the improvements for the lower-scoring F schools were not a great deal larger than for the higher-scoring F schools, one can safely rule out regression to the mean as the cause of the improvement in F schools and attribute it to a voucher threat. He found that both halves of the distribution of F-rated schools showed statistically significant effects and argued that regression to the mean was not the cause for this effect because the higher-scoring F schools had either larger or similar effects as the lower-scoring F-schools.

A replication of this analysis using Texas data from 1999 to 2000 indicated that, for reading, much of the improvement in the low-performing category may be due to regression to the mean. The lower half of the distribution of low-performing schools had a larger effect than the upper half of the low-performing campuses. Gains in the upper half of the low-performing campus group did not achieve statistical significance.

However, the upper half of the low-performing campus group did show statistically significant effects on the math exam, although the coefficient for the upper half was smaller than the lower half. This may indicate that, while regression to the mean accounts for some of the low-performance effect in math, there may be some other explanation as well. **Table B-5** shows the regression results for the upper half and lower half of low-performing campuses using the 1999 to 2000 data.

A replication of these analyses using 1996 and 1997 data indicate stronger evidence of a low-performance effect. For these years, both the higher and lower halves of the distribution of low-performing campuses showed statistically significant performance gains for both tests. **Table B-6** shows the regression results for the upper and lower halves of the low-performing campuses using the 1996 to 1997 data.

Camilli and Bulkley (2001) propose a different method for correcting for regression to the mean. In this approach, the difference between actual and expected scores is regressed on campus ratings, and, in Texas, low-performing

Table B-5 Regression of TLI change on low performance, 1999-2000

	Reading effect	P value	Math effect	P value
Constant	2.329	.000	2.966	.000
Prior performance	-0.027	.000	-0.040	.000
Upper half of low performing	0.018	.816	0.157	.007
Lower half of low performing	0.235	.012	0.614	.000

Source: Author's analysis.

Table B-6 Regression of TLI change on low performance, 1996-97

	Reading effect	P value	Math effect	P value
Constant	2.555	.000	3.336	.000
Prior performance	-0.028	.000	-0.038	.000
Upper half of low performing	0.243	.000	0.274	.000
Lower half of low performing	0.367	.000	0.644	.000

Source: Author's analysis.

schools gain more ground than do schools with other ratings for both exams and during both time periods; consistent with the other tests, the effect in math is larger than in reading.

As shown in **Table B-7**, for the 1999 to 2000 data, all campus ratings show statistical significance, and low-performing campuses show stronger gains than other campuses on both tests, but especially in math. Interestingly, acceptable campuses actually seem to lose some ground once the effect of regression to the mean is removed.

As indicated in **Table B-8**, for the 1996 to 1997 data there is a much more practically significant effect for low-performing campuses, which show growth of almost half of a standard deviation in math.

Summary

The above results indicate that schools rated low performing in Texas do grow at a faster rate than other schools even without the threat of vouchers, casting doubt on Greene's claim that a voucher threat was the impetus for such growth in Florida. Low-performing schools in Texas show stronger growth rates than other campuses both when using a replication of the method that Greene used and when employing a different and more commonly used correction for regression to the mean. Furthermore, when using data from 1996 to 1997, which

Table B-7 School rating effects correcting for regression to the mean, 1999 to 2000

Campus rating	Reading effect	P value	Math effect	P value
Exemplary	.125	.000	.111	.000
Recognized	.066	.000	.053	.000
Acceptable	-.042	.000	.037	.000
Low performing	.145	.016	.274	.000

Source: Author's analysis.

Table B-8 School rating effects correcting for regression to the mean, 1996 to 1997

Campus rating	Reading effect	P value	Math effect	P value
Exemplary	-.033	.181	-.039	.138
Recognized	-.038	.012	-.020	.214
Acceptable	.006	.435	.000	.997
Low performing	.280	.000	.404	.000

Source: Author's analysis.

probably more accurately reflect the current climate in Florida, the effects for low-performing campuses in Texas show gains that appear comparable to the F-rated campuses in Florida.

APPENDIX C

A replication of Jay Greene's voucher effect study using North Carolina data

by Helen F. Ladd and Elizabeth J. Glennie, Sanford Institute of Public Policy, Duke University

Jay Greene of the Manhattan Institute has recently used the large gains in student test scores in Florida's lowest-performing schools to argue that the threat of a voucher leads to school improvement. Because the lowest-performing schools (those rated F in Florida) are the only ones subject to the threat of a voucher, he attributes the larger achievement gains in those schools compared to the gains in the schools rated D to fear of the voucher program. He then refined the approach by comparing the gains in the top half of the F group of schools to those in the bottom half of the D group.

We believe that Greene has inappropriately attributed the differential gains to the voucher program rather than to the other effects of being labeled a failing school, such as shame, increased scrutiny, and possibly additional resources. To provide evidence to support this interpretation, we have replicated his study as closely as possible for North Carolina, a state that rates schools but does not have a voucher program. The logic of our approach is identical to his. The analysis differs only in that the North Carolina ABC's accountability program uses a somewhat different rating system. We base our analysis on test scores in math and reading in grades 3-8 and writing where appropriate.

In North Carolina, the four main categories of schools are exemplary, meets expectations, no recognition and low performing. We view the low-performing schools as comparable to Florida's F-rated schools. In contrast to Florida, North Carolina puts much more emphasis on the gains in scores from one year to the next in ranking the schools. In fact, it is the size of the gains relative to expected gains that essentially determine the top three categories. Each school's expected gain is based on predicted statewide gains by subject and grade level, with small and partially offsetting adjustments for regression to the mean and the initial proficiency of the students.

Exemplary schools meet their expected gains in test scores by more than 10%, and schools meeting expectations are those that have gains at least as large as the gains expected for them. No-recognition schools exhibit gains in student performance below their expected gains. Finally, low-performing schools meet neither a growth nor a performance standard. Such schools do not meet their expected growth and the percent of students at grade level falls short of the 50% performance standard.

TABLE C-1A Changes in performance composite, North Carolina, 1997-98

1997 school evaluation	Average difference in composite	Number of schools
Exemplary	2.37	514
Expected growth	3.43	380
No recognition	4.94	551
Low performing	10.68	114

The change for low-performing schools compared to schools with higher evaluations is statistically significant at p<.0001.

TABLE C-1B Changes in growth composite, North Carolina, 1997-98

1997 school evaluation	Average difference in composite	Number of schools
Exemplary	2.11	514
Expected growth	4.66	380
No recognition	7.48	551
Low performing	10.98	114

The change for low-performing schools compared to schools with higher evaluations is statistically significant at p<.0001.

Source: Authors' analysis.

Method and results

Relying on publicly available data, we first grouped schools by the classification they received from the state in 1997 (exemplary growth, expected growth, no recognition, low performing) and examined two different measures of changes in student performance from 1997 (the first year of the ABCs accountability program) to 1998: the change in the performance composite and the change in the growth composite. The performance composite is the percent of students at or above grade level in each year based on reading, math, and writing scores. The growth composite is the gain in test scores minus the expected gain in test scores in each year. Thus, the change in the performance composite is a change in levels, namely the change in the percent of the students scoring at grade level or above from one year to the next. The change in the growth composite is the difference between the gains in test scores relative to expected gains during the 1997-98 school year and the 1996-97 school year.

For both measures of change, low-performing schools had a greater positive change than any other school type, and significance tests of the difference

TABLE C-2A Changes in performance composite, North Carolina, 1997-98, comparing 'most similar' schools

1997 school evaluation	Average difference in composite	Number of schools
Lower-scoring no recognition	5.91	272
Higher-scoring low performing	9.22	57

The difference between the gains for the two school types is statistically significant at p<.0001.

TABLE C-2B Changes in growth composite, North Carolina, 1997-98, comparing 'most similar' schools

1997 school evaluation	Average difference in composite	Number of schools
Lower-scoring no recognition	7.24	272
Higher-scoring low performing	9.46	57

The difference between the gains for the two school types is statistically significant at p<.0001.

Source: Authors' analysis.

between the change for low-performing schools and that of every other school type show that these differences are statistically significant. In other words, the gains in student achievement observed in the low-performing schools differed from those in other schools by an amount that is very unlikely to have been produced by chance alone (see **Table C-1**.)

Following Greene's logic that high-performing F schools in Florida are very much like low-performing D schools in terms of incentives to improve their performance and challenges in doing so, we compared the high-scoring low-performing schools in North Carolina to low-scoring no-recognition schools. High-scoring low-performing schools are those in the top half of the 1997 performance composite score range, and low-scoring no-recognition schools are those in the bottom half of the 1997 performance composite score range. Repeating the above analysis for this subset of schools shows the same result: high-scoring low-performance schools had a higher gain did than low-scoring no-recognition schools, and this difference is statistically significant. This is true both for comparisons of gains in performance composite and growth composite. (See **Table C-2**.)

Conclusion

We conclude from this North Carolina analysis that the results that Jay Greene found for Florida probably have little or nothing to do with vouchers. If vouchers were the explanation for the gains in the F-rated schools in Florida, it is unlikely we would have found comparable patterns of gains in the low-performing schools in North Carolina.

Endnotes

1. In his latest book, *The Black-White Test Score Gap* (co-edited with Meredith Phillips, 1998), Jencks seems to argue that the biggest cause of the persistent gap is differences in family characteristics over which schools, public or private, have very little control. But he does suggest that some school improvements, like smaller classes or better-prepared teachers, might make a difference.

2. By the fourth year of the program, there were only 40 children left in the unsuccessful applicant comparison group, making their average test score highly sensitive to either high or low scores (Witte 1997).

3. Only 8,000 pupils had taken vouchers by the year 2000. Private schools not in operation when vouchers were offered had to be approved, limiting the supply of new schools. About 90 private schools, 80% religious, took voucher students in 2000-01. According to recent applications, 22 new private schools, still mainly religious, should be approved in 2001-02, enrolling another 2,000 children. Even so, the supply of schools to take advantage of a fairly large voucher ($5,300) is slow in materializing.

4. Many of the children receiving vouchers were scheduled to enter kindergarten in 1995, just as Cleveland abolished full-day kindergarten. This change could have influenced parents to take vouchers.

5. Between year 1 and year 2, 26 third graders in non-Hope Schools did not return to the program (approximately 20%). These were students who achieved significantly lower than other voucher students in the third grade even though all had statistically similar second-grade test results. All left the Cleveland school district.

6. This tuition arrangement differs from the Milwaukee experiment. Private schools in Milwaukee were not allowed to charge tuition over and above the voucher.

7. A 0.10 level of statistical significance indicates that there is a 20% probability that the gain reported is not different from zero. The reporting standard in such studies considers this to be a relatively "high" chance that the gain is indeed not different from zero.

8. These calculations use simple arithmetic means of test score gains for the three cities weighted by the number of observations in each year in each city. The Peterson group paper estimated averages using more complex weights, but it did not provide the data that would have enabled others to test or replicate the conclusions. However, the differences between the averages in Figure B and averages using the Peterson group's weights would be small.

9. Many of the students who applied for vouchers and who were initially tested were already attending private schools (up to 45% in Dayton). Up to one-third (again, in Dayton) of voucher recipients were also already attending private schools. However, the results in all the cities are reported only for those students initially in public school.

10. We would *not* expect a similar effect for children assigned to normal classes in the Tennessee experiment because there both satisfied and dissatisfied parents participated in the random draw.

11. In effect, this turns out to be similar to dividing the unbiased estimate of receiving a voucher by the probability of using the voucher. For example, the estimated reading gain for black students of receiving a voucher was 3.5 percentiles, and the gain of switching to a private school 6.1 percentiles. The ratio of 3.5 to 6.1 is 57%, approximately the probability of voucher recipients actually using the voucher.

12. In fact, the higher-scoring F schools had slightly higher average test scores from the previous year than lower-scoring D schools. This outcome can result because the state grade assigned to schools depends on the percentage of students above a certain threshold on the test score, not by the average test score for the school.

13. As Camilli and Bulkley note (in their footnote 2), this is specified under Rule 6A-1.09981 of the State Board of Education Administrative Rules.

14. In this first round of analyses, Browson does not make the overall correction for regression to the mean because she is trying to make a direct comparison to Greene's results.

15. In duplicating Greene's analysis, Brownson used the same assumptions he made regarding regression-to-the-mean effects. When she re-estimates the relative gains for failing schools in Texas using the Camilli-Bulkley method rather than the Greene method of comparing high-performing F schools to low-performing D schools, the relative gains to F schools increase for reading in both 1999-2000 and 1996-97, and for math in 1996-97.

16. According to Ladd, "[E]xemplary schools meet their expected gains in test scores by more than 10 percent and schools meeting expectations are those that have gains at least as large as the gains expected for them. No recognition schools exhibit gains in student performance below their expected gains. Finally, low-performing schools meet neither a growth nor a performance standard. Such schools do not meet their expected growth and the percent of students at grade level falls short of the 50 percent performance standard."

17. Hoxby's characterization of randomized field trials appears in Howell et al. 2000b, p. 2.

18. Good school management can also run out quickly, particularly since, as private schools expand in inner cities, they face the same fundamental problems as do public schools (see Leovy 2000).

19. One problem with this recent analysis, by Howell and Peterson, however, is that most of the positive result for African Americans in New York City is located in the single fifth/sixth-grade cohort. There might exist a particular cohort effect in that group that has little to do with private schools attended. The fifth/sixth-grade cohort of African Americans is a small proportion of the total African American sample. The difference in test scores between the rest of the African American sample attending private schools and Latinos attending private schools is probably very low to start with, so explaining an already small difference in test scores with school variable differences is unlikely to produce significant results.

20. The authors thank officials at the Florida Department of Education for this and other useful information.

21. According to documentation obtained from the Florida State Board of Education, the original low rating was given when, for two consecutive years, fewer than 33% of students had scored above the 50th percentile on reading and math, and fewer than 33% scored 3 or above on the Florida state writing test. As stated earlier, the newer standards are more stringent, but are difficult to compare because the testing instrument changed.

22. Different standard deviations are required for the 1995-97 data compared with the

1998-2000 data. In the latter case, the sample standard deviations calculated by Camilli and Bulkley are used. In the former case, it is assumed that the relationship between the cross-school standard deviation and the cross-student standard deviation is similar to that calculated by Camilli and Bulkley. Specifically, they found that the cross-student variation is 3.5 times larger in math and reading and 2.2 times larger in writing.

23. All of the regression-to-the-mean adjustments made throughout this appendix automatically adjust for trends; therefore, no additional corrections are necessary to find the real effects.

24. This was accomplished by regressing the raw scores on the percentages. The independent variables included percentages, percentages squared, and percentages cubed. The estimated equation was then applied to the data for earlier years.

25. Greene used two approaches. The one used here is the one he used in his response, subtracting the second-year score from the expected value of the second-year score, based on a regression that excludes the F schools.

26. Florida assigns campus ratings A through F based primarily on the Florida Comprehensive Assessment Test (FCAT).

27. It may also be that the publication of a low accountability rating by itself, or in conjunction with other sanctions besides vouchers, could cause such improvement.

28. See the Texas Accountability Manual at http://www.tea.state.tx.us/perfreport/account/2000/manual/ for a description of the Texas system.

29. While vouchers are not available to Texas families, the option of attending a public charter school does exist. In 1997-98, Texas had 19 charter schools. In the 1999-2000 school year, Texas had 160 charter schools serving about 35,000 students (roughly 0.7% of Texas public school enrollment). In most areas of the state, public charter schools do not offer a meaningful alternative to parents of students attending low-performing schools. First, there are relatively few charter schools from which to choose (they represent 2% of Texas public school campuses). In addition, the distinctive nature of charter schools (over half have programs specifically designed to serve students at risk of school failure and dropout) means that they are not appropriate for the typical student. Most important, student performance in Texas public charter schools is disappointingly low. In 2000, 11.4% of charter schools were rated low performing, compared with 2.1% of traditional public schools rated low performing. Another 13% of the public charter schools were rated "needs peer review"—an indication that the school has other difficulties.

30. For a description of the Texas Learning Index, see the *Technical Digest,* which can be found online at http://www.tea.state.tx.us/student.assessment/resources/techdig/index.html.

31. Camilli and Bulkley (2001) argue that aggregation across grade levels is inappropriate because diagnostic information regarding different effects in different grades may be lost, and because scales of different instruments should not be combined. This analysis aggregates across grade levels in order to replicate Greene's analysis as closely as possible, in spite of some possible limitations.

32. The standard deviation for the average reading TLI in 1999 was 6.6647. The standard deviation of the average math TLI was 6.3705

33. The standard deviation for the average reading TLI in 1996 was 5.61. The standard deviation of the average math TLI in 1996 was 5.36.

References

American Federation of Teachers (AFT). 1997. "Miracle or Mirage? Behind the Cleveland Hope Schools Voucher Students Study." Washington, D.C.: AFT.

Camilli, Gregory, and Katrina Bulkley. 2001. "Critique of an 'Evaluation of the Florida A-Plus Accountability and School Choice Program." *Educational Policy Analysis Archives* 9(7), March 4. <http://epaa.asu.edu/v9n7/>

Friedman, Milton. 1955. "The Role of Government in Education." In Robert Solo, ed., *Economics and the Public Interest*. New Brunswick, N.J.: Rutgers University Press.

Greene, Jay P. 2000. "The Effect of School Choice: An Evaluation of the Charlotte Children's Scholarship Fund." Civic Report No. 12. New York, N.Y.: Manhattan Institute for Policy Research, Center for Civic Innovation.

Greene, Jay. 2001. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." New York, N.Y.: Manhattan Institute for Policy Research, Center for Civic Innovation. <http://www.manhattan-institute.org/html/cr_aplus.htm>

Greene, Jay. 2001. "A Reply to 'Critique of an "Evaluation of the Florida A-Plus Accountability and School Choice Program,"' by Gregory Camilli and Katrina Bulkley." New York, N.Y.: Manhattan Institute for Policy Research.

Greene, Jay P., William G. Howell, and Paul E. Peterson. 1998. "Lessons From the Cleveland Scholarship Program." In Paul E. Peterson and Bryan C. Hassel, eds., *Learning From School Choice*. Washington, D.C.: Brookings Institution.

Greene, Jay P., Paul E. Peterson, and Jiangtao Du. 1996. "The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data From the Program's Evaluation." Paper prepared for the Panel on the Political Analysis of Urban School Systems, American Political Science Association, San Francisco, Calif., August 30.

Howell, William, Patrick Wolf, Paul Peterson, and David Campbell. 2000a. "Test Score Effects of School Vouchers in Dayton, Ohio, New York City, and Washington, D.C.: Evidence From Randomized Field Trials." Paper prepared for the American Political Science Association meeting, September.

Howell, William, Patrick Wolf, Paul Peterson, and David Campbell. 2000b. "The Effect of School Vouchers on Student Achievement: A Response to Critics." Cambridge, Mass.: Harvard Program on Educational Policy and Governance. <ksg.harvard.edu/pepg/>

Jencks, Christopher. 1966. "Is the Public School Obsolete?" *The Public Interest* 2, Winter, 18-27.

Jencks, Christopher, and Meredith Phillips, eds. 1998. *The Black-White Test Score Gap*. Washington, D.C.: Brookings Institution Press.

Kupermintz, Haggai. 2001. "The Effects of Vouchers on School Improvement: Another Look at the Florida Data." *Educational Policy Analysis Archives* 9(8), March 9. <http://epaa.asu.edu/v9n8/>

Leovy, Jill. 2000. "School Voucher Program Teaches Hard Lessons." *Los Angeles Times*, October 9.

Metcalf, Kim K., et. al. 1998. "Evaluation of the Cleveland Scholarship Program: Second Year Report." Bloomington: Indiana Center for Evaluation, Indiana University.

Metcalf, Kim K., et al. 1999. "Evaluation of the Cleveland Scholarship and Tutoring Grant Program, 1996-1999." Bloomington: Indiana Center for Evaluation, Indiana University.

Mosteller, Frederick. 1995. *The Tennessee Study of Class Size in the Early School Grades.* Somerville, Mass.: American Academy of Arts and Sciences.

Peterson, Paul E., Jay P. Greene, and William G. Howell. 1998. "New Findings for the Cleveland Scholarship Program: A Reanalysis of Data From the Indiana University School of Education Evaluation." Cambridge, Mass.: Harvard University Program in Education Policy and Governance. Mimeo.

Peterson, Paul E., William G. Howell, and Jay P. Greene. 1999. "An Evaluation of the Cleveland Voucher Program After Two Years." Cambridge, Mass.: Harvard University Program in Education Policy and Governance. Mimeo.

Peterson, Paul E., and William Howell. 2001. "Exploring Explanations for Ethnic Differences in Voucher Impacts on Student Test Scores." Paper presented at a conference cosponsored by the Brookings Institution and Edison Schools, "Closing the Gap: Promising Approaches to Reducing the Achievement Gap," February 1-2.

Reich, Robert B. 2000. "The Case for 'Progressive' Vouchers." *Wall Street Journal*, September 6.

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

Rouse, Cecilia. 1998b. "Schools and Student Achievement: More Evidence From the Milwaukee Parental Choice Program." *Federal Reserve Bank of New York Economic Policy Review*, 4(1):61-76. <www.ny.frb.org>

Safire, William. 2000. "Are School Vouchers the Answer?" *New York Times*, August 31.

Steele, Claude M., and Joshua Aronson. 1998. "Stereotype Threat and the Test Performance of Academically Successful African Americans." In Christopher Jencks and Meredith Phillips, eds., *The Black-White Test Score Gap.* Washington, D.C.: Brookings Institution Press.

Williams, Joe. 2000. "Choice May Draw 10,000 Students in Fall; 22 New Schools Will Join Voucher Program for Next School Year, DPI Announces." *Milwaukee Journal Sentinel,* May 16.

Witte, John F. 1993. "The Milwaukee Parental Choice Program." In M. Edith Rasell and Richard Rothstein, eds., *School Choice: Examining the Evidence.* Washington, D.C.: Economic Policy Institute.

Witte, John F. 1997. "Reply to Greene, Peterson and Du: 'The Effectiveness of School Choice in Milwaukee: A Secondary Analysis of Data From the Program's Evaluation.'" <http://hdc-www.harvard.edu/pepg/op/evaluate.htm>.

Witte, John F., Troy D. Sterr, and Christopher A. Thorn. 1995. "Fifth Year Report: Milwaukee Parental Choice Program." Madison, Wis.: Department of Political Science and the Robert M. La Follette Center for Public Affairs, University of Wisconsin.

About EPI

The Economic Policy Institute was founded in 1986 to widen the debate about policies to achieve healthy economic growth, prosperity, and opportunity.

Today, despite recent rapid growth in the U.S. economy, inequality in wealth, wages, and income remains historically high. Expanding global competition, changes in the nature of work, and rapid technological advances are altering economic reality. Yet many of our policies, attitudes, and institutions are based on assumptions that no longer reflect real world conditions.

With the support of leaders from labor, business, and the foundation world, the Institute has sponsored research and public discussion of a wide variety of topics: trade and fiscal policies; trends in wages, incomes, and prices; education; the causes of the productivity slowdown; labor market problems; rural and urban policies; inflation; state-level economic development strategies; comparative international economic performance; and studies of the overall health of the U.S. manufacturing sector and of specific key industries.

The Institute works with a growing network of innovative economists and other social science researchers in universities and research centers all over the country who are willing to go beyond the conventional wisdom in considering strategies for public policy.

Founding scholars of the Institute include Jeff Faux, EPI president; Lester Thurow, Sloan School of Management, MIT; Ray Marshall, former U.S. secretary of labor, professor at the LBJ School of Public Affairs, University of Texas; Barry Bluestone, University of Massachusetts-Boston; Robert Reich, former U.S. secretary of labor; and Robert Kuttner, author, editor of *The American Prospect,* and columnist for *Business Week* and the Washington Post Writers Group.

For additional information about the Institute, contact EPI at 1660 L Street NW, Suite 1200, Washington, D.C. 20036, (202) 775-8810, or visit www.epinet.org.

tampabay.com Know it now.

School vouchers study finds little difference between public schools, private schools

By Ron Matus, Times Staff Writer

Published Monday, June 29, 2009

Supporters often say school vouchers are lifelines to low-income students trapped in subpar public schools.

But academically, students using vouchers to attend private schools in Florida are doing no better and no worse than similar students in public schools, says a study ordered by the state Legislature.

"We consider the report a validation of what we've always said," said Mark Pudlow, a spokesman for the state teachers union. "There is no quick fix for struggling students."

Northwestern University economics professor David Figlio compared test scores of students in the voucher program, which served 23,259 students last year, to eligible public school students who opted not to participate.

Figlio said it's too early to draw hard-and-fast conclusions, and outlined some technical complications he expects to resolve with another year's worth of data.

But he said more data isn't likely to change the bottom line.

"I'm confident that it's highly unlikely that we're going to see huge differential positive test score gains from this program" or negative results either, he said after the report was released Monday. "My hunch is, when all is said and done ... it's going to be a wash in terms of test scores."

But voucher supporters said the findings prove private schools are educating voucher students as well as public schools, and for a lot less money. Per-pupil spending in Florida is about $7,000 a year. A voucher costs taxpayers $3,950.

"The fact that we're keeping (pace) and we're spending 57 cents on the dollar is a good first step," said Doug Tuthill, president of Step Up For Students, the Tampa nonprofit group that oversees the voucher program.

But critics said the findings show vouchers have failed to deliver. And to a point, one voucher researcher agreed.

Backers "promised the moon, and public policy almost never delivers the moon," said Jay Greene, a University of Arkansas professor who has studied Florida vouchers. "It doesn't mean that these programs can't be a lifeline." But don't expect to see that instantaneously.

The report comes as the tide in Florida's decade-long vouchers war seems to be turning.

In the last two legislative sessions, GOP supporters have successfully expanded the program, which offers businesses a dollar-for-dollar tax credit in return for donations. And a growing number of Democrats have signed on.

Voucher students don't have to take the Florida Comprehensive Assessment Test. But since 2006 the state has required that they take a comparable test.

Figlio compared test results from 2006-07 and 2007-08.

Vouchers are available to any student who qualifies for free or reduced lunch. But he found that students who accept them are more likely to be minorities, and be poorer. And while in public schools, they tended to be among the lowest performers. The latter finding undermines an argument by anti-voucher critics.

"Opponents have always said that you're picking off the best and brightest," said Rep. Will Weatherford, R-Wesley Chapel, who spearheaded last year's voucher expansion.

Figlio does not offer an exact comparison. The 2006-07 data was incomplete, and the applicant pool appears to have been skewed by a high number of families who applied for free or reduced lunch even though they did not qualify. Higher-income families usually correlate to higher-performing students.

With those caveats, Figlio concluded that voucher students made slightly smaller gains than similar nonvoucher students in reading and math, though the difference was not statistically significant.

Monday's report did not dissuade voucher parents, or provoucher lawmakers.

"I know my daughter," said Trinette Hicks, whose 9-year-old daughter, Tekoa, attends Southside Christian in St. Petersburg on a tax credit voucher. In private school, Hicks said, "She gets more attention."

Parents "are passionate about the value of their choice," said Sen. Don Gaetz, R-Niceville.

But the new study's results shouldn't be ignored, said Sen. Dan Gelber, D-Miami Beach, a longtime voucher opponent.

"If the kids aren't doing better, then we really ought to reconsider why we're doing it," he said.

Fla. tax credit vouchers

(2008-2009 school year)

	# of schools	# of students
Pinellas	47	572
Hillsborough	62	1,088
Pasco	19	195
Hernando	6	103
Florida	988	23,259

School vouchers: Recent findings and unanswered questions

Lisa Barrow and Cecilia Elena Rouse

Introduction and summary

Many people would argue that U.S. elementary and secondary public schools need to improve, and they would like to see U.S. students perform better in international comparisons. In addition, many people would like to do more to close the achievement gaps within the U.S. between lower-income and minority students and their counterparts. These concerns are shared by parents, employers, policymakers, and society more generally. There is less agreement on the root causes of the problems and how best to tackle them. For example, while some would argue that insufficient funding is the primary factor, others would say that this is not supported by the evidence, since per pupil spending has increased faster than achievement, as reflected in scores on the National Assessment of Educational Progress (NAEP).[1]

One strategy for improving school performance that has received a lot of attention by all those interested in education policy is increased competition. The idea is that just as competition can enhance efficiency and value in the marketplace for goods more generally, it can do the same for education. Namely, if schools must compete for students, then they will take steps to ensure that the educational experiences they offer are valued by parents and students. The primary mechanism proposed by those who favor more competition in elementary and secondary education in the public sector is an education voucher—a coupon redeemable for a maximum dollar amount per child if spent to attend a private school. In this way, voucher programs remove the monopoly power of local public schools. Instead of having to attend a neighborhood public school, students can use the voucher to attend a private school.

In the 2007–08 school year, roughly 55,000 students in three states—Florida, Ohio, and Wisconsin—and the District of Columbia were using publicly funded education vouchers to attend private schools (as well as other higher-performing public schools);[2] see table 1. Several other states have considered voucher programs, and some private organizations have helped create privately funded voucher opportunities. But are voucher programs effective? Do they improve the educational outcomes of the students who use them and do they improve the quality of the public schools? In this article, we review the existing empirical evidence on the impact of school vouchers on student achievement.[3] After reviewing the research, we conclude that expectations about the ability of vouchers to drastically improve student achievement, at least as measured by test scores, should be tempered by the results of the studies to date. That said, many questions remain. For example, no studies have examined the longer-run impact of vouchers on outcomes such as graduation rates, college enrollment, and future wages. Similarly, the research designs for identifying the potential impacts on students who remain in the public schools are far from ideal. Finally, we have little understanding of whether vouchers would represent a cost-neutral alternative to our current system of public education provision at the elementary and secondary school levels.

Lisa Barrow is a senior economist at the Federal Reserve Bank of Chicago. Cecilia Elena Rouse is director of the Education Research Section, director of the Industrial Relations Section, and Theodore A. Wells '29 Professor of Economics and Public Affairs at Princeton University; she is also a research associate at the National Bureau of Economic Research. *The authors thank Clive Belfield, David Figlio, Bhashkar Mazumder, Patricia Muller, Anna Paulson, and Jonathan Plucker for helpful comments and conversations. Emily Buchsbaum and Mitta Isley provided expert research assistance. They also thank the Ohio Department of Education and Washington Scholarship Fund for responding to their queries. Parts of this* Economic Perspectives *article are drawn from the authors' forthcoming article in the* Annual Review of Economics.

In the next section, we discuss the theoretical reasons for why education vouchers should improve student achievement followed by a discussion of the empirical approaches used for identifying the effects of vouchers. We then review the best evidence from studies of publicly and privately financed school voucher programs on the short-run impact on student achievement, discuss the evidence on the potential impact on students who remain in the public schools, consider additional voucher outcomes, and conclude.

Why should competition improve our educational system?

The idea of injecting competition into the public school system is not new; for example, Milton Friedman (1962) argued in the early 1960s for separating the financing and provision of public schooling by issuing education vouchers. The rationale behind school vouchers is that competitive markets allocate resources more efficiently than do monopolistic ones and that public schools in the U.S. have "monopoly" power because children are assigned to attend their local neighborhood school. Parents can always choose to send their children to a private school, but that means paying for schooling twice—once through property taxes (for the public schooling they are not using) and again through private school tuition. If parents had more publicly funded options for their children's schooling—once they had selected a residential location—then schools would have to compete for students. Further, more options may improve the match between the educational interests and needs of students and their schools. Importantly, schools in this model would have an incentive to improve in the areas valued by parents. Thus, if parents select schools based on their academic quality, then schools will compete for students by providing better academics; alternatively, if parents value religious education or sports, then one would expect to see schools compete to serve these interests.

There are two hypothesized ways by which increased school choice would improve student educational outcomes. The first is a "direct" effect for those students who actually exercise choice. Assuming that students would only choose to attend a school other than their neighborhood school if that school were academically *better* (or a better match), then the academic achievement of students who opt for a different school should improve relative to what their performance would have been had they stayed in the neighborhood public school. In addition, there is an "indirect," or "general equilibrium," effect on students remaining in the public schools. Competition should induce the public schools to improve in an effort to attract (or retain) students. Thus, not only should the achievement of those who choose to attend private schools increase, but so should the achievement of those who do *not* choose as well. In other words, competition should increase the efficiency of public schools. Of course, expansion of the private sector is a critical component of increasing competition. Without new school entries and/or increases in the size of current private schools, vouchers would have limited ability to increase choice.

Many empirical studies find that students in private schools have higher educational achievement levels than those in public schools.[4] The findings from such studies are presented by voucher advocates as prima facie evidence that vouchers would improve student achievement for all. Namely, voucher advocates argue that private schools outperform public schools because their existence depends on providing a good product. Educational vouchers are intended to make public schools compete in this same way; so, only schools (either public or private) providing a good product would survive. However, this literature is not conclusive because of the difficulty (described later) in identifying the impact of schools on student achievement. Not surprisingly, critics argue that the observed superiority of private schools in these studies arises because the students who attend private schools differ from the students who attend public schools rather than because private schools are more effective than public schools.[5] If the observed relative superiority of the private sector is due more to the particular background characteristics of its students than the greater effectiveness of its schools, the achievement of current public school students would not necessarily improve in private schools.

While the debate continues on whether private schools, in general, are better at educating children than are public schools, researchers have since turned to more direct evidence on the impact of vouchers by studying actual school voucher programs. We begin with an overview of the challenges of testing whether vouchers improve student outcomes before reviewing the evidence to date.

Empirical approaches to studying school vouchers

To study whether educational outcomes in the presence of vouchers are better than educational outcomes in the absence of vouchers, ideally, one would begin with a group of students and educate them for a period of time under the current public school system. At the end of the period, one would assess various

TABLE 1

Description of publicly funded voucher programs in 2007–08

Voucher program	Start of program	Number of scholarship students in 2007–08	Other information
EdChoice Scholarship Program (Ohio)	2006	6,580	Students are eligible to apply if they currently attend or will be entering an EdChoice-designated public school in the upcoming year or if they currently attend a charter school but would otherwise be assigned by the school district to an EdChoice-designated school. There is currently a maximum of 14,000 scholarships that can be awarded across the state, and there are no income requirements to receive a scholarship. If more than 14,000 students apply, then students who are renewing their scholarships followed by students who are at or below 200 percent of the poverty level receive priority. The program pays the minimum of a school's tuition or $4,375 per student from kindergarten through eighth grade and $5,150 per student from ninth through 12th grades.
Cleveland Scholarship and Tutoring Program (Ohio)	1996–97	6,017	Any student living within the boundaries of the Cleveland Metropolitan School District (CMSD) and entering kindergarten through eighth grade is eligible to apply. Low-income students are given priority, and scholarships are awarded by lottery drawings. A student is not eligible to apply once he or she has entered high school, although scholarships have been made available to program participants once they reach high school. All children currently attending Cleveland Metropolitan Schools in kindergarten through 12th grade are eligible for the tutoring program. The scholarship program pays either 75 percent or 90 percent of a school's tuition (depending on family income) not exceeding $3,450 for the 2007–08 school year.
A+ Opportunity Scholarship Program (Florida)	2000	1,305 in public schools (see next column)	Students are eligible for the A+ Opportunity Scholarship Program through the highest grade of their public school if their school is currently "failing" (graded an F) for its second year in a four-year period. The scholarships could continue into high school if the high school was assigned a grade below C. The program made an average scholarship payment of $4,206 per student in 2005–06. These scholarships (vouchers) for private schools were declared unconstitutional by the Florida Supreme Court in January 2006. Thereafter students could no longer use these vouchers to attend participating private schools; they are, however, still able to use the vouchers to attend higher-graded public schools.

outcomes for the students and administer a test, the results of which would perfectly reflect what the students know. Then one would turn back time so that the same group of students was reverted to their conditions at the beginning of the experiment. That is, the students would be the same age, have the same living conditions, and so on. This time, one would educate the students in a system with education vouchers. At the end of the same period of time, one would again assess the students' outcomes and what they know. The difference between the students' outcomes under the current and voucher systems would isolate the impact of vouchers because vouchers would be the only difference at the beginning of each experimental period.[6] If implemented on a small scale, this experiment would allow one to estimate the direct effect of vouchers; if implemented on a large scale, this would uncover the potential effects on all students, including those who remain in the public schools.

Obviously, such an evaluation is not possible. So, to study the direct effect of vouchers, researchers must rely on comparing the achievement (or other outcomes) of students who were offered a voucher (or actually used a voucher to attend a different school) with the outcomes of students who were denied a voucher, were ineligible for a voucher, or remained in the public schools for other reasons.[7] The empirical challenge is that the outcomes of the nonvoucher students may not provide a valid approximation of what

TABLE 1 (CONTINUED)

Description of publicly funded voucher programs in 2007–08

Voucher program	Start of program	Number of scholarship students in 2007–08	Other information
McKay Scholarships for Students with Disabilities Program (Florida)	1999–2000	19,439	To be eligible for this program, students must have been part of the Florida public school system for at least one year, have been counted in the prior school year's October and February enrollment surveys, and have an individual education plan. For the 19,439 students currently enrolled, there has been a total payment amount of $99,212,622.04 so far for the 2007–08 school year. There was a total payment amount of $119,092,631.54 across the state for the 2006–07 school year, assisting 18,273 students. Scholarship amounts are awarded on an individual basis.
Milwaukee Parental Choice Program (Wisconsin)	1990	18,882	Milwaukee Parental Choice Program students have family incomes at or below 175 percent of the federal poverty level ($35,843 for a family of four in 2007–08). Once a student is in the program, family income may rise to 220 percent of the federal poverty level ($45,057 for a family of four in 2007–08). The voucher was worth a maximum of $6,501 in 2007–08.
DC Opportunity Scholarship Program (District of Columbia)	2004	1,903	Eligibility requirements give preference to students in kindergarten through 12th grade who are eligible for free or reduced lunch and who are enrolled in public schools that fail to make sufficient yearly progress as defined by the No Child Left Behind Act. The program offers scholarships for up to $7,500 to cover the costs of tuition, fees, and transportation for students to attend participating DC private schools. Scholarships are renewable for up to five years as long as students remain eligible and in good academic standing at their private schools.

Sources: Ohio Department of Education, http://EdChoice.ohio.gov and www.ode.state.oh.us/gd/gd.aspx?page=2&TopicRelationID=672; Florida Department of Education, www.floridaschoolchoice.org; Wisconsin Legislative Fiscal Bureau, www.legis.state.wi.us/lfb/Informationalpapers/29.pdf; Washington Scholarship Fund, www.washingtonscholarshipfund.org; and Wolf et al. (2007).

would have happened to the voucher students had they not been offered or used a voucher. The voucher students may not have done as well (or as poorly!) as the nonvoucher students in the absence of the voucher program because their individual characteristics differ substantially. For example, the students with parents who are very educationally focused and motivated may be more likely to apply for a school voucher, and yet these students may have done better than their nonvoucher classmates even in the absence of the voucher program because of their higher level of parental support and encouragement. Unless the research design can take these individual characteristics fully into account, the estimated impact of vouchers will not likely generate the true impact of vouchers on student achievement (in statistical terms the estimated impact will be "biased"). As a result, researchers have relied on analytical strategies that attempt to control for all differences between the two groups of students, observed and unobserved.

The first strategy is to control for students' achievement (typically a test score) prior to using the school voucher in order to adjust for nonschool factors (such as having educationally focused parents or low family income) that affect this achievement and that might also be correlated with a student's likelihood of applying for a voucher. This strategy amounts to comparing the change in achievement of students before and after participation in the voucher program with the change in achievement of students who did not participate in the program. The assumption that must hold for this estimate to generate the true (unbiased) effect of vouchers is that there are no other differences that would explain changes in the test scores between the two groups of students except for the use of a voucher. Although this approach has some appeal, one might reasonably be concerned that students who were not doing well (or were doing very well) in the public schools were more interested in a different schooling experience and that prior test scores do not perfectly reflect academic "ability," achievement, or motivation.

A second, more compelling strategy to generate estimates of the effect of vouchers on student outcomes is to use a random assignment design. In this "experimental" research design, students are randomly assigned to either a "treatment group" that is offered a school

voucher or a "control group" that is not. In this case, there are no differences in the observed or unobserved individual characteristics, on average, between the two groups because the offer of a voucher was determined not by one's family's income or motivation, but purely by the flip of a coin. As a result, one need not control for other student characteristics. Properly implemented, such a strategy is viewed as the "gold standard" for estimating a causal relationship between vouchers and student outcomes. In practice, however, nonrandom differences can emerge between the treatment and control groups to the extent that the researchers conducting the study are not able to adequately follow up with both groups of students.

A simple comparison of the outcomes of students in the treatment group (those offered vouchers) and those in the control group (those not offered vouchers) will generate a true (unbiased) estimate of the impact of being offered a voucher on student outcomes—a parameter known as the "intention to treat" in the research literature. This impact reflects two factors that are important for proper evaluation of a voucher program: the rate at which students who are offered a voucher actually use one and the relative achievement of students in private schools. As such, the intention to treat has two appealing properties: It is the only unambiguously true (unbiased) estimate that one can obtain using typical statistical methods, such as ordinary least squares regression, and it reflects the overall potential gains from offering the vouchers as a policy (since those who are offered vouchers cannot be compelled to use them).

Many are also interested in whether students who actually use vouchers experience academic gains as a result—an effect known as the "treatment on the treated" among researchers. Because actual use of a voucher is not randomly determined, analysts must resort to nonexperimental methods to generate consistent estimates of the treatment-on-the-treated gains by those who actually use vouchers to attend private school. A common approach is to use an instrumental variables strategy: Whether a student was randomly offered a voucher is used as an instrumental variable for the student attending a private school. This type of analysis would generate a consistent estimate of whether the schools that the voucher students attended were more, less, or equally as effective as the schools that the nonvoucher students attended.

Evidence on the direct impact of school vouchers on students

In the U.S., two types of school voucher programs have been studied—those financed by the government (publicly funded school vouchers) and those provided by the private sector (privately funded school vouchers). From a public policy perspective, the evidence from publicly funded programs is most relevant as these programs incorporate some of the design features that might be built into a larger school voucher program, such as limitations on which students are eligible to receive a voucher and the provision or reimbursement of transportation (if any at all). That said, some of the most compelling evidence (from a methodological perspective) comes from the privately funded vouchers, so we review that evidence here as well. We begin with evidence from publicly funded programs.

We translate the estimated impacts for all of the programs into "standardized effect sizes" (σ) in order to compare estimates across studies. In particular, the estimated difference in test scores between voucher and comparison (nonvoucher) students has been divided by the standard deviation of the test score from a national sample of students. The reason for dividing by the standard deviation is to account for the fact that studies have used different tests to assess the students. The problem is that an assessment of whether a gain is "big" or "small" depends on the shape of the underlying distribution of the test. Thus, for example, a five-point gain using a test that has a narrow bell shape (a small standard deviation) implies a larger gain in student learning than does an eight-point gain using a test that has a wide bell shape (a large standard deviation). Thus, researchers "standardize" the test score gain by the spread of the distribution to account for its underlying shape and often report "effect sizes" in standard deviation units.

Once standardized, however, one must still judge whether an estimated effect size is large or small. Recently, Hill et al. (2007) attempted to review effect sizes from many studies of educational interventions. While they caution that it is only valid to compare effect sizes when using comparable populations, contexts, and interventions, as well as the outcomes being measured, they report that effect size estimates from randomized studies average 0.33σ for elementary schools, the typical grade level for the studies of voucher programs we review here.

Table 1 (pp. 4–5) briefly describes publicly funded voucher programs in the U.S. Since the launch of the Milwaukee Parental Choice Program in Wisconsin in the early 1990s, several other publicly financed voucher programs have been started, including one in Washington, DC, in which the vouchers are allocated on a randomized basis. Not only is the Milwaukee Parental Choice Program one of the oldest publicly funded voucher programs in the U.S., it has also been subject to numerous studies. The program is open to

low-income students who may use a voucher to attend any participating school (including religious schools) worth approximately $6,501 in the 2007–08 academic year. Nearly 19,000 students and 120 schools participated that academic year.

Most of the studies regarding potential achievement impacts of the Milwaukee program were conducted when the program had only been in operation for about four years and vouchers could only be used at nonreligious schools. At that time, about 12 schools and 800 students participated. Because the schools participating in the program were required to take all students who applied and to randomly select among applicants in the event of oversubscription, researchers had two potential comparison groups available—unsuccessful applicants and a random sample of low-income students from the Milwaukee Public Schools. Using both comparison groups, Rouse (1998) reports mixed results of the "direct" effect of the program: She estimates intent-to-treat effect sizes ranging from 0.06σ to 0.11σ in math and from -0.03σ to 0.03σ in reading, although the impacts in reading are never statistically different from zero, meaning that the difference may have arisen by chance.[8]

Evidence from the Cleveland Scholarship and Tutoring Program (CSTP) suggests even smaller impacts on student outcomes. The voucher program is open to all students living within the boundaries of the Cleveland Metropolitan School District, with preference given to students in low-income families.[9] Students are permitted to use the vouchers at both nonreligious and religious schools. (The tutoring program provides tutors to interested students from kindergarten through 12th grade.) The CSTP data allow researchers to identify three groups of applicants: voucher recipients who use the voucher, voucher recipients who do not use the voucher, and nonrecipients.[10] Additionally, Cleveland Metropolitan School District and test score data are available for a (nonrandom) sample of public school students.[11]

Analyzing data from the cohort of students who entered kindergarten in 1997, Belfield (2007) compares voucher winners and rejected applicants with the available sample of Cleveland public school students. He also estimates the effect of attending private school by comparing voucher users with the rejected applicants. In the third year of the program (when the cohort is in second grade), he finds that voucher winners scored significantly lower in math (-0.08σ) and lower in reading (-0.05σ) than those in the public school sample. Further, he finds that voucher users scored significantly lower in both math and reading (-0.11σ and -0.13σ, respectively) than the rejected applicants.[12] In the fifth year of the program (when the cohort was in fourth grade), the results are more mixed with estimated effects ranging from -0.08σ in math to 0.07σ in reading for the effect of using a voucher, but neither estimate is statistically different from zero.[13]

While the studies of both Milwaukee and Cleveland attempt to construct valid comparison groups and thereby identify causal impacts of the voucher programs on student outcomes, all of them rely on observational data and therefore may not fully account for preexisting differences between the voucher and comparison groups. This leads to biased estimates of the impact of vouchers. In the case of Milwaukee, the bias could be either positive (in that the students who participated in the voucher program were more motivated) or negative (in that the random sample of low-income students in the public schools was *too* advantaged relative to the voucher participants). While Rouse (1998) attempts to determine the extent of any such bias (and concludes it is likely minimal), it remains an untestable assumption. Belfield (2007) is subject to the same general research design concern.[14]

This methodological concern could, in theory, be addressed in the study of the relatively new DC Opportunity Scholarship Program (DC OSP) in Washington, DC, which is being evaluated using a random assignment program design.[15] In the first two years of the program (2004 and 2005), 2,038 eligible public school students participated in the lotteries; 1,387 of them were awarded scholarships, and the remaining 921 students became the "control group." Wolf et al. (2007) estimate that after one year, intent-to-treat effect sizes for the first two cohorts of students ranged from -0.01σ to 0.08σ in math and from -0.01σ to 0.03σ in reading. After two years, Wolf et al. (2008) report that the impacts ranged from -0.02σ to 0.01σ in math and from 0.05σ to 0.08σ in reading. Not only do these ranges include negative impacts but none of them are statistically different from zero at the 5 percent level.

Thus far, the evidence from the publicly funded voucher programs suggests, at best, mixed improvement among either students who were selected for a voucher (the intent to treat) or students who used a voucher (the treatment on the treated). The largest estimates, from the Milwaukee Parental Choice Program, suggest potential (intent-to-treat) gains of 0.11σ in math and gains of 0.14σ for those who actually use a voucher to attend a private school; most of the other estimates are much smaller or even negative. However, with the exception of the studies on the DC OSP, the studies suffer from potentially unsatisfactory comparison groups. As such, we now turn to evidence from the privately funded programs.

Although a fairly recent U.S. General Accounting Office (2002) report found 78 privately funded voucher programs to review, only a handful have been subject to any evaluation.[16] Three privately funded voucher programs—based in New York City; Dayton, Ohio; and Washington, DC—had randomized study designs, making them the best suited for rigorous evaluation. As in the publicly funded DC OSP, the privately funded programs in each city had greater numbers of applicants than vouchers available; therefore, applicants were randomly selected to receive or not to receive a voucher offer. For the New York City program (School Choice Scholarships Foundation), for example, the number of applicants was so large that the "control group" is made up of a random sample of applicants not selected to receive a voucher. We briefly describe each of the privately funded voucher programs with a randomized study design in table 2.

Across all three cities, Howell, Peterson, et al. (2002) find that using a voucher has no overall impact on student test scores. Indeed, after three years the estimated impact of attending a private school is only 0.02σ. Similarly, both Mayer et al. (2002) and Krueger and Zhu (2004) report very small impacts (at most a 0.06σ impact for using a voucher) in any year for the program in New York City, and none of the estimates are statistically distinguishable from zero.

Nevertheless, a widely publicized result from these privately funded programs is that there may have been differences across subgroups of students. Indeed, Howell, Peterson, et al. (2002) and Mayer et al. (2002) report statistically significant positive effects of private school attendance on test scores for African American students alone. Also, after three years, those African American students who used a voucher are estimated to have experienced a 0.23σ gain in achievement across the three cities; those African American students who used a voucher from the New York City program are estimated to have gained 0.26σ.[17]

However, these results are not robust. In their reanalysis of the data from the New York City program, Krueger and Zhu (2004) report that the results by race are particularly sensitive to two analytical decisions. First, Krueger and Zhu include all students, whereas Mayer et al. (2002) include baseline test scores in all of their specifications, which leads them to exclude the students who were missing baseline test score information; most of the excluded are first grade students who were not administered a baseline test. Because students were randomly chosen to receive or not to receive a voucher, baseline characteristics such as test scores should have been identical for the two groups (on average). The primary reason for including baseline characteristics would be to improve the precision of the estimates. However, Krueger and Zhu find very little difference in the precision of the estimated impact of vouchers when using the larger sample without baseline test scores. As a result, they argue that the gain in terms of statistical precision is not great enough to warrant the cost in terms of not generating estimates that are representative of the original target population.

The second substantive difference between the studies is how the researchers define a student's race. Mayer et al. (2002) identifies a student as African American if the mother's race is reported as African American and non-Hispanic (*irrespective* of the race or ethnicity of the father). Krueger and Zhu (2004) use alternative identifications. They identify a student as African American if either parent is African American and non-Hispanic; also, in their definition of African American, they include the group of students whose parents responded "other" to the survey but indicated that they (the parents) were "black" in the open-ended response. With the larger sample and the broadest identification of students as African American, Krueger and Zhu report that the estimated impact of being offered a voucher (intent-to-treat impact) for African American students falls to 0.05σ after three years and the estimated impact of using a voucher (treatment on the treated) falls to 0.03σ; neither estimate is statistically different from zero.

In sum, there is little evidence of overall improvement in test score outcomes for students offered an education voucher from privately funded voucher programs. Although there may be evidence that some subgroups of students benefit from being offered a voucher, the evidence is not robust to sensible alternative ways of constructing the analysis sample. In addition, the results of these experiments may not be valid for thinking about the average benefits of offering vouchers to all students. Namely, all participants in the experiment—both voucher recipients and nonrecipients—had expressed an interest in vouchers.

Evidence of public school response to competitive pressure

As we have emphasized, the studies discussed previously are based on relatively small voucher programs such that there was likely little competitive pressure to which the public sector would have responded. As such, the estimates are primarily of the direct effect of vouchers for those who use them. However, the true prize of a voucher system—or any significant increase in the competitive pressure experienced by the public schools—is overall improvement in the

TABLE 2

Description of privately funded voucher programs

Voucher program	Evaluation period	Number of scholarship students in program evaluation	Other information
School Choice Scholarships Foundation (New York City, NY)	1997–99	1,200	Students were eligible to apply if they were entering first through fifth grades, currently attending a public school, and qualified for the National School Lunch Program. The program began in 1997, paid up to $1,400 annually, and initially guaranteed three years of receipt. The scholarships were later extended beyond the initial three years.
Parents Advancing Choice in Education, or PACE (Dayton, OH)	1998–99	530	Students in kindergarten through 12th grade whose family income was less than 200 percent of the federal poverty line were eligible. (Students currently enrolled in private school were eligible for the program but not included in the evaluation. The evaluation focuses on students in the first through seventh grades at baseline.) At most, the scholarship was worth $1,200 or 60 percent of tuition, whichever was less. However, voucher awards were smaller for higher-income families. The program began in 1998, and while the evaluation followed students for only two years, PACE continued to award new scholarships through 2008. In 2008 the average scholarship is worth $1,800, and students are guaranteed a scholarship for at least four years.
Washington Scholarship Fund (Washington, DC)	1998–2001	1,000	Students entering kindergarten through eighth grade whose family income was less than 270 percent of the federal poverty line were eligible. (Students currently enrolled in private school were eligible for the program but not included in the evaluation. The evaluation focuses on students in the first through seventh grades at baseline.) At most, the scholarship was worth $1,700 or 60 percent of tuition, whichever was less. However, voucher awards were smaller for higher-income families. The voucher program began in 1993 and continued to offer scholarships in 2008. In 2008–09, the scholarships are worth up to $3,000 per child each year.

Sources: Howell, Wolf, et al. (2002); Parents Advancing Choice in Education, www.pacedayton.org; and Washington Scholarship Fund, www.washingtonscholarshipfund.org.

performance of the affected education system. Unfortunately, to develop a study that would generate unbiased estimates of any such systemic impacts is extremely difficult. One cannot simply compare the outcomes of students who use a voucher (or who were offered a voucher) to the outcomes of students who remained in the public schools (either by choice or from "bad luck" in a lottery) as this would likely underestimate the general equilibrium impact. The problem is that, in theory, the public schools should improve in response to the increased competition and this improvement should be reflected in the achievement of the public school students. As a result, the control (or comparison) students would not adequately represent what would have happened to the voucher students in the absence of the voucher program.

Rather, one would ideally gather a large group of education "markets" (assuming that any general equilibrium impacts remain within a market and there are no spillovers to others) and randomly assign some markets to a treatment group—in which the students would be eligible for school vouchers—and randomly assign the remaining markets to a control group—in which there would be no vouchers. After some period of time, the researcher would then compare the average outcomes of students in the voucher markets with those of students in the control markets. A simple comparison of the outcomes would yield an unbiased estimate of the general equilibrium impact of vouchers because, on average, the markets would have been similar ex ante. While such an experiment is possible in theory, in practice it would be extremely difficult to implement mostly because it would require the coordination and cooperation of so many different stakeholders. As a result, researchers have turned to other research designs to try to get an estimate of the potential impact of a large-scale voucher program.

Evidence from the expansion of the Milwaukee Parental Choice Program

After the experimental phase of the Milwaukee Parental Choice Program ended in 1995, the program was expanded to allow for a maximum of 15 percent of the public school enrollment; further, in 1998 the Wisconsin Supreme Court ruled that the vouchers could be used in religious schools. These two events led to a dramatic increase in participation in the program by both students and schools. In fact, the program was so popular that in 2006, participation was expanded to 22,500 voucher students. Researchers have attempted to analyze these last two expansions to estimate the potential impact of a large-scale voucher program on student achievement in the public sector (see Hoxby, 2003; Carnoy et al., 2007; and Chakrabarti, 2008). While some of the details differ, the basic strategy of all three studies is to attempt to identify those schools within the Milwaukee Public School District that face differing competitive pressure because of the mix of income levels among their students. (Those schools with a high proportion of low-income students who are eligible for the voucher program presumably face more competitive pressure than those with a low proportion of low-income students who are eligible.) The basic strategy of all three studies also identifies observably comparable districts elsewhere in Wisconsin in which there are no publicly funded vouchers. The following would be evidence of a positive impact of competition on school efficiency, as reflected in student test scores: Disproportionate gains among students attending schools facing competitive pressure compared with their peers at schools within Milwaukee facing relatively little pressure and at schools outside of Milwaukee (facing no pressure from vouchers).

All three studies find evidence that with the expansion of the voucher program in 1998, student performance improved in the first few years, especially in schools that were most likely to be affected by the increased competition. For example, Hoxby (2003) estimates that the fourth grade test scores of students attending schools likely facing the most competitive pressure improved by 0.12σ per year in math and by 0.07σ per year in reading relative to students attending comparison schools outside of Milwaukee.

While interesting, these results must be interpreted as being only suggestive. The identifying assumption is that there are no unobserved changes before and after the voucher program was implemented when comparing the schools with many voucher-eligible students to schools with few or no voucher-eligible students. However, within the Milwaukee Public School District, all schools were potentially affected by the vouchers. Further, outside of the Milwaukee Public School District, the demographic composition of the schools is quite different (specifically, the students are less likely to be minority and more likely to come from wealthier families) such that it is not clear researchers can adequately account for differences between the students. In addition, Carnoy et al. (2007) present some results that are not consistent with a simple interpretation that performance in the Milwaukee public schools improved because of increased competition. For example, they also find that there was little improvement after 2002 despite the fact that interest in the voucher program increased (as proxied by the number of applications). Further, they find no evidence of a general equilibrium impact when they employ other direct measures of competition (such as the number of nearby private schools or the relative number of voucher applications from a school).

Evidence from Florida's A+ Opportunity Scholarship Program

In order for a voucher program to spur improvement within the public schools, there need not be a substantial number (or proportion) of students who use a voucher to attend a private school. Rather, if public school administrators perceive there is the potential that the students will do so, they may have an incentive to improve the education in their schools. Thus, researchers have attempted to gain some insight into the potential response of public schools to increased competitive pressure a second way: by studying the schooling outcomes of students attending schools that were under the "threat" of becoming voucher-eligible—that is, schools with a high probability of their students becoming eligible to use a voucher. Researchers have done so by taking advantage of the design of Florida's school accountability system—its A+ Plan for Education. Specifically, since 1999, schools in Florida are given a grade of A through F, largely depending on the performance of the students. Schools that receive high grades and are improving receive bonuses. In contrast, low-performing schools (graded either D or F) are subject to increased administrative oversight. (These poor performers are also provided with some additional financial assistance.) In addition, if a school received an F in two out of four years and had an F in the current year, students became eligible for vouchers called Opportunity Scholarships.[18] While the other features of Florida's A+ Plan for Education remain in effect, the voucher program was declared unconstitutional by the Florida Supreme Court in January 2006. Thereafter students

could no longer use a voucher to attend a participating private school; they are, however, still able to use a voucher to attend a higher-graded public school.

Under Florida's A+ Plan for Education, school grades are determined by assigning "grade points" based on student test score performance.[19] Grades are then assigned based on whether the school is above or below the predetermined cut points for each of the letter grades. Arguably, schools earning just above the number of grade points needed to receive an overall grade of D are no different than schools receiving just below the number of grade points needed to receive a D grade. As a result, many of the schools that received an F grade are quite similar to many of those that received a D grade. Figlio and Rouse (2006), West and Peterson (2006), Rouse et al. (2007), and Chiang (2008) therefore compare student outcomes from schools earning D and F grades while controlling for the number of grade points earned so that they can recover the causal effect of the policy on educational achievement.

All of the papers find that test scores of students improve following a school's receipt of an F grade. For example, Rouse et al. (2007) and Chiang (2008) report gains ranging from 0.12σ to 0.14σ in math and about 0.10σ in reading. Further, these two studies also find evidence that the improvements persist even once the students leave the voucher-threatened school, particularly in math. In addition, Rouse et al. (2007) report finding evidence that the F-graded schools responded in educationally meaningful ways. For example, following receipt of an F grade, schools were more likely to focus on low-performing students, lengthen the amount of time devoted to instruction, and increase resources available to teachers. As such, these studies may provide some evidence that increased competitive pressure can generate some improvement in public schools.[20]

One should note, however, that the F-graded schools in Florida were also stigmatized as "failing" (one of the intents of the public announcements of the grades). So another possibility is that the stigma of being identified as a failing school (and perhaps the subsequent parental pressure to make changes) led the schools to improve. As such, one cannot strictly distinguish a "voucher effect" from a "stigma effect." That said, Figlio and Rouse (2006) indirectly assess the impact of stigma by comparing student achievement following the implementation of Florida's A+ Plan for Education—which enlisted both the threat of vouchers and stigma—with student achievement following the placement of schools on a critically low performers list in 1996, 1997, and 1998 that involved public stigma but no threat of vouchers. They estimate that the student gains in reading were nearly identical under the two regimes and were actually larger in math following placement on the critically low performers list, suggesting that the relative improvements among the low-performing schools may have been due more to stigma than to the threat of vouchers.

There is some evidence from the expansion of the Milwaukee Parental Choice Program and from the threat of vouchers created by Florida's A+ Plan for Education suggesting that the achievement of students attending schools facing increased competition improves. However, the research strategies do not allow one to definitively rule out other explanations for the improvements. As such, we conclude that the jury is still out on the potential for vouchers to spur public schools to improve.

Other potential social gains from vouchers

There may be other reasons why providing school vouchers may be appealing from a public policy standpoint. One might argue in favor of vouchers as a way to increase equity by giving poor families more opportunities to choose private schools over their neighborhood public schools. Also, based on parents' reports for the publicly funded DC voucher program (DC OSP), the schools that are chosen (private schools) may be safer. Parents of students offered a voucher reported a significantly lower level of perceived school danger than parents of students not offered a voucher.[21]

In a related fashion, student achievement may not be the only criterion by which to judge the success of voucher programs. If school choice means that parents are more satisfied with the education their children are receiving and if voucher programs are no more expensive than our current system, then a voucher program may be a cost-neutral way to increase social welfare. Importantly, one consistent finding in this literature is that voucher parents report being more satisfied with their current schooling than do nonvoucher parents. For example, in the DC OSP, parents of students offered a voucher gave their children's schools a significantly higher overall grade on a five-point scale (grades A through F) and were significantly more likely to give their children's schools a grade of A or B. Further, they reported significantly greater satisfaction with their children's schools on all aspects asked, including location, class sizes, discipline, academic quality, and the racial mix of the students (Wolf et al., 2007). These results have generally been reported for other voucher programs, such as those in New York City (Mayer et al., 2002) and Milwaukee (Witte, Sterr, and Thorn, 1995).[22]

Yet, the potential net improvement in social welfare depends on both the general equilibrium effects of vouchers and the cost advantage over current public schools—two issues that are not well understood. While small-scale voucher programs indicate that parents offered a voucher are more satisfied with their children's schools than those not offered a voucher, a large-scale voucher program might result in some parents who are more satisfied and some who are less satisfied. In order for social welfare to be increased with a cost-neutral voucher program, the gains to the parents who benefit must be large enough to outweigh the losses to parents who do not benefit.

In addition, there is not much information about whether a well-developed voucher program would, indeed, be cost-neutral. On its face an education voucher system should be no more expensive than the current system as the state (or some other public entity) would simply send a voucher check to participating schools for each participating child rather than to the local public school or district. However, if truly implemented on a large scale, there may be other, less obvious costs that would depend critically on the actual design of the program. Levin and Driver (1997) caution that, depending on a number of factors, the cost of a voucher system could actually exceed those of the current geographically based system. These factors include the transportation of children to and from school, recordkeeping, and the monitoring of student enrollment. Two additional concerns are how a program deals with students currently attending private schools and how disputes are adjudicated (particularly if there are differing voucher amounts). While Levin and Driver's estimates are rough, based on hypothetical voucher programs and crudely estimated costs, their analysis suggests, at a minimum, that we should not assume a voucher program would be cost-neutral. Further, there may be large costs associated with the transition to a voucher system that should be considered.

Finally, the studies to date necessarily focus on short-run effects of vouchers when in fact there may be longer-run impacts on high school graduation, college enrollment, or even future earnings. For example, Altonji, Elder, and Taber (2005a) study the effect of Catholic education on a variety of outcomes and find little evidence that Catholic schools raise student test scores. At the same time, their results suggest that Catholic schools increase the probability of graduating from high school and potentially the probability of enrolling in college. These longer-run effects have yet to be credibly examined in studies of school vouchers.

Conclusion

The best research to date finds relatively small achievement gains for students offered education vouchers, most of which are not statistically different from zero, meaning that those gains may have arisen by chance. Further, the very little evidence about the potential for public schools to respond to increased competitive pressure generated by vouchers also suggests that one should remain wary that large-scale improvements would result from a more comprehensive voucher system.

So why has it been so difficult for researchers to observe large improvements in student achievement with school vouchers in the U.S.? One explanation may be that schools already compete for students through residential choice such that the public sector does not operate as poorly as perceived by many. Another explanation may be that the education sector does not meet the conditions for perfect competition (Garner and Hannaway, 1982). For example, information on school quality may be costly and difficult for parents to obtain, so having more choice may generate less additional competitive pressure on schools than one would expect in a perfect information environment. Further, education is not a homogenous good. Therefore, while competition for students may make schools more responsive to parents, this may be achieved through changes in other areas of school life, such as religious education or sports, rather than academic achievement.

Despite the heretofore lackluster empirical findings, the theoretical rationale behind school vouchers remains compelling: If parents choose schools based on academic performance and if we allow them more choice, then the schools will need to improve academically in order to attract students. In addition, others have endorsed vouchers to promote greater equity: If rich families have the means to opt out of the public school system, should not poor families have a similar opportunity? It is perhaps for these reasons—combined with frustration that other approaches to improve the U.S. education system have proven weak or futile—that school vouchers remain high on the agenda for many policymakers.[23] However, expectations about the ability of vouchers to drastically improve student achievement, at least as measured by test scores, should be tempered by the results of the studies to date, and arguments for vouchers as a cost-neutral alternative should be subject to more careful analysis of the full costs.

NOTES

[1]The National Assessment of Educational Progress is the only nationally representative and continuing assessment of what students in the U.S. know and can do in various subject areas, such as mathematics and reading. The commissioner of education statistics, who heads the National Center for Education Statistics in the U.S. Department of Education, is responsible by law for carrying out the NAEP project; for further details, see http://nces.ed.gov/nationsreportcard/. See also Hoxby (2003).

[2]The Florida A+ Opportunity Scholarship Program—a publicly funded voucher program initially created for students to attend private schools—was declared unconstitutional by the Florida Supreme Court in January 2006. After this ruling, students could no longer use the voucher to attend a participating private school; they are, however, still able to use the voucher to attend a higher-rated public school. For further details, see the discussion on the Florida A+ Opportunity Scholarship Program later in the text and in table 1.

[3]A less efficient public sector and a less competitive (public schooling) environment may explain the larger impacts of school vouchers that have been estimated in other countries, such as Columbia (see, for example, Angrist et al., 2002). In the U.S., elementary and secondary public schooling has largely depended on local financing, meaning that choice between local school districts may already generate strong competitive pressure. As a result, there may be less potential for vouchers to generate large efficiency gains (see, for example, Barrow and Rouse, 2004).

[4]See, for example, Coleman, Hoffer, and Kilgore (1982a, 1982b); Evans and Schwab (1995); Neal (1997); and Altonji, Elder, and Taber (2005b).

[5]See, for example, Goldberger and Cain (1982), Cain and Goldberger (1983), and Altonji, Elder, and Taber (2005a).

[6]Of course, even if vouchers improved outcomes in the long run, there might be a transition period in which the full benefits were not realized. A more complex version of this hypothetical experiment would be needed to identify both the transitional costs and long-run effects of a voucher program.

[7]Ironically, this also means that this literature bears striking similarity to that of the differential effectiveness of private and public schools.

[8]The range reflects estimates from different model specifications. Other studies using these early data from Milwaukee include Witte (1997) and Witte, Sterr, and Thorn (1995), as well as Greene, Peterson, and Du (1999). Using only the sample of low-income students from the Milwaukee Public Schools as a comparison group, Witte (1997) and Witte, Sterr, and Thorn (1995) estimate no impact of the program on student achievement. Greene, Peterson, and Du (1999) only use the unsuccessful applicants as a comparison group and estimate a positive impact in both math and reading. See Rouse (1998) for further discussion of the differences between the studies.

[9]The voucher is also progressive in that it pays 90 percent of tuition up to $3,450 for those with family income below 200 percent of the poverty line and only 75 percent of tuition up to $3,450 for those from families earning above 200 percent of the poverty line. The original program paid tuition up to a maximum of $2,250 (Metcalf et al., 1998). The Cleveland Metropolitan School District changed its name from the Cleveland Municipal School District in 2007.

[10]The nonrecipient group potentially contains both students who did not win the voucher lottery and students not entered into the lottery due to the preference given to students from low-income families (Metcalf, 2001).

[11]The public school sample was generated by using the first grade classmates of voucher recipients who did not use their voucher, as well as the first grade classmates of program applicants who were not awarded a voucher (Metcalf, 2001).

[12]Although Belfield (2007) only reports results for the third and fifth years of the program, he notes that the results are similar for the fourth year when the cohort was in third grade.

[13]Belfield (2007) finds a statistically significant -0.06σ difference in math between voucher winners and the public school sample.

[14]In addition, Belfield (2007) includes some measures in his empirical specifications that are arguably outcomes of the voucher program, namely, class size and teacher's years of experience. That said, his results are largely similar when these controls are excluded.

[15]See Wolf et al. (2007) for more details. Students attending low-performing public schools were given a better chance of winning the lottery. Although private school students were eligible for the vouchers, they were excluded from the study.

[16]The U.S. General Accounting Office's legal name became the U.S. Government Accountability Office on July 7, 2004. For further details, see www.gao.gov/about/namechange.html.

[17]In contrast, Howell, Peterson, et al. (2002) estimate a negative impact for African American students after three years in the privately funded voucher program in Washington, DC, although the impact is not statistically different from zero. Results for the third year of the privately funded programs apply only to those in Washington, DC, and New York City because the Dayton, Ohio, program was evaluated for only two years.

[18]Currently Florida has two other voucher programs as well: an income tax credit for corporations to fund vouchers for low-income students and the McKay Scholarship for Students with Disabilities Program. Greene and Winters (2008) study the impact of the McKay Scholarships on the achievement gains of students with disabilities who remain in the public schools. Because their estimation strategy identifies the general effect of vouchers by using students whose disability status changes, the extent to which these results generalize to overall improvements in the public schools is unclear.

[19]Literally speaking, school grades were not assigned using "grade points" before 2002 when Figlio and Rouse (2006) studied the system. Nevertheless, their strategy is quite similar in spirit.

[20]A statistical issue with which all of the researchers wrestle is whether the disproportionate gains by students in the F-graded schools resulted from mean-reverting measurement error or reflected actual changes in response to Florida's A+ Plan for Education. Mean-reverting measurement error occurs when gains the year after a school scores unusually low—and is thereby labeled as F—reflect the measurement error in test scores. That is, the test scores of students might have increased in many of the F-graded schools even in the absence of Florida's education plan simply because they were transitorily low in the prior year. The reliance on a regression discontinuity design (one that compares the D-graded and

F-graded schools while also controlling for the grade points) helps to mitigate against the presence of mean-reverting measurement error, although the researchers employ other strategies as well.

[21]See Wolf et al. (2007), table H-3. While student perceptions also suggest that the chosen schools are safer on average, the difference was not statistically significant (see table H-4 of the same study).

[22]At the same time, not all parents are satisfied with the voucher schools. Focus groups among parents of DC OSP participants found that they believed a few schools misrepresented aspects of their programs and that there was a need for an evaluation of participating schools (Stewart et al., 2007). Similarly, in the early years of the Milwaukee Parental Choice Program, 43 percent of the parents who took their children out of the voucher schools cited the poor quality of the voucher school as one of the primary reasons they withdrew their children from the program. More specifically, they cited being unhappy with the staff, the education their children were receiving, and the lack of programs for special needs; they also noted that the teachers were too disciplinarian. Thirty percent cited the poor quality of the overall Milwaukee program—including hidden school fees, difficulties with transportation, and the limitation on religious instruction—as the primary reason for withdrawing their children (Witte, Sterr, and Thorn, 1995).

[23]Most recently the George W. Bush Administration proposed the strengthening of the choice provisions in the reauthorization of the federal No Child Left Behind Act, and there were (unsuccessful) ballot initiatives in California and Utah to create statewide voucher programs open to all students.

REFERENCES

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, 2005a, "An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling," *Journal of Human Resources*, Vol. 40, No. 4, Fall, pp. 791–821.

_____, 2005b, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, Vol. 113, No. 1, February, pp. 151–184.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer, 2002, "Vouchers for private schooling in Columbia: Evidence from a randomized natural experiment," *American Economic Review*, Vol. 92, No. 5, December, pp. 1535–1558.

Barrow, Lisa, and Cecilia Elena Rouse, 2004, "Using market valuation to assess public school spending," *Journal of Public Economics*, Vol. 88, Nos. 9–10, August, pp. 1747–1769.

Belfield, Clive R., 2007, "Achievement effects of the Cleveland Scholarship and Tutoring Program," City University of New York, Queens College, unpublished mimeo.

Cain, Glen G., and Arthur S. Goldberger, 1983, "Public and private schools revisited," *Sociology of Education*, Vol. 56, No. 4, October, pp. 208–218.

Carnoy, Martin, Frank Adamson, Amita Chudgar, Thomas F. Luschei, and John F. Witte, 2007, *Vouchers and Public School Performance: A Case Study of the Milwaukee Parental Choice Program*, Washington, DC: Economic Policy Institute.

Chakrabarti, Rajashri, 2008, "Can increasing private school participation and monetary loss in a voucher program affect public school performance? Evidence from Milwaukee," *Journal of Public Economics*, Vol. 92, Nos. 5–6, June, pp. 1371–1393.

Chiang, Hanley, 2008, "How accountability pressure on failing schools affects student achievement," Harvard University, mimeo, January.

Coleman, James, Thomas Hoffer, and Sally Kilgore, 1982a, *High School Achievement: Public, Catholic, and Private Schools Compared*, New York: Basic Books.

_____, 1982b, "Cognitive outcomes in public and private schools," *Sociology of Education*, Vol. 55, Nos. 2–3, April/July, pp. 65–76.

Evans, William N., and Robert M. Schwab, 1995, "Finishing high school and starting college: Do Catholic schools make a difference?," *Quarterly Journal of Economics*, Vol. 110, No. 4, November, pp. 941–974.

Figlio, David, and Cecilia Elena Rouse, 2006, "Do accountability and voucher threats improve low-performing schools?" *Journal of Public Economics*, Vol. 92, Nos. 1–2, January, pp. 239–255.

Friedman, Milton, 1962, *Capitalism and Freedom*, Chicago: University of Chicago Press.

Garner, W., and J. Hannaway, 1982, "Private schools: The client connection," in *Family Choice in Schooling: Issues and Dilemmas*, Michael E. Manley-Casimir (ed.), Toronto: Lexington Books, pp. 119–133.

Goldberger, Arthur S., and Glen G. Cain, 1982, "The causal analysis of cognitive outcomes in the Coleman, Hoffer, and Kilgore report," *Sociology of Education*, Vol. 55, Nos. 2–3, April/July, 103–122.

Greene, Jay P., Paul E. Peterson, and Jiangtao Du, 1999, "Effectiveness of school choice: The Milwaukee experiment," *Education and Urban Society*, Vol. 31, No. 2, February, pp. 190–213.

Greene, Jay P., and Marcus A. Winters, 2008, "The effect of special education vouchers on public school achievement: Evidence from Florida's McKay Scholarship Program," Manhattan Institute for Policy Research, civic report, No. 52, April, technical version available at www.manhattan-institute.org/pdf/Effect_of_Vouchers_for_SE_Students_on_Public_School_Achievement_2-19-08.pdf.

Hill, Carolyn J., Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey, 2007, "Empirical benchmarks for interpreting effect sizes in research," MDRC (Manpower Demonstration Research Corporation), working paper, July.

Howell, William G., and Paul E. Peterson with Patrick J. Wolf and David E. Campbell, 2002, *The Education Gap: Vouchers and Urban Schools*, Washington, DC: Brookings Institution Press.

Howell, William G., Patrick J. Wolf, David E. Campbell, and Paul E. Peterson, 2002, "School vouchers and academic performance: Results from three randomized field trials," *Journal of Policy Analysis and Management*, Vol. 21, No. 2, Spring, pp. 191–217.

Hoxby, Caroline M., 2003, "School choice and school productivity: Could school choice be a tide that lifts all boats?," in *The Economics of School Choice*, Caroline M. Hoxby (ed.), Chicago: University of Chicago Press, pp. 287–341.

Krueger, Alan B., and Pei Zhu, 2004, "Another look at the New York City voucher experiment," *American Behavioral Scientist*, Vol. 47, No. 5, January, pp. 658–698.

Levin, Henry M., and Cyrus E. Driver, 1997, "Cost of an educational voucher system," *Education Economics*, Vol. 5, No. 3, December, pp. 265–283.

Mayer, Daniel P., Paul E. Peterson, David E. Myers, Christina Clark Tuttle, and William G. Howell, 2002, "School choice in New York City after three years: An evaluation of the School Choice Scholarships program," Mathematica Policy Research Inc., report, No. 8404-045, February 19.

Metcalf, Kim K., 2001, "Cleveland Scholarship Program evaluation: 1998–2000 technical report," Indiana University Bloomington, Indiana Center for Evaluation, mimeo, September.

Metcalf, Kim K., William J. Boone, Frances K. Stage, Todd L. Chilton, Patty Muller, and Polly Tait, 1998, "A comparative evaluation of the Cleveland Scholarship and Tutoring Grant Program: Year one: 1996–97," Indiana University Bloomington, Indiana Center for Evaluation, report, March 16.

Neal, Derek, 1997, "The effects of Catholic secondary schooling on educational achievement," *Journal of Labor Economics*, Vol. 15, No. 1, part 1, pp. 98–123.

Rouse, Cecilia Elena, 1998, "Private school vouchers and student achievement: An evaluation of the Milwaukee Parental Choice Program," *Quarterly Journal of Economics*, Vol. 113, No. 2, May, pp. 553–602.

Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio, 2007, "Feeling the Florida heat?: How low-performing schools respond to voucher and accountability pressure," National Bureau of Economic Research, working paper, No. 13681, December.

Stewart, Thomas, Patrick J. Wolf, Stephen Q. Cornman, and Kenann McKenzie-Thompson, 2007, "Satisfied, optimistic, yet concerned: Parent voices on the third year of the DC Opportunity Scholarship Program," Georgetown University Public Policy Institute, School Choice Demonstration Project, report, No. 0702, December.

U.S. General Accounting Office, 2002, "School vouchers: Characteristics of privately funded voucher programs," report, Washington, DC, No. GAO-02-752, September 10.

West, Martin R., and Paul E. Peterson, 2006, "The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments," *Economic Journal*, Vol. 116, No. 510, March, pp. C46–C62.

Witte, John F., 1997, "Achievement effects of the Milwaukee voucher program," University of Wisconsin–Madison, Department of Political Science and Robert M. La Follette Institute of Public Affairs, mimeo, January.

Witte, John F., Troy D. Sterr, and Christopher A. Thorn, 1995, "Fifth year report: Milwaukee Parental Choice Program," University of Wisconsin–Madison, Department of Political Science and Robert M. La Follette Institute of Public Affairs, mimeo, December.

Wolf, Patrick, Babette Gutmann, Michael Puma, Brian Kisida, Lou Rizzo, and Nada Eissa, 2008, *Evaluation of the DC Opportunity Scholarship Program: Impacts after Two Years*, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, report, No. NCEE 2008-4023, June.

Wolf, Patrick, Babette Gutmann, Michael Puma, Lou Rizzo, Nada Eissa, and Marsha Silverberg, 2007, *Evaluation of the DC Opportunity Scholarship Program: Impacts after One Year*, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, report, No. NCEE 2007-4009, June.

Voucher study finds parity

Students achieve about as well as those at MPS

By Alan J. Borsuk, Journal Sentinel, Inc
Feb. 26, 2008

The first full-force examination since 1995 of Milwaukee's groundbreaking school voucher program has found that students attending private schools through the program aren't doing much better or worse than students in Milwaukee Public Schools.

The study, released Monday in Madison, is the first from a five-year project aimed at providing a comprehensive evaluation of the voucher program, which this year is allowing more than 18,000 Milwaukee children from low-income families to attend private schools, 80% of them religious schools.

The authors caution repeatedly that stronger conclusions will come only when trends over several years can be examined, and not much should be read into this year's results.

But the early findings, based on examining standardized test results for voucher students and comparing them to those of a matched set of MPS students, are unlikely to be seen as good news by advocates of the program that was launched in 1990 with hopes of being a powerful step to increase educational success among the city's children.

The Milwaukee program is the largest, oldest and arguably most significant school voucher effort in the United States. As Patrick J. Wolf, the lead researcher in the project, wrote, "When one thinks of school choice, one thinks of Milwaukee."

"We have displayed a rough and limited snapshot of the average performance of Choice (Milwaukee Parental Choice Program) students in certain grades that suggests they tend to perform below national averages but at levels roughly comparable to similarly income-disadvantaged students in MPS," Wolf, a professor at the University of Arkansas, concluded.

At one point in the reports, researchers use the phrase "relative parity" in describing the small differences between the performance of MPS students and voucher students.

They say there is little evidence that voucher schools are "skimming the cream" by taking the best students from MPS, as some critics have claimed. What they conclude is that the performance of both MPS and voucher students is fairly typical for low-income students nationally, pointing at the broader American dilemma of how to achieve widespread educational success among poor children, minority

children and children from homes where there is little history of educational success.

Apples-to-apples effort

The researchers' conclusions are based on test results from the 2006-'07 school year, when they gave a sample of voucher students the same tests given to public school students in Wisconsin and compared the results to those of a scientifically matched group of MPS students.

• Overall, they found, fourth-grade voucher students scored "somewhat lower" than MPS students but eighth-grade voucher students scored "somewhat higher."

• At all grades, both MPS and voucher students had overall test scores well below the 50th percentile nationally, and generally around the 33rd percentile, meaning they were generally scoring lower than two-thirds of students.

Results for individual voucher schools were not released as part of the study, despite calls from several legislators and others to see the private school results.

The study was conducted by the School Choice Demonstration Project, part of the Department of Education Reform at the University of Arkansas. The main researchers included John Witte, a University of Wisconsin-Madison professor who conducted studies of the Milwaukee voucher program from 1990 to 1995, before the Legislature dropped the requirement for such studies.

Since Witte's last study, the program has grown enormously, but there has been a minimal amount of research on its effectiveness.

The program provides up to $6,501 per student to private schools in the city. State officials expect about $120 million in voucher payments to be made in this school year.

As part of a deal in 2006 between Democratic Gov. Jim Doyle and Republican legislative leaders, the voucher program was allowed to grow to as many as 22,500 students, but the private schools were required for the first time to administer nationally accepted standardized tests, and the School Choice Demonstration Project was authorized to launch its study.

In his summary, Wolf calls the research "the most comprehensive evaluation of a school choice program ever attempted."

Some surprises

Researchers released the first year's analysis in the form of four reports dealing with finances of the voucher program, characteristics of the schools involved, student performance, and parent and student opinion of both MPS and voucher schools.

Some of the findings confirm assumptions about the program - for example, that religion is a major reason why parents enroll children in private schools. Other findings are more surprising - for example, that MPS parents who were surveyed were more likely than voucher parents to help their children with homework, and that teachers in voucher schools had more experience on average than MPS teachers.

One trend the researchers found is that the variation in scores on tests among MPS schools tended to be much narrower than the variation among private schools. In other words, it could be that the range of quality among voucher schools is much wider - from very weak to outstanding - while the range of schools in MPS tends to stick closer to the system averages. That would square with observations of classes in both MPS and the voucher schools made by education reporters for the Journal Sentinel, especially in a major reporting project on the voucher program in 2005.

The researchers emphasize that the results are a snapshot that does not address many possible factors behind the differences. "It would be a mistake for readers to draw conclusions about the effectiveness of the MPCP based on these simple annual descriptive statistics," the report says.

How families compare

Surveys of students and parents in MPS and the voucher schools found that voucher and MPS students are about equally likely to be living with both parents (38% and 36% respectively).

Voucher parents are notably more likely to be involved in activities at their children's schools, such as doing volunteer work or taking part in parent/teacher organizations, but MPS parents are more likely than voucher parents to help their children with homework or read books with them, the study found.

Both MPS and voucher parents express high levels of satisfaction with their children's schools, but voucher parents are much more likely to say they are "highly satisfied," while MPS parents are more likely to pick the "satisfied" response.

Witte, one of the most experienced school choice researchers in the U.S., said the high level of parental satisfaction with the voucher schools is an important aspect of answering the question of whether the program is successful.

The researchers found that voucher schools are, on average, much smaller than those in MPS overall and have smaller class sizes, a factor that clearly appealed to the parents who were surveyed.

Of 120 schools examined by the researchers, 95 identified themselves as religious and seven were classified as non-religious but operating within a religious tradition. Thirty-six of the schools were Catholic, 26 Lutheran and 22 were from other Christian denominations.

The reports say that 43% of MPS teachers have a master's degree or higher, compared with 29% in the voucher schools. But 66% of voucher school teachers have at least five years of experience, compared with 56% in MPS.

The reports do not include figures on what percentage of voucher school teachers have state certification or college degrees, both points of debate about the program. State law does not require private school teachers to meet those standards - it requires only that they have high school diplomas - but a relatively new requirement that the voucher schools become accredited by independent agencies is expected to result in the near-elimination of teachers without college degrees.

While cautioning that the figures are a bit uncertain, the researchers came up with student/teacher ratios of 13.6 to 1 for voucher schools and 16.6 to 1 for MPS.

While there is not much difference between MPS and voucher parents when it comes to the percentage who say their children have physical disabilities, there is a significant difference when it comes to other special education needs.

"The percentage of respondents who said that their child had a learning disability is twice as large in the MPS sample (18.2%) than in the MPCP sample (8.7%)," the researchers wrote. They say some of the difference might be due to differences between the two streams of schools in labeling children with special needs.

MPS officials have said frequently in recent months that the public schools are shouldering a far larger portion of special education students than the private schools are, and that the trend is causing major stresses on MPS.

Fiscal impact

Among other findings, the researchers' report on finances concludes that the way the voucher program is funded puts a greater burden on Milwaukee property-tax payers, while actually providing financial help to property-tax payers in the rest of the state and some reduction in state income taxes.

In simplified form, the reasons are that voucher students receive less public money than MPS students, while the formula for how to come up with the public money puts a larger load on Milwaukee property taxes than MPS students put.

Milwaukee Mayor Tom Barrett and others have argued for fixing the "funding flaw" in the voucher program to help city taxpayers. Action by the Legislature last fall provided partial relief.

The research project is being paid for with private funds, primarily from major foundations, including at least three that are firmly identified with advocacy for school choice programs such as Milwaukee's. They are the Walton Foundation, funded by Wal-Mart heirs and based in Arkansas; the Lynde and Harry Bradley Foundation, based in Milwaukee; and the Kern Family Foundation, a relative newcomer to the scene, based in Waukesha.

Plans call for reports in future years that will examine such questions as how much progress students in the voucher program are making from year to year, and how that compares with a comparable MPS group. Ninth-graders identified in the 2006-'07 school year will be followed to determine graduation rates and other outcomes several years from now.

"Stay tuned," Wolf said.