Analysis of School Choice Programs and Corresponding Evaluations for Policy Development: New York, Cleveland, Florida, and Milwaukee

By

Shawntel B. Hines
B.S. Urban Studies and Planning
Massachusetts Institute of Technology, 2006

SUBMITTED TO THE DEPARTMENT OF URBAN STUDIES AND PLANNING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER IN CITY PLANNING

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2007

© 2007 Shawntel B. Hines. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author: ____________________________

Department of Urban Studies and Planning
May 22, 2007

Certified by: ____________________________

Frank Levy
Professor of Urban Economics
Thesis Advisor

Accepted by: ____________________________

Langley Keyes
Chair, Master of City Planning Committee
ABSTRACT

Although many Americans assume that the education provided by public schools will prepare them for higher learning and/or the workforce, recent studies have shown that American students’ test scores lag behind their counterparts around the world and a growing number are not prepared for graduation. School choice, in the form of publicly funded vouchers for low-income students, has caused significant debate as a form of education reform. Proponents argue that vouchers will induce competition between schools, help low-income students obtain a better education, and increase parental satisfaction. Opponents of school vouchers argue that publicly funded vouchers will drain public schools of much needed resources, leave the most difficult to educate students in public schools, and violate the constitution by funding sectarian institutions.

School voucher programs have been implemented in New York, Florida, Cleveland and Milwaukee yet there is no consensus in terms of the effect of vouchers on achievement scores. This thesis looks at four school choice programs, evaluates and compares the design/methodologies of the program evaluations, and draws conclusion about which results are the most reliable and why. Once the methodologies and findings of the evaluations were analyzed, I found that most programs did not have a significant effect on achievement test scores of voucher recipients and did not induce competition between public and private schools.

Thesis Advisor: Frank Levy
Title: Daniel Rose Professor of Urban Economics
## Table of Contents

Abstract ........................................................................................................................................... 3

Chapter I: Introduction .................................................................................................................. 7
  Background .................................................................................................................................... 11

Chapter II: Methodology ................................................................................................................ 17

Chapter III: A Look at Current School Voucher Programs ....................................................... 19
  * Milwaukee Parental Choice Program ....................................................................................... 19
  * Cleveland Scholarship and Tutorial Program ........................................................................ 33
  * New York School Choice Scholarships Program .................................................................... 42
  * Florida A-Plus Accountability Program .................................................................................. 48

Chapter IV: Conclusion .................................................................................................................. 56

Appendix 1: Summaries of School Choice Programs .................................................................. 62
Chapter I: Introduction

In the United States, an education is considered every citizen’s right. While the quality of the education is not guaranteed in any part of the US Constitution, most citizens have come to expect that the government should provide a high quality education. Each year, parents send their children off to school expecting that these students will be prepared for either college or the workforce.

Yet, several trends show that the nation’s public schools are not measuring up to expectations. According to Richard J. Murnane, author of “The Role of Markets in American K-12 Education,” studies show that American students do not score as well as students in many other countries on comparable tests of math and science. US eighth graders scored significantly below the cross-national mean on science and math tests taken by students in many OECD countries in 1999. In addition, only seventy percent of all students in public high schools graduate and only thirty-two percent of all students leave high school qualified to attend a four year college. Nonetheless, the number of students enrolled in America’s public schools is on the rise; enrollment in public elementary and secondary schools rose twenty percent between 1985 and 2001.

Given that some of America’s public schools are failing to provide an adequate education, what can parents do to ensure that their children do receive a quality education? One frequently discussed proposal is the creation of a market-based school choice initiative. This initiative calls for transforming the public education system into a free market system that will create competition among schools by allowing families to send their children to schools other than the public schools in their assigned districts. School choice programs include several policy alternatives: charter schools, open
enrollment, intra-district choice, dual enrollment and tuition tax credits. One of the most controversial of school choice options is education vouchers. Vouchers allow parents to select the school of their choice - public or private - and use a government allotment to pay for part or all of the tuition.

Proponents of school choice argue that these programs will force schools to compete with each other for students and the resulting market pressures will stimulate responsiveness, innovation, and improvements in school performance. By allowing parents to act as rational consumers in a free-market education system, schools will be forced to increase the quality of education they provide and the efficiency with which they provide it, or risk going out of business. However, opponents of school choice question whether school choice will indeed induce competition between schools and contend that the program, particularly vouchers, will drain financial resources from public schools. In addition, vouchers and tax credits may benefit middle and upper class families the most, leaving low-income families trapped in failing public schools.

Despite the debate surrounding school choice programs, voucher programs are being implemented in school districts across the United States, including programs in Milwaukee, Charlotte, the District of Columbia, and most notably, Cleveland. In 2002, the United States Supreme Court ruled that Ohio’s voucher program does not violate the establishment clause of the First Amendment. Ohio’s Pilot Project Scholarship Program provides tuition vouchers for low-income students in Cleveland City School District to attend religious and secular public or private schools participating in the school choice program. The Supreme Court held that because Ohio’s program is part of the state’s general effort to provide improved educational opportunities to children, religious
institutions can receive government aid through the deliberate choices of individual voucher recipients; the advancement of any religious mission is therefore attributable to the voucher recipients, not the government.

This Supreme Court ruling has spurred more school districts to seriously consider implementing school choice programs. But how effective are these programs in improving the education of America’s students? Are all students benefiting from school choice programs or just those from a particular income group? Have school choice programs been successful in inducing competition between schools? The effectiveness of school choice programs is important to policymakers, school administrators, teachers, and parents, especially as more states are proposing implementation of school choice programs.

Several school choice programs have been evaluated to answer similar questions, yet little consensus exists. John Witte’s five-year evaluation of the Milwaukee Parental Choice Program was unable to find statistically significant differences in the achievement of students attending schools participating in the choice program and the achievement of a comparable group of students attending the Milwaukee Public Schools. Witte’s research was attacked by Harvard professor Paul Peterson in 1995, who asserted that Witte’s methods were flawed and had therefore produced findings that understated the positive academic impact of the Milwaukee Parental Choice Program. The debate in Wisconsin and elsewhere continues to be driven more by competing social and political philosophies than research data because of the lack of clear evidence showing a positive effect on student achievement as a result of participation in a school choice program.

Similarly, Caroline Hoxby’s study on school competition, which found that cities
with more streams tended to have schools with higher test scores, has been recently
criticized by Princeton professor, Jesse Rothstein, for methodological flaws. Hoxby
noticed that cities with more streams had more school districts than cities with few
streams and later found that metro areas with more streams tended to have more districts,
more competition, and higher student achievement. Because some streams can affect
more than school-district borders, Hoxby divided streams into larger and smaller ones
and entered them into her equations separately to make the distinction clear. However,
Rothstein argues that Hoxby did not clearly explain precisely how she measured a
stream’s width, which could lead to variations in the number of streams in a school
district, making Hoxby’s larger-streams variable subjective and unverifiable. Even when
he tried other ways to count streams, Rothstein only found a statistically insignificant link
between streams and school performance. The critiques of Witte’s and Hoxby’s studies
show how small variations in methodologies can lead to significant differences in
findings.

Approaching school vouchers from the perspective of a governor’s aid helping to
shape education policy, I want to determine how policymakers can determine if vouchers
are successful in fulfilling their promises. Do vouchers induce the competition they
claim, forcing public schools to improve their services or be pushed out of the market?
Do students who use vouchers to attend private schools score higher on achievement tests
than their public school counterparts? Because evaluations arrive at different results, we
have to take a deeper look at the design of program evaluations in order for governments
to be able to implement effective education reform.

The purpose of this thesis is to analyze attempts by researchers to evaluate school
choice programs throughout the United States in order to better understand what the evidence does say about the effectiveness of school choice programs from a policy perspective. Here, I will examine school choice programs, their evaluations and corresponding findings to determine what a governor or mayor should know and consider when looking at school choice programs and their effectiveness at raising achievement scores. This will be done by first outlining the details of four school choice programs: Milwaukee Parental Choice Program, Cleveland Scholarship and Tutorial Program, New York School Choice Scholarship Program, and Florida A-Plus Accountability Program. These are some of the longest-running school choice programs, all of which have been evaluated at least once since their implementation.

Next, I will present the methodologies and evaluation findings of each school choice program, highlighting the particular statistical method used to determine the effect of vouchers on achievement test scores as well as whether the evaluators found the effect of vouchers on achievement to be significant. Finally, I will provide an analysis of each program and corresponding evaluations, outlining methodological strengths and weaknesses along with a conclusion of which findings are the most reliable.

Background

Initially, education in the United States was distributed as a private good to be purchased, though parents did not have as many options then as they do now through school choice plans. In the first 50 years of the republic, most children who attended school were either in private schools or in town-sponsored schools funded by a mix of taxes on property and fees and donations from parents. In small towns and rural areas,
“district” schools were formed on a neighborhood-by-neighborhood basis. These semi-public institutions were financed by a combination of tuition charges and local taxes, which allowed some of the poorer students to attend for free while those who could afford to, paid their own way. In more populous areas, many schools were entirely independent. Teachers would solicit students through word of mouth and advertisements in the local press. Some teachers attempted to appeal to the widest possible clientele by offering a plethora of classes for students of all ages, while others offered intensive and specialized instruction in areas such as applied mathematics\textsuperscript{xii}.

Toward the middle of the 19\textsuperscript{th} century, school reformers such as Horace Mann and Henry Barnard convinced legislatures to embrace what became known as the common school. According to Murnane, reformers felt strongly that the common public school would provide critical public benefits to the emerging nation. However, critics of the reform argued that they should not have to pay for the education of other parents’ children. Others resented state infringement on local communities’ decisions about the content and governance of their children’s schooling. Still others, including a growing number of Catholics, resented the use of the Protestant Bible in the public schools and the way their children were treated in these schools\textsuperscript{xiii}.

Today, a return to market-based education initiatives, i.e. school choice, has left parents with significantly more choices in how and where their children are educated. In the idealized free market, the law of supply and demand influences prices toward an equilibrium that balances the demand for the products against the supply. In this type of market, all exchanges are voluntary. Although no national economy can naturally exist as a free market, the term “free market economy” is used to describe economies that
approximate the ideal by virtue of having a government that engages in little or no interventionist regulation\textsuperscript{xiv}.

Milton Friedman, noted free market economist, argues that unregulated market forces of supply and demand can do a better job than government programs in helping meet society's educational needs. Friedman goes on to argue that the government has a monopoly over public education, with little incentive to keep costs low or quality high because there is no fear that their patrons will be lured away. Because of mandatory assignment policies that make it necessary for families to move to another neighborhood in order to change schools, most families are unable to consider alternative public education providers. In addition, most parents are restricted from exercising their option to switch to alternative, private providers because of the cost of tuition in addition to a continued legal obligation to support the public education system through their taxes. Friedman argues that because public monopolies are supported by tax revenues, the public school system will not suffer any financial loss if consumers opt out\textsuperscript{xv}. Joseph Klesney, author of "The Competitive Edge," agrees, stating that due to compulsory attendance, government schools rarely need to worry about operating efficiently, attracting students, or being accountable to the public.

According to the market-based education theory, by inducing competition between private and other schools, school choice programs will force all schools to become more cost-efficient, more attractive to students, and more accountable if they are to remain competitive. If a school fails to accomplish these goals, students will have the ability to leave that school and attend another institution that will provide a better education, creating the incentive for schools to offer innovative services, excellent
academics, and quality facilities. Klesney argues that the free market nature of school choice plans will ensure that superior educational products will be rewarded while inferior ones are weeded out.

However, critics of a market-based education system argue that the market will not solve the problems of America's public education system. According to Helen F. Ladd, author of *Market-Based Reforms in Urban Education*, public schools that are not meeting the educational needs of their students cannot be shut down unless there are alternative schools for the children to attend because of government compulsory attendance policies. Therefore, a key mechanism of a typical market, the potential for firms to fail, does not work effectively when applied to education. In addition, opponents of market-based education initiatives contend that a free market would only be effective in improving the education system if there was real competition, which does not exist between public and private schools. Public schools are required to fund students' transportation whereas private schools do not. Moreover, public schools face restrictions and regulations that private schools do not, including treatment of children and building maintenance, which further reduces competition.

Despite the debate surrounding the ability of school choice plans to induce competition amongst schools, some aspects of school choice are considered viable options for improving the education system. According to Jeffrey Henig, author of *Rethinking School Choice*, choice plans can be based on administrative procedures, vouchers, or tuition tax credits. Some are limited to public schools and some plans include parochial private schools, while others include all schools. School choice plans may be district-wide or cross district boundaries; they may be locally initiated, mandated
by state law, encouraged by state incentives, or stimulated by federal grants. Most options can be combined to create distinct combinations, and all may be further distinguished by their emphasis on other objectives, including racial balance or empowering neighborhoods\textsuperscript{xix}.

There are several policy alternatives of school choice programs that can be implemented to provide families with a wider variety of schools to choose from. The most controversial option of the market based school choice initiative is publicly funded vouchers. Vouchers are publicly funded government allotments given directly to parents that allow them to send their children to the school of their choice. The voucher is then cashed by the receiving school. Depending on the stipulations of the school choice program, vouchers can go to either public or private schools and may or may not cover the estimated cost of a child’s education.

The controversy around school vouchers stems from the use of public funds to potentially fund private and parochial schools, and to cause the loss of funds from schools that need them the most. According to the Center for Education Reform, in school choice programs where families can opt to send their children to public schools in other school districts, public funds are simply transferred from one district to another. Yet even in these programs, funds are transferred with students who choose to leave failing schools to attend better schools, possibly leaving those schools with even less money for improvement\textsuperscript{xix}. However, “School Vouchers: Settled Questions, Continued Disputes” contends that when students opt to attend private schools through school vouchers, although state funds follow students if they choose to attend a school in another district, local revenues, which stays in the district regardless of the movement of students,
will then be spread over a smaller number of students\textsuperscript{xxi}.

Opponents also argue that school vouchers will only benefit the most promising students. Murnane argues that voucher programs should be designed to minimize the likelihood that public schools are left with a disproportionate share of students who are especially expensive to educate or are perceived as negative influences on the learning of their classmates\textsuperscript{xxii}. "School Vouchers" explains that several studies show that there are few signs that interest in vouchers are limited to the most talented. In addition, the CER contends that the "best" students are most likely to remain in the schools that helped them succeed, while the ones most in need of help tend to leave. In addition, annual reports of Milwaukee's school voucher system show that it's not the highest performing students, but the struggling ones who are most likely to exercise choice in school choice programs. Rather than "creaming" the best students away from the public school system, school vouchers provide an alternative education environment for students who are not doing well in public schools\textsuperscript{xxiii}. 
Chapter II: Methodology

The lack of consensus amongst evaluation results could be the result of flaws in the design of either the school choice program or the evaluation. To understand what the evidence says about the effectiveness of school choice programs, an analysis of school choice programs as well as evaluations of these programs will be performed, which will allow students, parents, and policymakers to get a better grasp of what results of school choice program evaluations are reliable.

First, details of each school choice program will be outlined, including: 1) when the program was implemented; 2) the amount of each private school voucher; 3) student/family eligibility requirements; 4) voucher selection process; and 5) any changes in the program since its initial implementation. All programs examined in this thesis use a lottery to randomly select students to participate in the voucher program. However, some programs simply require students to qualify for the federal free lunch program as an income eligibility marker. Others hold eligibility sessions where families of applicants are required to complete paperwork and surveys concerning their income and satisfaction with their current school, while students take approved reading and mathematics tests.

Following the outline of the school choice program are the corresponding evaluations. For each evaluation, I will outline the methodology used by the author, including how they divided up students (one vs. multiple treatment and control groups), what statistical method they used to analyze student data, how the evaluation handled missing baseline characteristics. In addition, I will explore how the evaluation handles students who withdraw from the program. Although some evaluations attempt to address students enrolled in the program for different lengths, many evaluations classify students
into either the treatment group or the control group, or excludes them from the evaluation altogether. An evaluation is flawed if it only analyzes the effects of the school choice program on students who completed the program; this only tells part of the story. In addition, I will include any critiques the evaluation has of other evaluations of the same school choice program.

As part of the evaluation description, I will also include findings as well as any critiques of other evaluations of the same program.

Finally, I will compare evaluations and provide an analysis of which methodologies and results are the most reliable. Not only will critiques from evaluations be taken into account, but I will also consider what methodology is the best considering the design of the program, as well as weaknesses. In addition, I will look at parts of the evaluation design that contribute to bias in the results, including but not limited to: inclusion or exclusion of students with missing baseline characteristics, how students who withdraw from the program are handled, and the availability of data. This analysis will outline evaluation strengths and weaknesses, which will help policymakers understand whether any results found are reliable.
Chapter III: A Look at Current School Voucher Programs

We must first look at the details of several voucher programs and their evaluations to begin to understand how school voucher programs are implemented and evaluated and how differences in program design and evaluation can affect the lack of consensus concerning the effectiveness of school choice programs.

**Milwaukee Parental Choice Program**

The original Milwaukee Parental Choice Program was established in 1990 to give low-income families in Milwaukee an opportunity to attend private schools. Participating families were required to live in Milwaukee and have incomes equal to 175 percent of the poverty line or less. Students currently enrolled in private schools or other public school districts were not eligible. Admission into the choice program could not be based on gender, religion, race or prior school records. For each student enrolled in the choice program, a private school received a payment equal to the Milwaukee Public Schools per-student state aid. Private schools that chose to participate in the choice program were required to limit the number of choice students enrolled to 49% of their total student population. In the program's first four years, the total number of choice students in any year was limited to one percent of the Milwaukee public school membership. In the fifth year, it was increased to 1.5%. If there were more choice applicants than seats available in any grade, the schools had to select students at random to participate in the program.

In order for a private school to participate in the Milwaukee school choice program, it must have been registered non-sectarian private school which meant that it
had to have a sequential curriculum, meet applicable health and safety codes and instruct students for a minimum number of minutes per year. In order to continue participation in the school choice program, private schools had to have at least 70 percent of the students in the program advance one grade level each year. At least 80 percent of school choice students had to have demonstrated significant academic progress. Additionally, the private school’s average attendance rate for students participating in the school choice program has to be at least 90 percent. Finally, at least percent of the families of choice students had to meet the parental involvement criteria.

The student application process for Milwaukee’s school choice program only required parents to complete an application at the participating private school they wished their child to attend. This application determined the family’s eligibility by requesting family size, where the student had attended school the previous year, and family income. Family background information was solicited from all participating students and a random sample of non-participating public school students. Applicants were chosen randomly by grade and all students were placed in rank order, thereby creating a randomly selected waiting list for those students who were not initially selected. The Iowa Test of Basic Skills was administered to program participants and non-participants in grades K through 8.

John F. Witte

When the Milwaukee Parental Choice Program was established in 1990, the Milwaukee legislature required that it be evaluated. The Department of Public Instruction asked John Witte, a political science professor at the University of Wisconsin to head the
Witte conducted a five year evaluation that allowed the tracking of individual students. Each year, Witte made six comparisons between the choice and public schools of Milwaukee: 1) scores on achievement tests for all choice students, all Milwaukee Public Schools (MPS), and all low-income MPS students; 2) choice and public students’ achievement scores by observing the effects of several independent variables on one dependent variable (multiple regression analysis); 3) changes in scores on achievement tests of choice students and all MPS students; 4) turnover rates in choice schools and Milwaukee public schools; 5) scores of choice students and low-income MPS students and 6) opinions of parents. To control for prior achievement, Witte’s evaluation only included choice students and a random sample of MPS students who had prior measures of achievement (i.e. standardized test scores). Witte argues that his evaluation required random selection from among applicants, which allowed for a natural experiment and controlled for unmeasured selectivity bias.

Witte used three experimental groups to determine the effect of the voucher program on test scores: Choice, Rejects, and MPS random sample. The Choice group is comprised of students chosen to receive a voucher and enrolling in a private school for some period of time. If these students leave the program, only their period in the program is considered part of the experiment. The Rejects group is composed of choice applicants who were not selected and did not enroll in a private school. The MPS random sample is composed of randomly selected MPS students who did not apply to the Choice program.

---

1 Unmeasured selectivity means that after evaluators measure and try to control for all variables that might affect learning, some factors are still not measured. In addition, the possibility that unmeasured factors are correlated to students applying to the program or selection by private schools of Choice applicants cannot be dismissed.
Witte finds that, when comparing Choice students with MPS students, there are no consistent differences between the two groups\textsuperscript{xxxvi}. It could be argued that the rejected students were potentially a valuable control group because they provided a natural control on the selection bias. These students had committed to choice and therefore had all the unmeasured characteristics of choosers, but were randomly not selected to participate in the program. In the experimental sense, comparing the Choice students with the Rejects is inadequate for several reasons. Random selection was used only for students in particular schools where there was oversubscription and then, only in certain grades. In addition, subsequent test information is absent for a significant number of rejected students. There is no data on who was admitted under the sibling rule and how this may have affected randomization\textsuperscript{xxxvii}. Again, with much missing data and varying sample sizes, Witte argues that there are no consistent differences between Choice students and randomly selected MPS students or rejected applicants\textsuperscript{xxxviii}.

\textit{Paul Peterson, Jay P. Greene, and Jiangtao Du}

In his critique of Witte's evaluations of Milwaukee's school choice program, Paul Peterson analyzes not only the methodological flaws of Witte's evaluations but also the flaws of the program itself. In exploring the flaws of Witte's methodology, Peterson describes how an evaluation of the choice program was supposed to be scientific and based on the basic fundamentals of program evaluation. Students applying for participation in the program were to be accepted at random by choice schools and those who could not be placed in a choice school were to make up the control group against which choice students would be compared. These students would be the ideal group for
comparison because they would identical to choice students in all respects (desire to participate, eligibility, etc).

However, Peterson argues that several factors make it difficult to assign students randomly to the test group and control groups, which partially contribute to the inaccuracies of Witte’s results. Because students apply to particular choice schools and not to the choice program as a whole, some schools may reject several applicants because they had few openings while others may only reject few applicants because they had several openings. In addition, most choice schools have more openings for students applying for kindergarten and first grade than upper grades which leads to age being a significant factor in whether or not a child is admitted. Peterson argues that the number of applicants does not greatly exceed the number of openings, which means that the differences between groups must be large to be statistically significant.

While these factors negatively affected Witte’s ability to accurately evaluate the choice program, methodological flaws also contributed to the inaccuracy of Witte’s evaluations. Witte compares annual achievement levels in reading and math for each of four years of choice students with those of a cross-section of MPS students as well as a cross section of low-income MPS students. Peterson argues that this is meaningless because choice students in every year had decidedly lower reading and math achievement scores before entering the choice schools.

Furthermore, Witte’s regressions on student performance on math and reading tests for each of four years include a student’s status in a choice school, ethnicity, gender, low-income status and prior achievement test performance. However, Witte’s regressions omit controls for parental education, an adequate measure of parental income, parental
occupation, student’s native language, a student’s social adjustment, family structure, and family’s dependency on welfare. Peterson concludes that a simple comparison between the two groups does not provide an adequate test of the effect of treatment on achievement scores because choice students were not similar to low income students remaining in public schools.

Paul Peterson, Jay P. Greene, and Jiangtao Du performed their own evaluation of the Milwaukee Parental Choice Program that differed from Witte’s in that it avoids selection bias by randomly assigning students to treatment and control groups. With random assignment, the members of the two groups can be assumed to be similar in all respects other than whether or not they receive treatment. Therefore, differences in average outcomes could be attributed to treatment the groups received. Because students applied each year for a seat in a specific grade in a specific school as opposed to the choice program as a whole, they were selected (or not) randomly by school and by grade. Peterson et al’s analysis used a fixed effects model that took into account the year of the application and the grade to which the student applied, controlled for gender, and was able to estimate the effects of enrollment in choice schools on test scores.

Separate ordinary least squares regressions produced an estimate of the effect of one, two, three, and four years of treatment on math and reading scores. Peterson et al admit that their data is limited in that it was only available for 78 percent of those randomly assigned to the treatment group and 72 percent randomly assigned to the control group. By the fourth year of the program, the percentage of test scores available decreased to 40 percent of the treatment group and 48 percent of the control group. In

---

2 A fixed effects model takes into account different initial test scores (prior to the start of the treatment program) by including a dummy variable for each student that is only activated when regressing for that particular student.
addition, Peterson et al assume that the missing cases did not differ appreciably by
observing characteristics of students in the treatment and control groups and concluding
that the background characteristics of both groups do not differ in important respects,
including prior achievement scores and parental involvement.

In their main analysis, Peterson et al found that the estimated effects of choice
schools on math achievement were slight for the first two years of the program. After
three years of enrollment, students scored 5 percentile points higher than the control
group. At the four year mark, students in the treatment group scored 10.7 percentile
points higher. Peterson et al argue that the differences between the two groups are .24
and .51 standard deviation of the national distribution of math test scores and are
statistically significant. On the reading test, differences were between 2 and 3 percentile
points for the first three years and increased to 5.8 percentile points in the fourth. When
the results of the third and fourth years are jointly estimated, Peterson et al. also found
that these results were statistically significant.

However, because the main analysis has several missing cases, it is possible that
the two groups (the treatment and control groups) were not similar in relevant respects.
To uncover whether this possibility contaminated the results, a fixed effects analysis that
took into account mother’s education, gender, parent’s marital status, income, time spent
with the child and educational expectations was performed. However, because this
information was gathered through a written questionnaire which most parents did not
complete, the number of cases available for analysis was reduced and the point estimates
are less reliable. In addition, the main analysis did not control for the students’ test
scores before entering into the choice program. However, Peterson et al found that
controlling for this control was unnecessary because the average pretest scores at the time of application for the two groups was essentially the same\textsuperscript{xlvix}.

To understand the effect of the choice program on all those admitted, regardless of their subsequent enrollment decision, Peterson et al conducted an analysis identical to the main analysis, except that this analysis compared all students initially assigned to treatment and control groups regardless of the schools they actually choice to attend\textsuperscript{3}. The disadvantage to this approach is the inclusion of students who were accepted to the private schools but did not attend as well as those who attended private schools for less than the full period under study in the treatment group. However, the advantage of this type of analysis is that it may have better captured what might happen if the choice between public and private schools were generalized (students moving back and forth before between public and private schools)\textsuperscript{xlvi}. This analysis found slight positive effects for the first three years after application to the program and moderately large effects after four years. Additionally, students who were given a choice of schools performed better than the control group, regardless of the public or private school they chose to attend\textsuperscript{xlvii}

\textit{Witte's Reply to Peterson}

In 1996, Witte published a reply to Peterson et al’s critique of his first evaluation of the school choice program. Witte argues that it is difficult to determine whether Peterson et al’s use of randomized block design was appropriate because Peterson et al do not describe nor cite the technique in their analysis. The randomized block design, according to Witte, might be an appropriate method to use in circumstances where it is

\textsuperscript{3}In medical research, this type of analysis is known as an intention to treat analysis. In medical experiments, subjects may be more or less faithful in complying with the treatment. This type of analysis uncovers whether treatment is effective when compliance is less than 100 percent.
clear what relevant factors are used to create these blocks. However, Peterson et al create blocks based on race and grade, but not on income, parent education or gender. In addition, the number of students in the treatment and control group in each block varies significantly. Sixty percent of the blocks have less than four control students, which means that the amount of information available for the control group students is significantly less than that of the treatment group, thereby biasing Peterson et al’s results.

Further bias comes from the use of non-selected choice students as a control group. Witte agrees that the use of these students as a control group could provide for a unique opportunity to control for unmeasured selectivity bias, but points out that this depends on the assumption that the missing cases do not differ significantly from those remaining in the sample. However, Witte argues that Peterson et al fail to analyze the differences in the non-selected students who “continue” in the program by returning to Milwaukee Public Schools and those who go to other schools outside the program. The differences between students who continue in Milwaukee Public Schools (for which there is subsequent test data) and those who do not point to a residual control group which comes from poorer families. Witte argues that this residual control group has less educated parents who are less involved in their child’s education, which creates a major bias which will favor choice students when they are compared to this group.

Witte also argues that all students not in the choice program were not tested with the same frequency as those in the treatment group. Non-selected students who did not return to MPS were not tested every year, as the choice students were. In addition, income determined which MPS students were tested every year in that students in the
Chapter 1 program were required to be tested every year, while those not in Chapter one were not. Therefore, a larger proportion of the control group having test data were likely to be poor, which adds to the need to control for income in all estimates. Furthermore, Witte argues that it is illogical to base the success of the program on the “survivors” after three of four years, given attrition rates of 30% per year and the fact that three private schools from 1990-1995 closed in the middle of the year.

Cecilia Rouse

Cecelia Rouse’s evaluation of the Milwaukee Parental Choice Program in 1998 differs from Witte’s and Peterson et al’s in that she compares test scores of successful applicants with those of both unsuccessful applicants and a random sample of Milwaukee public school students. Rouse argues that, while unsuccessful applicants are a tempting control group because they are likely to have parents who are similarly motivated to the parents of successful applicants, parents of unsuccessful applicants may have been more likely to enroll their children in a private school not registered with the choice program instead of enrolling them in a Milwaukee public school. If post-application data were available on unsuccessful applicants who enrolled in other private schools, this would not be a problem. However, the data available did not track students enrolled in public schools outside of Milwaukee Public Schools or non-choice private schools. In addition, by the fourth year of the program, there were fewer than forty unsuccessful applicants to use in evaluating the program; such a small sample leaves the estimated effects of the school choice program sensitive to extremely high or low test scores.

---

4 This was made easier by the Partners for Advancing Values in Education (PAVE) program, a privately funded initiative that provided scholarships to students interested in attending (primarily) Catholic schools.
Rouse goes on to argue that the achievement of successful applicants could be compared to that of a random sample of students from the Milwaukee public schools and this group is less subject to non-random attrition because they were not overtly interested in leaving public schools. However, these students may have refrained from applying to the school choice program because their parents were less motivated, which would lead to an exaggeration of the effects on achievement of the program. In addition, Witte's findings that choice schools had no effect on test scores are flawed because students who qualified for the Milwaukee choice program were disadvantaged and scored lower than wealthier, more advantaged students. Therefore, these disadvantaged students would have continued to score below average even if they were in the choice program. Rouse argues that this is why family background and student ability must be controlled. Rouse further criticizes Witte's comparison of test scores of the control and treatment groups. Although controlling for prior test scores has the advantage of accounting for student ability which changes over time, this strategy only allows for the inclusion of students who have prior test scores in the analysis.

Rouse contends that her evaluation is more complete because, in measuring the choice program's effects on student achievement, it compares not only whether private schools are better than public schools but also whether students who were selected to attend a choice school enrolled and remained there. She uses a fixed effects model as well, and acknowledges that this model is based on the assumption that the students in the control group and the treatment group, even if they have different prior test scores, have test scores gains at about the same rate. If students in the treatment group (those who enroll in choice schools) have faster test score gains than unsuccessful applicants and
those who chose not to apply to the choice program, the fixed effects analysis will attribute the faster achievement growth to the choice program rather than the student (who would have made faster gains anyway). Rouse attempts to address this by analyzing pre-application test score trajectories of students in the choice program and those in the Milwaukee public schools, which indicated that the results using the individual fixed-effects model were not overstated.

Rouse found that students selected for the program made gains each year in math, particularly after the second year of application. Both the unsuccessful applicants and the sample of Milwaukee public school students experienced large declines in their math test scores in the third and fourth years. However, there are no detectable differences in the reading test scores between the three groups. Because Rouse’s math results agree with those reported by Peterson et al, she argues that with these data, if one adequately controls for student characteristics, there is not a significant difference in whether one defines choice students those who were selected to attend a choice school or as those who are enrolled in a choice school. However, Rouse acknowledges that these data are not ideal for an evaluation of the choice program because students who were not enrolled in either a choice school or a Milwaukee public school were not included.

Analysis

There are several issues that arise throughout the Witte, Peterson and Rouse evaluations. The concept of an ideal control group is one that is contested amongst the three evaluations. Peterson argues that the control group should be made up of students who applied to the choice program but were not accepted because they are identical to
students who were accepted to the choice program in all respects except whether they received treatment. Both Witte and Rouse agree that, ideally students who applied but were rejected from the control group would be the best control group, but because of the missing post-application information, the most reliable control group is non-applicants.

However, Witte contends that Peterson does not analyze the differences in the non-selected students who “continue” in the program by returning to Milwaukee Public Schools and those who go to other schools outside the program. The students who continue in Milwaukee Public Schools make up a residual control group, with less educated parents who are less involved in their child’s education creating a major bias. Data is not available for students who were rejected from the choice program but chose to enroll in other private schools. Rouse and Witte argue that rejected students cannot serve as an adequate control group because, by the fourth year, the small number of unsuccessful applicants relative to the number of choice students can lead to either an overstatement of the treatment effect.

Rouse and Witte do not find the same results, however, because of the lack of controlling for student demographics and the inclusion of the students without baseline test scores. Rouse argues that Witte’s findings are flawed because, the students who qualified for the choice program were already poor and disadvantaged and therefore, would score lower on achievement tests regardless of whether or not they were enrolled in the choice program. Hence, it is crucial that student demographics are included to ensure that students are truly identical in all respects except for whether or not they received treatment. Additionally, Rouse argues that Witte’s findings are flawed because, although he controlled for prior test scores, this limits the students included in the
evaluation because of no prior test scores. In the New York School Choice Scholarship Program, Krueger and Zhu found that omitting baseline scores had a trivial effect on achievement scores.

The fixed-effects model is the best proposed method of determining the effect of vouchers on achievement scores. Rouse uses the fixed effects model to predict the trajectories of treatment and control group scores, assuming that students in both groups made similar gains at a similar rate prior to treatment. However, this method could make Rouse’s findings somewhat questionable because she includes students who do not have prior test scores, which makes it difficult to develop a full test score trajectory. Yet, Krueger and Zhu included students who had missing test scores and found that this had little effect on the intent-to-treat effect. Using a fixed effects model to predict a trajectory means that any additional increases in gains once treatment begins can be attributed to the choice program.

For the Milwaukee School Choice Program, Rouse’s findings are the most reliable, based on her methodology and comparison groups. Using the fixed-effects methodology and comparing choice students to both unsuccessful applicants and Milwaukee Public School students (with both control groups having demographics similar to the treatment group), she found math gains similar to those found by Peterson et al and no detectable differences in reading for all three groups. We have to keep in mind, however, the assumption that choice students, unsuccessful applicants and non-applicants made gains at about the same rate, regardless of differing prior test scores.
Cleveland Scholarship and Tutorial Program

In 1995, the state of Ohio established the Cleveland Scholarship Program, which provided families with up to $2,250 toward the cost of attending the private school of their choice. The program, begun in September 1996, was originally limited to children in grades kindergarten through grade three. Almost 2,000 students participated in the first year of the scholarship program, with most coming from public schools or just beginning school. Only about 25 percent of the students had previously been attending private schools.

Several changes to the program were instituted that had an effect on program outcomes, beginning with the expansion of the program. During the summer of 2003, the Ohio Legislature decided to expand the program into high school by offering grade nine scholarships to students who had used a scholarship to attend private school during the previous year. In the 2004-2005, grade ten scholarships were made available to students who had used a scholarship to attend private school in ninth grade the previous year. In addition, the Ohio legislature increased the funding amount for the scholarships. In 2003-2004, the maximum amount of each scholarship was increased from $2,500 to $3,000 for students in grades kindergarten through eight and $2,700 for students in grades nine and ten.

Further changes included adjustments to how available scholarships were awarded. Prior to 2004, three applicant categories based on the federal poverty index were used to group applicants by income. The first group of applicants included families who earned less than 100% of the federal poverty index, while the second group earned between 100% and 200% of the federal poverty index. The third group was composed of
applicants whose families earned more than 200% of the federal poverty index. The award process was designed such that when the number of applicants in the first two groups exceeded the number of available vouchers, a lottery occurred among applicants in the first two groups. If the number of vouchers available outnumbered the applicants in the first two groups, the remaining vouchers were offered to the third group. However, in 2004, the award process changed such that available scholarships were first awarded to eligible kindergarten students, with priority given to families who were in the lowest income category. Remaining scholarships were awarded to families of students in first through eighth grades using a lottery, again with priority given to families in the lowest income category. Families were given thirty days to use the scholarship. After that point, the lottery and allocation process was repeated with unused scholarships and distributed to remaining eligible applicants\textsuperscript{lviii}.

In the Hope Academy and Hope Ohio City Schools, the California Achievement Test was administered on three occasions: Fall 1996, Spring 1997, and Fall 1997. The Hope schools were established in response to the adoption of the Cleveland Scholarship Tutorial Program. These schools announced they would accept all students who applied for admission, allowing many of the poorest and most educationally disadvantaged students participating in the choice program to attend Hope schools\textsuperscript{lix}. The Hope schools represent approximately one-quarter of scholarship users.

\textit{Indiana University School of Education}

In 1998, the Indiana University School of Education released its evaluation of the Cleveland Scholarship and Tutorial Program, which found that the first year of the
scholarship program had no effect on the test scores of 94 third-grade students attending choice schools other than the Hope schools. In 2001, a second report was released by the Indiana University School of Education, evaluating the second phase of the program. The primary data sources used to determine student achievement was the Terra Nova, a standardized test produced by CTB/ McGraw-Hill. This test was chosen because none of the schools in which data were collected use the Terra Nova as their primary off-grade testing tool. In addition, scale scores are provided that allow comparisons across time. Each year, this test was administered to targeted students by representatives of the Indiana Center for Evaluation.

CSTP office records were used to identify students as scholarship winner-users, applicant non-recipients, and scholarship winner-non-users. Through the process of elimination, the evaluation team was also able to identify non-applicants as well as former scholarship winner-users. Additionally, the office records provided information on the names of the private schools in which scholarship students are enrolled, which allowed the evaluation team to track students' schools of attendance and school changes across time. This information was further supplemented by records from Cleveland Municipal School District, which provided information for students whose most current school was not available from CSTP records as well as for public school non-applicants.

To initiate a longitudinal evaluation, the evaluation team obtained the broadest possible sample of participating and non-participating students during their first grade year in 1998-1999. Students were identified based on five groups: 1) scholarship students attending private schools, 2) scholarship applicant non-recipients- students who...
applied for but did not receive a scholarship and who attend public schools, 3)
scholarship winner non-users- students who applied for and received a scholarship but
did not use the scholarship and attend public schools, 4) former scholarship recipients –
students who received and used a scholarship for one or more years, later withdrew from
the CSTP and now attend public schools, and 5) non-applicants- public school students
whose families never applied for a scholarship\textsuperscript{iii}.

Groups were then divided into subgroups. The scholarship recipients were divided
into subgroups that varied from one to four year scholarships recipients (one year
scholarship recipient being a student who entered the program in third grade and have
only participated for one school year). Former scholarship recipients were divided into
three subgroups, from three year former scholarship recipients to one-year former
scholarship recipients, with three year former scholarship recipients being students who
participated in the program for one school year during kindergarten, subsequently
withdrew from the program after kindergarten and have attended public schools for three
years and so on\textsuperscript{iii}.

A mixed-design covariance\textsuperscript{5} model was used to examine: 1) differences in
academic achievement among the different groups; 2) differences in academic
achievement among each of the four testing episodes, regardless of group membership;
and 3) the interaction between testing episode and group membership. To compare the
academic achievement of students in the various groups, group membership was used as
a between-subjects variable and testing episode was used as a within-subject variable\textsuperscript{iv}.

Because student demographic characteristics accounted for a statistically significant

\textsuperscript{5} A mixed-design covariance model is similar to the difference-in-difference method of evaluating data.
The difference-in-difference estimator compares the difference in outcomes after and before treatment for
groups affected by it to the difference for untreated groups.
portion of the variance in each of the four testing episodes, student sex, student eligibility for subsidized school meals, and minority status were included as covariates in the analysis. This provided for the opportunity to investigate the impact of program participation across time while controlling for initial differences among the comparison groups.

This model also provided for the examination of possible differences in achievement among scholarship students based on different entry into the program as well as differences between students who continuously remained in the program and those who chose to withdraw from the program and return to public schools. Metcalf et al found that, although a significant initial gap existed in achievement between scholarship recipients and public school students at the beginning of first grade, the gap was later closed by public school students. Throughout second and third grade, scholarship and public school students continued to perform and improve at statistically similar levels.

*Program on Education Policy and Governance (PEPG)*

The initial evaluation of the Cleveland Scholarship Program was undertaken by Harvard University’s Program on Education Policy and Governance. This evaluation compared the satisfaction levels of a cross-section of parents participating in the scholarship program with the satisfaction of those who initially applied but whose children remained in public schools. PEPG found that parental satisfaction was significantly greater with Cleveland’s private schools than with its public schools.

The PEPG team was also able to obtain test score data from 263 students attending the Hope schools. Most of the students who PEPG was able to get test score
data for had never before attended a private school\textsuperscript{viivii}. The test scores from the Hope schools provided information about new learning environments created when choice programs are introduced. These scores showed gains in both reading and math. Although there were moderate gains in reading and large gains in math, there was a decline in language test scores\textsuperscript{6}.

\textit{Center for Evaluation and Education Policy}

In 2006, Jonathan Plucker, Patricia Muller, John Hansen, Russ Ravert and Matthew Makel at the Center for Evaluation and Education Policy collaborated to perform an evaluation of the Cleveland program to examine the impact of participation in the CSTP on students’ academic achievement. The primary achievement outcome data source for the study is the Terra Nova, a standardized test administered to students each year by representatives of the evaluation team.

According to Plucker et al, the broadest possible sample of participating and non-participating students was selected during their first grade year in the 1998-1999 school year to provide the most valid and meaningful evaluation design\textsuperscript{ivviii}. This longitudinal cohort was tracked from first through six grades. In addition, because comparing achievement outcomes in longitudinal and non-randomized interventions is challenging, they use five comparison groups to best examine the impact of the voucher program. The first group was composed of those students who received a scholarship and used it to attend a private school (scholarship recipient-users). The next group of students applied for but did not receive a scholarship through the lottery and attended public schools at the time of the evaluation (scholarship applicant non-recipients).

\textsuperscript{6} The language test is not normally administered in the Cleveland Public Schools.
The public recipient non-users group was composed of students who applied for and received a scholarship but did not use the scholarship and attended public schools during the five years of the evaluation. The former scholarship users group consisted of students who received and used a scholarship for one or more years, but later withdrew from the CSTP and resumed attending public schools. Finally, public school students whose families never applied for a scholarship made up the public non-applicants comparison group.

To compare the academic achievement of students who have used a scholarship continuously from kindergarten through sixth grade with that of scholarship applicant non-recipients attending public school, as well as non-applicants attending public school, Plucker et al used analyses of covariance. Because the available data do not result from a randomized controlled trial design, the following covariates were included to statistically control for differences between students in the comparison groups before the program began: student mobility, minority status and prior achievement. Fall 1998 test scores from the beginning of first grade do no account for program differences that may have occurred during the students’ participation in the program in kindergarten. However, Plucker et al argue that this is the best available indicator to account for pre-program differences in achievement.\textsuperscript{aix}

After controlling for differences in minority status, student mobility, and prior achievement, Plucker et al found that there were no statistically significant differences in overall achievement scores between students who have continuously used a scholarship throughout their academic career and students in the two public school comparison groups. However, there were statistically significant differences in three specific subject
areas, with sixth grade scholarship students who had been in the CSTP since kindergarten outperformed both public school comparison groups: social studies, language and science\textsuperscript{ix}. In addition, when scholarship users were compared to groups of students who had used scholarships for various numbers of years before returning to Cleveland Public Schools, the former had higher levels of achievement than those who left the CSTP across all subject areas\textsuperscript{xix}.

\textit{Analysis}

Both the Metcalf et al and Plucker et al evaluations use the same methodologies and arrive at the same findings. Having five different comparison groups allowed the evaluators to compare students who used the vouchers, were offered vouchers but did not use them, applied for vouchers but did not receive an offer and those who never applied. In addition, looking at achievement longitudinally allows both sets of evaluators to determine any differences in achievement between students who use vouchers for the entire six years and those who use vouchers but later return to Cleveland Public Schools.

The concept of using several comparison groups has been explored in other school choice evaluations, though most of the time with the evaluator attempting to determine whether there is a difference in using students who have been offered vouchers or students who have been awarded and actually used vouchers as the treatment group. In addition, while most evaluations have looked at overall differences in achievement as a result of vouchers, many do not include students who accept the voucher but later return to public schools. Leaving these students out of an evaluation could potentially overstate the effects of the voucher program. Students may have left the program because they
were not doing well; therefore, their test scores could make the effect of voucher less significant.
New York School Choice Scholarships Program

The New York School Choice Scholarships Foundation program was begun in February 1997, when the School Choice Scholarships Foundation (SCSF) announced that it would provide 1,300 scholarships in the amount of $1,400 annually for at least three years to children from low-income families who are currently attending public schools. The scholarship could be applied toward the cost of attending a private school, religious or secular. To be eligible for the scholarship, children had to be entering grades one through five, attending a public school at the time of application, living in New York City and be a member of a family with an income that qualified the children for the Federal School Lunch Program. In order to determine eligibility, students and an adult member of the family had to attend verification sessions to document family income and the child’s public-school attendance.

The Mathematica Policy Research Institute (MPRI) randomly selected families for scholarships through a two-stage procedure, due to a surplus of applications. As families applied for scholarships, they were invited to eligibility assessment and data-collection sessions. Although all families were initially invited to the eligibility assessment and data-collection sessions, it became clear that more families would attend the sessions than could be accommodated. Therefore, applicants were randomly selected for the sessions. During these eligibility verification sessions, students were asked to take the Iowa Test of Basic Skills (ITBS) in reading and mathematics. Students in kindergarten applying for a scholarship for the first grade were exempted from the test requirement. Parents completed questionnaires that asked about their satisfaction with their child’s current school, their involvement in their child’s education, and their
demographic characteristics. The eligibility sessions were held at private schools where students could take tests in a classroom setting on Saturday mornings and vacation days during March, April and early May 1997. Parents completed the questionnaires, available in both English and Spanish, in a separate room as their children took the test. This procedure gave parents sufficient time to complete the questionnaire and the opportunity to ask clarifying questions.

Once the first stage was completed, families that attended the sessions and met the eligibility requirements were randomly selected either for either the scholarship or control group. To make sure all families from the different sessions had the same chance of being selected for the scholarship group, we adjusted the second-stage selection probabilities to reflect the different chances of being invited to the verification.

The final lottery was administered by MPRI in mid-May 1997 and SCSF announced the winners. SCSF, along with the evaluation team, decided in advance that 85 percent of the scholarships would be allocated to applicants from public schools whose average test scores were lower than the citywide median. Applicants from these schools, who represented about 70 percent of all applicants, were assigned a higher probability of winning a scholarship.

The lottery created two statistically equivalent groups of families: a scholarship group with 1,000 families and a control group with 960 families. Families in both groups were invited to attend sessions in April, May, and June 2000 in which students took the ITBS in reading and mathematics. At the same time, parents completed surveys that asked questions about the educational experiences of their oldest child within the age range eligible for a scholarship.
The Mathematica Policy Research Institute and the Program on Education Policy and Governance evaluated the School Choice Scholarship Foundation program based on the impact of an offer of a private school voucher and actual private school attendance on academic performance. The impact of a scholarship offer is reported as the effect on student national percentile rankings. For each student, the average impact of the offer of a scholarship on the student’s combined test score is determined as well as separate estimates on reading and mathematics. To compute the impact of a scholarship offer on a child’s test score, Mayer et al estimated statistical models that accounted for whether a student was offered a scholarship, baseline reading and mathematics test scores, and variables used to define the randomization process. To compute the effects of attending private school on students’ test scores, Mayer et al estimated a statistical model that accounted for baseline test scores, variables used to define the randomization process and whether students attended private or public school.

This analysis proves estimates of the impact of ever switching to private school and the impact of attending private school for three years. The impacts of switching to a private school were estimated by using an instrumental variable estimator, which requires the assumption that the offer had no impact on the year-three test scores for students who were selected for the scholarship and never used it. Use of the instrumental variable estimator to estimate the impact of attending private school for three years requires the assumption that never attending a private school or switching back and forth between public and private schools had neither a positive or negative effect. The SCSF evaluation shows that, after three years, the offer of a scholarship had
no overall impact on student academic performance. Students offered scholarships had about the same test scores as students in the control group\textsuperscript{lxvi}. Among all students ever switching to a private school, which includes students who may have attended a private school for one, two or three years, the switch had no significant effect on year-three test scores\textsuperscript{lxviii}. To compute the impact of attending a private school for three years, Mayer et all assumed that there was no harm or benefit in attending private school for one or two years, then returning to public school. For students attending private school for three years, there was no impact on their test scores\textsuperscript{lxix}.

\textit{Alan B. Krueger and Pei Zhu}

In April 2003, Alan B. Krueger and Pei Zhu re-evaluated the New York School Choice Scholarships Program and found different results by including students with missing baseline scores in their analysis. Because of random assignment, estimates are unbiased even without conditioning on baseline information. For the subsample with baseline scores, omitting the baseline score only trivially affects the estimated treatment effect. Including students with missing baseline test data increases the sample size by 44 percent in the last two follow-up years. Krueger and Zhu mainly focus on intent-to-treat estimates, i.e. the impact of offering students a voucher on their test performance, as opposed to the effect of attending private schools on test performance. The offer of vouchers, as opposed to compelling students to switch to private schools, is the policy decision that is most relevant. In addition, there is a cleaner statistical interpretation of the intent-to-treat estimates. Krueger and Zhu were also interested in the effect of attending a private school for varying lengths of time\textsuperscript{lx}.  

45
Krueger and Zhu found that omitting the baseline scores had a trivial effect on the ITT estimate. When students initially in Kindergarten and those initially in grades 1-4 with missing baseline scores are included, the ITT effect on the reading test is declines even more while the treatment effect for the mathematics test is larger though still not significant at the .05 level.

Analysis

Evaluations of the New York School Choice Scholarship Program highlight the difference between looking at the voucher offers and actual voucher use on achievement scores. The Mathematica evaluation looked at the impact of a voucher offer as well as actually attending private school on achievement scores, while Krueger and Zhu’s evaluation focuses solely on the intent-to-treat effect. One of the primary arguments of proponents of school choice is that competition will force public schools to improve their educational services. Perhaps this is the reason that both evaluations choose to look at the intent-to-treat effect, or the threat of vouchers. When students attending public schools are offered vouchers, all evaluators of the New York School Choice Scholarship Program find that there is no improvement in test scores.

The decision of whether or not to include students with missing baseline scores is one that evaluators of school choice programs across the country are facing. For the New York School Choice Scholarship Program, the inclusion and exclusion of students with missing baseline scores had no effect on the overall conclusion that vouchers have little effect on test scores. However, for most evaluations, the inclusion (or lack thereof) of baseline scores can either undermine or overstate the treatment effect. That two different
evaluations, each with differing methodologies in terms of what students are included, both found that the intent-to-treat had no effect on test scores confirms that the offer of vouchers sincerely did not affect test scores.
Florida A-Plus Accountability Program

In 1999, Florida adopted the “A-Plus” accountability system, which included a provision that allowed students in certain low-performing schools to receive school vouchers. Florida’s A-Plus Accountability and School Choice Program assigns each public school a grade based on the performance of its students on the Florida Comprehensive Assessment Tests in writing, math, and reading. The reading and writing FCATs were administered in fourth, eighth, and tenth grades, while the math FCAT was administered in fifth, eighth and tenth grades. The results from these tests were divided into five categories. If a school receives two F grades in a four-year period, its students are offered the opportunity to attend a better performing public school or vouchers that they can be used to attend a private school[xxx].

The FCAT was administered in the spring of 1998. After the second administration of the exam in 1999, only two schools in Florida received two failing grades. Both of those schools had vouchers offered to their students. Almost 50 students chose to attend one of the nearby private schools, most of which were religiously affiliated. No additional schools had vouchers offered to their students in 2000 when the FCAT was administered again because none failed for a second time.

Jay P. Greene

Jay P. Greene evaluated the A-Plus Accountability and School Choice Program with the goals of determining whether the FCAT is a valid test of student performance and whether the prospect of having to compete to retain students who are given vouchers inspires schools to improve their performance. In order to determine whether the test that
is used to determine school grades is a valid test of student performance, students were required to take a nationally accepted Stanford 9, which had no effect on whether or not students would receive vouchers. If the results of the Stanford 9 correlated with the results of the FCAT, then the FCAT would be proven to be a valid measure of student performance. Greene found that schools with the highest scores on the FCAT also had the highest scores on the Stanford 9 tests that were administered around the same time.

Greene argues that one way to isolate the impact of the threat of voucher offers is to compare the improvements achieved by higher-scoring F schools to those realized by lower-scoring D schools. This is because, while similar in performance and challenges in improving, they are faced with different outcomes if there is no improvement. Schools that received F grades in 1999 and were in danger of losing their students to private schools if their failure was repeated showed the largest gains between their 1999 and 2000 FCAT results. Schools that received D grades and were close to the failing grades that could precipitate vouchers appear to have achieved somewhat greater gains than those achieved by the schools with state grades of A, B or C. However, these gains were still not larger than those achieved by schools receiving F grades. For schools that received A, B, or C grades, year to year changes in FCAT results did not differ significantly.

Gregory Camilli and Katrina Bulkley

Gregory Camilli and Katrina Bulkley argue that the evidence does not support Greene's argument that the improved scores in schools that previously had an "F" grade...
are meaningful improvements and linked to the threat of vouchers. Camilli and Bulkley first argue that although the two tests have substantial correlations, the correlation coefficients computed on aggregate scores typically have much higher values than those computed with student scores.

In addition, Camilli and Bulkley debate Greene's sample selection. Greene uses the school averages of "standard curriculum" students to obtain school-level gains scores. However, "standard curriculum" students tend to score higher on the FCAT; Greene's evaluation does not include certain types of students with disabilities. Camilli and Bulkley argue that, for the purposes of the evaluation it would be preferable to look at the potential impact of the A-Plus program on all curriculum groups. In addition, in their description of the five weaknesses of the "One-Group Pretest-Posttest Design", Stanley and Campbell assert that experimental units chosen on the basis of extreme scores tend to drift toward the mean during the posttest: high scores drift downward and low scores drift upward. The trend for all three FCAT subjects is that higher achievement schools gain less and lower achievement schools gain more.

Greene rejects regression to the mean as a possibility, arguing that scores for the F schools were nowhere near the bottom of the scale for possible results. Camilli and Bulkley argue that "extremeness" should be evaluated in terms of the number of standard deviation units (or in z-score units) below the overall group mean, rather than relative to the lowest possible score. Camilli and Bulkley found that, after calculating the z-scores of the lowest performing school in 4th, 8th and 10th grade reading, and 5th, 8th, and 10th grade mathematics, there was a strong likelihood of obtaining a regression artifact (regression to the mean) in difference scores.
Camilli and Bulkley felt that Greene’s results were affected by regression to the mean and an incorrect definition of net effect and therefore, reanalyzed the data using the technique of residual gain scores. This technique consists of administering parallel forms of the achievement tests before and after treatment and predicting post-test scores from pretest scores and use the deviation as a measure of gain. Glass and Hopkins argue that simple posttest-minus-pretest measures alone are contaminated by the regression effect. The regression effect is removed because the predicted score takes into account any movement toward the mean. In addition, the predicted value accounts for the average state gain, which will lead to unique net effects for any particular accountability grades. Camilli and Bulkley found that, in terms of standard deviation units, the effects of vouchers on reading and mathematics achievement were small.

Greene argues that another way to test whether the effect of vouchers is the result of regression to the mean is to isolate the gains achieved by the schools with the lowest scores from the previous year. If improvements made by “F” schools were concentrated among those “F” schools with the lowest previous scores, then the improvements were more of an indication of regression to the mean than the effect of vouchers. However, Greene argues that, because gains achieved by low-scoring “F” schools are not greater than the gains achieved by higher-scoring “F” schools, the outcomes are not the result of regression to the mean.

Camilli and Bulkley also contest the way Greene computes the effect size. Greene computed the effect sizes relative to the standard deviation of schools, while the typical practice is to use student individual variation for the standard interpretation. Using the standard deviation of schools shows an effect size for reading that is 350% larger than
if the individual-level standard deviation\textsuperscript{lx}. Using residual scores, Camilli and Bulkley repeated Greene’s analysis and found estimates of the effect were small and non-significant\textsuperscript{xvii}.

\textit{David N. Figlio and Cecilia Rouse}

Figlio and Rouse argue that one of the key issues with Greene’s evaluation is whether the observed effects of voucher threat on school performance were true reflections in student learning, and if they were “real”, whether they reflect the threat of vouchers or other elements of the accountability system. Like Camilli and Bulkley, Figlio and Rouse argue that gains made by the “F”-rated schools could be largely the result of “mean reversion” or regression to the mean. Another explanation for the large effects of the threat of vouchers is that the composition of students changed.

Like Camilli and Bulkley, Figlio and Rouse argue that gains made by the “F”-rated schools could be largely the result of “mean reversion” or regression to the mean. Another explanation for the large effects of the threat of vouchers is that the composition of students changed\textsuperscript{xvii}. There are rumors that districts have attempted to redraw school attendance area boundaries to improve the student characteristics of low-rated schools. The final hypothesis is that, while the improvements were “real”, the schools only focused on “high-stakes” grades and taught to the test. Therefore, evaluators and policymakers cannot infer that overall student “learning” improved as a result of the threat of vouchers.

Rather than using school-level data like Greene, Figlio and Rouse use student
level data from a subset of school districts in Florida to address the question of whether
the threat of vouchers and stigma have an immediate effect on public school
performance. They include scores on the FCAT and norm-referenced examinations prior
to 2000 as well as basic student demographic attributes, including information on student
race, poverty status, limited English proficiency status, ethnicity, and disability status.
Not only do they attempt to determine if observed gains were “real” or due to other
behaviors by schools, Figlio and Rouse study whether schools focused on the entire
spectrum of children or on certain subgroups in response to the new accountability
system.

The authors find that the unadjusted test scores of students in low-performing
schools appear to have increased relative to higher-rated schools after the passage of the
A+ plan on both high stakes and low stakes examinations. However, the relative gains in
reading appear to be explained by student characteristics while positive gains on relative
math test scores appear limited to students in the high stakes grades\textsuperscript{xxxviii}. Figlio and
Rouse argue that this bias could occur because of the normal mobility of students
between schools, deliberate changes in school attendance area boundaries, changes in
student and parent school choice when a school becomes voucher threatened or
stigmatized. They attempt to account for changing student characteristics by including
student demographics, English proficiency status, student’s lagged test score, and socio-
economic status\textsuperscript{xxxix}.

Figlio and Rouse also find modest evidence that teachers may have focused
attention on the higher-stakes grades. Since then, the plan has made all grades from three
to ten high stakes, thereby spreading effort among a larger number of grades. In addition,
Figlio and Rouse find modest gains in low-performing schools and somewhat larger gains, especially in math, across the board for lower-performing students. The authors conclude that because low-performing schools in Florida tended to focus attention on lower-performing students, it is other aspects of the A+ Plan, rather than the threat of vouchers, that led to the improvements in student test scores.

Analysis

There are several flaws with Greene’s evaluation that make his findings unreliable. Camilli and Bulkley, as well as Figlio and Rouse, argue that Greene’s findings are not reliable because of the potential of regression to the mean. When Camilli and Bulkley use a different methodology to control for regression to the mean, they find that the effect of the threat of vouchers is small and statistically insignificant.

In addition, the students Greene uses in his evaluation are “standard curriculum,” so it is difficult to tell whether the threat of vouchers has a significant positive effect on non-standard curriculum students. Finally, Greene uses the standard deviations of the schools to determine effect size. However, Richard Murnane agrees with Figlio and Rouse, arguing that evaluators and policymakers must be sure that gains made are the result of the threat of vouchers and not changing student composition. To do this, the evaluation should use the standard deviation of students, not schools, unless you could be sure that the students are all the same.

When determining the intent-to-treat, or in this case, the effect of the threat of vouchers, evaluators have to consider factors outside of the threat of vouchers that may
result in increases in test scores. Figlio and Rouse’s evaluation finds that there are some improvements in higher stakes students’ test scores and lower performing students. However, Figlio and Rouse argue that this is the result of teachers focusing on higher stakes grades. Figlio and Rouse also mention changing student composition and regression to the mean as possible causes of the achievement gains.

Figlio and Rouse’s evaluation is the most thorough in that attempts to not only uncover whether test score gains are “real,” but also whether these gains are the result of the threat of vouchers. It is not enough to determine the most accurate way to determine the effect of the threat of vouchers on achievement scores in failing schools. Figlio and Rouse’s evaluation goes a step further by controlling for student characteristics to determine if student composition changed. In addition, they look at improvements in low stakes and high stakes grades. This is the best way to present findings to policymakers, so that they can determine which results occurred because of the threat of vouchers, and which were the result of other factors.
Chapter IV: Conclusion

When considering school choice programs, particularly private school vouchers, policymakers must look deeply and comprehensively at the evaluations of other programs to determine whether vouchers are truly having a significant effect on achievement test scores. An issue that arose in evaluations of the New York School Choice Scholarship Program is whether to control for baseline characteristics when a treatment is randomly assigned. Excluding students with missing baseline characteristics reduces sample size, but it can also increase precision. On the other hand, including students with missing baseline characteristics increases the sample size and expands the population for which results can be generalized. In addition, some evaluations look at the effect of intent-to-treat on achievement test scores while others observe the effect of actually taking advantage of awarded vouchers and attending private schools. There is an important distinction between the two, particularly when deciding the purpose of the school choice program. Is it to induce competition and thereby pose a threat to failing schools and force them to improve? Or is it to actually put students who currently attend failing public schools in better, private schools to improve scores on achievement tests?

There are additional issues that were not raised in the evaluations covered in this thesis. Some school choice program evaluations look at the effect of vouchers on certain ethnic groups, particularly African Americans. In making decisions about the purpose of vouchers, we have to determine who will be the targeted group in the implementation of the program. Aiming the voucher program at a particular ethnic group runs the risk of being discriminatory.

Let's assume that a reliable evaluation has proven vouchers to be effective in
raising test scores of African American students. Even in an attempt to close the current black-white achievement gap, proposing a plan to offer vouchers to low-income minority students is likely to bring reprimand from taxpayers who are also parents of low-income, non-minority children. Further limiting vouchers to low-income African American students reduces the number of students benefiting from vouchers, though if vouchers are only proving effective for low-income African American students, this may not be an issue.

Nonetheless, limited funding prevents school systems from offering vouchers to all low-income students at failing schools. Offering vouchers only to low-income minorities leaves more students in failing public schools, many who do not have the financial resources to attend private schools. Further minimizing the availability of vouchers dissipates the purpose of vouchers because the program can not adequately focus on improving test scores of a limited number of low-income African American students, or on posing a threat to failing schools to hopefully improve the test scores of remaining students.

The unexplored areas of the program evaluations cannot be discussed without considering the limitations of the evaluations. Most evaluations do not consider the ratio of private schools to public schools, which would not only have an effect on the number of voucher slots available but the reality of competition existing between the two. In addition, there is no control for teachers. Significant changes in teacher numbers as well as composition (e.g. a significant shift from or to teachers with master’s degrees or a large disparity in salaries). Furthermore, none of the evaluations account for changes in curriculum. Changes in curriculum after the voucher program is put into effect could
have been implemented as the result of losing students; however, if this is not explained in the program or evaluation design, the effect of vouchers on student achievement could be understated.

Additionally, most evaluations do not account for changes in student composition or school district characteristics. In Florida, there were rumors that boundaries for student attendance districts were being redrawn after the implementation of the accountability and school choice program. When there are significant changes in public schools that could significantly effect achievement scores but cannot be controlled by the evaluators, we are still left questioning what effect, if any, vouchers (use or offer) are having on achievement scores and inducing competition.

From many of the evaluations covered in the thesis, we see that vouchers, if they do have a significant effect on achievement scores, often have a more significant effect on mathematics scores. This fact must be taken into account when considering implementing a voucher program. Again, we have to address the issue of the purpose of the program. Policymakers have to answer the question of whether it is worth funding a voucher program that is likely to only improve mathematics skills. With these types of findings, we have to consider alternative education reforms that have shown to consistently increase mathematics and reading scores.

**Recommendations**

Discussing which subjects are seeing significant gains as the result of vouchers leads us back to the larger question of the purpose of vouchers. All of the evaluations analyzed in this thesis attempt to determine the effect of vouchers on
achievement scores. While some evaluations look at the effect of voucher use, others analyze the effect of the offer of vouchers. The purpose of vouchers, as explained by its supporters, is to not only give students in failing public schools the opportunity to obtain a better education (evidenced by increased test scores) by attending a private school, but also to induce competition between public and private schools. Unfortunately, the evaluations covered in this thesis focus on improvements in test scores but do not look at changes in public schools that are the result of the implementation of the program. By only looking at vouchers’ effect on achievement scores, evaluators are only answering half of the question.

Therefore, it is my recommendation that you do include vouchers as part of future educational policy. Only the Milwaukee Parental Choice Program has proven to have a significant effect on test scores, and even then, only in math. The other three programs analyzed in this thesis do not have a significant effect on the achievement scores of students offered vouchers, regardless of whether or not they used them. Parental satisfaction increased in all of these programs, because of smaller class sizes, increased sense of security, better facilities, more student support for students and parents, a more diverse curriculum, and a wider variety of services provided. However, one of the main purposes of vouchers is to provide a better education to students by allowing them to choose to go to a better school. Providing funding for choice has not lead to any tangible improvements, as in improvements in test scores, in the education received.

In addition, because none of the evaluations controlled for changing characteristics of public schools, we cannot assess whether vouchers caused public schools to improve the services they provide. None of the evaluations considered
variables such as teacher salary, student-teacher ratio, and curriculum before the implementation of the voucher program, so it is difficult to determine whether public schools made any changes to the services it offers. Additionally, the number of public school students who are using vouchers, in many cases, is not significant enough to give public schools a sense of urgency concerning the loss of students and state funds. Thus, not only have voucher programs have failed to fulfill their purpose of improving test scores for students who use them to attend private schools, they have not induced competition between public and private schools.

Education reform is always a controversial issue, even more so with the consideration of implementing school voucher programs. Opponents argue that publicly funded school vouchers drain public schools of financial resources as well as leave failing schools with the most difficult to educate students. Proponents contend that local revenues stay with public schools, regardless of how many students choose to take advantage of the offer of vouchers. In addition, the threat of vouchers will induce competition between schools, forcing public schools to improve their facilities and services provided.

As policymakers, we have to use an objective eye when we look at evaluations of school choice programs. Methodological flaws, such as the omission/inclusion of students with missing baseline scores and regression to the mean, can lead to an over- or underestimate of the effect of vouchers. Before implementing a voucher plan, I advise
you to strongly consider whether vouchers are intended to raise scores of low-income students by allowing them to afford to attend better schools or to induce competition among failing public schools and private schools. From there, you must look closely at the evaluations of other school choice programs to truly determine if it is worth funding such a program and if it will fulfill our educational reform purposes of improving the quality of education in our public schools.
## Appendix 1: Summaries of School Choice Programs

<table>
<thead>
<tr>
<th>Program</th>
<th>Voucher Use/ Voucher Offer</th>
<th>Control Group</th>
<th>Treatment Group</th>
<th>Significant Effect on Achievement Scores (As the Result of Vouchers)</th>
<th>Effect on Black-White Achievement Gap (in Standard Deviations)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milwaukee Parental Choice Program</td>
<td>John Witte</td>
<td>Voucher Use</td>
<td>Non-applicants</td>
<td>Scholarship Winner-Users</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Scholarship Winner-Users</td>
<td>Yes: Math</td>
<td>Reduces by .24-.5</td>
</tr>
<tr>
<td></td>
<td>Paul E. Peterson, Jay P. Greene, and Jiangtao Du</td>
<td>Voucher Use</td>
<td>Applicant Non- recipients</td>
<td>Scholarship Winner-Users</td>
<td>Yes: Math</td>
</tr>
<tr>
<td></td>
<td>Cecilia Rouse</td>
<td>Both</td>
<td>Non-applicants</td>
<td>Scholarship Winner-Users, Scholarship Winner Non-Users, Former Scholarship Recipients</td>
<td>No</td>
</tr>
<tr>
<td>Cleveland Scholarship and Tutorial Program</td>
<td>Indiana University School of Education</td>
<td>Both</td>
<td>Non-applicants, Applicant, Non-recipients</td>
<td>Scholarship Winner-Users, Former Scholarship Recipients</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Program on Education Policy and Governance</td>
<td>N/A</td>
<td>Non-recipients</td>
<td>Scholarship Winner-Users</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Center for Evaluation and Education Policy</td>
<td>Both</td>
<td>Non-applicants, Applicant, Non-recipients</td>
<td>Scholarship Winner-Users, Former Scholarship Recipients</td>
<td>Yes: Science, Language, Social Studies</td>
</tr>
<tr>
<td>New York School Choice Scholarship Program</td>
<td>Mathematics Policy Research Institute &amp; Program on Education Policy and Governance</td>
<td>Both</td>
<td>Non-applicants</td>
<td>Scholarship Winner-Users</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Alan B. Krueger and Pel Zhu</td>
<td>Voucher Offer</td>
<td>Non-recipients</td>
<td>Scholarship Winner Non-Users</td>
<td>No</td>
</tr>
<tr>
<td>Florida A-Plus Accountability and School Choice Program</td>
<td>Jay P. Greene</td>
<td>Voucher Offer</td>
<td>Schools with Grades D and above</td>
<td>Schools with Grade F</td>
<td>Yes: Reading</td>
</tr>
<tr>
<td></td>
<td>Gregory Camilli and Katrina Buitkei</td>
<td>Voucher Offer</td>
<td>Schools with Grades D and above</td>
<td>Schools with Grade F</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>David N. Figlio and Cecilia Rouse</td>
<td>Voucher Offer</td>
<td>Students in Schools with Grades D and above</td>
<td>Schools with Grade F</td>
<td>No</td>
</tr>
</tbody>
</table>
Murnane, Richard J. "The Role of Markets in American K-12 Education." 6


The State of U.S. Education: 2001 iv


Henig, 4 vi


Murnane 2 xix


Murnane .3 xix

Free Market http://en.wikipedia.org/wiki/Free_market xvi

Henig 59 xx


Henig 4 xxi

The Center for Education Reform xxi

"School Vouchers: Settled Questions, Continued Disputes" 22 xxi

Murnane 31 xxi

The Center for Education Reform xxvii

Witte 44 xxvi

Goldhauer et al 261 xxvi

Goldhauer et al 261 xxvii

Witte 44 xxviii

Witte 46 xxix

Witte 55 xxx

Goldhauer et al 261 xxxi

Witte 55 xxxi

Goldhauer et al 261 xxxii

Peterson 5 xxxii

Peterson 14 xxxiv

Witte 13 xxxv

Witte 132 xxxvi

Witte 137 xxxvii

Witte 143 xxxviii

Peterson 12 xxxix

Peterson 13 xli

Peterson 21 xlii

Peterson 338 xliii

Peterson 343 xliii

Peterson 345 xliiv