Evaluating Three Programs Using a School Effectiveness Model: Direct Instruction, Target Teach, and Class Size Reduction[1]

Bruce Thompson
Milwaukee School of Engineering

Abstract

Value-added models, which rate schools for effectiveness while taking into account the poverty and other socioeconomic status of the students, are generating increased interest. This paper describes the use of one such model to evaluate whether school ratings changed when three new programs were introduced: the "Target Teach" curriculum alignment, direct instruction, and the SAGE class-size reduction. Average ratings for the schools introducing curriculum alignment and direct instruction underwent statistically significant increases. Those implementing the SAGE class-size reduction did not. The paper discusses possible reasons for the findings and the strengths and limitations of this approach to program evaluation.

Introduction

The search for programs that raise student achievement, particularly among low-income and minority students, has become a national priority. The No Child Left Behind Act amendments to the federal education code raise the stakes of this search. Too often, however, new programs have been introduced with little prior analysis, only to be replaced after a few years when disillusionment over results set in.

In an ideal method for evaluation, two or more programs would be compared while holding all other factors constant. Unfortunately for researchers, many other factors can affect educational outcomes, including students, their families, their environment, their teachers, and the schools.

The most widely accepted model in educational research is one in which two or more groups of students are randomly selected from a larger population and given different programs. Any statistically significant differences between the outcomes are then credited to differences in the programs. The US Department of Education seems to have largely adopted this model in implementing the No Child Left Behind Act's mandate that research be scientifically sound. (Coalition for Evidence-based Policy, 2002)

Randomized experiments have proven difficult to implement in education. Cost and timing are obvious barriers, as administrators resist using time and resources for multiple programs, particularly if they have already had made up their minds as to which is the best. Teachers and parents may prefer one program over another and resist being assigned, or having their children assigned, to the one they consider less effective. Most educational programs do not lend

[1] The author wishes to acknowledge the generous and kind comments and advice of two anonymous, expert reviewers, plus Richard P. Phelps, who took responsibility for editing this article. Any errors remain the responsibility of the author.

themselves to the double-blind design common for drug testing, where neither the patient nor the doctor knows which is which.

Because of these and other obstacles, true randomized experiments are rare in education, although they will probably increase with the new federal requirements that programs be "scientifically based." These and other difficulties are reflected in the results posted on the What Works Clearinghouse (WWC) web site.[1] Among middle school mathematics programs, only eight studies met their standards for research validity. Several of those that survived the WWC screen have been criticized by others on the basis that they were sponsored by publishers or others with an interest in the outcomes. Similarly, an analysis sponsored by the National Research Council of 147 studies of mathematics programs concluded that the studies did "permit one to determine the effectiveness of individual programs with a high degree of certainty" (Mathematical Sciences Education Baord, 2004).

Because of barriers to random assignment of students, investigators often resort to some form of design that attempts to match existing groups of students in a way that minimizes the danger of bias. In these quasi-experiments, they attempt to match classes that are similar in such characteristics as income, ethnic mix, and prior educational success. Such studies are often treated with less respect than experiments because of the fear that one of the uncontrolled factors may influence the results.

The reliability of both educational experiments and quasi-experiments may suffer if participants have a stake in the outcome. Teachers and others, if they prefer one of the programs being compared, may feel an obligation to make it successful and work extra hard so their students achieve.

Both true experiments and quasi-experiments may suffer from selection bias. If parents consider one of the programs to be more desirable, they may lobby to have their children placed in it. As a result, one of the programs being compared may end up with a disproportionate number of children of more educationally-aware parents.

Method

In this paper, I describe the evaluation of program effectiveness by comparing school effectiveness ratings before implementation of a program with ratings of the same schools following implementation. This approach makes use of data already collected on schools, their student performance and demographics. This school effectiveness model is designed to avoid penalizing schools serving low-income populations.

As described in Thompson (2004) the school effectiveness ratings start by finding regression equations relating average school test scores to the poverty level of each school.[2] These regression equations are used to predict each school's test scores and to calculate the residuals, the differences between the actual and predicted results. Schools are rated by their standardized residuals (also called effect sizes or z-scores) averaged over all tests given in a year. The

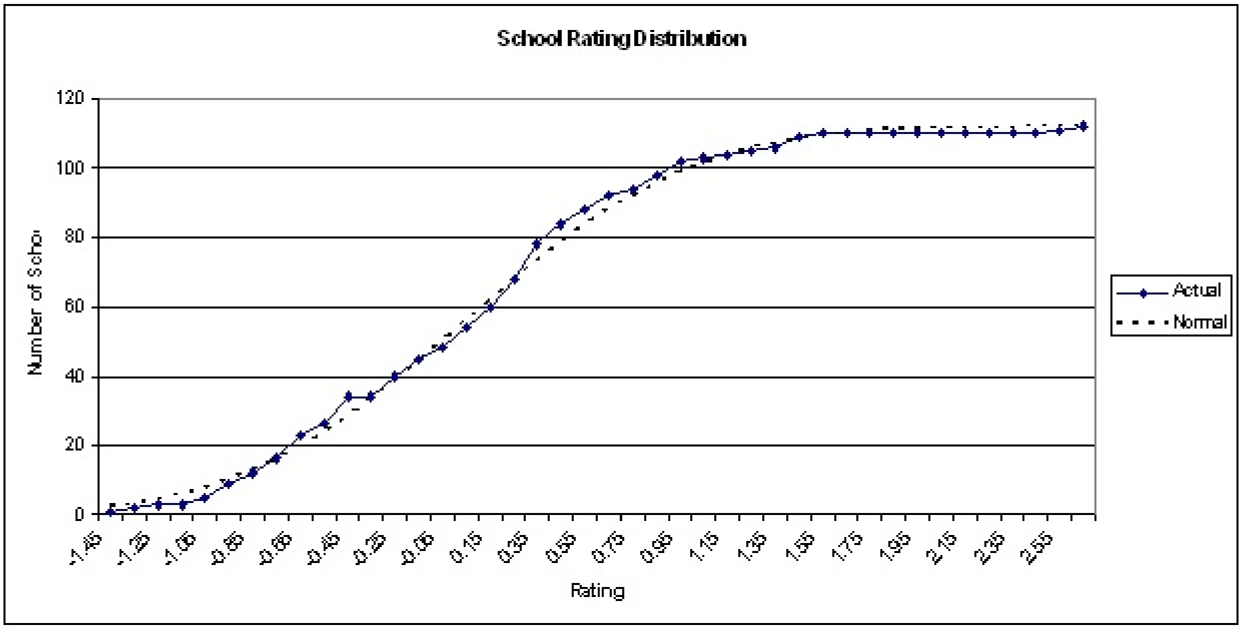

Figure 1. Distribution of School Ratings

resulting rating for each school incorporates all tests given at the school adjusted for the poverty level of the school.

The result is a set of school ratings, as shown in Figure 1. The ratings appear to follow a cumulative normal distribution with a mean around zero and a standard deviation of about .75.[3]

It should be emphasized that the school ratings are relative. A school's rating shows how it did, adjusted for poverty, compared to all other schools. Standing alone, the ratings say nothing about absolute changes in performance. Thus, a decline in a school's ratings could reflect poorer performance by the school, improved performance by the average school, or a combination.

While originally developed to identify high-performing schools for study and replication, this rating system also offers a tool for program evaluation under certain conditions. The hypothesis is that schools adopting an effective program will see a significant increase in their ratings over the course of time.[4] This approach depends on several conditions. First is that the program under study be introduced during a period for which ratings are available. Second, the number of schools introducing a program must be sufficient for statistically significant results. Finally and conversely, the program cannot be introduced in so many schools as to leave no comparison group.

Using this rating system to evaluate program effectiveness has several advantages over the more common approach of judging schools based upon one or two test scores. Because all achievement test scores given systemwide in grades three through five are incorporated into the ratings, the effect of random variation on individual tests is reduced. And because ratings are adjusted for poverty, it is not necessary to limit the comparison to schools at the same poverty level.

The ratings cover about 112 elementary schools in the Milwaukee Public Schools (MPS) starting in the 1996-97 school year.[5] Three programs met the criteria listed above: Target Teach, Direct Instruction, and the SAGE class-reduction program. Others, such as Robert Nash's Pure Phonics, were implemented within individual schools or classrooms, but not in sufficient numbers for statistical analysis.

For each of the schools, I calculated the average annual change in the rating using regression coefficients. From these, I averaged the changes for the schools involved in each of the three programs (and in some cases for subsets of those programs). I then examined whether those schools moved up, moved down, or stayed the same in the ratings.

To analyze the statistical significance of any change, I first calculated the standard deviation of the schools' changes. I then calculated the standard error of the change by dividing the standard deviation by the square root of the number of schools in each of the three programs.[6] I used the ratio of the change to the standard error to calculate the p-value of the change. Unless otherwise stated, I looked for statistical significance at a 95% or better confidence level.

Two kinds of errors can appear in such analysis: random error and bias. Random error appears because of variations in individual students. Bias appears if the two groups were selected in such a way that there would be differences in the changes between the two groups even without the programs under study.

Bias is much more dangerous than random error. Statistical tests are developed to measure and compensate for random error. With bias, there may be an apparent effect from a program whereas the real cause is some other difference between the two sets of schools.

Student mobility, for example, tends to mask the impact of a program, since many of the students would not have received the full benefit of the program. Thus mobility could make it harder to show a statistically significant effect. But only if mobility has a different impact on changes in schools in a program than those outside would it introduce bias. Similarly other factors unique to the Milwaukee school system, such as the school choice programs, are relevant only if they are likely to have a differential effect depending on the programs being considered.

The most likely source of bias in a study such as this is selection bias. Perhaps schools were chosen to participate because of abnormally low test scores which would have rebounded (regressed to the mean) even without any intervention. Conversely they may have been chosen because of abnormally high scores, which were unlikely to be maintained at that level.

Results

The Target Teach Curriculum Alignment Program

Target Teach is marketed by Evans-Newton, Inc., and aims at aligning curriculum with state tests (see Evans-Newton, 2005). During the 1998-99 school year 25 low-performing Milwaukee public elementary schools were selected for the Target Teach Five Step Program in Reading and Language Arts. The following year, an additional nine schools implemented Target Teach. MPS chose schools based largely on previous low test scores.

Evans-Newton credits the underlying concept of Target Teach to Fenwick English (2000), a critic of standardized testing who nevertheless advocates schools adopt a strategy to improve test scores. Relatively little research has been done on the effectiveness of Target Teach. In an unpublished draft, Ryder (2000) compared test scores at MPS Target Teach schools to a control group and found no improvement in the first year.[7] The Evans-Newton (2005) website lists three unpublished studies.

As described in an MPS summary (Washington, 2002), the Target Teach program encompassed five steps:

1. Identify the goal by prioritizing state and district goals.
2. Align the teaching and testing curricula.
3. Identify gaps in the instructional curriculum.
4. Determine objectives and benchmarks to periodically assess student mastery.
5. Monitor student progress using computer software.

The MPS summary further describes the goal as improving students' performance on the Wisconsin third and fourth grade reading tests. With that goal in mind, MPS examined the adopted reading texts and created a resource packet to fill the gaps (Washington, 2002).

According to the summary, teachers assessed student progress four times yearly using "short stories or reports that mirror the format of the state assessments" and multiple choice questions. A computer program generated reports on student progress "to drive instructional decisions, re-teach objectives, identify student strengths and weaknesses, identify instructional strengths and weaknesses, identify instructional strategies needed, and create flexible groups for skills instruction" (Washington, 2002).

Unlike with Direct Instruction, described in the next section, participation was mandatory. While there was some unhappiness at participating schools, it appears that they all implemented at least the basic elements of the program.

Table 1 shows results for the Target Teach schools. Average scores for these schools rose at the rate of .13 per year.[8] The standard error of these 34 schools is .03, giving a statistically significant p-value of .002. This group of schools started with ratings substantially below the average MPS school and had closed most of the gap by 2000-01.

Table 1. Change in Target Teach Schools

Program	No. of Schools	Annual Change	St. Error	p-value
Target Teach	34	0.11	0.03	0.002
Started 1998	25	0.13	0.04	0.002
Started 1999	9	0.04	0.07	0.260

These results should be interpreted cautiously. As discussed below, the selection process may have introduced bias. The improvement stems mainly from the 25 original schools.[9] However, the results are consistent with the hypothesis that the program does help student performance.[10]

Did the method of selecting schools bias the results in favor of Target Teach? The schools were chosen largely because poor test scores indicated help was needed in reading. To the extent that low scores in one year reflect random variation, some improvement could be expected absent any intervention as schools return to the mean. Such selection bias would be strongest if all schools randomly varied around the same mean. Picking the worst performing schools in one year would almost guarantee improvement in the next.

In practice, however, a number of factors intervene to minimize this effect. Most important, schools do not vary about the same mean. An examination of results over several years show that schools that are low tend to stay low; those high tend to stay high. Thus, the number of schools likely to enter the pool of those judged lowest due to random variation is limited.

The potential impact of selection bias on apparent school improvement can be estimated through simulation. The simulation model used assumed that each school has a true score, which the reported score varies randomly about. Because of this random variation in each year some schools whose true score would put them in the bottom group will show inflated scores that remove them from that group and other schools whose true score is slightly above the cutoff for the bottom group will have depressed scores that place them in the bottom. Thus it is useful to estimate the number of schools that could move into or out of the bottom group solely because of this random variation.

Based on more than 100 simulations, of the 34 schools with the lowest apparent scores in any year, five would be in this group only because of random variation. As a result, the average score of the 34 schools selected is .15 lower than given by the average of their base scores. If the schools returned to their base scores in the following years, the result of this rebound would be to inflate the apparent annual improvement by .03. This calculation suggests that the true annual improvement might be closer to .08 than the .11 shown in Table 1. While smaller, this gain is still significant (p=.016).

There are several considerations suggesting that this simulation represents a worst-case scenario. First it assumes schools were chosen solely because of poor ratings in the 1996-97 year. In practice, schools were chosen for their poor performance over several years. Second, only two test scores, the third and fourth grade reading tests, were use to select schools for assignment to Target Teach. By contrast, the school ratings in the present analysis used all ten scores reported in 1996-97.

If the Target Teach schools improved their scores over time, as appears from the discussion above, what was the mechanism? The Evans-Newton literature, as well as English (2000), implies that much, if not all, of the improvement comes from aligning curriculum to tests. To some critics, this approach can imply the simple substitution of material on the test for material not on the test, with no overall gain in learning. In this view, test scores improve simply because students have abundant practice with questions similar to those on the test.

If the gain results simply from alignment, one might expect tests in subjects not aligned to be unchanged or even decline if they received less attention. As mentioned, Target Teach aligned the curriculum to the third and fourth grade reading tests. MPS gave four other tests, in mathematics, language arts, science and social studies, to fourth graders in both 1996-97 and 2000-01. Table 2 shows a comparison of the ratings of the 34 Target Teach schools in these two years on the individual tests.

Table 2. Average Scores of Target Teach Schools on State Tests

	Grade 3	WSAS - Grade 4					
	Reading	Reading	Language	Math	Science	Soc Studies	Total Score
1996-97	-0.86	-0.51	-0.42	-0.50	-0.49	-0.59	-0.49
2000-01	-0.23	-0.13	-0.01	0.07	-0.03	-0.11	-0.07
Change	0.64	0.39	0.41	0.57	0.46	0.48	0.42

Despite the emphasis on alignment, Target Teach school results improved in all subjects. This suggests that factors other than alignment may be at work. Perhaps greater emphasis on reading helps students achieve in other subjects which depend on effective reading skills. Another possible explanation is that Target Teach's emphasis on measurement helped change the culture of schools so that teachers became more aware of student progress in a variety of subjects, not just those aligned. This explanation is consistent with an MPS study of high-performing schools (Milwaukee Public Schools, 2001) which described them as "data-driven," constantly using student assessment to drive instruction.

Direct Instruction

The term "direct instruction" is used generically to refer to programs that are highly structured, require specific student responses, are teacher-directed, and use phonics as the basis for reading. In Milwaukee, the term generally refers to the Engelmann Direct Instruction model distributed by

SRA/McGraw-Hill. Schug, Tarver, and Westen (2001) and White (2005) describe the program in Milwaukee in more detail.

Numerous observers have commented on the resistance to Direct Instruction among many educators. As the National Research Council's Committee on the Prevention of Reading Difficulties in Young Children commented, despite research suggesting "very positive results for the program, it has not been as widely embraced as might be expected (Snow, Burns, and Griffin[1998], page 176)." This resistance is reflected in the Milwaukee experience. In contrast to Target Teach, the first schools to adopt Direct Instruction did so largely on their own initiative without central office encouragement.[11]

Identifying the Direct Instruction schools was a challenge. There was no process of certifying that a school had an adequate implementation of Direct Instruction. In a survey, 45 Milwaukee elementary schools indicated they used Direct Instruction program for one purpose or another. In many cases its use was limited, however, focusing on particular grades or students judged learning disabled or at-risk. Some schools with school-wide implementation of Direct Instruction had nevertheless neglected important elements of the program, particularly adequate training, consultants, and coaches for teachers. (White, 2005, reported similar challenges.)

Based on this survey, as well as conversations with teachers, principals, and consultants, twenty-one schools that had implemented the program school-wide before 2000 were identified . Of these, ten appeared to have fully supported programs, with adequate training and coaches and consultants available to teachers. All these schools adopted Direct Instruction programs in reading and language. Some also added Direct Instruction modules in other areas, such as spelling or mathematics.

Table 3 shows the annual improvement in scores for those 21 schools. On average their scores increased by .06 per year. As shown by the p-value, this change is significant at the 90% level but not at the 95% level. For the ten schools with fully-supported programs, annual growth in ratings increased to .14. Despite the decreased sample size, statistical significance also increases. As Table 3 shows, the remainder show no improvement in ratings.

Table 3. Change in Direct Instruction School Scores

Program	No. of Schools	Annual Change	St. Error	p-value
Direct Instruction	21	0.06	0.04	0.084
Fully-supported	10	0.14	0.06	0.022
Other DI Schools	11	-0.01		

These results are consistent with research that finds Direct Instruction is effective in increasing student achievement--so long as it is well-implemented. In the words of one Direct Instruction supporter, "the missing coaching element is the slow death of D.I." The difference between the fully supported and the other Direct Instruction schools was significant with a p-value of .04. A school expecting a quick boost by using Direct Instruction materials as supplements may be disappointed.

To what extent could selection bias have affected the change in the direct instruction schools? In contrast to the other two programs examined here, these schools were self-selected. Their participation could indicate unusual initiative on the part of the principal or other staff members in the school. There seems to be no good way to measure any bias introduced as a result, but this suggests that results were Direct Instruction mandated, the result might not be as strong.

SAGE Class Size Reduction

The education research literature contains an ongoing debate about the impact of class size reduction on student achievement and whether spending money to reduce class size represents the best use of resources. Krueger (2000) summarizes the case for class size reduction and Hruz (2000) the case against. Bohrnstedt and Stecher (1999) also summarize studies of class size reduction.

To the extent that any consensus has emerged, it could be summarized as follows: children in classes of 15 gain more than their peers in larger classes in the first year of class reduction. The advantage is greater in mathematics than reading but is statistically significant in either case. Some studies show a greater advantage for low-income children and black children. In subsequent years, the small-class-size children maintain their advantage but do not increase it. This consensus stems mainly from studies of the Tennessee STAR program, usually considered the best available experiment (Ehrenberg, Brewer, Gamoran, and Willms, 2001).

Class size reduction is very popular with teachers and parents. Policy makers, on the other hand, recognize that its widespread implementation is very expensive and could cut into other programs and worsen teacher shortages. The California Class Size Reduction Consortium (Stecher & Bohrnstedt, 2002) found that state-wide implementation led to teacher shortages, particularly in schools serving low-income students, and to cutting back on other activities in order to cover the additional costs of low class size.

Starting in 1996, Wisconsin established a class reduction program for students in kindergarten through third grade under the acronym of SAGE (Student Achievement Guarantee in Education). Extra funding based on low-income student enrollment is given to schools that reduce class size ratios to 15 students to one teacher. Because of space constraints, in Milwaukee this ratio is often achieved by placing two teachers in a classroom with 30 students. Initially the number of schools allowed to participate was very limited. Starting with the 2000-2001 school year, the legislature removed the limits on participation, while keeping state funding based on enrollment of low-income students.

In Milwaukee seven schools started the program in the 1996-97 school year. The first cohort of SAGE students at these seven schools would have reached fifth grade in 2000-01. A additional seven Milwaukee schools started SAGE in 1998-99. Their students would have reached third grade in 2000-01.

Table 4 shows the average annual change in the SAGE schools' scores over time. The first group of schools shows a slight increase and the second a slight decrease. Neither change is statistically significant.

Table 4. Change in SAGE School Scores

Program	No. of Schools	Annual Change	St. Error	p-value
Sage-1996	7	0.03	0.08	0.356
Sage-1998	7	-0.02	0.08	0.376
All Sage	14	0.00	0.05	0.485

Can these results be reconciled with studies that find a statistically significant effect from reduced class sizes? First, it should be noted that comparing schools creates a substantially greater hurdle than the more common approach of comparing individual students or classes. A sample size of fourteen schools is much smaller than sample sizes of thousands when individual students are used.

The official study of the SAGE program was undertaken by a team at the University of Wisconsin-Milwaukee (UWM). This study includes students in the first group of Milwaukee schools as well as other schools around Wisconsin. This study is described in five annual reports (Maier, Molnar, Percy, Smith, & Zahorik, 1997; Molnar, Smith, & Zahorik, 1998; Molnar, Smith, & Zahorik, 1999; Molnar, Smith, & Zahorik, 2000; Molnar, Smith, Zahorik, Halbach, Ehrle, Hoffman, & Cross, 2001)

The UWM researchers administered the Comprehensive Test of Basic Skills (CTBS), Terra Nova edition, to students in the SAGE program and to a Comparison group of students. The battery included subtests in reading, language arts, and mathematics.

Table 5 shows the average cumulative advantage found in the UWM study for SAGE students compared to students in the Comparison group.[12] Overall, the UWM researchers found SAGE students gaining more than the Comparison students in first and second grades, then losing some of that advantage in third grade, nevertheless finishing ahead of the Comparison group. The SAGE advantage was greatest in mathematics, smallest in reading.

In an unpublished report, Smith and Kritek (1999) point out that while on average SAGE classrooms showed greater gains than Comparison schools some Comparison schools did better than many SAGE schools. They argue that while "small classes provide the opportunity for increased achievement," it is not clear "why some classrooms respond to this opportunity differently than others."

Table 5. Average Cumulative Gain in Scale Scores

Cumulative Gain	SAGE vs. Comparison		
	1st Grade	2nd Grade	3rd Grade
Reading	6.16	5.71	6.86
Language Arts	4.44	9.28	4.74
Mathematics	8.19	15.01	13.40
Total	6.31	9.61	7.61

In 2001-02, Milwaukee elementary students took the Terra Nova exam in reading, language arts, and mathematics in second grade through fifth grade. Fourth graders also took the Terra Nova science and social studies tests. I used the increases shown in Table 5 for the expected boost from SAGE in reading language arts, and math. I used the gains shown for Total scores in Table 5 to predict the expected gains on the science and social studies tests.

The UWM study did not address how well SAGE students perform after leaving the program at fourth grade. Using Tennessee STAR data, Finn (1998) found low-class-size students enjoyed a post-program advantage of .15 standard deviations. Because the average SAGE third grade advantage found by the UWM researchers was also around .15 standard deviations, I used the third grade gains shown from Table 5 to predict fourth and fifth grade gains.

Because of student turnover, some students in SAGE schools do not go through the full program. I used the average SAGE school mobility rate to estimate the percentage of SAGE students in each class. On average, 73% of the students at a school are still at the school one year later. For an upper estimate of improvement, I applied this mobility rate to fourth and fifth graders only, assuming students starting in second and third grade get the full benefits despite their late start. For a lower estimate, I applied the mobility factor to all grades.

For schools starting SAGE in 1998 I assumed no SAGE improvement in fourth and fifth grade test scores, as their first SAGE students would have only reached third grade in the spring of 2001. Finally, I assumed those scores on other tests would have increased proportionally, so that the total gain would reflect that calculated for the Terra Nova tests.

These calculations predict an average improvement in the ratings of SAGE schools between .25 and .42, for an average annual gain of .06 to .11. Table 6 shows a test of the null hypothesis that the improvement of SAGE schools was .06 or better and .11 or better, respectively.

At the lower limit, the null hypothesis cannot be rejected at a 95% confidence level. This raises the possibility that the divergence in results between my model and the UWM analysis could be explained by random error.

Another factor that might contribute to different results is that, except for the first group of seven Milwaukee schools, my analysis used different schools than the UWM study. As noted earlier, most MPS schools achieved their low student-teacher ratio by placing two teachers in a room with thirty students. Some class size reduction advocates argue that “large classes with two teachers are less likely to yield the same benefits” (Finn, 2002). Thus the two studies could be reconciled if the gains reported by the UWM researchers were due to larger gains at schools outside Milwaukee. Unfortunately, the UWM group has not published separate Milwaukee results. However, they do report gains by black students larger than those for all students. Given Wisconsin demographics, most of those students would be in MPS, undercutting this explanation.

Another possibility is selection bias. In the year before they entered the program, ratings for the first group of SAGE schools were already high, suggesting that those schools’ participation may have been a reward for outstanding past performance.

Changes in the test schedule could also affect the results. In 2000-01 the number of tests given in second and third grade climbed from one to seven, increasing the influence of students still within SAGE classrooms on the school ratings. However, this change should have helped the results for the SAGE schools, since they would have more heavily weighted by students currently within the program.

Conversely, some of the difference could stem from limitations in the UWM study. Gain calculations involve the process of taking the differences of differences. First, scores from one year are subtracted from those for the succeeding year to get a gain. Then the gains for the comparison group are subtracted from the gains for the SAGE group. As a result, small variations in test scores result in large variations in comparative gains.

The resulting variability is reflected in the gain comparisons given in each of the annual reports. There are significant variations in the gains shown from one report to another, even when the calculations are made for the same cohort of students.

Table 6 Predicted vs. Actual Improvement in Sage Schools

Limit	No. of Schools	Actual	Null	Difference	St. Error	p-value
Lower	14	0.00	0.06	-0.06	0.05	0.134
Upper	14	0.00	0.11	-0.10	0.05	0.036

So long as the student populations are the same, it should make no difference whether one first calculates the gain of each student and then averages those gains or whether one first calculates the average scores of all students at the beginning and end of a period and then takes the differences of those gains. Yet, the SAGE advantage is cut roughly in half when the second method is used compared to the first. Apparently the averages included students who were excluded from the differences, because they left the program or missed a test.[13] But it is not clear why those leaving should have hurt the SAGE scores relative to the Comparison scores.

The sensitivity of the results to inclusion or exclusion of a few students underlies the need to apply firm and consistent rules on which students to include. Otherwise, it is easy to unconsciously bias the results. Researchers likely must decide whether to include students with poor attendance, who temporarily transfer to another school, or whose health problems interfere with full participation. If such students are also doing poorly academically, the researcher may be tempted to exclude them from the study.

The study also had some difficulty maintaining its Comparison group (Molnar, et.al., 2000, page 13). Lacking any incentive to participate, several Comparison schools dropped out. In fact, two Comparison Milwaukee schools were among the seven schools starting SAGE in 1998.[14] This result underlines the concern that the experiment may have been biased if teachers in the SAGE schools were more motivated to show good results than those in the Comparison schools.

The UWM study collected valuable data on the effect of class-size reduction and perhaps other early intervention strategies. Because of strong political opposition to standardized testing of young children, there is often resistance to collecting consistent data on early learning. Yet the UWM researchers were able to collect consistent test data on several thousand children starting at the beginning of first grade. It seems desirable, therefore, that other researchers examine the UWM data and the assumptions made.

Discussion

In this section, I make some final observations, first about the three programs analyzed and then about the usefulness of this approach to assess educational policies and programs.

The Target Teach results support the value of encouraging teachers to systematically measure student progress using standardized tests allowing for comparison of their students with a control group. The apparent benefits in subjects beyond those covered by the program underlines the benefits of using data to drive education.

The study joins a large and growing body of research indicating a positive relationship between Direct Instruction and student achievement. In addition, it underlines the importance of the quality of its implementation. I would take this further and recommend that the federal government require that well-implemented Direct Instruction be used as the control program for evaluations of the effectiveness of other educational programs. In all too many studies purporting to show a program is effective, little or no information is given on the program or programs used for students in the control group.

This study suggests a need for a more nuanced and less ideological approach to class size reduction, recognizing that reducing classes may take resources from other educational initiatives. More needs to be done to identify how teaching must change to take advantage of low class sizes and which students will best benefit.

More generally, this study suggests a broadening of the approaches to research on educational effectiveness. Currently, the preferred approaches are experimental and quasi-experimental comparisons of a program with some comparison. Often the analogy to double-blind studies in medicine is used to support this position. Yet even in medicine, many major questions do not lend themselves to this approach, notably looking at the connection between health and such lifestyle decisions as diet, smoking, and exercise.

Particularly motivated by the No Child Left Behind act, school systems are collecting massive amounts of information on student achievement and demographics. So far, it appears that little is being done to mine these data to learn more about what works in education. The present study suggests one way to make use of all these data.

Often school systems introduce new programs without an accompanying evaluation system. Under the right circumstances, models such as the one used here allow an *ex post facto* evaluation of program effectiveness. Compared to educational experiments, a clear advantage is cost. This study used data that was already collected and all analysis used common spreadsheets. By contrast, educational experiments may be deferred due to their high cost and heavy burden on teachers and students, who often must change their behavior especially if they are assigned to the treatment they do not prefer. By contrast, this analytical approach required no change in behavior.

A final advantage is flexibility. The rating system used here is able to easily accommodate new tests or changed reporting systems. By contrast, experiments often require a particular measurement scheme and can be derailed by a change in district or state policy.

Along with these advantages come limitations and potential pitfalls. As mentioned earlier, it is desirable that the number of schools implementing a program be sufficiently large for statistical analysis, but not so great as to leave no control group. For example, this approach would not be useful in measuring class size reduction in Milwaukee now that all schools are able to participate in the SAGE program. A possible danger is biases introduced by the testing scheme itself. In Milwaukee, fourth grade is a heavily tested year. Thus this approach may be biased toward programs that target fourth graders and miss programs that have more effect at another grade. Likewise changes in the years or subjects emphasized by tests may create an apparent change in school performance where none exists. As with all program analysis, there is the constant risk that some outside change may bias the results.

With these caveats, the use of a value-added model for program evaluation joins a useful array of tools that helps understanding. Results that confirm other research, as with Direct Instruction, may encourage educators to try a program. Where the results conflict with some other research, as with class size, they may encourage further exploration or perhaps a redefinition of the issue. Where there is little other research, as with Target Teach, results may encourage further exploration.

But to be most useful the data already being collected on student achievement and demographics should be joined by reliable data on schools practices. Currently there is no systematic tabulation of programs or other educational attributes of schools and classrooms. In a few cases, it is relatively straight-forward to get a list of schools participating in a program, as here with Target Teach and SAGE in this study. But the difficulty in getting a reliable list of Direct Instruction schools is probably more typical of the situation in most cases. Every time a school changes its math or reading program, there is a lost opportunity to measure a change in results, if such changes are not systematically recorded.

Could districts collect data that categorizes the practices of the schools? Here is where I think qualitative analysis might make a contribution. A knowledgeable and neutral analyst visiting classrooms and recording their approach would help immensely in teasing out the connections between practice and student achievement.

References

Bohrnstedt, G. W. & Stecher, B. M. (1999) *Class Size Reduction in California: Early evaluation findings, 1996-98*. Palo Alto, CA: American Institutes for Research

Coalition for Evidence-based Policy (2002). *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User Friendly Guide*. Washington: U.S. Department of Education. Available at http://www.ed.gov/rschstat/research/pubs/rigorousevid/index.html .

Ehrenberg, R. G.; Brewer, D. J.; Gamoran, A.; and Willms, J. D. (2001), Class Size and Student Achievement. *Psychological Science in the Public Interest*. American Psychological Society.

English, F. W. (2000) *Deciding What to Teach and Test: Developing, Aligning, and Auditing the Curriculum*. Newbury Park CA: Corwin Press, Inc.

Evans-Newton, Inc. (2005) Website, http://www.evansnewton.com/home.htm

Finn, J. D., (1998) Class size and students at risk: What is known? What is next? Washington, DC: U.S. Department of Education.

Finn, J. D., (1998) Class size reduction in grades K-3. In Molnar, A. (Ed.) *School Reform Proposals: The Research Evidence.* Tempe AZ: Education Policy Studies Laboratory, Arizona State University.

Heywood, J. S., Thomas, M., & White, S.B. (1997). Does Classroom Mobility Hurt Stable Students? An Examination of Achievement in Urban Schools. *Urban Education*, 32, 354-372.

Hruz, T. (2000) *The Costs and Benefits of Smaller Classes in Wisconsin: A Further Evaluation of the SAGE Program*. Thiensville WI: Wisconsin Policy Research Institute, Inc., available at www.wpri.org

Krueger, A. (2000) *Economic Considerations and Class Size, Working Paper #447.* Princeton NJ: Princeton University Industrial Relations Section, available at www.irs.princeton.edu/pubs/working_papers.html

Maier, P., Molnar, A., Percy, S., Smith, P., Zahorik, J. (1997). *The 1996-1997 Evaluation Results of the Student Achievement Guarantee in Education (SAGE) Program.* Milwaukee WI: University of Wisconsin-Milwaukee. *Available:* http://www.uwm.edu/Dept/CERAI/sage.html

Mathematical Sciences Education Board, Committe for a Review of the Evaluation Data on the Effectiveness of NSF-Supported and Commercially Generated Mathematics Curriculum Materials (2004), On evaluating curricular effectiveness: judging the quality of K-12 mathematics Evaluation. Washington: National Academies Press.

Milwaukee Public Schools, Department of Curriculum and Instruction (2001). *Revised Elementary School Survey, 2000/2001* Author

Milwaukee Public Schools, Dept. of Research. (1998). *Characteristics of Effective Schools: An Analysis of Eight High Performing Elementary Schools*. Milwaukee WI: Author.

Molnar, A. (1998). *Smaller Classes Not Vouchers Increase Student Achievement*. Harrisburg PA: Keystone Research Center.

Molnar, A., Smith, P., & Zahorik, J. (1998). *The 1997-1998 Evaluation Results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee WI: University of Wisconsin-Milwaukee

Molnar, A., Smith, P., & Zahorik, J. (1999). *The 1998-1999 Evaluation Results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee WI: University of Wisconsin-Milwaukee

Molnar, A., Smith, P., & Zahorik, J. (2000). *1999-2000 Evaluation Results of The Student Achievement Guarantee In Education (SAGE) Program*, *CERAI-00-3*4. Milwaukee WI: University of Wisconsin-Milwaukee. *Available:* http://www.uwm.edu/Dept/CERAI/sage.html.

Molnar, A., Smith, A., Zahorik, J., Halbach, A., Ehrle, K., Hoffman, L., & Cross, B. (2001). *2000-2001 Evaluation Results of The Student Achievement Guarantee In Education (Sage) Program*. Milwaukee, WI: School of Education, University of Wisconsin—Milwaukee. *Available:* http://www.uwm.edu/Dept/CERAI/sage.html

Raynor, K., Foorman, B.R., Perfetti, C.A., Pesetsky, D., and Seidenberg, M.S. (2001). How Psychological Science Informs the Teaching of Reading. *Psychological Science in the Public Interest,* 2, 31-74.

Ryder, R. J. (2000). *Milwaukee Public Schools Reading Assessment*, Unpublished draft, February, page 50.

Schug, M., Tarver, S., Westen, R (2000). *Direct Instruction and The Teaching of Early Reading: Wisconsin"s Teacher-Led Insurgency*. Thiensville WI: Wisconsin Policy Research Institute.

Smith, P.L., and Kritek, W. (1999). *The Effects of Class-size on Achievement: a Closer Look at Conventional Wisdom*. Unpublished report. Milwaukee: Institute for Excellence in Urban Education, University of Wisconsin-Milwaukee School of Education.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children Washington*, DC: National Academy Press.

Stecher, B. M. & Bohrnstedt, G.W. (2002). *Class Size Reduction in California: Findings from 1999-00 and 2000-01*. Palo Alto, CA: American Institutes for Research.

Thompson, B. R. (2004). Equitable Measurement of School Effectiveness, *Urban Education*, Mar 2004; 39: 200 - 229.

Washington, D. R. (2001). *Target Teach Reading Alignment Program*. Milwaukee WI: Milwaukee Public Schools.

White, S. (2005). *Education That Works in the Milwaukee Public Schools: The Benefits from Phonics and Direct Instruction.* Thiensville, WI: Wisconsin Policy Research Institute. Available at: http://www.wpri.org/Reports/Volume18/Vol18no4.pdf.

Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. (2000). Smaller Classes, Better Teaching? Effective Teaching in Reduced-Size Classes. In S.W.M. Laine & J.G. Ward (Eds.) *Using What We Know* (pgs 53-73). Oak Brook IL: North Central Regional Educational Laboratory.

Endnotes

[1] The research is at http://www.whatworks.ed.gov/

[2] The earlier study examined a number of possible independent variables in addition to poverty, including student mobility and the ethnic mix of the school. Multiple regression using all available variables did not notably improve the fit compared to poverty alone, mainly reflecting high correlations between the variables. As discussed in the earlier study the percentage of students qualifying for free and reduced lunch is not a perfect measure of poverty, but it is the best available.

[3] To test the normal distribution hypothesis, cumulative scores were superimposed on the normal cumulative curve and a chi-square test was run. The mean can differ slightly from zero if some of the schools used to calculate the mean in any particular year are not included in the comparison, because they are missing scores for one of the years being compared. For each test individually the standard deviation is exactly 1, but when results are averaged the standard deviation declined to about .75. This is still high compared to what would be expected if the scores were randomly distributed by school and reflects the correlation between scores within a school. In a year with 10 tests for example the expected standard deviation would be closer to .31 if scores were randomly distributed among schools.

[4] In statistical terms, an increase in ratings is the alternative hypothesis. The null hypothesis is that ratings stay the same or go down.

[5] The actual number varied slightly from one year to another.

[6] An argument could be made that these data represent populations–all the schools in the three programs–so that tests of significance are not relevant. Our interest, however, is in what the results from these schools may suggest about a broader (hypothetical) population of schools potentially adopting these programs, so I would argue that tests of significance supply useful information.

[7] It is unclear how Ryder chose his control group. On average, the scores for the control group were somewhat higher than the Target Teach school prior to the start of the program.

[8] In these and the following calculations, the standard deviation used was calculated from the average of the standard deviations over time of all schools. This value likely overestimates the true random variation, since some of the change from one year to the next is likely to reflect underlying changes in school performance.

[9] The second group of schools had a significant dip in their ratings for the 1997-98 year. To the extent that this dip reflected a decline in reading scores, it was probably the most important factor in selecting those schools for the program. The baseline year for this study was 1996-97. To the extent that the dip reflected not an abnormal low in the second year but an abnormal high in the first, that could help account for the relatively weak performance of this second group.

[10] More precisely stated, the null hypothesis that there is no effect can be rejected.

[11] Since then, MPS support for Direct Instruction has grown and the number of schools adopting it has increased substantially.

[12] I calculated these values from the numbers given in the reports for Persisters, those students that stayed in the schools from the first-grade pre-test through the third-grade test. The values are the differences in Terra Nova scale scores gains. The UWM study reports on three cohorts of students, starting first grade in the fall of 1996, 1997, and 1998. I averaged the gains for all three groups. The gains calculated for all students are substantially less than those for persisters.

[13] Betsy Ann Schoeller, 21 Feb. 2000, Re: SAGE Report [Internet, e-mail to the author]; Phil Smith, 28 Feb. 2000, SAGE Report [Internet, e-mail to the author]; Phil Smith, 16 March 2000, SAGE Report [Internet, e-mail to the author].

[14] Questions have also been raised as to the desirability of having a principal investigator (Molnar) who headed the task force that originally proposed the SAGE program and is a well-known advocate of class-size reduction.