

Running head: EVALUATING THREE PROGRAMS

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

Bruce Thompson

Rader School of Business

Professor of Management

Milwaukee School of Engineering

(414) 277-7378

FAX: (414) 277-7479

thompson@msoe.edu

Abstract

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

Evaluating Three Programs Using a School Effectiveness Model:

Direct Instruction, Target Teach, and Class Size Reduction

Introduction

The search for programs that raise student achievement, particularly among low-income and minority students, has received increasing attention. Recent 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 set in.

An ideal program evaluation would entail sending identical students through different programs and comparing the outcomes. Unfortunately for researchers, identical students are not available and students are changed by education. Thus even when outcome measures are well-defined, it is often difficult to assess whether differences in outcomes stem from program differences or from student differences.

Lacking identical students, many researchers consider randomized experiments the ideal research model. In these experiments, students are assigned randomly to programs under study. Progress among the participating students is compared to that of the nonparticipant, or control, group. Ideally, students in the two groups are similar except for their participation in the program.

In practice, randomized experiments are often very difficult to implement effectively. They can be quite costly. Experiments may run into political resistance from opponents who decry using students as guinea pigs. If teachers or parents view a program as desirable, they may lobby to be

included and, to the extent successful, bias the experiment. Researchers' desire for randomization may conflict with parental choice in education.

Timing can also be a barrier. Administrators, under pressure to improve performance, may be reluctant to delay implementing a program until an experimental plan can be developed, funded, and implemented.

Because of these and other obstacles, true randomized experiments are rare in education. Even when well-designed in concept, they may suffer in implementation. As a result, those who disagree with the results can usually find reasons to reject them.

Experiments can fall victim to unexpected events. A principal may leave, a school's curriculum may change its focus, or the local economy may decline, forcing program cuts. Similarly, students may have experiences that affect their results. Researchers are then faced with deciding whether to include the results from these schools and students.

Experiments may also suffer from the fact of their being experiments. Teachers and others, particularly if they like the program under investigation, may feel an obligation to make it successful and work extra hard so their students achieve.

Finally, there may be a problem scaling up the results from individual students and classrooms to predict overall school performance. Much current concern centers about the typical low average performance of students in urban schools. Theoretically if a program shows positive results with

individual students, it should be possible to use those results to project how much a school's results would improve if it were to adopt the program. But this scaling up is seldom done.

Thus, in many cases, it is not possible to call on randomized experiments to evaluate programs.

And when available, the results may be open to interpretation.

### Method

In this paper, I propose an alternative approach to program evaluation. It makes use of data already collected on schools, their student performance and demographics, using a school effectiveness model designed to not penalize schools serving low-income populations.

My model starts by finding regression equations relating average school test scores to the poverty level of each school. It then uses these regression models to predict each school's test scores and calculates the residuals. Schools are rated by their standardized residuals (also called effect sizes or z-scores) averaged over all tests given in a year. (For more information on this model see Thompson, 2002.)

While I originally developed this model to identify high-performing schools for study and replication, the model also offers an alternative approach to 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.<sup>1</sup> This approach depends on several conditions. First is that the program under analysis be introduced during the 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 control group.

I have developed ratings for about 112 elementary schools in the Milwaukee Public Schools (MPS) covering the five-year period from 1996-97 to 2000-01. I identified three programs that meet 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 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. 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.

## 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, 2001). During the 1998-99 school year 25 low-performing elementary schools adopted 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. In his writings, English (2000) laments the loose coordination between what is taught and what is assessed. He advocates improving the match between curriculum content and test content. It appears that Target Teach has been the subject of little research on its effectiveness. A search of the ERIC educational database turned up no references to the program. 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.

As described in an MPS summary (Washington, 2002), the Target Teach program encompasses 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 programs 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, including "short stories or reports that mirror the format of the state assessments" (Washington, 2002).

According to the summary, teachers assess student progress four times yearly using short stories or reports and multiple choice questions. A computer program generates 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).

Table 1 shows results for the Target Teach schools. Average scores for these schools rose at the rate of .13 per year. I calculated the standard error of these 34 schools to be .03, giving a statistically significant p-value of .002. This group of schools started with ratings substantially

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

below the average MPS school and had closed most of the gap by 2000-01.

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.<sup>2</sup> However, the results are consistent with the hypothesis that the program does help student performance.<sup>3</sup>

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 returned 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 importantly, schools do not vary about the same mean. 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 schools judged lowest due to random variation is limited.

To measure the potential impact of selection bias on apparent school improvement, I simulated the ratings. I first constructed a distribution of 112 school scores meant to approximate the base distribution using the mean and standard deviation of the actual scores. I superimposed an error function that varied randomly using the average standard deviation of school scores across the years. I calculated average ratings over 100 simulations.

If a group of 34 schools is chosen with the lowest scores, on average five schools appear on the list which would not appear if only the base scores had been used, displacing five other schools. 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 tests, the third and fourth grade reading tests, were used to select schools. School ratings used in my analysis used ten tests 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) imply 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. Does Target Teach bring about an improvement in the measure of learning, the test, while neglecting the underlying learning being measured?

If the gain results simply from alignment, one would expect tests of other subjects that were not aligned to be unchanged or even decline if they received less attention. As mentioned, Target

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

	Grade 3	WSAS - Grade 4					Total Score
	Reading	Reading	Language	Math	Science	Soc Studies	
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

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.

Despite the emphasis on alignment, Target Teach school results improved in all these subjects. This suggests that factors other than alignment may be at work. Perhaps greater emphasis on reading helps students achieve in subjects such as science 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. An MPS study of high-performing schools (Milwaukee Public Schools, 2001) 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, et.al. (2001) describe in more detail the program in Milwaukee.

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, et.al. [1998], page 176)." This resistance is reflected in the Milwaukee experience. In contrast to Target Teach, schools adopt Direct Instruction largely on their own initiative without central office encouragement.

In a survey 45 Milwaukee elementary schools indicated they used Direct Instruction program for one purpose or another. In many cases, however, the use was limited, for example focusing on certain grades or students judged learning disabled or at-risk. Some schools which had implemented throughout the school had neglected important elements of the program, particularly adequate training, consultants, and coaches for teachers.

Based on this survey, as well as conversations with teachers, principals, and consultants, I identified 21 schools that had implemented the program school-wide before 2000. Of those 21, ten appeared to have fully supported programs, with adequate training and coaches and consultants available to teachers.

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		

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, the annual growth in ratings increases to .14. Despite the decreased sample size, the statistical significance also increased. As Table 2 shows, once the ten are removed the remaining eleven schools show no growth in ratings.

These results are consistent with research that identifies Direct Instruction as effective in increasing student achievement. They also suggest the importance of effective implementation. In the words of one Direct Instruction supporter, “the missing coaching element is the slow death of D.I.” A school expecting a quick boost by using Direct Instruction materials supplements may be disappointed.

### 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) and Hruz (2000) summarize the research from opposing viewpoints. Bohrnstedt and Stecher (1999) 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 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).

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

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 usually 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. State funding is still based on the enrollment of low-income students.

Seven Milwaukee 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 second seven 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.

Can these results be reconciled with studies that find a statistically significant effect from reduced class sizes? First, it should be noted that my approach of comparing schools creates a substantially greater hurdle than the more common approach of comparing individual students or

classes. The sample size using 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.<sup>4</sup> Overall, the UWM researchers found SAGE

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

students gaining more than the Comparison students in first and second grades, then losing some

of that advantage in third grade, to nevertheless finish ahead of the Comparison group. The SAGE advantage was greatest in mathematics, smallest in reading.

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.

Table 6 Predicted vs. Actual Improvement in Sage Schools

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

At the group of schools starting SAGE in 1998 I assumed no SAGE improvement in fourth and fifth grade test scores, as their first SAGE students would have reached third grade in the spring of 2001. Finally, I assumed that scores in other tests would have also 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.

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, the two studies employed different schools. As noted earlier, most MPS schools achieved the 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 that possible explanation.

It is possible that small sample size, nonrandom selection of schools, or any of the many other changes going on simultaneously in the schools and their student populations could swamp the SAGE effects. Before starting the program, the first group of SAGE schools already scored high, suggesting that there was something unusual about those schools. 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, if anything this change would improve the results for the SAGE schools.

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.

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. In some cases, the averages may have included students who were not included in the differences, because they left the program or missed a test.<sup>5</sup> But it is not clear why those leaving would 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.<sup>6</sup>

The UWM study collected very 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 duplicate the results.

### 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.

My study represents one of the first attempts, perhaps the first, to measure the effectiveness of the Target Teach program. The results indicate a positive relationship between that program and student achievement. They further suggest that the impact is not simply the result of curriculum alignment.

My 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 implementation.

However, this study does not show a positive relationship between reduced class size and school-wide achievement. While reduced class sizes may be helpful to student achievement, they are not a panacea. Many other changes in schools can swamp its effect. Particularly with the squeeze on resources, it may be more productive to target specific students or classes where reduced size brings significant benefits.

As school systems collect more information on student achievement and demographics, they are likely to recognize the advantages of building so-called “value added” ratings systems to rate schools while controlling for student characteristics, similar to the model used in this analysis. One advantage of these models is that they can incorporate all student performance data, not just one or two tests.

Often school systems introduce new programs without an accompanying evaluation systems. Under the right circumstances, value-added models 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.

Related is the burden on teachers and students. Experiments often require them to change their behavior. They may be assigned to a treatment they do not prefer. By contrast, this analysis required no change in behavior.

A final advantage is flexibility. This model 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 probably not be useful in measuring class size reduction in Milwaukee now that all schools may 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 towards 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.

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.

### **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
- 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. (2001) Website, [www.evansnewton.com/about/about.htm](http://www.evansnewton.com/about/about.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 School Reform Proposals: The Research Evidence, Molnar, A., ed. 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](http://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](http://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>

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-00-34. 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>
- 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.

- 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. (2002) Equitable Measurement of School Effectiveness, Milwaukee: Milwaukee School of Engineering. Manuscript submitted for publication.
- Washington, D. R. (2001), Target Teach Reading Alignment Program Milwaukee WI: Milwaukee Public Schools.
- Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. Smaller Classes, Better Teaching? Effective Teaching in Reduced-Size Classes, in Using What We Know by North Central Regional Educational Laboratory.

### Endnotes

<sup>1</sup> Technically, an increase in ratings is the alternative hypothesis. The null hypothesis is that ratings stay the same or go down.

<sup>2</sup> 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.

<sup>3</sup> More precisely stated, the null hypothesis that there is no effect can be rejected.

<sup>4</sup> 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, those first grade starting 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.

<sup>5</sup> 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].

<sup>6</sup> Questions have also been raised as to the desirability of having a principal investigator who headed the task force that originally proposed the SAGE program and is a well-known advocate of class-size reduction (Molnar, 1998). In the interest of full disclosure, I should note that while president of the Milwaukee School Board, I lobbied the governor and state legislators for expansion of the SAGE program. (Successfully, I might add.)