February 9, 2004

Participation in the Student Achievement Guarantee in Education (SAGE) Program and Performance on State Assessments at Grade 3 and Grade 4 for Three Cohorts of Students—Grade 1 Students in 1996–97, 1997–98, and 1998–99

Norman L. Webb and Robert H. Meyer
with Adam Gamoran and Jianbin Fu

Wisconsin Center for Education Research
School of Education University of Wisconsin–Madison

Participation in the Student Achievement Guarantee in Education (SAGE) Program and Performance on State Assessments at Grade 3 and Grade 4 for Three Cohorts of Students—Grade 1 Students in 1996–97, 1997–98, and 1998–99

February 9, 2004

Norman L. Webb and Robert H. Meyer
with Adam Gamoran and Jianbin Fu

Wisconsin Center for Education Research

The research in this report was supported by the Wisconsin Department of Public Instruction via Purchase Order # PAC00000173 to the Wisconsin Center for Education Research, School of Education, University of Wisconsin. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the supporting agencies.

Table of Contents

Acknowledgements

Cecilia Pohs	Researcher
Margaret Powell	Editor
Lynn Lunde	Program Assistant

We appreciate the assistance we received from Philip Smith and John Zahorik, University of Wisconsin–Milwaukee, in helping us to understand the structure of the existing database. Janice Zmrazek, Wisconsin Department of Public Instruction, was very helpful in providing us with information about the SAGE program. Jean Whitcomb and other DPI staff from the Reading and Student Achievement/Office of Educational Accountability were cooperative in providing the data on the Grade 3 Wisconsin Reading Comprehension Test (WRCT) and the Grade 4 Wisconsin Knowledge and Concepts Examination (WKCE).

Executive Summary

This is the report of an analysis to determine the sustained effect of reduced class size three to four years after students began the Student Achievement Guarantee in Education (SAGE) program as kindergarten or 1st-grade students. In addition, the analysis produced data on the cumulative effect on student performance of the number of years participating schools spent in the SAGE program. Data used in the analysis were collected by researchers at the School of Education of the University of Wisconsin–Milwaukee over a five-year period from 1996 to 2001 and were acquired from the Wisconsin Department of Public Instruction for the grade 3 Wisconsin Reaching Comprehension Test (WRCT) and the grade 4 Wisconsin Knowledge and Concepts Examination (WKCE). Data records from the three data sets were matched and combined for three cohort of students—those students who were in grade 1 in 1996–97 (Cohort 1), students who were in grade 1 in 1997–98 (Cohort 2), and students who were in grade 1 in 1998–97 (Cohort 3). Some students in Cohorts 2 and 3 began the SAGE program in kindergarten. The achievement data used in the analysis included Comprehensive Test of Basic Skills (CTBS) TerraNova scores in reading, language arts, and mathematics for Fall grade 1, Spring grade 1, Spring grade 2, Spring grade 3; WRCT grade 3 scores; and WKCE grade 4 scores in reading, language arts, and mathematics. The experimental group included the 30 original SAGE schools and 10 to 17 comparison schools each year.

From grade 1 through grade 3, 40% to 50% of the students in the experimental schools left during the course of the study. With some effort to overcome missing data, we were able to successfully match about 80% of the students in the three cohorts to the grade 3 WRCT and the grade 4 WKCE scores. Once students left a SAGE school or a comparison school, very few returned to the same school or moved among any of the study schools. Thus, variation among students by years in SAGE was limited to students with consecutive years of enrollment—those in grade 1 only, those in grade 1 and 2 only, and those in grades 1, 2, and 3. The relatively few students who joined the SAGE program after the fall of grade 1—late entrant students—were excluded from the analysis because they lacked a pretest score.

Focusing on the pool of students in all three cohorts who stayed in the same school from grade 1 to grade 4, we found a positive and statistically significant effect of the SAGE program for all three content areas through the end of grades 1, 2, and 3. This confirms the findings from the UW–Milwaukee study of a statistically significant gain by SAGE students, compared to the contrast group over grade 1, that was sustained over the next two grades (Smith, Molnar, & Zahorik, 2003).

We found conflicting results when we considered the effect on student performance on the grade 4 WKCE. There was no cumulative effect for the SAGE program on the grade 4 WKCE for the three content areas—reading, mathematics, and language arts. The effect detected for the SAGE program was not statistically different from zero. This could imply that the effects observed in the earlier grades, 1st grade to 3rd grade, were not sustained. However, when we considered the pool of students remaining in the same schools and their performance on the grade 3 WRCT, we also found no cumulative effect for the SAGE program. Thus, with the same group of students, those in grade 3 and in the same school year, the results on the grade 3 WRCT and the grade 3 TerraNova reading produced conflicting results on the effectiveness of the SAGE program. The WRCT results indicated no significant effect, whereas the TerraNova results indicated significant cumulative effects. These inconsistencies indicate that with the available data, the true effects of the SAGE program cannot be determined with any sufficient degree of certainty. There are other factors contributing to the results on the tests than just the SAGE program. We were unable to control for all of these factors, and with the relatively large standard errors, we were unable to determine whether or not there were sustaining effects that could be attributable to the SAGE program.

The absence of significant differences on the state assessment instruments when strong SAGE effects were seen on instruments administered for the experiment is puzzling. We do not rule out the possibility that there is an effect due to the program, but only conclude that this would be less likely. The statistical model used to compute the effect of the SAGE program addressed school effects along with program effects. The large variation among schools produced large standard errors, increasing the difficulty of achieving statistically significant results. However, grade 3 students in the SAGE program who took the WRCT in February did not vary greatly from grade 3 students in the contrast schools, but did show a significant difference in May on the CTBS TerraNova reading scale. One possible explanation is that students in the two types of schools had different incentives for performing well on the two assessments. Teachers and others in the SAGE program had a greater incentive than those in the contrast schools for students to perform well on the May test, which was given to study the effectiveness of the program, whereas students in both groups had equally high incentives to do well on the WRCT. There also could be alternative explanations that are not as apparent. This analysis did not take into consideration the different approaches schools and teachers used to reduce class size. Such an analysis may produce evidence that some approaches are more effective than others for increasing performance on the state assessments; it may also provide evidence that when the SAGE program is administered in specific ways and accompanied by professional development, it does indeed give students in primary grades an advantage.

The data do raise the possibility that the SAGE program has a more positive effect on African American students than on White students. Although the 4th-grade WKCE data did not show a significant difference between African American students and White students, the cumulative differences were in the direction that suggests the possibility of a greater impact of the SAGE program on African American students. The same finding was not evident when students from low-income families were compared to students from high-income families.

The number of schools included in this analysis was limited to the 30 original SAGE schools and a small contrast group. With fewer than 50 schools in the sample, this study cannot be considered the definitive study on class size in Wisconsin. It does raise questions about the sustainability of the effects of the program and counters some of the enthusiasm from previous studies. However, only with additional investigations can we produce more solid findings on the efficacy of the SAGE program.

Participation in the Student Achievement Guarantee in Education (SAGE) Program and Performance on State Assessments at Grade 3 and Grade 4 for Three Cohorts of Students—Grade 1 Students in 1996–97, 1997–98, and 1998–99

Norman L. Webb and Robert Meyer
with Adam Gamoran and Jianbin Fu

Introduction

In June, 1995, the Student Achievement Guarantee in Education (SAGE) program was signed into Wisconsin law. The SAGE program is intended to help improve student academic achievement by requiring that participating schools do the following:

- Reduce the student-teacher ratio in classrooms to 15:1 in grades K–3;
- Remain open extended hours (creating "lighted schoolhouses");
- Develop rigorous academic curricula; and,
- Implement plans for staff development and professional accountability.

During 1995–96, any school district in Wisconsin that had at least one school serving 50% or more children living in poverty was eligible to apply for participation in SAGE. One school (with an enrollment of at least 30% or more children living in poverty) in each eligible district could participate. In Milwaukee, up to ten schools could participate.

In September, 1996, the SAGE program was implemented in 30 schools in 21 school districts throughout Wisconsin. SAGE implementation was phased in over three years: K–1 in 1996–97, K–2 in 1997–98, and K–3 in 1998–99. SAGE served kindergarten through 3rd–grade students in 78 schools in 46 school districts in the 1999–2000 school year. Nearly 500 additional schools joined the program in the 2000–2001 school year. Participating districts received $2,000 per student and were required to meet specific "contractual" requirements with the Department of Public Instruction.

In 1996, SAGE was created as a five-year pilot program with the requirement that participating school districts take part in an annual evaluation. This evaluation was conducted by the Center for Urban Initiatives and Research and the School of Education at the University of Wisconsin–Milwaukee. Evaluation of the first three years of the program showed that students in the SAGE schools scored significantly higher than students in the contrast schools in reading, language arts, and mathematics (Maier, Molnar, Percy, Smith, & Zahorik, 1997; Molnar, Smith, & Zahorik, 1999, 2000, 2001; Smith, Molnar, & Zahorik, 2003; Zahorik, Molnar, Ehrle, & Halbach, 2000).

In the 2000–2001 school year, the state provided an additional $37 million to allow approximately 500 more schools to join SAGE. A $3 million categorical aid program was also created to reimburse school districts for 20% of debt service costs related to the construction of new classrooms for the SAGE program. The Wisconsin biennial 2003–2005 budget that was approved by the legislature and signed by the governor maintained the total SAGE aid available at $95 million. The calculation of this aid continued to be based on $2,000 per low income K–3 pupil. In the 2002–2003 school year, nearly 550 Wisconsin elementary schools participated in the SAGE program.

Statement of the Research Questions

A viable issue concerning the effect of reduced class size on student learning is the relationship between the number of years a student spends in classrooms with a lower teacher-student ratio and student achievement. It is reasonable to expect that if a student benefits from being in the SAGE program for one year, the student will benefit more from being in the SAGE program for multiple years. However, it is also possible that the benefits a student receives from being in a class with a lower teacher-student ratio in one grade, such as kindergarten or grade 1, will continue even if the student joins classes with a larger number of students in the higher grades. That is, the gains a student realizes in achievement and confidence in one class may be enough to accelerate the learning process for that student in later grades, even in classes of larger sizes. Of course, there are many factors that can contribute to student performance, including socioeconomic status, home conditions, teacher effectiveness, and the curriculum. These factors make it extremely difficult to ascertain the true impact of the number of years spent in SAGE on student achievement.

A related issue is how and when student achievement is measured. It is possible for reduced class sizes to have an immediate impact on student achievement as measured by end-of-year achievement tests, or by annual gains as measured by pretests and posttests. If so, then it would be reasonable to expect that these effects may be sustained over time. If students are tested two or three years later, then it is not unreasonable to expect that the students' scores will be higher because of the improved achievement experienced in early grades. However, tests are developed for specific purposes—some tests are not as sensitive to measuring student achievement in some content areas as others, or to measuring effects that may be associated with reduced class sizes.

This study of the SAGE program focused on two major questions:

1. What is the relationship between the number of years students spend in SAGE and student achievement?

2. What is the effect of SAGE on students' state test performance after they have been in SAGE up to three or four years?

Description of Data

Three main sources of data were used in this study. The first source of data was the student test score data collected by Alex Molnar and his colleagues at the University of Wisconsin–Milwaukee as a part of their evaluation of the SAGE program. Molnar's group collected data on three cohorts of students, beginning in Fall, 1996, and ending in Spring, 2001 (Molnar et al., 1999, 2000, 2001). The students assessed were from 30 schools that participated in the SAGE program and 23 schools identified as a contrasting group. The number of schools in the contrasting group varied over the course of the study from as few as 10 to as many as 17 schools at any one time. From 1996 through 2001, a total of 23 schools participated in the contrasting group. However, only 9 schools in the contrasting group over the five years of the study participated fully in all of the data gathering.

The 30 SAGE schools were located in 21 school districts that launched the SAGE program in 1996–97 in kindergarten and 1st grade. Second grade was added in 1997–98, and 3rd grade in 1998–1999. Contrast schools, initially 17, were selected from the same districts as the schools participating in SAGE in order to match those schools as closely as possible on contextual variables. In the first year of the study, students in the SAGE schools and contrast schools were given the Comprehensive Test of Basic Skills (CTBS) Complete Battery, TerraNova edition Level 10 in October, 1996, and Level 11 in May, 1997. This was repeated in each of the next two years. Level 12 was administered to students at the end of grade 2 in 1998, 1999, and 2000. Level 13 was administered at the end of grade 3 in 1999, 2000, and 2001. The SAGE evaluation conducted by Molnar's group collected data on three cohorts of students (Table 1).

Table 1
Testing Times of Three Cohorts of Students Included in the SAGE Evaluation

	Initial SAGE YEAR	Grade 1 Fall	Grade 1 Spring	Grade 2 Spring	Grade 3 Spring
Cohort 1	Grade 1 96–97	1996	1997	1998	1999
Cohort 2	K 96–97, Gr 1 97–98	1997	1998	1999	2000
Cohort 3	K 97–98, Gr 1 98–99	1998	1999	2000	2001*

* Grade 3 students in Milwaukee Public Schools took the Basic Multiple Assessments Plus test administered by the district. This test contained a greater number of subtests than the Comprehensive Test of Basic Skills (CTBS). However, both tests were Level 13 Form A, with scores reported on the same scale.

The same three cohorts of students from the SAGE schools and the contrast schools were the focus of investigation in the current study. We used the scores of the tests administered by UW–Milwaukee's group and stored in its database (Molnar, 2001). This database contained data from achievement test files, student profiles, teacher demographics, teacher log, and teacher questionnaire. In addition, we used scores of the students in each of the three cohorts on the

Wisconsin Reading Comprehension Test (WRCT) administered in 1999, 2000, and 2001. The WRCT is an assessment of primary-level reading that is administered in the Spring to nearly all grade 3 students in Wisconsin. This test provided a second measure of students' reading performance in the Spring of grade 3. We also used the scores of students in each of the three cohorts on the grade 4 Wisconsin Knowledge and Concept Examination (WKCE) in reading, language arts, and mathematics, administered in 2000, 2001, and 2002. At the time of this study, the WKCE was administered to nearly all grade 4 students in Wisconsin in the Spring.

Table 2
Testing Data for Three Cohorts of Students Included in the SAGE Evaluation

	Grade 1 Fall	Grade 1 Spring	Grade 2 Spring	Grade 3 Spring	Grade 4 Spring
Cohort 1	CTBS R, LA, M Lv10	CTBS R, LA, M Lv11	CTBS R, LA, M Lv12	CTBS R, LA, M Lv13, WRCT	WKCE R, LA, M
Cohort 2	CTBS R, LA, M Lv10	CTBS R, LA, M Lv11	CTBS R, LA, M Lv12	CTBS R, LA, M Lv13, WRCT	WKCE R, LA, M
Cohort 3	CTBS R, LA, M Lv10	CTBS R, LA, M Lv11	CTBS R, LA, M Lv12	CTBS R, LA, M Lv13, WRCT	WKCE R, LA, M

R represents reading
LA represents language arts
M represents mathematics

Data Preparation and Cleaning

In order to match the data files from the University of Wisconsin–Milwaukee study, the grade 3 WRCT for 1999, 2000, and 2001, and for grade 4 WKCE for 2000, 2001, and 2002, a considerable effort was made to include all students possible. Students were matched using last name, first name, and date of birth when that information was available on all three data sets. However, in a few cases, one or more of these descriptors for a student was missing, or a first name was used as a last name, or some other anomaly was found. In general, across the three cohorts, we were able to match from 75% to 80% of the students included in the analysis. A more detailed description of the data preparation and cleaning is included as Appendix A of this report.

Review of the Literature on Class Size Research

The five-year study of the Wisconsin Student Achievement Guarantee in Education (SAGE) program conducted by School of Education researchers of the University of Wisconsin–Milwaukee concluded that the SAGE program, which focused on reduced class size, produced a third or a half of one school year's growth, depending on the content area, when compared to the norm group averages (Smith, Molnar, Zahorik, 2003). SAGE students compared to the contrast students made the greatest gains in 1st grade. These gains were sustained over 2nd and 3rd grades with a slight increase in some content areas. The UW–M study indicated that African American students profited more from the SAGE experience than White students in comparison with non-SAGE students. The differences in achievement between African American students and White students in the SAGE program narrowed in 1st grade, contrary to those of non-SAGE students, and were maintained over the next two years. In the contrast, or non-SAGE, classrooms, the achievement gap between these two groups of students increased. Findings from the five- year study of SAGE are reported in a number of sources and publications (Maier, Molnar, Percy, Smith, & Zahorik, 1997; Molnar, Smith, & Zahorik, 1998; Molnar, 1999; Molnar et al., 1999; Molnar et al., 2000; Molnar et al., 2001).[1]

Findings on the SAGE program generally support the findings produced from studies of the most prominent class-size experiment, Tennessee's Project STAR, and numerous other, smaller studies (Word et al., 1990; Ehrenberg, Brewer, Gamoran, & Willms, 2001; Pate-Bain, Boyd-Zaharias, Cain, Word, & Binkley, 1997; Finn, 1998; McRobbie, Finn, & Harman, 1998; Nye, Hedges, & Konstantopoulos, 2001). Students in small classes showed statistically higher achievement, by about one-fifth of a standard deviation, than students in larger classes. These effects occurred in kindergarten and 1st grade and were generally sustained in latter grades for as long as six years (Nye, Hedges, & Konstantopoulos, 2001). The effects were significantly larger for minority students.

Studies on the effect of reduced class sizes, however, have shown failings in design that warrant some caution. The Tennessee STAR experiment, in which students and teachers were randomly assigned to small classes or contrast classes, did not randomly select schools and could suffer from a Hawthorne effect—i.e., teachers with reduced class sizes had a greater incentive to improve their students achievement. Students in the experiment also were not administered a pretest to verify the comparability of the treatment and the contrast groups at the beginning of the study. Designs of other studies can be challenged on such issues as failure to use true randomization, conducting the study over too short of a period of time, small scale, and not subjecting the study to an independent or rigorous evaluation.

Studies of the SAGE program and other attempts at reducing class sizes have considered teachers' practices and how these change, if they do in fact change, when working with smaller classes. Teachers who used active methods of teaching, such as explaining, modeling, checking, evaluating, and who took a teacher-directed approach, had classes with higher student achievement than teachers who used a more student-directed approach (Zahorik, Molnar, Ehrle, & Halbach, 2000). Hanushek (1998) emphasized the importance of teacher quality and its relationship to student achievement. If new hires resulting from a class-size reduction policy are above the average quality of existing teachers, average student performance is likely to increase. However, California's policy of reducing class sizes, which was associated with declines in teacher qualifications and an inequitable distribution of credentialed teachers, produced inconclusive evidence of improvement in student achievement (Bohrnstedt & Stecher, 2002).

[1] An annotated bibliography of studies is included as Appendix D in this report.

In summary, the preponderance of studies on small class size indicates there is some effect particularly at lower grades and for minority students. However, in the research in this area, it has proved to be extremely difficult to tease out the effect of small class size from other possible confounding factors, including school effects, teacher practices, and variations. In most cases, researchers use data analyzed on the basis of the number of students enrolled in a class and ignore the variations among schools. Such variation can produce a considerable measurement error, resulting in a smaller estimate in the absolute value of the true relationship between class size and outcomes (Ehrenberg, Brewer, Gamoran, & Willms, 2001).

Findings

Characterizing Exposure to the SAGE Intervention

Since one of the objectives of this study is to evaluate the effects on students of greater and lesser exposure to the SAGE intervention (that is, the number of years they spend in SAGE), it is useful to consider the attendance patterns that give rise to variations in this factor. Table 3 (Column 2) reports the total number of students in all three cohorts classified by their attendance status in grades 1, 2, and 3. In the 3–digit sequence code, the left (first) digit indicates attendance status in 1st grade, the second digit indicates status in 2nd grade, and the right (third) digit indicates status in 3rd grade. The status codes are defined as follows: 1 = attended a SAGE school, 2 = attended a contrast school, 3 = attended a nonexperimental school.[2] For example, a student who was enrolled in a SAGE school in 1st and 2nd grade and then transferred to a nonexperimental school (neither a SAGE or contrast school) would have an attendance sequence code equal to 113.

The attendance patterns are grouped into two categories in Table 3: Group A, students who enrolled in an experimental school in kindergarten or 1st grade, and Group B, late entrants, students who enrolled in an experimental school later than 1st grade. The latter group was excluded from our analysis because data on student achievement growth were not available for the Fall of 1st grade and, for most students, the Spring of 1st grade. Hence, it was not possible to contrast differences among students in initial achievement. This left us with a sample of 7,841 students as the total number of students enrolled in SAGE classrooms, or in a contrast school. The difference between the total number of students identified under the Total Column in Table 3 and the number in the columns representing the time at which students were tested indicates the number of students not tested, or that their test score could not be found in the database. For example, of the 3,031 students enrolled in SAGE schools for three grades (sequence number 111), 2,703 (89%) of the students were tested in the Fall of 1st grade. As indicated in Table 3 (and reproduced below in Table 4), the vast majority of students (92%) in Group A had one of the following six attendance patterns: 111, 113, 133, 222, 223, or 233.

[2] A student was classified as being in a treatment or contrast school in a given grade only if enrolled in both semesters. If this condition was not met, the student was classified as being in a nonexperimental school (code 3).

Table 3

Number of Students With Test Data in Each Grade by Treatment/Contrast/Other

(Grade 1-3 Sequence)							
Sequence Code: 1= Treatment, 2 = Contrast, 3 = Other School							
Sequence	Total	1st/Fall	1st/Spring	2nd	3rd	3rd: WRCT	4th: WKCE
Group A **Students Enrolled in an Experimental (Treatment or Contrast) School in First Grade or Earlier**							
111	3,031	2,703	2,675	2,629	2,614	2,682	2,574
112	21	18	17	15	14	18	17
113	905	835	691	473	38	425	470
121	2	2	2	2	2	2	1
122	22	22	18	10	7	18	18
123	12	10	8	2	0	4	5
131	62	46	39	18	35	46	45
132	22	18	18	1	7	20	18
133	828	700	505	6	17	385	398
Sub-Total	4,905	4,354	3,973	3,156	2,734	3,600	3,546
211	20	18	13	9	8	17	14
212	3	3	2	0	0	2	2
213	12	11	8	6	0	5	5
221	17	16	13	9	13	12	13
222	1,642	1,518	1,511	1,372	1,253	1,515	1,462
223	561	518	443	294	45	293	319
231	10	8	4	0	4	9	8
232	131	115	116	12	30	110	109
233	540	463	361	4	10	251	286
Sub-Total	2,936	2,670	2,471	1,706	1,363	2,214	2,218
Total	7,841	7,024	6,444	4,862	4,097	5,814	5,764
Group B **Late Entrants: Students Enrolled in an Experimental (Treatment or Contrast) School Later Than First Grade**							
311	550	0	115	449	440	434	404
312	9	0	1	9	6	7	6
313	274	0	47	227	15	109	112
321	9	0	0	7	4	6	6
322	567	0	54	430	418	488	494
323	245	0	34	199	11	115	127
331	662	0	12	125	576	538	491
332	738	0	10	108	581	654	621
333	818	0	223	140	402	487	459
Total	3,872	0	496	1,694	2,453	2,838	2,720
Grand Total	11,713	7,024	6,940	6,556	6,550	7,652	8,484

Note: In the 3-digit Sequence code, the left (first) digit indicates status in 1st grade, the second digit from the left indicates status in 2nd grade, etc. A student was classified as being in a treatment or contrast school only if enrolled for both semesters.

The result that students in the study either remained in their school after 1^{st} grade or transferred to a nonexperimental school was not unexpected. It is due to the fact that the number of SAGE treatment and contrast schools was very small relative to the total number of schools within the attendance zones of the experimental sample. For the most part, students who transferred out of experimental schools (SAGE and contrast schools) attended schools that were not part of the experiment. This has several implications.

First, variation in exposure to the SAGE program is the result of student mobility, not the result of random assignment. This suggests that the number of students with one and two years of program exposure—partial program exposure—might be insufficient to yield a statistically precise effect estimate of the effectiveness of partial program exposure (versus no or full exposure). Furthermore, if the mobility process differs between students initially enrolled in treatment and those in contrast schools, the estimates of partial and full program exposure could be biased. One way to assess the latter issue is to compare the proportion of the students with dominant attendance patterns for students initially enrolled in the treatment and contrast schools. This is given in the right column in Table 4. As indicated, 97% of the students initially enrolled in SAGE schools had one of the three dominant attendance patterns. The comparable number for students initially enrolled in contrast schools is 93%. Thus, the mobility rate is somewhat lower for SAGE as opposed to contrast-group students, but the numbers are not too dissimilar. This suggests that it is legitimate to draw on the variation in years of SAGE exposure due to mobility to investigate the consequences of differences in exposure.

Second, it is not possible to address what could be termed "second-order" evaluation questions such as: which is more effective, participating in the SAGE program in 1^{st} or 2^{nd} grade? Instead, we are limited to comparing the effectiveness of participating in the SAGE program in: (1) 1^{st} grade only, (2) 1^{st} and 2^{nd} grade only, (3) 1^{st}, 2^{nd}, and 3^{rd} grade only, and (4) not at all.

Third, test data could not be obtained in the original UW–Milwaukee study after students transferred to nonexperimental schools. This means that it was (and is) impossible to assess the effectiveness of participating in the SAGE program only in 1^{st} grade, or only in 2^{nd} grade, or only in 2^{nd} grade and 3^{rd} grade, respectively. One of the great benefits of merging the 3^{rd} grade WRCT and 4^{th} grade WKCE test data with the original UW–Milwaukee data is that it is possible to evaluate the effectiveness of partial exposure to the SAGE intervention in later grades.

In summary, given the study design, it is possible to evaluate the immediate and delayed consequences of variation in years of SAGE exposure. In order to use the simplest possible evaluation models, our analysis was limited to students with the attendance patterns listed in Table 4. The analysis sample was further restricted (eliminating 289 students) to exclude students who moved from a SAGE school to a contrast school during the school year, or from 3^{rd} to 4^{th} grade. This left us with an analysis sample of 7,218 students. The number of students in this data set is presented in Tables 5 and 6 (see discussion below). Many of our analyses were also based on an analysis sample that excluded a small number of students who moved between

Table 4
The Number of Students by Major Attendance Patterns

Major Attendance Patterns	Years Enrolled in Experimental School*	Years Enrolled in SAGE School	Number of Students in All Three Cohorts	Percent
Students in SAGE Schools				
111	3	3	3,031	62
113	2	2	905	18
133	1	1	828	17
Sub-Total	NA	NA	4,764	97
All in SAGE Schools	NA	NA	4,905	100
Students in Contrast Schools				
222	3	0	1,642	56
223	2	0	561	19
233	1	0	540	18
Sub-Total	NA	NA	2,743	93
All in Contrast Schools	NA	NA	2,936	100
Total with Consecutive Years in Experimental Schools	NA	NA	7,507	96
Total	NA	NA	7,841	NA

* Excludes enrollment in kindergarten.

SAGE schools or contrast schools. Appendix B shows the number of students in each SAGE and contrast school by cohort.

Students who entered an experimental school after the Fall testing in 1st grade were not included in the analyses. These “late entrants” students, Group B in Table 3, are designated by a 3 as the first digit of the sequence number. Of the 3,872 students in Group B, 1,504 (39%) were enrolled for at least some time in a SAGE school and 1,568 (40%) were enrolled for at least some time in a contrast school. Since there were a larger number of SAGE schools and number of students from a SAGE school who participated in the study, it would not be expected that the late entrants would be evenly distributed between the SAGE schools and contrast schools. Rather, it is reasonable to expect that there should be more late-entrant students in SAGE schools than in contrast schools. These figures suggest that there was a higher proportion, about 35%

more, of late-entrant students in the contrast schools, or students not tested in the Fall of 1st grade, than would be expected. The higher percentage of late-entrant students in the contrast schools is compatible with a higher mobility rate at these schools. The difference in the proportion of late entrants of experimental schools implies that these two groups of schools were not matched on all variables, but were fairly close on demographic variables (Table 7).

It should be noted in Table 3 that a total of 818 students (21% of Group B students) were identified as students who were enrolled in nonexperimental schools for all three grades (sequence number of 333), but were tested at least once for the study. It is likely that these students spent some time in the experimental schools, at least enough time to be tested, but spent the majority of their time in nonexperimental schools.

The Availability of Test Data: Test Taking and the Outcome of the Merge Process

Tables 5 and 6 report the number of students with available test data by attendance pattern for all of the tests used in our analysis:

- 1st grade Fall TerraNova (Reading, Language Arts, and Math)
- 1st grade Spring TerraNova (Reading, Language Arts, and Math)
- 2nd grade Spring TerraNova (Reading, Language Arts, and Math)
- 3rd grade Spring TerraNova (Reading, Language Arts, and Math)
- 3rd grade March Wisconsin Reading Comprehension Test (WRCT)
- 4th grade Spring WKCE/TerraNova (Reading, Language Arts, and Math)

The total number of students for each attendance sequence in Tables 5 and 6 is a little less than those given in Table 3 because students who attended more than one SAGE school or contrast school were excluded from the analysis sample. So the difference of 18 students from the total of 3,031 students (Table 3) who attended SAGE schools for three years (111 sequence) and the analysis sample total of 3,013 students (Table 5) with a sequence pattern of 111 is due to the fact that 18 students attended more than one SAGE school.

In Table 5, the attendance sequence code is the same 3-digit sequence used in previous tables. In Table 6, a 4-digit sequence is used. The fourth digit represents status in 4th grade. It is helpful to view the data categorized by the 4-digit sequence because our capacity to merge test data from the 4th grade WKCE clearly depends on whether a student moved to a different school in 4th grade. Tables 5 and 6 indicate two major characteristics of the data. First, with respect to the data collected by the UW–Milwaukee team, the percentage of students with available test data was generally high. Secondly, the match rate for the WRCT and WKCE was very high, particularly for the students with no school mobility. For example, for the WKCE, 2,342 out of 2,462 (95%) observations (Table 6) were matched for the group that was enrolled in a SAGE school from 1st to 4th grade. The comparable numbers for students enrolled in a contrast school from 1st to 4th grade were 1,322 out of 1,361 (97%). For all of the students who were enrolled at least in an experimental school, 74.2% of the WRCT observations were successfully matched and 73.0% of the WKCE observations were successfully matched.

Table 5

Number of Students With Test Data by Grade, Treatment Status, and School Status Through Grade 3: Analysis Sample

(Grades 1-3 Sequence)

Sequence Code: 1 = Treatment, 2 = Contrast, 3 = Other School

Number of Students

Sequence	Total	1st/Fall	1st/Spring	2nd	3rd	3rd: WRCT	4th: WKCE
111	3,013	2,686	2,658	2,613	2,601	2,669	2,557
113	824	765	623	419	2	368	399
133	776	659	474	0	6	354	361
222	1,630	1,507	1,502	1,360	1,244	1,503	1,450
223	482	451	376	238	0	238	250
233	493	422	334	0	0	224	249
Total	7,218	6,490	5,967	4,630	3,853	5,356	5,266

Percentage of Students With Test Data

Sequence	Total	1st/Fall	1st/Spring	2nd	3rd	3rd: WRCT	4th: WKCE
111	100	89.1	88.2	86.7	86.3	88.6	84.9
113	100	92.8	75.6	50.8	0.2	44.7	48.4
133	100	84.9	61.1	0.0	0.8	45.6	46.5
222	100	92.5	92.1	83.4	76.3	92.2	89.0
223	100	93.6	78.0	49.4	0.0	49.4	51.9
233	100	85.6	67.7	0.0	0.0	45.4	50.5
Total	100	89.9	82.7	64.1	53.4	74.2	73.0

Note 1: In the 3-digit Sequence code, the left (first) digit indicates status in 1st grade, the second digit from the left indicates status in 2nd grade, etc. A student was classified as being in a treatment or contrast school only if enrolled for both semesters.

Note 2: The analysis sample includes students enrolled in an experimental school (treatment or contrast) in 1st grade or earlier and with acceptable school mobility patterns.

Table 6

Number of Students With Test Data by Grade and Treatment Status With School Status Through Grade 4: Analysis Sample

(Grade 1-4 Sequence)

Sequence Code: 1= Treatment, 2 = Contrast, 3 = Other School

Number of Students

Sequence	Total	1st/Fall	1st/Spring	2nd	3rd	3rd: WRCT	4th: WKCE
1111	2462	2246	2223	2220	2254	2300	2342
1113	551	440	435	393	347	369	215
1133	824	765	623	419	2	368	399
1333	776	659	474	0	6	354	361
2222	1361	1258	1250	1159	1064	1299	1322
2223	269	249	252	201	180	204	128
2233	482	451	376	238	0	238	250
2333	493	422	334	0	0	224	249
Total	7218	6490	5967	4630	3853	5356	5266

Percentage of Students With Test Data

Sequence	Total	1st/Fall	1st/Spring	2nd	3rd	3rd: WRCT	4th: WKCE
1111	100	91.2	90.3	90.2	91.6	93.4	95.1
1113	100	79.9	78.9	71.3	63.0	67.0	39.0
1133	100	92.8	75.6	50.8	0.2	44.7	48.4
1333	100	84.9	61.1	0.0	0.8	45.6	46.5
2222	100	92.4	91.8	85.2	78.2	95.4	97.1
2223	100	92.6	93.7	74.7	66.9	75.8	47.6
2233	100	93.6	78.0	49.4	0.0	49.4	51.9
2333	100	85.6	67.7	0.0	0.0	45.4	50.5
Total	100	89.9	82.7	64.1	53.4	74.2	73.0

Note 1: In the 4-digit Sequence code, the left (first) digit indicates status in 1st grade, the second digit from the left indicates status in 2nd grade, etc. A student was classified as being in a treatment or contrast school only if he or she was enrolled for both semesters.

Note 2: The analysis sample includes students enrolled in an experimental school (treatment or contrast) in first grade or earlier and with acceptable school mobility patterns.

Simple Statistics in SAGE and Contrast Schools

Following the quasi-experimental design of the SAGE evaluation, contrast schools were selected in order to be comparable to SAGE schools with respect to family income, reading achievement, K–3 enrollment, and racial composition. Table 7 indicates that across all three cohorts, the students in the SAGE and contrast schools were quite similar. Statistical tests indicate that for most variables, the mean differences between the two groups were not statistically significant. Nonetheless, the SAGE group had somewhat higher proportions of students on free or reduced-price lunch (3.8% greater). The racial/ethnic distributions also differed somewhat. The proportion of students who were Native American was higher (by 9.6%) in SAGE schools and the proportion of students who were African American or Hispanic was lower (by over 3%). In addition, the contrast schools had a slightly higher mobility rate.

Table 7
Difference in Student Demographic Variables Between SAGE and Contrast Groups

Variable	**SAGE**		**Contrast**		**Total**		**Difference**
	Percent	N	Percent	N	Percent	N	Percent
Male	51.5	5,100	51.9	3,141	51.7	8,241	-0.4
Asian	5.5	4,925	6.1	3,093	5.7	8,018	-0.6
African American	27.3	4,925	30.5	3,093	28.5	8,018	-3.2
Hispanic	7.1	4,925	11.0	3,093	8.6	8,018	-3.9
Native American	11.2	4,925	1.6	3,093	7.5	8,018	9.6
Other	1.7	4,925	2.6	3,093	2.1	8,018	-0.8
English Disadvantaged	7.0	5,114	6.7	3,167	6.9	8,281	0.2
English Status Missing	9.8	5,114	12.4	3,167	10.8	8,281	-2.6
Reduced-Price Lunch	12.3	5,114	11.0	3,167	11.8	8,281	1.4
Lunch Status Missing	8.9	5,114	7.2	3,167	8.3	8,281	1.7
Free Lunch	51.5	5,114	47.7	3,167	50.0	8,281	3.8

With respect to test scores, the number of students tested differed, as expected, across grades (Table 8). Since our data set is limited to students who were enrolled in a SAGE or contrast school from the Fall of 1st grade (or earlier), the number of enrolled students dropped steadily from 1st to 2nd to 3rd grade. For example, in reading, the total number of students tested in both SAGE and contrast schools dropped from 7,024 in the Fall of 1st grade to 4,109 in the Spring of 3rd grade. The number of students with WRCT and WKCE test scores (from 3rd and 4th grade, respectively) was about 6,000, or 73% of the total number of students, almost as large as the number of students tested in the Spring of 1st grade. As indicated earlier, the process used to match student data from the SAGE study with state WRCT and WKCE data was highly successful.

Average test scores at the baseline (Fall of 1st grade) are somewhat lower for SAGE students than contrast-group students, a difference of 2.52 points in reading. Table 8 also

Table 8

Simple Statistics for Students in SAGE and Contrast Groups

Variable	**SAGE** Mean	**SAGE** Standard Deviation	**SAGE** N	**Contrast** Mean	**Contrast** Standard Deviation	**Contrast** N	**Total** Mean	**Total** Standard Deviation	**Total** N	**Difference** Mean
Reading										
Baseline: 1st Grade Fall	532.85	37.18	4,359	535.37	39.04	2,665	533.80	37.92	7,024	-2.52
1st Grade Spring	580.90	39.82	3,977	575.67	41.82	2,467	578.90	40.67	6,444	5.23
2nd Grade Spring	609.59	34.90	3,166	605.10	36.73	1,709	608.02	35.61	4,875	4.49
3rd Grade Spring	634.05	37.91	2,742	628.65	38.87	1,367	632.25	38.31	4,109	5.40
3rd Grade WRCT	50.57	12.67	3,676	50.98	12.40	2,338	50.73	12.56	6,014	-0.40
4th Grade Spring WKCE	641.59	36.56	3,626	643.22	36.08	2,354	642.23	36.38	5,980	-1.63
Language Arts										
Baseline: 1st Grade Fall	530.35	43.86	4,359	530.95	45.18	2,664	530.58	44.37	7,023	-0.60
1st Grade Spring	583.81	42.30	3,978	576.37	44.79	2,467	580.97	43.42	6,445	7.44
2nd Grade Spring	612.66	40.53	3,165	602.78	39.55	1,710	609.20	40.46	4,875	9.88
3rd Grade Spring	630.35	34.93	2,742	625.77	35.99	1,367	628.83	35.35	4,109	4.58
4th Grade Spring WKCE	642.27	33.39	3,626	643.89	34.08	2,354	642.91	33.67	5,980	-1.62
Mathematics										
Baseline: 1st Grade Fall	491.97	41.61	4,330	493.85	40.53	2,654	492.68	41.21	6,984	-1.88
1st Grade Spring	541.48	41.35	3,981	531.97	42.66	2,448	537.86	42.11	6,429	9.52
2nd Grade Spring	574.36	41.23	3,184	561.26	40.35	1,695	569.81	41.39	4,879	13.10
3rd Grade Spring	609.22	35.46	2,758	596.76	38.10	1,361	605.10	36.82	4,119	12.46
4th Grade Spring WKCE	628.95	32.69	3,685	628.47	32.95	2,386	628.77	32.79	6,071	0.48

includes average test scores in reading, language arts, and mathematics for the Spring of 1st grade through 4th grade. Since the number of students varied by grade, thus changing the confidence interval, the mean differences in test scores between the SAGE and contrast students do not necessarily reflect the effectiveness of the SAGE program. Nonetheless, the simple statistics are suggestive. The simplest summary of the data is that there appear to be positive mean differences in favor of the SAGE students between SAGE and contrast-group students with respect to test scores in 1st, 2nd, and 3rd grade and essentially no mean differences between the two groups of students with respect to the state WKCE and WRCT test scores.

Data Sets Used to Estimate the Effectiveness of the SAGE Program

As indicated above, the number of students tested by grade differs quite dramatically. This poses a challenge in interpreting effect estimates because if the maximum number of students is utilized to estimate each model, the samples differ substantially from grade to grade. One option is to assume that the students who are in some sense missing from the database (because they moved to another school or were absent the day of testing) are "missing at random." This assumption means that the SAGE and contrast schools did not systematically lose high- or low-growth students. Since we cannot know whether it is legitimate to make this assumption, our strategy is to provide estimates for three different samples:

- The "Maximum Sample" is the largest sample—5,394 students. It includes all students who were tested at a given grade, had baseline test data and demographic information, and who did not change schools while enrolled in a SAGE or contrast school. (The latter restriction eliminated a small number of students who moved between different schools participating in the study.) This sample includes students who attended a SAGE or contrast school for 1, 2, 3, or 4 consecutive years.
- The "Common Sample" is the smallest sample—2,940 students. It includes only students who had test data in every grade and met the other data restrictions listed above. This data set automatically excludes students who participated in the study (in a SAGE or contrast school) for less than three years because these students were not tested in 2nd or 3rd grade. The advantage of this data set is that we can compare estimates from different grade models without worrying that any differences in the results are due to differences in the samples. The disadvantage of this data set is that it provides no information on the effects of participating in the SAGE program for one or two years.
- The "Full Treatment Sample of Students Who Participated in the SAGE study for 3 Years" is a compromise between the other two data sets—3,824. As in the case of the Common Sample, it is limited to students who participated in the SAGE study for a full 3 years. But, it does not eliminate students who changed schools after 3rd grade and it does not eliminate students who were absent from school in any of the grades.

Estimates of the Effectiveness of the SAGE Program

We focus our discussion primarily on the results obtained using the Maximum Sample (reported in Table 9). As is evident from comparing Tables 9, 10, and 11, the SAGE effectiveness estimates are remarkably similar across the different samples. The values in the rows labeled “Effect” represent the number of points on the test that the SAGE students out performed the contrast students. A negative number indicates that the contrast students performed higher than the SAGE students. The three tables contain estimates of the effectiveness of the SAGE Program in reading, language arts, and mathematics for all three cohorts combined, along with the significance level. The effect is judged as statistically significant if the significance level is 0.05 or less.

As discussed in the methods section (Appendix C), we report estimates of the cumulative effect of the SAGE Program. Thus, using the test scores data collected during the SAGE program in grades 1–3, we report estimates of the effectiveness of the SAGE Program in 1st grade (row 1 of Tables 9, 10, and 11), 1st and 2nd grade (row 2), and 1st, 2nd, and 3rd grade (row 3). Using the original study data, we cannot, for example, report on the effect of being in the program for 2 years on 3rd grade achievement. Students who left the study were not tested after they left except as part of the state assessment program.

Our estimates of the effectiveness of the SAGE Program based on the original study data are very consistent with estimates reported by the original study team.[3] The estimates indicate that reading scores were 5 to 6 points higher for SAGE students compared to contrast-group students after 1st grade and these differences persisted (and increased slightly) through 3rd grade on the TerraNova administered for the experiment. In Table 9, the effect favoring SAGE schools on reading was 5.363. This effect estimate was more than double the standard error (2.434) and was significant at the .028 level. This indicates that the 1st grade gain by SAGE students compared to students in the contrast schools was over 5 points on the test in reading. The effects in language arts and mathematics were even larger. In language arts, SAGE students outperformed contrast students at the end of 1st grade by 6.492 points (Table 9). This was significant at the 0.008 level. For mathematics, the SAGE program had a 1st grade effect of 9.482 points at a significance level of 0.001. It is important to note that these estimates (obtained from a fully specified multilevel model) are very similar to those implied by a simple comparison of means (see Table 8). This is due to the fact that the SAGE and contrast groups are very similar. The primary advantage of using a multilevel model is to obtain correct estimates of precision and significance.

The second and third rows of data in Tables 9–11 show the cumulative effect of the SAGE program when compared to the performance of students at the contrast schools. The second row, labeled 1st to 2nd, indicates the SAGE effect from the beginning of 1st grade to the end of 2nd grade. In reading, the cumulative for the two grades was 3.372 (Table 9). This value is lower than the SAGE effect of over 5 points over grade 1, indicating that the contrast students gained about 2 points over the SAGE students in 2nd grade. However, the cumulative effect over

[3] It appears that the precision of estimates and significance levels reported by the original study team may be somewhat high. This may be due to the fact that we use a multilevel model to obtain estimates. We found that OLS estimates were substantially inflated.

1st and 2nd grades in language arts and mathematics were both positive and significant. The point increase for 2nd grade. The cumulative effect of SAGE in mathematics, 12.250, represents a cumulative effect of SAGE in language arts for 1st and 2nd grade was 7.782, indicating over a 1-nearly 3-point increase in grade 2. In language arts and mathematics, the SAGE students not only retained the advantage they had gained in 1st grade, but they increased the effect over 2nd grade, although not to as great an extent as in the prior year.

The cumulative effect from 1st to 3rd grade, the third row of data in Tables 9-11, continued to increase for reading and mathematics, but stayed the same for language arts. For reading in 3rd grade SAGE students regained the effect size they had in 1st grade and improved on it by about 1.7 points. Thus, the cumulative effect for the SAGE program from 1st to 3rd grade was 7.096 (Table 9) with a significance level of 0.009. In mathematics, the SAGE students gained an additional 1 point in 3rd grade, resulting in a cumulative effect for the SAGE program from 1st grade to 3rd grade of 13.320 points, which was highly significant.

In re-analyzing the data collected and analyzed by the University of Milwaukee research team, we replicated the findings that the SAGE program had a significant cumulative effect from the beginning of 1st grade through 3rd grade in all three content areas. The greatest cumulative effect was in mathematics, over 13-point difference in favor of the SAGE students. The cumulative effects in reading and language arts were less, but still significant at the 0.01 level. These data indicate that there is a benefit for being in SAGE classrooms over multiple years. The benefit was the greatest in mathematics, but also was observed in reading and language arts.

In this study, we incorporated additional data using student scores on tests administered as required by the state of Wisconsin, the 3rd grade WRCT and the 4th grade WKCE. We fully expected to see the results on the 3rd grade WRCT test administered in the spring of the 3rd grade—the same semester as the 3rd grade TerraNova was given to the experimental schools—to produce results similar to those we found on the TerraNova. However, on the 3rd grade WRCT, we found no effect between the SAGE program and the contrast schools (Tables 9, 10, and 11). This finding was in direct contrast with the results on the TerraNova given to the same students at the end of the 3rd grade. In Tables 9, 10, and 11, in the fourth row of data, the effect size for one year of SAGE on the WRCT was 0.650. The effect size for two years of SAGE was negative, -0.233, indicating that the results slightly favored the contrast students. The effect size for three years of SAGE on the WRCT was 0.328, slightly favoring the SAGE students. All of these effect sizes were less than one standard error and were not significant. Thus, our study produced contradictory results with one measure showing strong positive effects of the SAGE program, while a second measure administered for state purposes to the same students showed no significant effects.

It is not unusual for two measures to produce different results. Tests are designed for different purposes, have different item types, and are administered under different incentives. The 3rd-grade WRCT test is designed to identify students who are considered to have reached a specific level of reading proficiency while not showing the full range of reading ability of all the

Table 9
The Cumulative Effect of the SAGE Program by Grade in All Subjects for the Maximum Sample: Students in the SAGE Program for 1, 2, or 3 Years

Grades Evaluated	Test	Years in SAGE School		Reading	Language Arts	Math
1st	TerraNova	1	Effect	5.363	6.492	9.482
			SE	2.434	2.425	2.866
			Significance	0.028	0.008	0.001
			N	5,393	5,394	5,376
1st to 2nd	TerraNova	2	Effect	3.372	7.782	12.250
			SE	2.351	3.023	3.401
			Significance	0.152	0.010	0.000
			N	4,192	4,193	4,188
1st to 3rd	TerraNova	3	Effect	7.096	7.548	13.320
			SE	2.717	2.561	2.644
			Significance	0.009	0.003	0.000
			N	3,459	3,459	3,462
1st to 3rd	WRCT	1	Effect	0.650	NA	NA
			SE	1.286		
			Significance	0.613		
			N	409		
		2	Effect	-0.233	NA	NA
			SE	1.226		
			Significance	0.850		
			N	491		
		3	Effect	0.328	NA	NA
			SE	0.875		
			Significance	0.708		
			N	3,282		
			Total N	4,182		
1st to 4th	WKCE/TerraNova	1	Effect	-3.070	-2.543	0.529
			SE	3.347	3.185	2.997
			Significance	0.359	0.425	0.860
			N	428	428	439
		2	Effect	-4.725	-0.029	-3.539
			SE	3.165	3.021	2.853
			Significance	0.136	0.992	0.215
			N	521	521	530
		3	Effect	-0.653	-0.240	1.186
			SE	2.002	1.995	1.904
			Significance	0.744	0.904	0.534
			N	3,604	3,939	3,641
			Total N	4,553	4,888	4,610

Table 10

The Cumulative Effect of the SAGE Program by Grade in All Subjects for the Full Treatment Sample of Students Who Participated in the SAGE Program for 3 Years

Grades Evaluated	Test	Years in SAGE School		Reading	Language Arts	Math
1st	TerraNova	1	Effect	5.477	5.757	8.239
			SE	2.559	2.504	2.863
			Significance	0.032	0.022	0.004
			N	3,823	3,824	3,808
1st to 2nd	TerraNova	2	Effect	4.237	9.241	12.852
			SE	2.346	3.060	3.408
			Significance	0.071	0.003	0.000
			N	3,602	3,603	3,602
1st to 3rd	TerraNova	3	Effect	7.119	7.528	13.382
			SE	2.749	2.599	2.661
			Significance	0.010	0.004	0.000
			N	3,441	3,441	3,444
1st to 3rd	WRCT	3	Effect	0.078	NA	NA
			SE	1.024		
			Significance	0.940		
			N	3,759		
1st to 4th	WKCE/TerraNova	3	Effect	-0.840	-0.621	0.647
			SE	2.132	2.078	2.149
			Significance	0.694	0.765	0.764
			N	3,604	3,604	3,641

Table 11
The Cumulative Effect of the SAGE Program by Grade in All Subjects for the Common Sample: Students Who Participated in the SAGE Program for 3 Years and With No Missing Test Data

Grades Evaluated	Test	Years in SAGE School		Reading	Language Arts	Math
			Common Sample N	**2,907**	**2,940**	**2,928**
1st	TerraNova	1	Effect	6.306	5.847	7.226
			SE	2.784	2.940	3.060
			Significance	0.024	0.047	0.018
1st to 2nd	TerraNova	2	Effect	4.796	10.641	14.798
			SE	2.309	3.053	3.606
			Significance	0.038	0.001	0.000
1st to 3rd	TerraNova	3	Effect	6.841	7.665	13.326
			SE	2.672	2.569	2.638
			Significance	0.011	0.003	0.000
1st to 3rd	WRCT	3	Effect	0.227		
			SE	1.038		
			Significance	0.827		
1st to 4th	WKCE/Terra Nova	3	Effect	-0.125	-1.529	0.456
			SE	2.346	2.283	2.429
			Significance	0.958	0.503	0.851

students. The TerraNova is designed to place students on a scale based on their reading performance and to compare to a national norm. So the differences between the two 3rd – grade reading tests could be explained. However, when we considered scores on the test the students took in 4th grade, the findings again indicated no significant differences between the SAGE students and the students in the contrast schools, supporting the findings found on the 3rd grade WRCT and not the findings on the TerraNova administered as part of the experiment.

No significant differences were found in the results on the 4th-grade WKCE tests between students who had been in the SAGE program and students who had been in the contrast group. On reading, the effects favored the contrast students when compared to students who were in the SAGE program for only one school year, for two school years, and three school years (Table 9). The effect size in reading of -0.653 for students with three years of SAGE compared to students with three years in the contrast schools was very small, with a significance level of 0.744. The effect size for the same group of students in language arts was -0.240, and in mathematics 1.186. None of these effect sizes were significant and all of them were less than one standard error from zero. If we considered only the 4th-grade test results and not the 3rd-grade WRCT results, then it would be easy to conclude that the SAGE effect did not sustain itself beyond 3rd grade. However, because the 3rd-grade WRCT results also showed no significant effects in the Spring of 3rd grade, while the 3rd-grade TerraNova administered in May of that year did show significant SAGE effects, the contradictory results indicate that something more complex than simply measuring the effects of the SAGE program was involved. Some other factors were contributing to the insignificant results between the SAGE program and the contrast schools when the state-mandated tests were used, while significant results were being obtained on the experimental tests. The contradictory results were even more puzzling because the 4th-grade WKCE is a version of the TerraNova developed by the same commercial testing company.

To explore the test results further, we analyzed the data on the basis of student ethnicity and SES (Tables 12 and 13). When comparing African American students in the SAGE program with African American students in the contrast schools, the effect sizes favoring the SAGE program were more dramatic than the effect sizes for White students (Table 12). The cumulative effect sizes for African American students from 1st grade to 3rd grade on the TerraNova were 13.134 for reading, 11.038 for language arts, and 19.037 for mathematics (see the forth row of data in Table 12). All of these effect sizes favored the SAGE group and were significant at the 0.01 level. The cumulative effect sizes for White students for each of the three content areas favored the SAGE program, but only the effect for mathematics was statistically significant (at the 0.01 level). This would indicate that African American students gained the most from participating in the SAGE program. The differences in results between the African American students and White students (see the third column of results by the three content areas in Table 12) favored the African American students, but none were statistically significant. The greatest difference in cumulative effect for African American students when compared to White students was in reading with an effect of 10.217 (at a level of 0.065) followed by an effect in mathematics of 8.953 (at a level of 0.095) and then an effect in language arts of 6.366 (at a level of 0.235).

When we considered the results on the state mandated tests, the 3rd-grade WRCT and the 4th-grade WKCE, the trend indicated some benefit for African American students, but the significant effects in favor of the SAGE program were not sustained. African American students who were in the SAGE program performed better on the 4th-grade WKCE for all three content areas than did African American students who attended the contrast schools—0.674 in reading (significance level of 0.867), 2.654 in language arts (significance level of 0.485), and 3.866 in mathematics (significance level of 0.243). None of these differences were significant (Table 12). On the 3rd-grade WRCT, the effect size between the African American students in the SAGE program and those in the contrast schools was very small and favored the contrast students (-0.413 with a significance level of 0.844). The White students in the SAGE program scored lower than White students from the contrast schools on both the 3rd-grade WRCT and the 4th-grade WKCE, but again the difference was not statistically significant. The cumulative effect of three years of the SAGE program was essentially the same for African American students and White students on the 3rd grade WRCT (0.005) and favored the African American students on the 4th grade WRCT on all three content areas, but not a significant level.

We also compared the difference between the performance of students from low income families with those from high income families (Table 13). The results are similar to those for the difference between African American students and White students. The cumulative effects from 1st grade to 3rd grade for low-income students on the TerraNova was higher for low-income students when compared to high-income students (9.228 in reading compared to 5.737, 11.598 in language arts compared to 4.108, and 15.848 in mathematics compared to 10.904) (see the fourth row of data in Table 13). The results for the low-income students were all significant, but none of the cumulative effects for high-income students was significant. The 3-year cumulative difference between the low- and high-income students was most significant in language arts (4.011 at a significance level of 0.062). There was essentially on difference on income status between the SAGE students and the contrast students on the 3rd-grade WRCT with effect sizes all nearly zero. There also was no significant effect size related to income status between the SAGE students and the contrast students on the 4th-grade WKCE. For all three content areas, all of the significant levels were at least 0.5 or greater. Thus, there was some evidence of the SAGE program had a greater effect on for students from low-income families, more so than for students from high-income families. This cumulative effect, however, was not apparent on the two state mandated tests.

Conclusions

In extending the experiment conducted by the UW–Milwaukee research group, we were successful in locating the scores on a high percentage of the students on the 3rd-grade WRCT and the 4th-grade WKCE tests not administered as part of the experiment. Our analysis confirms the results from the experiment that show a significant cumulative SAGE effect in all three content areas, but more so in mathematics than in reading and language arts. The greater effects appear in 1st grade; subsequent grades show some effect, but not as great as that in grade 1.

The effects were not sustained into 4th grade on the WKCE in reading, language arts, or mathematics tests. This would suggest that whatever gains students achieved in participating in the SAGE program through 3rd grade were not continued into 4th grade. However, it became evident that this interpretation of the results was too simplistic, given that the results on the 3rd-grade WRCT were very similar to the results on the 4th-grade WKCE, but were in stark contrast with the 3rd-grade TerraNova administered to nearly the same students within three months of each other. Whereas the cumulative effect from 1st grade to 3rd grade on the TerraNova in reading was 7.119 ($p < .01$), the cumulative effect over three years on the 3rd-grade WRCT was 0.078 and not significant. This large difference in results on two tests, both designed to measure students' reading skills, although with somewhat different purposes, indicates that there are other factors coming into play than simply the SAGE program.

Table 12
The Cumulative Effect of the SAGE Program by Grade in All Subjects in the Maximum Sample of Students Who Participated in the SAGE Program for 3 Years (African American vs. white students)

				African American			White			African American-White		
Grades Evaluated	**Test**	**Years in SAGE School**		**Reading**	**Language Arts**	**Math**	**Reading**	**Language Arts**	**Math**	**Reading**	**Language Arts**	**Math**
			Effect	-0.931	-1.837	-1.640	1.399	7.214	2.747	-2.330	-9.051	-4.386
			SE	3.902	4.953	5.078	2.560	3.239	3.263	4.667	5.918	6.036
Baseline	Terra Nova	1	Significance	0.811	0.711	0.747	0.585	0.026	0.400	0.618	0.126	0.467
			N	844	844	842	2253	2253	2259	844/2253	844/2253	842/2259
			Effect	7.934	8.969	10.158	3.540	3.360	5.515	4.394	5.609	4.643
			SE	5.756	5.325	5.118	2.313	2.358	3.208	6.204	5.823	6.041
1st	Terra Nova	1	Significance	0.169	0.093	0.048	0.126	0.154	0.086	0.479	0.335	0.442
			N	780	781	764	2141	2141	2145	780/2141	781/2141	764/2145
			Effect	5.221	10.907	13.716	1.943	4.561	10.080	3.278	6.346	3.637
			SE	4.343	4.486	4.680	2.737	3.463	3.856	5.134	5.667	6.064
1st to 2nd	Terra Nova	2	Significance	0.230	0.015	0.004	0.478	0.188	0.009	0.523	0.263	0.549
			N	714	714	709	2020	2020	2022	714/2020	714/2020	709/2022
			Effect	13.134	11.038	19.037	2.917	4.672	10.084	10.217	6.366	8.953
			SE	4.940	4.477	4.816	2.484	2.943	2.349	5.530	5.357	5.358
1st to 3rd	Terra Nova	3	Significance	0.008	0.014	0.000	0.241	0.113	0.000	0.065	0.235	0.095
			N	696	696	690	1892	1892	1898	696/1892	696/1892	690/1898
			Effect	-0.413	NA	NA	-0.418	NA	NA	0.005		
			SE	2.094			0.987			2.315		
1st to 3rd	WRCT	3	Significance	0.844			0.672			0.998		
			N	739			2140			739/2140		
			Effect	0.674	2.654	3.866	-1.380	-2.527	-0.487	2.055	5.181	4.353
			SE	4.029	3.802	3.311	2.558	2.367	2.540	4.773	4.479	4.172
1st to 4th	WKCE/Terra Nova	3	Significance	0.867	0.485	0.243	0.590	0.286	0.848	0.667	0.247	0.297
			N	671	671	677	2087	2087	2104	671/2087	671/2087	677/2104

Table 13

The Cumulative Effect of the SAGE Program by Grade in All Subjects in the Maximum Sample of Students Who Participated in the SAGE Program for 3 Years (students from low income family vs. high income family)

				Low Income			High Income			Low Income-High Income		
Grades Evaluated	**Test**	**Years in SAGE School**		**Reading**	**Language Arts**	**Math**	**Reading**	**Language Arts**	**Math**	**Reading**	**Language Arts**	**Math**
Baseline	Terra Nova	1	Effect	-0.492	1.593	1.714	0.144	5.521	1.079	-0.635	-3.928	0.635
			SE	2.576	3.299	3.430	2.710	3.516	3.339	3.739	4.821	4.787
			Significance	0.849	0.629	0.617	0.958	0.117	0.747	0.865	0.415	0.894
			N	1785	1784	1781	2088	2088	2086	1785/2088	1784/2088	1781/2086
1st	Terra Nova	1	Effect	4.871	7.670	8.535	4.699	3.963	7.255	0.172	3.707	1.280
			SE	3.490	3.357	3.183	2.443	2.471	2.993	4.260	4.169	4.369
			Significance	0.163	0.023	0.007	0.055	0.109	0.015	0.968	0.374	0.770
			N	1632	1633	1616	1994	1994	1996	1632/1994	1633/1994	1616/1996
1st to 2nd	Terra Nova	2	Effect	6.855	10.335	14.304	1.902	7.518	11.802	4.953	2.817	2.502
			SE	2.318	3.155	3.768	2.618	3.373	3.645	3.496	4.619	5.243
			Significance	0.003	0.001	0.000	0.468	0.026	0.001	0.157	0.542	0.633
			N	1515	1516	1515	1886	1886	1883	1515/1886	1516/1886	1515/1883
1st to 3rd	Terra Nova	3	Effect	9.228	11.598	15.848	5.737	4.108	10.904	3.490	7.490	4.943
			SE	3.114	2.750	3.171	3.101	2.920	2.524	4.395	4.011	4.052
			Significance	0.003	0.000	0.000	0.065	0.160	0.000	0.427	0.062	0.223
			N	1471	1471	1475	1819	1819	1818	1471/1819	1471/1819	1475/1818
1st to 3rd	WRCT	3	Effect	-0.001	NA	NA	-0.020	NA	NA	0.019		
			SE	1.304			1.036			1.665		
			Significance	1.000			0.985			0.991		
			N	1565			1994			1565/1994		
1st to 4th	WKCE/Terra Nova	3	Effect	-1.712	-0.749	0.824	-0.792	-0.750	-0.150	-0.920	0.001	0.973
			SE	2.688	2.344	2.344	2.324	2.421	2.431	3.554	3.369	3.377
			Significance	0.524	0.750	0.725	0.733	0.757	0.951	0.796	1.000	0.773
			N	1482	1482	1510	1932	1932	1941	1482/1932	1482/1932	1510/1941

One obvious factor that could contribute to the inconsistent results could be that the contrast students had different incentives for performing well on the state mandated tests than on the experimental tests. When the stakes were the same, as they would be on the state tests, the contrast students could have performed better than when they were given a test at the end of the year that had no consequences for them. We have no evidence to indicate that the tests given as part of the experiment were administered differently in the SAGE schools compared to their administration in the contrast schools. There could be other factors that come into play. The TerraNova tests may be different enough from the WRCT and the WKCE to result in some variation in results. Another very important factor to consider is that the relatively small number of schools in the study resulted in a fairly high standard error, making it more difficult to detect differences.

The estimates from the data raise the possibility that the SAGE program has a more positive effect on African American students than on White students. Although the 4th-grade WKCE data did not show a significant difference between African American students and White students, the cumulative differences were in the direction that suggests the possibility of a greater impact of the SAGE program on African American students. The same finding was not true when students from low income families were compared to students from high income families.

The results in this study are in no way definitive. They raise more questions than they answer. The study only considered the effects on the performance of students, as in contrast to the performance of schools, and was restricted to the nearly 50 schools that participated in the experiment. This included the first 30 SAGE schools. Now the SAGE program has been expanded to nearly 500 schools. The popularity of the SAGE program alone indicates that a number of schools are finding value in it. A more definitive study of the SAGE program should analyze the effects on the performance of the full number of SAGE schools, comparing these to match schools. Such an analysis would not be restricted by such high standard errors as the current study and would be more sensitive to possible SAGE effects. The results of the current studies were too inconsistent to serve as a basis for drawing any strong conclusions.

References

Bohrnstedt, G. W., & Stecher, B. M. (Eds.), CSR Research Consortium. (2002, September). *Capstone Report. What we have learned about class size reduction in California.* Sacramento: California Department of Education.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, Vol. 2*, No. 1, 1-30.

Finn, J. D. (1998, April). *Class size and students at risk: What is known? What is next?* A Commissioned paper prepared by Jeremy D. Finn. National Institute on Education of At-Risk Students, Office of Educational Research and Improvement, U.S. Department of Education.

Hanushek, E. A. (1998). *The evidence on class size*. Occasional Paper No. 98-1. W. Allen Wallis Institute of Political Economy. Rochester, NY: University of Rochester.

Maier, P., Molnar, A., Percy, S., Smith, P., & Zahorik, J. (1997, December). *First-year results of the Student Achievement Guarantee in Education Program*. Submitted by the SAGE Evaluation Team Center for Urban Initiatives and Research. University of Wisconsin–Milwaukee.

McRobbie, J., Finn, J. D., & Harman, P. (1998, August). *Class size reduction: Lessons learned from experience.* WestEd's Policy Program Policy #23, August, 1998.

Molnar, A., Smith, P., & Zahorik, J. (1999, December). *1998–99 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: Unviersity of Wisconsin–Milwaukee, School of Education.

Molnar, A., Smith, P., & Zahorik, J. (2000, December). *1999–2000 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: University of Wisconsin–Milwaukee, School of Education.

Molnar, A., Smith, P., & Zahorik, J. (2001). *2000–2001 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: University of Wisconsin–Milwaukee, School of Education.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2001, Spring). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. *Journal of Experimental Education, (69)*3, 245–257.

Pate-Bain, H., Boyd-Zaharias, J. B., Cain, V. A., Word, E., & Binkley, M. E. (1997, September). *The Student/Teacher Achievement Ratio (STAR) Project: STAR Follow-up Studies,1996–1997.* Prepared by Health and Education Research Operative Services, Inc. (HEROS).

Smith, P., Molnar, A., & Zahorik, J. A. (2003). *Class size reduction in Wisconsin: A fresh look at the data.* An Education Policy Research Unit Study, Education Policy Studies Laboratory. Tempe, AZ: Arizona State University, College of Education.

Word, E., Achilles, C. M., Bain, H., Folger, J., Johnston, J., & Lintz, N. (1990, June). *Project STAR final executive summary report: Kindergarten through third grade results (1985–1989).* Nashville, TN: Project STAR Office, Tennessee State Department of Education. Retrieved July 16, 2002 from http://www.cde.ca.gov/classsize/eval/projstar.htm

Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. (2000). *Effective teaching in reduced-size classes.* Milwaukee: Center for Education Research, Analysis, and Innovation, University of Wisconsin–Milwaukee.

Appendices

A. Data Preparation and Cleaning in the SAGE Study

B. Enrollment in SAGE and Contrast Schools by Cohort, Grade, and School Year

C. SAGE Evaluation Model

D. Annotated Bibliography: Class Size and Student Achievement

Appendix A

Data Preparation and Cleaning in the SAGE Study

Data Preparation and Cleaning in the SAGE Study

The working data set is the result of the following steps of data merging and cleaning.

Step 1 Combine and clean UW–Milwaukee data

1. The grade 3 student achievement files (public version) of three cohorts (96–99, 97–00, 98–01) in the SAGE Evaluation Data CD from UW–Milwaukee's study were combined into one data set that contains all of the students' test scores on reading, language and mathematics on four occasions from grade 1 to grade 3 across three cohorts. Furthermore, the student IDs, as well as the test scores left in the grade one or two student achievement files but not in the grade three files, were picked up and merged into this data set too. Before the merger, the very few cases in the student achievement files where a student ID has been used for more than one student had been cleaned up: only one record with maximum information was retained, or a randomly picked one, if the records with the same student ID contained the same amount of information.

2. The profile data of students in the score data set from (1) were obtained from the 2000–2001 student profile file (public version, and confidential version just for students' full names and birthdays) in the SAGE Evaluation Data CD, and combined into the data set from (1). The 2000–2001 student profile file (public version) contains some student records, which were not in the score data from (1), were also merged into (1). It turns out that the variable indicating a student's grade in the 1997–98 school year from the 2000–2001 student profile file (public version) has serious mistakes. Therefore, this variable was instead taken from 1999–2000 student profile file (public version): only the missing values were filled in by the data from the 2000–2001 student profile file (public version) if they were not missing there. There are four student IDs in the 2000–2001 student profile file (public version) and 2 student IDs in the 1999–2000 student profile file (public version) that have been used in two student records. Those problematic cases had been taken care of before the merger: one of the two records with a same student ID had been reassigned a new ID.

3. The data set from Step (2) still contained a few dirty cases, which had the following problems:

    A. An identical student ID assigned to two different students. Three such cases had been found and cleaned up: one record was deleted based on the profile information.

    B. Two different student IDs assigned to one student. Two such cases had been cleaned up: the information in the two student records had been merged into one record, and one student ID had been discarded. We also found some other suspicious cases as such, but we did not clean them because we were not 100 percent sure about the judgment, or how to clean them correctly; more importantly, their influence on the later analyses is negligible.

The resultant UW–Milwaukee dataset had a total of 16,516 records, where 4,803 records did not have any data from student achievement files, while 477 records did not have any data from the 2000–2001 (or 1999–2000) student profile. Most of those incomplete records were excluded from the later analyses because they lacked the information necessary for the analysis models. In addition, the problematic cases as well as the cleaning procedure mentioned above had little effect on the later analyses because of their very small number with an even smaller number used in later analyses.

Step 2 Merge UW–Milwaukee dataset with WRCT (99–01) and WKCE (00–02)

The WRCT and WKCE scores of the students in the UW–Milwaukee dataset from Step 1 were taken from the WRCT dataset (99-01) and WKCE dataset (00-02). The basic criteria to match student records between UW–Milwaukee data and WRCT and WKCE data were students' first names, last names, and birthdays. Unfortunately, some missing values, typographical errors, and truncations in these fields made the match process a little challenging. The approach that we developed here was a multistage, semiautomatic match process, where the data sets were gone through multiple times to find the matched cases based on the all or part of the three matching criteria, and then the partially matched cases were scanned by eye to determine whether they were the real matches or not based on other available information, such as gender, race, school ID, as well as the similarity of first name, last name, or birthday.

The match process had three stages. In Stage 1, all three criteria were used to match records via PROC SQL in SAS 8.0, and the results were considered as the real matches. In Stage 2, the rest records (excluding those matches in Stage 1) were matched based on any two of the three criteria, and then the matches were inspected by two persons to judge whether they were the real matches or not. In Stage 3, the remaining records (excluding the real matches in Stages 1 and 2) were matched again, but this time based on any one of the three criteria. And again, the matches were judged by two persons as to be real matches or not.

The match summaries (by cohort), after applying the above match process on UW–Milwaukee dataset with WRCT (99–01) and WKCE (00–02) are shown in Tables A-1 through A-6. The match rates are all around 80% (of the records in UW–Milwaukee dataset), although the rates with WRCT are a little higher than those with WKCE. Note that (1) WRCT was matched first with UW–Milwaukee dataset and, then, the information from WRCT was also used in the match process between UW–Milwaukee dataset and WKCE to increase the match hits; (2) most of the matches are from Stage 1, which are based on the strongest evidence (three criteria). This is especially true for the matches with WKCE, which is due to the fact described in (1).

Step 3 Create attendance variables

In longitudinal research such as the research used in this study, the attendance variables are the first key information required by the analysis models. The attendance variables (school ID in each semester) in the working dataset represent the combined information from different sources as well as our guessing. In particular, the creation of the attendance variables was gone through the following steps:

(1) The school IDs in all fall semesters and in the spring semesters of grade 1 for Cohort 2 and 3 students and grade 2 for all cohort students were primarily from the corresponding school ID variables in the 2000–2001 student profile (public version). The missing values were filled in by the inferred school IDs from the class ID variables in the same file, if they were not missing there. For the spring semesters, the missing values were further filled in by the combined school IDs from the achievement files as described in the next step (2).

(2) The school IDs in spring semesters other than those mentioned in (1) were primarily from the corresponding school ID variables in the achievement files. Again, the missing values were filled in by the inferred school IDs from the class ID variables in the achievement files. In addition, the still missing values were filled in by the school ID information from the 2000–2001 student profile file (public version), which included the combined information from the school ID variables and class ID variables just as described in step (1) above.

(3) If a student took the test in the fall of grade 1, but the school ID was missing, the school ID for the spring of grade 1 was used.

(4) If school IDs were missing in the fall of grade 2 (or grade 3), they were filled in using the school IDs in the spring of grade 2 (or grade 3).

(5) If a student took the test but had a missing school ID in both semesters in a given academic year, the school ID was filled in using those in the previous or later semesters if there were school IDs in those semesters.

Match Summaries Between UW–Milwaukee Dataset and WRCT (99–01) by Cohort

Table A-1

Match Cases Summary (WRCT, Cohort 1)

Matching Criteria	**Same School**		**Missing School Name[a]**		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	
Last Name First Name Birthday	1685	NA	648	NA	27	NA	2360
Last Name First Name	127	302	74	113	2	0	618
Last Name Birthday	269	0	92	0	1	0	362
First Name Birthday	72	0	26	0	3	0	101
Last Name	25	37	4	2	0	0	68
First Name	5	14	2	3	0	0	24
Birthday	15	0	2	0	0	0	17
Total	2198	353	848	118	33	0	3550
Total Cases of UW-Milwaukee, Cohort 1 Excluding Missing on All Three Criteria (% Match)							4255 (83.4%)
Total Cases of UW-Milwaukee, Cohort 1(% Match)							4440 (80.0%)

Table A-2
Match Cases Summary (WRCT, Cohort 2)

Matching Criteria	**Same School**		**Missing School Name**[a]		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	
Last Name First Name Birthday	1528		518		9		2055
Last Name First Name	84	324	61	171	0	0	640
Last Name Birthday	250	1	76	0	0	0	327
First Name Birthday	69	0	18	0	2	0	89
Last Name	8	73	1	0	0	0	82
First Name	10	10	1	4	0	0	25
Birthday	31	0	0	0	0	0	31
Total	1980	408	675	175	11	0	3249
Total Cases of UW-Milwaukee, Cohort 2 Excluding Missing on All Three Criteria (% Match)							3926 (82.8%)
Total Cases of UW-Milwaukee, Cohort 2(% Match)							4135 (78.6%)

Table A-3

Match Cases Summary (WRCT, Cohort 3)

Matching Criteria	**Same School**		**Missing School Name**[a]		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	No Missing Criteria	Missing Criteria[b]	
Last Name First Name Birthday	1427		371		9		1807
Last Name First Name	60	438	36	261	1	0	796
Last Name Birthday	240	0	67	0	1	0	308
First Name Birthday	108	0	15	0	1	0	124
Last Name	8	58	0	0	0	0	66
First Name	3	33	0	1	0	0	37
Birthday	15	0	3	0	0	0	18
Total	1861	529	492	262	12	0	3156
Total Cases of UW-Milwaukee, Cohort 3 Excluding Missing on All Three Criteria (% Match)							3957 (79.8%)
Total Cases of UW-Milwaukee, Cohort 3(% Match)							4040 (78.1%)

Note a. Missing school name or/and district name in either or both (WRCT and UW-Milwaukee) data sets.

b. For two-criteria matching, missing criteria means the third criterion misses in either or both (WRCT and UW-Milwaukee) data sets. For one-criterion matching, missing criteria means either or both the other two criteria misses in either or both (WRCT and UW-Milwaukee) data sets.

Match Summary Between UW–Milwaukee Dataset and WKCE (99–01) by Cohort

Table A-4

Match Cases Summary (WKCE, Cohort 1)

Matching Criteria[a]	**Same School**[b]		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[c]	No Missing Criteria	Missing Criteria[c]	
Last Name First Name Birthday	2346	NA	420	NA	2766
Last Name First Name	49	21	36	22	128
Last Name Birthday	300	3	87	0	390
First Name Birthday	25	0	13	0	38
Last Name	9	0	1	0	10
First Name	0	0	0	0	0
Birthday	4	0	1	0	5
Total	2733	24	558	22	3337
Total Cases of UW-Milwaukee, Cohort 1 Excluding Missing on All Three Criteria (% Match)					4255 (78.4%)
Total Cases of UW-Milwaukee, Cohort 1(% Match)					4440 (75.2%)

Table A-5

Match Cases Summary (WKCE, Cohort 2)

Matching Criteria[a]	**Same School**[b]		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[c]	No Missing Criteria	Missing Criteria[c]	
Last Name First Name Birthday	2238		446		2684
Last Name First Name	37	23	41	28	129
Last Name Birthday	181	0	79	0	260
First Name Birthday	15	0	17	0	32
Last Name	8	0	4	0	12
First Name	1	0	1	0	2
Birthday	5	0	1	0	6
Total	2485	23	589	28	3125
Total Cases of UW-Milwaukee, Cohort 2 Excluding Missing on All Three Criteria (% Match)					3926 (79.6%)
Total Cases of UW-Milwaukee, Cohort 2(% Match)					4135 (75.6%)

Table A-6

Match Cases Summary (WKCE, Cohort 3)

Matching Criteria[a]	**Same School[b]**		**Different School**		**Total**
	No Missing Criteria	Missing Criteria[c]	No Missing Criteria	Missing Criteria[c]	
Last Name First Name Birthday	2122		454		2576
Last Name First Name	35	33	28	29	125
Last Name Birthday	175	0	91	0	266
First Name Birthday	44	0	33	0	77
Last Name	5	0	3	0	8
First Name	2	0	0	0	2
Birthday	11	0	4	0	15
Total	2394	33	613	29	3069
Total Cases of UW-Milwaukee, Cohort 3 Excluding Missing on All Three Criteria (% Match)					3957 (77.6%)
Total Cases of UW-Milwaukee, Cohort 3(% Match)					4040 (76.0%)

Note a. The criteria from both UW-Milwaukee and WRCT are used to match WKCE. Match is based on the matching criteria between either UW-Milwaukee or WRCT and WKCE.
b. Same school is defined as the school name in grade 3 in either UW-Milwaukee or WRCT, same as the school name in grade 4 in WKCE.
c. For two criteria matching, missing criteria means the third criterion misses in either UW-Milwaukee and WRCT or WKCE, or all three data sets. For one criteria matching, missing criteria means either or both the other two criteria misses in either UW-Milwaukee and WRCT or WKCE, or all three data sets.

Appendix B

Enrollment in SAGE and Contrast Schools by Cohort, Grade, and School Year

Table B-1

Enrollment in SAGE and Contrast Schools by Cohort, Grade, and Year

	Cohort I Number of Students							Cohort II Number of Students							Cohort III Number of Students							Percent of Students By Cohort					
Cohort	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	1	2	3	1	2	3
Grade	1	1	2	2	3	3	4	1	1	2	2	3	3	4	1	1	2	2	3	3	4	4/1	4/1	4/1	4/3	4/3	4/3
School YR	96-97	96-97	97-98	97-98	98-99	98-99	99-00	97-98	97-98	98-99	98-99	99-00	99-00	00-01	98-99	98-99	99-00	99-00	00-01	00-01	01-02						
Sch YR (1-6)	1	1	2	2	3	3	4	2	2	3	3	4	4	5	3	3	4	4	5	5	6						
Fall/Spring	F	s	f	s	f	s	s	f	s	F	s	f	s	s	f	s	f	s	f	s	s						
SAGE ID																											
101	74	74	65	64	60	59	53	75	74	65	66	66	64	54	70	76	65	63	64	63	57	0.72	0.72	0.81	0.88	0.82	0.89
102	47	47	42	37	33	34	35	49	48	41	41	32	32	29	46	46	43	43	38	38	34	0.74	0.59	0.74	1.06	0.91	0.89
103	55	55	41	40	37	36	32	32	33	24	26	24	25	22	54	56	45	43	38	39	32	0.58	0.69	0.59	0.86	0.92	0.84
104	43	43	42	41	41	41	38	35	35	32	33	32	32	28	26	30	25	24	23	23	21	0.88	0.80	0.81	0.93	0.88	0.91
105	24	24	20	20	16	17	15	15	15	15	15	14	14	13	21	20	19	16	16	16	15	0.63	0.87	0.71	0.94	0.93	0.94
106	32	31	17	18	11	10	10	29	28	20	21	26	20	16	34	34	25	22	21	20	15	0.31	0.55	0.44	0.91	0.62	0.71
107	72	72	51	38	34	34	29	70	68	50	51	47	40	32	48	44	38	32	25	24	26	0.40	0.46	0.54	0.85	0.68	1.04
108	28	28	20	16	14	15	12	33	34	24	24	18	18	15	30	21	16	15	10	11	10	0.43	0.45	0.33	0.86	0.83	1.00
109	59	59	49	49	50	51	44	52	52	52	51	48	48	43	46	45	41	41	40	40	39	0.75	0.83	0.85	0.88	0.90	0.98
110	36	36	31	31	28	28	24	31	33	28	28	26	26	24	31	31	25	24	17	17	18	0.67	0.77	0.58	0.86	0.92	1.06
111	40	41	22	23	20	20	14	48	46	38	38	33	30	19	45	42	36	32	31	29	29	0.35	0.40	0.64	0.70	0.58	0.94
112	27	27	24	24	23	23	22	25	25	25	25	24	24	19	25	25	22	22	21	19	17	0.81	0.76	0.68	0.96	0.79	0.81
113	67	67	40	37	30	28	21	46	43	39	37	31	28	22	56	49	31	35	32	33	27	0.31	0.48	0.48	0.70	0.71	0.84
114	92	92	80	81	73	71	62	83	83	78	78	67	67	58	102	99	86	83	78	76	65	0.67	0.70	0.64	0.85	0.87	0.83
115	102	102	75	75	61	63	49	80	77	61	61	55	55	44	76	73	56	54	45	44	40	0.48	0.55	0.53	0.80	0.80	0.89
116	79	79	14	13	41	41	31	75	71	55	56	51	45	32	77	77	62	59	50	51	37	0.39	0.43	0.48	0.76	0.63	0.74
117	140	142	76	71	57	57	44	146	138	107	104	78	76	46	120	118	95	91	81	72	60	0.31	0.32	0.50	0.77	0.59	0.74
118	130	133	86	85	58	54	33	109	108	98	95	85	77	35	139	116	95	75	59	57	46	0.25	0.32	0.33	0.57	0.41	0.78
119	91	91	65	65	53	50	29	85	85	63	61	49	53	28	79	79	58	55	35	34	21	0.32	0.33	0.27	0.55	0.57	0.60

Table B-1 (Continued)

Enrollment in SAGE and Contrast Schools by Cohort, Grade, and Year

	Cohort I Number of Students							Cohort II Number of Students							Cohort III Number of Students							Percent of Students By Cohort					
Cohort	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	1	2	3	1	2	3
Grade	1	1	2	2	3	3	4	1	1	2	2	3	3	4	1	1	2	2	3	3	4	4/1	4/1	4/1	4/3	4/3	4/3
School YR	96-97	96-97	97-98	97-98	98-99	98-99	99-00	97-98	97-98	98-99	98-99	99-00	99-00	00-01	98-99	98-99	99-00	99-00	00-01	00-01	01-02						
Sch YR (1-6)	1	1	2	2	3	3	4	2	2	3	3	4	4	5	3	3	4	4	5	5	6						
Fall/Spring	f	s	f	s	f	s	s	f	s	f	s	f	S	s	f	s	f	s	f	s	s						
SAGE ID																											
120	58	57	37	32	22	22	19	42	44	26	25	21	22	18	34	33	22	22	18	19	16	0.33	0.43	0.47	0.86	0.86	0.89
121	126	129	73	67	53	52	41	122	109	78	77	57	54	38	113	113	91	83	63	59	38	0.33	0.31	0.34	0.77	0.67	0.60
122	66	67	46	46	39	37	30	68	69	61	61	44	50	35	57	57	57	53	48	47	34	0.45	0.51	0.60	0.77	0.80	0.71
123	24	24	24	24	25	23	20	15	15	16	16	16	16	14	28	28	26	25	24	23	19	0.83	0.93	0.68	0.80	0.88	0.79
124	21	21	21	21	20	18	19	18	18	17	17	15	15	14	19	19	19	18	16	16	13	0.90	0.78	0.68	0.95	0.93	0.81
125	41	41	35	37	32	32	29	26	26	25	24	26	27	23	41	39	36	35	31	32	32	0.71	0.88	0.78	0.91	0.88	1.03
126	9	9	7	7	6	6	6	3	3	2	2	3	3	2	6	6	3	3	3	3	3	0.67	0.67	0.50	1.00	0.67	1.00
127	72	74	55	58	52	47	40	60	60	47	45	44	42	40	67	67	53	51	50	47	43	0.56	0.67	0.64	0.77	0.91	0.86
128	60	58	48	50	42	43	44	84	82	75	80	73	69	60	67	65	55	54	51	51	44	0.73	0.71	0.66	1.05	0.82	0.86
129	12	12	9	9	7	8	6	11	10	9	9	8	8	3	16	16	15	14	14	16	13	0.50	0.27	0.81	0.86	0.38	0.93
130	74	74	66	64	65	65	59	58	57	52	54	52	53	49	52	55	47	46	44	46	45	0.80	0.84	0.87	0.91	0.94	1.02
All SAGE Schools	1801	1809	1281	1243	1103	1085	910	1625	1589	1323	1321	1165	1133	875	1625	1579	1307	1233	1086	1065	909	0.50	0.54	0.56	0.82	0.75	0.84

Table B-1 (Continued)

Enrollment in SAGE and Contrast Schools by Cohort, Grade, and Year

	Cohort I Number of Students							Cohort II Number of Students							Cohort III Number of Students							Percent of Students By Cohort					
Cohort	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	1	2	3	1	2	3
Grade	1	1	2	2	3	3	4	1	1	2	2	3	3	4	1	1	2	2	3	3	4	4/1	4/1	4/1	4/3	4/3	4/3
School YR	96-97	96-97	97-98	97-98	98-99	98-99	99-00	97-98	97-98	98-99	98-99	99-00	99-00	00-01	98-99	98-99	99-00	99-00	00-01	00-01	01-02						
Sch YR (1-6)	1	1	2	2	3	3	4	2	2	3	3	4	4	5	3	3	4	4	5	5	6						
Fall/Spring	f	s	F	s	f	s	s	f	s	f	s	f	s	s	f	s	f	s	f	s	s						
Comparison Schools ID																											
201	23	23	17	17			14	23	23					17							1	0.61	0.74				
202	55	55					39						1	2							5	0.71					
203	69	69	47	44	32	29	24	70	69	50	52	50	46	35	49	49	39	36	29	29	24	0.35	0.50	0.49	0.75	0.70	0.83
204	59	58					24							1							2	0.41					
206	73	73	49	46			36	63	60				1	42							4	0.49	0.67				
207	70	70	60	60	57	55	53	83	85	63	63	59	58	56	70	69	66	63	59	58	54	0.76	0.67	0.77	0.93	0.95	0.92
208	83	83	51	48	46	44	41	75	67	52	51	44	46	40	49	43	36	35	26	26	23	0.49	0.53	0.47	0.89	0.91	0.88
209	86	86	62	55	5	34	45	116	100	82	83	79	77	54	83	83	62	61	52	53	37	0.52	0.47	0.45	9.00	0.68	0.71
210	120	120	77	76	66	67	53	118	114	76	77	68	64	43	112	109	76	72	66	66	40	0.44	0.36	0.36	0.80	0.63	0.61
211	97	98	45	40	40	39	27	98	80	62	61	44	42	33	94	91	59	56	43	41	37	0.28	0.34	0.39	0.68	0.75	0.86
212	57	57	48	45	42	41	38	55	52	49	48	29	38	33	54	56	40	42	33	32	26	0.67	0.60	0.48	0.90	1.14	0.79
213	109	109	86	85	80	76	69	116	115	85	85	67	65	48	106	103	89	86	80	80	68	0.63	0.41	0.64	0.86	0.72	0.85
214	53	53	31	29	24	20	17	55	52	37	42	39	37	29	56	2					17	0.32	0.53	0.30	0.71	0.74	
215	56	56	38	34	33	33	25	59	59	43	42	37	36	24	58	58	44	39	39	39	27	0.45	0.41	0.47	0.76	0.65	0.69
216	35	35	31	30			28	47	47					35							2	0.80	0.74				

Table B-1 (Continued)

Enrollment in SAGE and Contrast Schools by Cohort, Grade, and Year

	Cohort I Number of Students							Cohort II Number of Students							Cohort III Number of Students							Percent of Students By Cohort					
Cohort	1	1	1	1	1	1	1	2	2	2	2	2	2	2	3	3	3	3	3	3	3	1	2	3	1	2	3
Grade	1	1	2	2	3	3	4	1	1	2	2	3	3	4	1	1	2	2	3	3	4	4/1	4/1	4/1	4/3	4/3	4/3
School YR	96-97	96-97	97-98	97-98	98-99	98-99	99-00	97-98	97-98	98-99	98-99	99-00	99-00	00-01	98-99	98-99	99-00	99-00	00-01	00-01	01-02						
Sch YR (1-6)	1	1	2	2	3	3	4	2	2	3	3	4	4	5	3	3	4	4	5	5	6						
Fall/Spring	f	s	f	s	f	s	s	f	s	f	s	F	s	s	f	s	f	s	f	s	s						
Contrast Schools ID																											
217	33	33	29	24			23	36	35					32							3	0.70	0.89				
218					1	1	1			1	1	4	3	4	20	19	16	15	14	11	10			0.50	1.00	1.00	0.71
219												1	1		61	60	55	52	44	43	40			0.66		0.00	0.91
220										1	1	3	1		72	71	57	54	48	47	42			0.58		0.00	0.88
221										2		5	10		70	59	50	47	38	37	26			0.37		0.00	0.68
222					5	5	7			2	1	2	2	5	55	53	43	42	41	42	39			0.71	1.40	2.50	0.95
223					4	4	4			4	3			6	65	66				1	55			0.85	1.00		
224										1	1				46	46	34	31	35	33	34			0.74			0.97
All Contrast Schools	1078	1078	671	633	435	448	568	1014	958	610	611	531	528	539	1120	1037	766	731	647	638	616	0.53	0.53	0.55	1.31	1.02	0.95
All Schools	2879	2887	1952	1876	1538	1533	1478	2639	2547	1933	1932	1696	1661	1414	2745	2616	2073	1964	1733	1703	1525	0.51	0.54	0.56	0.96	0.83	0.88

Appendix C

The SAGE Evaluation Model

The SAGE Evaluation Model

Consider a two-level model of the effect of participating in SAGE during the 1st grade on student achievement measured at the end of that grade. Level one is the student level. Level two is the school level. (Below, we extend the model to three levels to explicitly incorporate teacher effects.) The model is defined for a given cohort. Thus, cohort and time subscripts are not needed. Later, we will extend the model to allow for the effects of participation in SAGE in 2nd and 3rd grade on student achievement in 2nd, 3rd, and 4th grade and will consider how to pool data and aggregate estimates across the three SAGE cohorts. Define the following indices, variables, and parameters:

$i =$ student

$j =$ school

$g =$ grade (1, 2, 3 and 4)

$Y^*_{gi} =$ true student achievement in a given subject area measured at the end of grade g

$Y_{gi} =$ measured student achievement in a given subject area measured at the end of grade g

$= Y^*_{gi} + v_{gi}$

$v_{gi} =$ the error in measured achievement

$X_i =$ a vector of student demographic characteristics (predictors of achievement growth)

$P_{gi} =$ a zero-one indicator that is equal to 1 if a student is enrolled in a SAGE school, 0 otherwise

$S_{gij} =$ a school indicator equal to 1 if a student is enrolled in a school j, 0 otherwise

$\alpha_{gj} =$ a fixed school effect for school j normed to have a weighted (by school size) mean of zero

$\omega_{gj} =$ a random school effect for school j with mean zero

$e_{gi} =$ a random student-level error component with mean zero

$\mu_g =$ the student-level intercept and γ_g and β_g are conformable parameters

$\eta_{g0} =$ the school-level intercept

$\eta_g =$ the effect on achievement growth of participating in the SAGE program in grade g.

Model of Student Achievement Growth in 1st Grade

The student and school levels of the two-level model of 1st grade achievement are given, respectively, by: [1]

$$\text{Student Level: } Y_{1i} = \mu_1 + \gamma_1 Y_{0i} + \beta_1 X_i + \sum_{j=1} \alpha_{1j} S_{1ij} + \varepsilon_{1i} \qquad (1)$$

$$\text{School Level: } \alpha_{1j} = \eta_{10} + \eta_1 P_{1j} + \omega_{1j} \qquad (2)$$

[1] One implicit assumption of the model is that students stay enrolled in a given SAGE or contrast school for the entire school year. Hence, the school indicators take on values of one or zero. This is a reasonable approach given the fact that most of the students who change schools drop from the sample rather than move from one study school to another.

This two-level model can be estimated using either a two-step or one-step procedure. The two-step procedure is useful because it yields value-added estimates of school effectiveness $\hat{\alpha}_{gt}$ from step one (and associated standard errors). As is demonstrated later in this report, these school effect estimates can be analyzed and graphed to obtain additional insights into the sources of school productivity. To implement the two-step procedure, estimate the student-level equation using ordinary least squares.[2] (This equation is a standard analysis of covariance (ANCOVA) model.)

In step two, estimate the school-level model using the estimated school effect as the dependent (left-hand-side) variable. To properly estimate this model, it is necessary to account for the fact that the observed school effect is an estimate rather than a true (population) variable. Let u_{gj} = the error in estimating α_{gt} so that $\hat{\alpha}_{gj} = \alpha_{gj} + u_{gj}$. Then, the observable school-level equation is given by

$$\begin{aligned} \hat{\alpha}_{1j} &= \eta_{10} + \eta_1 P_{1j} + \xi_{1j} \\ \xi_{1j} &= \omega_{1j} + u_{1j}. \end{aligned} . \qquad (3)$$

Note that the variance of the error due to estimation error u_j varies across schools (indexed by *j*).[3] Estimates of these variances are provided by step one (in the form of standard errors for the estimated school effects normed around zero). Given the nonconstant variance of the error term, inefficient, but unbiased, estimates of the observable school-level equation can be obtained using ordinary least squares (OLS) (although the reported standard errors are incorrect). Efficient and unbiased estimates can be obtained using weighted least squares (WLS).[4]

Alternatively, the model can be estimated in one step. To obtain a single equation suitable for estimation, substitute the school-level equation into the student-level equation (eliminating the fixed school effect α_j):

[2] To impose the restriction that the mean school effect is zero (and to obtain appropriate standard errors for the school effect estimates), the left- and right-hand-side variables (excluding the school indicators) are normalized so that the sample mean for each variable equals zero. This normalization is equivalent to subtracting the weighted mean school effect from the estimated school effects if the effects are estimated without imposing the mean zero restriction. The major advantage of imposing this restriction in the original estimation is that OLS automatically produces standard error estimates that can be used to produce optimal weights for step two.

[3] We assume, as is traditional in these models, that the variance of the in-equation error ω_j is constant across schools.

[4] The error terms could also be correlated across schools (a violation of OLS assumption that observations are independent). In our context (and because the schools effects are normed around zero), the correlation of estimation error across school observations is small and thus can be ignored. This is confirmed by the fact that the one- and two-step estimates are very similar.

Multilevel (Student and School) Equation:

$$Y_{1i} = \mu_1 + \gamma_1 Y_{0i} + \beta_1 X_i + \sum_{j=1} (\eta_{10} + \eta_1 P_{1j} + \omega_{1j}) S_{1ij} + \varepsilon_{1i} \quad (4)$$

$$Y_{1i} = (\mu_1 + \eta_{10}) + \gamma_1 Y_{0i} + \beta_1 X_i + \eta_1 P_{1i} + \sum_{j=1} \omega_{1j} S_{1ij} + \varepsilon_{1i}.$$

(In this equation, the SAGE program indicator is subscripted by *i* to indicate that it exists in the student-level dataset as a single student-level variable.) To obtain correct standard errors (and proper significance levels), it is necessary to estimate the multilevel equation using a procedure that handles the random school-level component $\omega_{1j.}$ We use the SAS procedure MIXED for this purpose.

It should be noted that if the multilevel equation is estimated using OLS, the reported standard errors and statistical significance levels are typically enormously biased. In our context, the OLS standard errors were biased downward by more than 50%.

Comments on the Experimental Approach

The above model provides a convenient framework for discussing the quasi-experimental approach used in the SAGE study. In our context, there are two assignment dimensions: (1) the assignment of treatment or contrast status to schools and (2) the assignment of students to schools. Pure random assignment with respect to both dimensions (if successful) guarantees that the program variable P_{1i} or P_{1j} is uncorrelated (in expectation) with the student-level components Y_{0i}, X_i and ε_{1i} and the school-level random component ω_{1j}. This powerful result implies that unbiased estimates of program effects can be obtained simply by differencing the mean outcomes for students in the treatment group and students in the contrast group. In other words, it is not even necessary to contrast for other determinants of student achievement (initial achievement Y_{0i} and demographic characteristics X_i). Furthermore, given random assignment, unbiased estimates of program effects can be obtained from a model that includes regressors Y_{0i} and X_i even if the variables are measured with error and/or endogenous and correlated with student and school-level random components. The advantage of including regressors of this type is that they typically explain much of the variance of achievement in 1st grade (and later grades) even if they are measured with error or otherwise endogenous. As a result, inclusion of these variables typically yields more precise estimates of program effects relative to estimates based on mean differences.

In this report, we assume that the quasi-experimental design used in the SAGE evaluation approximates the conditions of a randomized design. As a result, the models presented in this section should be robust to measurement error in baseline achievement and other sources of endogeneity of baseline achievement.[5] If this assumption is false, the estimates of the SAGE effect could be systematically biased upwards and downwards.

[5] As is discussed in the next section, we do not assume that randomization guarantees that student test scores *after* 1st grade are uncorrelated with assignment to a SAGE or contrast school. If the SAGE program is effective (either positively or negatively), student test scores after 1st grade must be correlated with program status.

This assumption can be checked in at least one respect. One implication of a randomized design is that the students in the treatment and contrast groups should not be systematically different. This implication can be tested with respect to baseline (Fall of 1st grade) achievement and the demographic characteristics of the SAGE and contrast-group samples. As indicated in Table 7 (in the main report), the SAGE and contrast group samples are quite similar with respect to average characteristics (see the discussion in the report). Unfortunately, we cannot check whether the SAGE and contrast schools differ with respect to unmeasured student (ε_{gi}) or school factors (ω_{gj}) that contribute to achievement growth.

A Model of Student Achievement Growth in 2nd Grade

Now, consider the almost identical model of achievement growth from the end of 1st grade to the end of 2nd grade:

$$Y_{2i} = (\mu_2 + \eta_{20}) + \gamma_2 Y_{1i} + \beta_2 X_i + \eta_2 P_{2i} + \delta_2 M_{2i} + \sum_j \omega_{2j} S_{2ij} + \varepsilon_{2i} \qquad (5)$$

where M_{2i} = a zero-one indicator of whether the student transferred to a different school from 1st grade to 2nd grade and δ_2 is the corresponding parameter. Estimating this equation raises several issues. First, it is possible that school mobility (as measured by M_{2i}) and program status in 2nd grade P_{2i} (generally different from P_{1i} for mobile students) are non-random events (even after controlling for demographic variables and baseline achievement). In particular, the unmeasured components of achievement growth for SAGE and contrast-group students could differ, thereby making it difficult (if not impossible) to obtain unbiased parameter estimates (including estimates of the effectiveness of the SAGE program).

In practice, this concern is overshadowed by a second factor: students who moved to a non-study school after 1st grade were not subsequently tested (at least, not until the state tests were administered in 3rd grade and 4th grade) and thus lost from the sample. Eliminating these students from the sample (plus the few students who moved between study schools) could produce biased parameter estimates for the same reasons discussed above. (This bias is referred to as attrition bias.) We speculate that the biases due to attrition may be rather small, given that the mobility rates of SAGE and contrast group students are similar.[6]

Finally, one of the interesting features of this model is that the SAGE participation variable P_{2i} will be correlated with 1st grade achievement if the SAGE program has a nonzero effect. As a result, it is no longer the case that the SAGE effect (for 2nd grade) can be estimated using a difference of program and contrast-group means or using a simple regression model, as in the case of a random experimental design, or a quasi-experimental design that approximates the conditions of a random design. At a minimum, it is necessary to address the problem of measurement error in the pretest

[6] We can also we can partially sidestep the attrition bias problem by estimating models from different years using a common restricted sample that eliminates all mobile students. As indicated later in the report, we find that our results are quite robust to sample changes.

variable. For this reason, it is advantageous to recast the 2nd grade model so that it captures cumulative achievement growth from the beginning of 1st grade to the end of 2nd grade—two years of growth.

A Model of Cumulative Student Achievement Growth in 1st and 2nd Grade

Such a model can be derived by substituting (4) into (5), which yields:

$$Y_{2i} = \mu_{20} + \gamma_1\gamma_2 Y_{0i} + \mathrm{B}_{20}X_i + \gamma_2\eta_1 P_{1i} + \eta_2 P_{2i} + \delta_2 M_{2i} + \gamma_2\sum_{j=1}\omega_{1j}S_{1ij} + \sum_{j=1}\omega_{2j}S_{2ij} + \varepsilon_{20i} \quad (6)$$

where $\mu_{20} = \gamma_2(\mu_1 + \eta_{10}) + \mu_2 + \eta_{20}$, $\mathrm{B}_{20} = \gamma_2\beta_1 + \beta_2$, and $\varepsilon_{20i} = \gamma_2\varepsilon_{1i} + \varepsilon_{2i}$. This equation is very similar to the 1st grade model except for the following new factors. First, it contains two program variables (one for each grade) and two corresponding treatment effects: (1) $\gamma_2\eta_1$ = the effect, as of 2nd grade, of being in the SAGE program in 1st grade, and (2) η_2 = the effect, as of 2nd grade, of being in the SAGE program in 2nd grade. With respect to the first effect, the parameter γ_2 captures the durability of the SAGE effect one year after receiving the treatment. Second, it includes an indicator of school mobility set to one for students who changed schools between 1st and 2nd grade. Third, it includes random school components for both 1st and 2nd grade (ω_{1j} and ω_{2j}).

Unfortunately, not all of the parameters in this equation can be estimated due to the loss of mobile students (sample attrition). In particular, the mobility variable drops out of the model and the 1st and 2nd grade program effects merge due to the fact that the program variables are identical if there is no school mobility. (The random school components also merge.) The merged program effect $\theta_{22} \equiv (\gamma_2\eta_1 + \eta_2)$ measures the cumulative effect of being in the SAGE program in 1st and 2nd grade at the end of 2nd grade.[7] Using the same logic, we can derive a model of cumulative achievement growth from 1st through 3rd grade. The cumulative program effect in this case is equal to $\theta_{33} \equiv (\gamma_2\gamma_3\eta_1 + \gamma_3\eta_2 + \eta_3)$.

The bottom line is that if we assume that the SAGE experimental design achieved conditions that approximate a random experimental design and we assume school mobility was essentially random (given demographic characteristics and initial achievement), the data collected by the original study team support estimating the following SAGE program effects:

[7] With respect to the cumulative effect parameter θ_{22} the subscript "22" indicates the effect of two years in the SAGE program at the end of the 2nd year.

Outcome	Grade Effects	Parameter	Effect Estimated
1st Grade Achievement Y_{1i}	1st	$\theta_{11} \equiv \eta_1$	Effect of being in SAGE program in 1st grade at the end of 1st grade.
2nd Grade Achievement Y_{2i}	1st & 2nd	$\theta_{22} \equiv (\gamma_2\eta_1 + \eta_2)$	Cumulative effect of being in the SAGE program in 1st and 2nd grade at the end of 2nd grade.
3rd Grade Achievement Y_{3i}	1st, 2nd & 3rd	$\theta_{33} \equiv (\gamma_2\gamma_3\eta_1 + \gamma_3\eta_2 + \eta_3)$	Cumulative effect of being in the SAGE program in 1st, 2nd, and 3rd grade at the end of 3rd grade.

Fortunately, these are the most important effects.

Interestingly, the SAGE design does not support estimation of specific grade effects, other than the 1st grade effect (as discussed above). Note, however, that given an estimate of durability of the SAGE intervention from 1st to 2nd grade (as captured by the parameter γ_2), an estimate of the 2nd grade SAGE effect is given by $\eta_2 = \theta_{22} - \gamma_2\theta_{11}$. Similarly, an estimate of the 3rd grade SAGE effect is given by $\eta_3 = \theta_{33} - \gamma_3\theta_{22}$. The key to obtaining grade-specific effect estimates is knowing the durability of the SAGE intervention. Below, we consider a strategy for estimating the durability of the SAGE program over time. This analysis makes use of the 4th grade test-score data assembled for this project by merging data collected for the original SAGE evaluation with 4th grade WKCE data (and the 3rd grade WRCT data).

A Model of Cumulative Student Achievement Growth in 1st, 2nd, 3rd, and 4th Grade

In this section, we derive a model of cumulative achievement growth for the period 1st through 4th grade. As above, the cumulative model is obtained by recursively substituting grade-specific growth equations. The model of growth from 1st through 3rd grade is given by:

$$\begin{aligned} Y_{3i} = \mu_{30} + \gamma_1\gamma_2\gamma_3 Y_{0i} + \mathrm{B}_{30} X_i + \gamma_2\gamma_3\eta_1 P_{1i} + \gamma_3\eta_2 P_{2i} + \eta_3 P_{3i} \\ + \gamma_3\delta_2 M_{2i} + \delta_3 M_{3i} \\ + \gamma_2\gamma_3 \sum_{j=1} \omega_{1j} S_{1ij} + \gamma_3 \sum_{j=1} \omega_{2j} S_{2ij} + \sum_{j=1} \omega_{3j} S_{3ij} + \varepsilon_{30i}. \end{aligned} \tag{7}$$

The model of growth from 1st through 4th grade is given by:

$$\begin{aligned} Y_{4i} = \mu_{40} + \gamma_1\gamma_2\gamma_3\gamma_4 Y_{0i} + \mathrm{B}_{40} X_i + \gamma_2\gamma_3\gamma_4\eta_1 P_{1i} + \gamma_3\gamma_4\eta_2 P_{2i} + \gamma_4\eta_3 P_{3i} + \eta_4 P_{4i} \\ + \gamma_3\gamma_4\delta_2 M_{2i} + \gamma_4\delta_3 M_{3i} + \delta_4 M_{4i} \\ + \gamma_2\gamma_3\gamma_4 \sum_{j=1} \omega_{1j} S_{1ij} + \gamma_3\gamma_4 \sum_{j=1} \omega_{2j} S_{2ij} + \gamma_4 \sum_{j=1} \omega_{3j} S_{3ij} + \sum_{j=1} \omega_{4j} S_{4ij} + \varepsilon_{40i}. \end{aligned} \quad (8)$$

Note that this model includes a program effect for 4th grade (η_4). Since the SAGE program operates in kindergarten through 3rd grade, but not in 4th grade, we assume that this parameter is equal to zero (and thus drop it from the model). This would be an unwarranted assumption if we believed that school participation in the SAGE program had (positive or negative) effects in the grades where class size is not reduced.

As currently structured, the SAGE effects in the above models are grade-specific rather than cumulative. It is helpful to modify the models to recast these effects as cumulative rather than grade-specific for the following reasons. One, a major objective of the study is to estimate the cumulative effects of participation in the SAGE program for one, two, or three years. Two, cumulative effects would be directly comparable with the cumulative effects estimated for earlier grades.[8] Three, if school mobility is relatively low, there may be too little variation in the program indicators P_{1i}, P_{2i}, and P_{3i} to precisely estimate the separate effects of each variable. Four, the model as presented above cannot be estimated in its current form due to the fact that information on student mobility and school enrollment is missing (due to the design of the experiment) for students who moved from SAGE or a contrast school immediately after 1st grade.

The restructured cumulative model of 4th grade achievement is given by:

$$\begin{aligned} Y_{4i} = \mu_{40} + \gamma_1\gamma_2\gamma_3\gamma_4 Y_{0i} + \mathrm{B}_{40} X_i + \theta_{14} D_{1i} + \theta_{24} D_{2i} + \theta_{34} D_{3i} \\ + \pi_{14} F_{1i} + \pi_{24} F_{2i} + \pi_{34} F_{3i} \\ + \gamma_2\gamma_3\gamma_4 \sum_{j=1} \omega_{1j} S_{1ij} + \gamma_3\gamma_4 \sum_{j=1} \omega_{2j} S_{2ij} + \gamma_4 \sum_{j=1} \omega_{3j} S_{3ij} + \sum_{j=1} \omega_{4j} S_{4ij} + \varepsilon_{40i}. \end{aligned} \quad (9)$$

where D_{1i}, D_{2i}, and D_{3i} are zero/one indicators that a student was enrolled in the SAGE program for a duration of one, two, or three years, respectively; θ_{dg} represents (as before) the effect of *d* consecutive years in the SAGE program at the end of grade *g*; F_{1i}, F_{2i}, and F_{3i} are zero/one indicators that a student attended a SAGE or contrast school for a duration of one, two, or three years, respectively, and then moved to another school; and π_{dg} represents the differential effect of student mobility compared to students who stayed in the same SAGE or contrast school for all four years. The cumulative SAGE effects that can be estimated by the four achievement outcomes are summarized below.[9]

[8] Recall that it was necessary because of sample attrition to recast the SAGE effects as cumulative rather than grade-specific in the models based on the outcome data from the original SAGE study. Here, we are choosing to recast the SAGE effects.

[9] We also present estimates based on the 3rd grade WRCT test. Since this test is scored using a scale that is not comparable with the scale used for the other tests, the parameter estimates cannot be strictly compared.

Outcome	Grade Effects	Parameter	Effect Estimated
1st Grade Achievement Y_{1i}	1st	$\theta_{11} \equiv \eta_1$	Effect of being in SAGE program in 1st grade at the end of 1st grade.
2nd Grade Achievement Y_{2i}	1st & 2nd	$\theta_{22} \equiv (\gamma_2\eta_1 + \eta_2)$	Cumulative effect of being in the SAGE program in 1st and 2nd grade at the end of 2nd grade.
3rd Grade Achievement Y_{3i}	1st, 2nd & 3rd	$\theta_{33} \equiv (\gamma_2\gamma_3\eta_1 + \gamma_3\eta_2 + \eta_3)$	Cumulative effect of being in the SAGE program in 1st, 2nd, and 3rd grade at the end of 3rd grade.
4th Grade Achievement Y_{4i}	1st	$\theta_{14} = \gamma_2\gamma_3\gamma_4\eta_1$	Effect of being in SAGE program in 1st grade at the end of 4th grade.
4th Grade Achievement Y_{4i}	1st & 2nd	$\theta_{24} = \gamma_2\gamma_3\gamma_4\eta_1 + \gamma_3\gamma_4\eta_2$	Cumulative effect of being in the SAGE program in 1st and 2nd grade at the end of 4th grade.
4th Grade Achievement Y_{4i}	1st, 2nd & 3rd	$\theta_{34} = \gamma_2\gamma_3\gamma_4\eta_1 + \gamma_3\gamma_4\eta_2 + \gamma_4\eta_3$	Cumulative effect of being in the SAGE program in 1st, 2nd, and 3rd grade at the end of 4th grade.

It is important to point out that although we have derived the cumulative model of student achievement presented above using grade-specific equations, the cumulative models do not restrict the cumulative effect parameters to follow the pattern dictated by the grade-specific patterns. The cumulative effect parameters—the θ's —are free (unrestricted parameters). Nonetheless, one major advantage of pursuing the structural-model building approach used above is that it provides a useful way of interpreting the change in effect estimates over time. In particular, the durability of the SAGE intervention can be readily estimated. An estimate of the durability of the cumulative effect of being in the SAGE program in 1st, 2nd, and 3rd grade from the end of 3rd grade (the last year of the program) to the end of 4th grade (one year later) is given by:

$$\gamma_4 = \frac{\theta_{34}}{\theta_{33}}. \qquad (10)$$

It is also possible to estimate the durability parameters γ_2 and γ_3.[10] One of the benefits of obtaining estimates of the durability parameters is that they would allow us (in principle)

[10] It may be very reasonable to impose the assumption that the durability parameters at different grade levels are the same. This assumption, if true, would enhance the efficiency of the estimates.

to construct estimates of the grade-specific SAGE effects—the η's. It will, of course, be possible to produce statistically precise estimates of grade-specific SAGE effects only if the cumulative SAGE effects for every grade interval are estimated with high precision. It seems very likely that the cumulative effects of being in the SAGE program for a full three years (θ_{33} and θ_{34}) can be estimated precisely, since most of the students in the study followed this attendance pattern. Given the smaller sample size, it is likely that the effects of participating in the SAGE program for only one or two years will be estimated with less precision.

Additional Comments on the Cumulative Achievement Model

The cumulative achievement model captures school mobility in a useful, but simple, way. Rather than modeling grade-specific mobility events, it distinguishes mobility *patterns*, as defined by the year in which a student first dropped out of the original SAGE or contrast school that the student attended (or equivalently, the number of years that a student was continuously enrolled in the student's original study school). This approach parallels our characterization of the cumulative SAGE intervention. It has several advantages. In addition to controlling for the causal (presumably negative) effect of school mobility, it controls for possible unmeasured differences between different groups characterized by their mobility patterns. Secondly, it does not require information on mobility events after a student dropped out of the SAGE study, information that is not available for students who left the program after the 1st grade (or for students who left after the 2nd grade if they could not be matched with the 3rd grade WRCT data). This approach makes the assumption that the incidence of mobility is similar for SAGE and contrast-group students in the years following their attrition from the original SAGE study. This issue does not, of course, pertain to the vast majority of students who participated in the SAGE study from 1st to 3rd grade.

The cumulative achievement model includes random school components for four grades: $\omega_{1j}, \omega_{2j}, \omega_{3j}$ and ω_{4j}.

Appendix D

Annotated Bibliography on Class Size and Student Achievement

Annotated Bibliography

Class Size and Student Achievement

CSR Research Consortium. (2002, September). *Capstone Report: What we have learned about class size reduction in California.* G. W. Bohrnstedt & B. M. Stecher, Eds.

The California Department of Education and a group of California foundations awarded contracts to the American Institutes for Research, who along with RAND headed up a consortium to evaluate the effects of class-size reduction on achievement, on the quality of the state's teaching corps, on special needs students, and on other practices. The Consortium, which also included Policy Analysis for California Education (PACE), West Ed, and EdSource, has produced three evaluation reports thus far. This is the fourth and final report. In it, the authors summarize previous findings and discuss new research done in the final year of the contract. They also include a set of policy recommendations and conclude with lessons learned. Major findings are: 1) Implementation of CSR occurred rapidly, although it lagged in schools serving minority and low-income students; 2) Our analyses of the relationship of CSR to student achievement was inconclusive; 3) CSR was associated with declines in teacher qualifications and a more inequitable distribution of credentialed teachers; 4) CSR had only a modest effect on teacher mobility; 5) CSR implementation did not affect special education identification or placement; 6) Students in reduced size third-grade classes received more individual attention, but similar instruction and curriculum; 7) parents liked reduced size classes; 8) Classroom space and dollars were taken from other programs to support CSR; and 9) In spite of budget shortfalls districts are not projecting CSR cutbacks for 2002-03.

Egelson, P., Harman, P. & Achilles, C. M. (1996). Does class size make a difference? Recent findings from state and district initiatives. SERVE (SouthEastern Regional Vision for education Associated with the School of Education, University of North Carolina at Greensboro.

(From the Conclusion): The results of reduced class size studies found in this document are positive and add an important replication to class size research studies. The consistency in the findings of these studies may be attributed, part, to the duration of these studies, the numbers of pupils in the studies, and the fact that the studies are implemented in the early primary grades—the first years of schooling for the pupils in the studies. Reduced class size gives a good early start in school, which is important for student achievement and later success. Smaller class sizes make sense for all children but appear to especially benefit minority children as evidenced in the Tennessee study, the results in Nevada, and the reduced class size plan in Wisconsin.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A. & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, vol. 2, No. 1, May,* pp. 1-30.

This article reviews major studies of class size and student achievement, the policy context, and reasons that class size does and does not impact student

achievement. According to this research, class size is not the determining factor in improving student achievement—teachers must change their teaching practices in order to have an effect on achievement. Examples of teaching characteristics found to improve student achievement in small classes are working with small groups, developing personal relationships with students, and utilizing hand-on projects. An important policy consideration is whether or not quality teachers are available to staff the extra classrooms generated by reducing class size. Money for hiring qualified teachers, as well as adequate space for the extra classrooms must be made available. It could be more cost effective to increase teacher compensation in order to have more quality teachers in place (more knowledgeable about subject matter, etc.) than to reduce class size, particularly if qualified teachers are not available. Retrieved November 14, 2002 from http://www.psychologicalscience.org/pdfreq.cfm?PATH_INFO=/pdf/pspi/pspi2_1.pdf&VARACTION=GO&CFID=798988&CFTOKEN=73217902

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001, November). Does class size matter? *Scientific American,* November, 2001.

Reports on the significance of class size to student learning. Includes an overview of class size in various countries, the importance of teacher adaptability, and the Asian paradox of large classes allied to high test scores.

Finn, J. D. (1998, April*). Class size and students at risk: What is known? What is next?* A Commissioned Paper prepared by Jeremy D. Finn. National Institute on the Education of AT-Risk Students, Office of Educational Research and Improvement, U. S. Department of Education. Retrieved 10/22/2002 from http://www.ed.gov/pubs/ClassSize/title.html

(From the Forward and Summary). This report is an overview of recent research on the effects of class size on the academic performance and behavior of students at risk. In several ways, it is not a conventional literature review. It emphasizes one recent large-scale investigation, Tennessee's Project STAR (Student-Teacher Achievement Ratio).

Project STAR demonstrated that small classes benefit students in grades kindergarten through 3 academically. That pupil behaviors are affected was shown clearly in the STAR grade 4 follow-up (i.e., the LBS). Ratings of specific engagement dimensions revealed improvements in the expenditure of effort, initiative taking, and reduced disruptive and inattentive behavior in contrast to students in regular classes. Both of these outcomes—enhanced performance and academic engagement—are likely to be beneficial especially to students at risk. Yet results for this population have not been examined closely enough to reveal the extent to which this is so.

Although STAR provides some answers about the effectiveness of small classes, to date it provides only hints about other related questions. The key questions that remain include the long-term consequences of attending a small class, the interactions of instructional processes with class size, and the particular impact of small classes on students at risk. Other past and current studies provide some answers to these questions and more than a few hypotheses. Yet there remains a tremendous amount of work to be done.

Hanushek, E. A. (1998). The evidence on class size. Occasional Paper Number 98-1, W. Allen Wallis Institute of Political Economy, University of Rochester.

(From p. 35): It appears that the ultimate effect of any large-scale program to reduce class size will depend much more importantly on the quality of new teachers hired than on the effects of class size reductions per se. Variations in teacher quality have been shown to be extraordinarily important for student achievement, and the econometric studies providing such results indicate that these variations completely dominate any effects of altered class size. Thus, if new hires resulting from a class size reduction policy are above the average quality of existing teachers, average student performance is likely to increase. If below, average student performance is likely to fall with class size reductions. From past experience, there is little reason to believe that teacher quality of new teachers will be significantly different from that of existing teachers unless incentives facing schools also change.

Maier, P., UW-Milwaukee, A., Percy, S., Smith, P., & Zahorik, J. (1997, December). First year results of the Student Achievement Guarantee in Education Program. Submitted by the SAGE Evaluation Team Center for Urban Initiatives and Research, University of Wisconsin-Milwaukee.

(From the Executive Summary): Evaluation of the 1996-97 results of the SAGE program suggests that students in SAGE first grade classrooms have a consistent achievement advantage over first grade students in contrast schools. As a group, SAGE students scored significantly higher on the post-test in reading, language arts, and mathematics. The total score of SAGE students was also significantly higher than the total score of contrast group students. The achievement advantage associated with participation in the SAGE program is revealed both in the analysis of individual student scores and in the analysis of averaged classroom scores.

African American students in SAGE first grade classrooms outperformed their counterparts in contrast school classrooms on the post test. It appears that African American males, in particular, may benefit from participation in the SAGE program.

As a group, first grade African American students scored lower than White students on the pre-test in both SAGE and contrast schools. The analysis of post-test results suggest that the gap in the achievement between African American students, as a group, and White students, as a group, widened in contrast school first grade classrooms during the 1996-97 school year. In contrast, African American students, as a group, and White students, as a group, appeared to increase their achievement by similar amounts in SAGE school first grade classrooms.

McRobbie, J. (1996, Fall). Smaller classes aim to launch early literacy. WestEd Policy Support Program. Focus Magazine—12 pages. Retrieved 10/17/2002 from http://web.wested.org/online_pubs/focus-fall96/csr.htm

Makes the point that class size reduction is not enough. Changing instruction is the key. Class-size reduction and teacher support must go hand-in-hand, and smaller classes must be part of a comprehensive approach. References a couple of studies: one of Austin where 15 schools with poor student performance received money for a five-year class size reduction program. After four years, only two schools showed dramatic gains in student achievement; those schools who significantly changed their instruction practices.

In another study of Nevada, a preliminary evaluation showed mixed results on testing but "improvements in other areas such as teacher-student interaction, monitoring of student work, feedback to students on progress, and small group and one-to-one instruction." (p. 5)

McRobbie, J., Finn, J. D., Harman, P. (1998, August). Class size reduction: Lessons learned from experience. WestEd's Policy Program. Policy Brief #23, August, 1998. Retrieved October 17, 2002 from http://www.wested.org/pub/docs/policy/class_red.htm

(From "Effect on Student Achievement" section):

1. STAR's small classes had 12-17 students, while the control "large" group had 22-26. Children who gained most from smaller classes were minority students and those in inner-city schools. And the benefits lasted, at least through 7th grade. Several recent smaller studies generally support STAR's findings, notably in terms of gains for urban minority students.
2. In Project STAR, Tennessee schools were "laboratories" for class size research. Certain conditions prevailed without which the positive effects of small classes may not occur: a) adequate supply of good teachers. In Tennessee, all STAR teachers were state certified and qualified to teach in their assigned grades; b) sufficient classroom space; c) a representative student mix in each class; d) teacher access to adequate materials and services.
3. Some studies have found that small classes allow teachers to spend more time on instruction and less on classroom management: a) more interaction between teachers and individual students; b) more "on task events; c) more small group instruction; d) faster coverage of curriculum in more depth; e) students are better engaged in learning
4. STAR researchers believe that the greater the class size beyond 17, the less the likelihood that the outcomes will be as positive. Earlier research suggested that the most dramatic gains accrue when class size is 15 or below.
5. Project STAR found that the defining feature of success is smallness itself. STAR analysts conclude that only smallness reduces the number of institutional events, creates an environment in which every student becomes engaged in learning, and allows the teacher to attend to every student.

6. The STAR research as well as a smaller study done in North Carolina suggests that the main benefits occur in the first year a student is in a small class and are sustained or increase slightly after that. Researchers as yet cannot say whether one year of small classes may be just as effective as three or four and, if so, which age or grad level should be the focus. STAR data are being re-analyzed to answer these questions.

Molnar, A. (1999, June). *Smaller classes and educational vouchers: A research update*. Harrisburg, PN: Keystone Research Center.

There is strong evidence that reducing class size in the primary grades increases student achievement, while voucher programs do not have a strong impact on achievement. He cites evidence from Tennessee's STAR program: a) advantages of having attended small classes increased as children reached higher grades; b) stronger evidence now exists that the benefits of smaller classes are cumulative; c) students who attended small classes in Tennessee took college entrance exams at significantly higher rates than their peers who attended regular-size classes; more small-class students graduated from high school on schedule.

He also cites evidence from SAGE: a) after two years, the impact of reduced class size in Wisconsin's SAGE program appears to be similar to the impact of smaller classes in Tennessee. b) In both 1996-97 and 1997-98, students in small first-grade classes achieved bigger increases in test scores in language arts, reading, and mathematics; c) From fall 1997 to spring 1998, first-grade African-American students in small classes reduced the achievement gap with white students by 19 percent (while in contrast schools, the gap grew 58 percent); d) In 1997-98, student achievement in SAGE first-grade classes with one teacher and 15 students was not significantly different from achievement in classes with two teachers and 30 students. This suggests that school districts may not need to construct new schools and classrooms to achieve the benefits of smaller classes.

Molnar, A., Smith, P., & Zahorik, J. (1998, December). 1997-98 Evaluation Results of the Student Achievement Guarantee In Education (SAGE) Program. Milwaukee, WI: University of Wisconsin-Milwaukee School of Education.

(From the Executive Summary):
<u>First Grade</u>

- Students in SAGE classrooms achieved significantly higher scores than students in contrast school classrooms in all tested areas: mathematics, reading, and language arts. The total scores of students in SAGE classrooms were also significantly higher than those of contrast school students
- African-American SAGE students scored lower than African-American students in contrast schools in the fall pre-test, but made significantly larger gains than contrast school students from pre-test to post-test. The test scores of African-American students in SAGE classrooms surpassed those of African American students in contrast school classrooms on the spring post-test.
- African-American students in SAGE classrooms achieved greater gains on the CTBS total score than white SAGE students from pre-test to post-test, reducing the achievement gap. In contrast, African-American students in

contrast school classrooms achieved lesser gains and the gap in achievement between African-American students and white students widened.

Second Grade

- The achievement advantage of students in SAGE first-grade classrooms in 1996-97 appears to be maintained in second grade in 1997-98. The advantage, however, does not appear to have increased significantly.

Classroom Configuration

- In comparing the performance of classrooms with a 15:1 student-teacher ratio to classrooms with a 30:2 student-teacher ratio, researchers found no achievement advantage for the one teacher to fifteen students form of classroom organization on any of the CTBS sub-tests.

Molnar, A., Smith, P., & Zahorik, J. (1999, December). 1998-99 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program. Milwaukee, WI: Unviersity of Wisconsin-Milwaukee, School of Education.

(From the Executive Summary):

First Grade

- Test scores of SAGE and comparison schools show statistically higher performance of SAGE students in language arts, mathematics and total scores on the post-test in 1998-99.
- In 1997-98 and again in 1998-99, African American SAGE students cored lower on the CTBS pre-test than African American comparison school students, but made significantly larger gains than comparison school students from pre-test to post-test, surpassing African American comparison school students on the post-test.
- In both 1997-98 and 1998-99 African American students scored significantly lower than white students on the pre-test total scale score for both SAGE and comparison schools. African American SAGE students achieved greater gains on the total scale score than White SAGE students from pre-to post-test, closing the achievement gap. At the same time, African Americans in comparison schools achieved lesser gains, and the achievement gap with their White comparison school classmates widened.

Second Grade

- African American SAGE students cored significantly higher than African American comparison school students on every sub-test and total scores in 1998-99
- When looking at gains made in 1998-99 from the first grade post-test to the second grade test, SAGE African American students made the same significant gains that the white SAGE students did, and did close the achievement gap between African American and White SAGE students although the relative gain was not significant.

Third Grade

- SAGE students scored significantly higher in reading, language arts, math, and total score than comparison students on the third grade test in 1998-99
- In 1998-99, test results suggest that statistically significant positive effects of SAGE occurred in first grade, were maintained in second and third grade

- In 1998-99, African American SAGE students performed significantly higher on every sub-test and total score over African American comparison students on the third grade test.
- When second grade is used for a baseline score, African American SAGE students outperform African American comparison students in reading, math, and total in 1998-99
- African American students gained significantly more than SAGE White students in third grade, closing the achievement gap. Comparison school African American students did not gain significantly from second to their grade. In comparison schools the gap between the performance of White and African American students widened.

Additional Analyses

- Classrooms with 30:2 student-teacher ratios achieved just as well as classrooms with 15:1 student-teacher ratios with the exception of language arts and mathematics sub-tests in second grade.
- In 1998-99 at the class level of analysis, smaller classrooms tended to score significantly higher in language arts, mathematics and reading, as well as total score after adjusting for individual pre-test results, socio-economic status and attendance. In other words, classrooms with fewer students are more likely to have higher class average achievement scores.

Molnar, A., Smith, P. & Zahorik, J. (2000, December). 1999-2000 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program. Milwaukee, WI: University of Wisconsin-Milwaukee, School of Education.

(From the Executive Summary):

Second Grade

- When adjusted for pre-existing differences, second grade SAGE students showed a significant achievement advantage over their comparison group counterparts in all areas. The exception was in the area of reading when first grade post-test was used to adjust for achievement differences.
- African American SAGE second graders scored significantly higher than African American comparison school students in mathematics, language arts, and total scores at the end of second grade. When examining gains made in 1998-00 from the first grade pre-test to the second grade, African American SAGE students made significantly larger gains than their comparison school counterparts on the total scale score as well as every sub-test except for language arts. When using the first grade post-test as the baseline, African American SAGE students again made larger gains in 1999-00 on every test except for reading, but the gains were not statistically significant.
- African American students, as a group, scored significantly lower than White students on total scale scores and on all sub-tests, regardless of whether they were in SAGE or comparison schools, although the gap between African Americans and Whites is larger in comparison schools.

Third Grade

- When adjusted for pre-existing differences in academic achievement, attendance, socioeconomic status and race, SAGE students showed significant improvement over their comparison school counterparts from the beginning of first grade to the end of third grade across all academic areas. From the beginning of second grade (first grade post test) significant additional differences in gain were seen in mathematics. From the beginning of third grade no significant additional differences in gain advantage were found.

- African American students continued to score significantly lower than White students on total scale score and on all sub-tests, regardless of whether they were SAGE or comparison school students. Gains made by African American versus White students were significantly better in SAGE schools from the beginning of first grade to the end of third grade. The opposite pattern was observed in comparison schools.

. Molnar, A. & Others (2001, December). 2000-2001 Evaluation Results of the Student Achievement Guarantee in Education (SAGE) Program. Milwaukee: WI: University of Wisconsin-Milwaukee, School of Education.

(From the Executive Summary)
<u>Third Grade, 2000-01</u>

- The SAGE achievement advantage persists. When scores are adjusted for pre-existing differences in socioeconomic status, ethnicity, attendance, and prior knowledge, a SAGE advantage from the beginning of first grade to the end of third grade is shown on all subtests. From the end of first grade to the end of third grade, a SAGE advantage is shown on all subsets. From the end of second grade to the end of third grade, a SAGE advantage is shown in the third-grade reading subtest.
- Additional students lower the average performance of classrooms. Each student added to a classroom beyond the 15:1 SAGE student-teacher ratio results in a decrease of approximately one scale score point in the class average in all academic scores
- No significant differences in achievement gains were found between 15:1 and 30:2 classrooms.

Molnar, A. and Others. (1999). Evaluating the SAGE program: A pilot program in targeted pupil-teacher reduction in Wisconsin. Educational Evaluation and Policy Analysis 21, 2, pp. 165-177.

Wisconsin's Student Achievement Guarantee in Education (SAGE program was designed as a 5-year K-3 pilot project that began in the 1996-97 school year. The program requires that participating schools implement four interventions including reducing the pupil-teacher ratio within classrooms to 15 students per teacher. The SAGE evaluation uses a quasi-experimental, comparative change design utilizing descriptive statistics, linear regression, and hierarchical linear modeling. In addition, qualitative analyses of life in SAGE schools and classrooms are conducted. Results of the 1996-97 and 1997-98 first grade data reveal findings consistent with the Tennessee STAR class size experiment. Also, individualization emerged as a key characteristic of instruction in SAGE classrooms.

Nye, B., Hedges, L. V., Konstantopoulos, S. (2001, Spring). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. Journal of Experimental Education, 69, 3, p. 245-257.

(Abstract): Reducing class size to increase academic achievement is a policy option currently of great interest. Although the results of small-scale randomized experiments and some interpretations of large-scale econometric studies point to positive short-term effects of small classes, some scholars view the evidence as ambiguous. Project STAR in Tennessee a 4-year, large-scale randomized experiment on the effects of class size—provided persuasive evidence that small classes have immediate positive effects on academic achievement. Unlike most other early education interventions, these effects persisted for several years after the children returned to regular-sized classes. The authors of the present article report analyses of a 6-year follow-up of the students in that experiment. Class size effects persisted for at least 6 years and remained large enough to be important for educational policy. The results suggest that small classes in early grades have lasting benefits and that those benefits are greater for minority students than for White students.

Pate-Bain, H., Boyd-Zaharias, J. B., Cain, V. A., Word, E., & Binkley, M. E. (1997, September). The Student/Teacher Achievement Ratio (STAR) Project: STAR Follow-up Studies, 1996-1997. Prepared by Health and Education Research Operative Services, Inc. (HEROS).

(Summary): Tenth graders who had participated in STAR small classes (K-3) appear to have maintained academic achievement advantages over their peers who attended regular or regular/aide STAR classes. Although there were no statistically significant differences in the TCE grade-10 test scores by class type, it was found that a significantly larger portion of small-class students than regular-and regular/aide-class students had already passed the TCE requirement at grade 8.

Data from the pilot study showed that over the years, the students from small classes were less likely to fail a grade level, or be suspended than their peers who were in regular and regular/aide classes. Small-class students were found to be making better grades in their high school courses and to be taking more advanced courses than students from the other two cohorts. This enabled investigators to compare the academic paths taken by STAR small, regular and regular-aide class students.

Pritchard, I. (1999). Reducing class size: What do we know? National Institute on Student Achievement, Curriculum and Assessment. U.S. Department of Education: Office of Educational Research and Improvement. Retrieved October 22, 2002 from http://www.ed.gov/pubs/ReducingClass/title.html

Reviews experimental studies of class size: Indiana's *Prime Time* project; Tennessee's *Project STAR;* Burke County, North Carolina; and Wisconsin SAGE. Concludes the following:

- A consensus of research indicates that class size reduction in the early grades leads to higher student achievement. Researchers are more cautious about the question of the positive effects of class size reduction in 4th through 12th grades. The significant effects of class size reduction on student achievement appear when class size is reduced to a point somewhere between 15 and 20 students, and continue to increase as class size approaches the situation of a 1 to 1 tutorial.

- The research data from the relevant studies indicate that if class size is reduced from substantially more than 20 students per class to below 20 students, the related increase in student achievement moves the average student from the 50th percentile up to somewhere above the 60th percentile. For disadvantaged and minority students the effects are somewhat larger.
- Students, teachers, and parents all report positive effects from the impact of class size reductions on the quality of classroom activity.

Reichardt, R. (2001, April). Reducing class size: Choices and consequences. Aurora, CO: Mid-Continent Research for Education and Learning, Policy Brief. EDRS Document # ED456-532

(Key Points):

- Smaller classes can raise student achievement and help narrow the achievement gap between minority and non-minority students.
- Reducing class size is costly—and more so for small schools.
- Giving districts and schools flexibility in how class size is measured can lower costs.
- Additional qualified teachers and classroom space are critical planning considerations.
- A large-scale initiative to reduce class size should include strategies for retaining qualified teachers in schools that serve high proportions of at-risk students.
- Class size reduction should be considered as part of a larger systemic approach to raising student achievement.

Scudder, D. F. (2000). Class-size reduction evaluation, 1999-2000. A Report to the North Carolina Department of Public Instruction. Raleigh, NC: Wake County Public schools System, Dept. of Evaluation and Research.

(Abstract): The effectiveness of the Wake County Public School System (WCPSS), North Carolina, plan to reduce class size was evaluated, assessing program implementation and the effects of class size reduction on academic achievement. For the 1999-2000 school year, North Carolina received federal funds targeting class size reduction under the Class Size Reduction Program (CSRP), and an allocation was made to the WCPSS of approximately $1.1 million. The objective was approached by hiring as many fully qualified teachers as possible, establishing implementation models, and determining the grade levels to target. Twenty-three teachers were supported by CSRP funds, and they were sent to 23 schools where between 21.6 and 51.1% of students received free or reduced-price lunches and between 50 and 117 students per school were considered low-achieving. District staff developed four implementation models, and schools were asked to implement class size reduction in grades 1 or 2, with the preferred model being the introduction of a new class of about equal size to other classes in the target grade. Reduced class sizes thus affected about 2,473 students. Students did show improvement in academic achievement, with improved growth greatest where class size was smallest. However, low-income students appeared to benefit less from class size reduction, even though their achievement improved to some extent. Some departures from the implementation plan were found, and some suggestions are presented for better implementation in the next school year.

Smith, P., Molnar, A., & Zahorik, J. (2003). Class size reduction in Wisconsin: A fresh look at the data. Education Policy Studies Laboratory, Arizona State University. (http://edpolicylab.org)

This report gives an overview of findings from the evaluation the Student Achievement Guarantee in Education (SAGE) program initiated by Wisconsin state statue in 1995. The five-year evaluation compared the performance of students in 30 SAGE schools with students in 14 to 17 comparison schools. The number of schools in the comparison group varied by year. In this reanalysis of the data, achievement of SAGE students on the Comprehensive Test of Basic Skills (CTBS) TerraNova edition was compared to national norms for comparable age groups and achievement data for students in the comparison schools. The researchers concluded that the size of the SAGE effect varied by subject matter, but was 1/3 to ½ of one school years' growth compared to the norm group averages. The greatest effect in achievement attributed to SAGE compared to the comparison groups occurred in first grade. African-American students profited more from participating in the SAGE program than did White students when compared to non-SAGE students. The SAGE program narrowed the achievement gap between African-American and White students in first grade and prevented the gap from widening in second and third grades. Finally, the SAGE effect may have helped to compensate for attendance problems because low attending students in SAGE performed comparably to the best attending non-SAGE students.

U. S. Department of Education. (1999, November). Local success stories: Reducing class size. Retrieved October 17, 2002 from http://www.ed.gov/offices/OESE/ClassSize/localsuccess.html

This brief discusses new class size reduction initiatives; lessons from early implementation; how the federal class size reduction program works; research update on benefits of class size reduction; and implementation challenges and opportunities.

U. S. Department of Education (2000, September). The Class Size Reduction Program: Boosting student achievement in schools across the nation: A first-year report. Jessup, MD: U.S. Department of Education.

Report gives examples of programs throughout the country, how school districts are utilizing funds from the program, research about the benefits, and how class size reduction is cost-effective.

WestEd. (1999, January). Class size reduction: Great hopes, great challenges. Policy Brief. San Francisco: WestEd.

Numerous states have enacted or are considering measures to reduce class size. Additionally, as part of a seven-year program to ensure an average class size of 18 for grades one through three, the federal government has committed more than $2.5 billion to a national class size reduction (CSR) initiative. These efforts stem from research findings on CSR's achievement benefits, as well as from its enormous popularity with parents, administrators, and teachers. This policy brief summarizes benefits, challenges, recommendations, and resources for implementing reduced class size programs.

Word and Others (1990, June). *Project Star final executive summary report: Kindergarten through third grade results (1985-1989).* Nashville, TN: Project START Office, Tennessee State Department of Education. Retrieved July 16, 2002 from http://www.cde.ca.gov/classsize/eval/projstar.htm

From the "Project STAR Fact Sheet":

The small classes made the highest scores on the Stanford Achievement Test (SAT) and Basic Skills First (BSF) test in all four years (K-3) and in all locations (rural, suburban, urban, inner-city). The greatest gains on the SAT were made in inner-city small classes. The highest scores on the SAT and BSF were made in rural small classes.

Inner-city (predominately minority) students in small classes always outscored inner-city students in regular and regular/aide classes. This suggests that small classes are very beneficial to minority students.

In every grade, every location, and every class type, non-free lunch students outperformed free lunch students.

Non-free lunch minorities in suburban small classes performed as well as non-free lunch Whites.

Zahorik, J. A. (1999, September). Reducing class size leads to individualized instruction. *Educational Leadership*, 54 (1). Retrieved July 18, 2002 from http://www.uwm.edu/Dept/CERAI/documents/sage/individualized.html

Wisconsin researchers have found that class-size reduction in a variety of formats increases attention to individual students. Teachers of smaller classes, however, are individualizing instruction—not content.

Zahorik, J., Molnar, A., Ehrle, K., & Halbach, A. (2000). *Effective teaching in reduced-size classes*. Milwaukee: Center for Education Research, Analysis, and innovation, University of Wisconsin Milwaukee.

Qualitative study comparing a group of teachers in the SAGE program whose classes showed higher-than-expected gains with a group of teachers whose classes showed lower-than-expected gains. Researchers found that teachers who use

active methods of teaching such as explaining, modeling, checking, and evaluating, and take a teacher-directed approach to teaching, have classes with higher student achievement than teachers who use a more student-directed approach.

References

Bohrnstedt, G. W., & Stecher, B. M. (Eds.), CSR Research Consortium. (2002, September). *Capstone Report. What we have learned about class size reduction in California.* Sacramento: California Department of Education.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001). Class size and student achievement. *Psychological Science in the Public Interest, Vol. 2*, No. 1, 1-30.

Finn, J. D. (1998, April). *Class size and students at risk: What is known? What is next?* A Commissioned Paper prepared by Jeremy D. Finn. National Institute on Education of At-Risk Students, Office of Educational Research and Improvement, U.S. Department of Education.

Hanushek, E. A. (1998). *The evidence on class size*. Occasional Paper No. 98-1. W. Allen Wallis Institute of Political Economy. Rochester, NY: University of Rochester.

Maier, P., Molnar, A., Percy, S., Smith, P., & Zahorik, J. (1997, December). First-year results of the Student Achievement Guarantee in Education Program. Submitted by the SAGE Evaluation Team Center for Urban Initiatives and Research. University of Wisconsin-Milwaukee.

McRobbie, J., Finn, J. D., & Harman, P. (1998, August). *Class size reduction: Lessons learned from experience.* WestEd's Policy Program Policy #23, August, 1998.

Molnar, A., Smith, P., & Zahorik, J. (1999, December). *1998-99 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: Unviersity of Wisconsin-Milwaukee, School of Education.

Molnar, A., Smith, P. & Zahorik, J. (2000, December). *1999-2000 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: University of Wisconsin-Milwaukee, School of Education.

Molnar, A., Smith, P, & Zahorik, J. (2001). *2000–2001 Evaluation results of the Student Achievement Guarantee in Education (SAGE) Program*. Milwaukee, WI: University of Wisconsin-Milwaukee, School of Education.

Nye, B., Hedges, L. V., Konstantopoulos, S. (2001, Spring). The long-term effects of small classes in early grades: Lasting benefits in mathematics achievement at grade 9. *Journal of Experimental Education, (69)*3, 245-257.

Pate-Bain, H., Boyd-Zaharias, J. B., Cain, V. A., Word, E., & Binkley, M. E. (1997, September). *The Student/Teacher Achievement Ratio (STAR) Project: STAR Follow-up Studies, 1996–1997.* Prepared by Health and Education Research Operative Services, Inc. (HEROS).

Smith, P., Molnar, A., & Zahorik, J. A. (2003). *Class size reduction in Wisconsin: A fresh look at the data.* An Education Policy Research Unit Study, Education Policy Studies Laboratory Tempe. AZ: Arizona State University, College of Education.

Word, E., Achilles, C. M, Bain, H., Folger, J., Johnston, J., & Lintz, N. (1990, June). *Project STAR final executive summary report: Kindergarten through third grade results (1985-1989).* Nashville, TN: Project STAR Office, Tennessee State Department of Education. Retrieved July 16, 2002 from http://www.cde.ca.gov/classsize/eval/projstar.htm

Zahorik, J., Molnar, A., Ehrle, K, & Halbach, A. (2000). *Effective teaching in reduced-size classes.* Milwaukee: Center for Education Research, Analysis, and Innovation, University of Wisconsin-Milwaukee.