Standards, Assessments—and What Else? The Essential Elements of Standards-Based School Improvement

CSE Technical Report 528

Diane J. Briars, Pittsburgh Public Schools

Lauren B. Resnick, CRESST / University of Pittsburgh

August 2000

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 3.1 Construct Validity: Understanding Cognitive Processes and Consequences
Lauren Resnick, Project Director, CRESST / University of Pittsburgh

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR / Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education, and in part by the National Science Foundation, ESI 96-34048.

STANDARDS, ASSESSMENTS—AND WHAT ELSE?

THE ESSENTIAL ELEMENTS OF STANDARDS-BASED SCHOOL IMPROVEMENT[1]

Diane J. Briars, Pittsburgh Public Schools

Lauren B. Resnick, CRESST/University of Pittsburgh

In the national search for ways of raising academic achievement, there seems to be widespread agreement that a "standards-based" education system (National Commission on Education Standards and Testing, 1992; Smith & O'Day, 1990) is the key to improvement. Core features of such a system, ones that most states and many school districts have now embraced in rhetoric and the beginnings of practice, are content and performance *standards* for each school discipline, along with *assessments* aligned to the standards. Standards and assessments are viewed as the foundation stones of a system in which educators determine the means by which they will meet publicly established expectations, but in which states and school districts may establish various systems of public *accountability* for meeting them. It is widely agreed that standards, assessments, and accountability can raise achievement only if they motivate and enable better teaching—presumably the result of *curriculum* that is aligned with the standards and assessments, along with improved *professional development* for teachers and administrators. There is less agreement, however, about who should—indeed, who has the right to—establish and monitor teaching and professional development programs. Some believe that approaches to teaching, curriculum, and professional development for teachers should be left to very local professional decision makers, individual schools, or, in an extreme view, individual teachers. Others believe that a standards-based strategy for raising achievement calls

[1] Special thanks to: Dr. Claudia Harper-Eaglin for her leadership in the Pittsburgh Public Schools (PPS) Unit of Teaching, Learning and Assessment; Dr. Cherri Banks for her work leading and coordinating the PPS CEIP planning and reporting process; Dr. Shula Nedley for her key role in the design of the PPS standards-based assessment system; Dr. James J. Staszewski for thoughtful comments on an earlier version of this paper and suggestions regarding data analyses; Dr. Jack Garrow for doing the data analyses; Ms. Mary Lynn Raith, co-director of the PRIME project; Ms. Deborah Saltrick-Friss, PRIME project manager, and the PRIME elementary team, Dr. Ruth Downs, Ms. Marilee Glick, Ms. Yvonne Comer-Holbrook, Ms. Anne McFeaters, and Ms. Cindi Muehlbauer, for their tireless support of teachers in the implementation of *Everyday Mathematics.*

for more active engagement of districts, or even states, in specifying curricula, textbooks, teaching methods, and approaches to professional development.

The elements of a standards-based system are coming into place unevenly in states and cities across the country. Most states now have content standards, although their quality varies and evaluators (such as the American Federation of Teachers, the Council on Basic Education, and the Fordham Foundation) do not always agree with one another. Only a minority of states have established true performance standards, that is, descriptions and illustrations of the kinds of work students are expected to be able to do. Many states and virtually all school districts administer tests, and many use the language and rhetoric of standards in communicating with parents and the public about the results of these tests. But it is still rare that the tests used have been systematically aligned to the officially adopted standards. In some jurisdictions, an off-the-shelf norm-referenced test is used as part of a nominally standards-based system, with score points being used to establish "standards." There is no true alignment in such a process although individual items in the tests can be matched with some of the standards.

It is even more rare to find instructional materials and strategies well aligned to standards and accompanied by systematic professional development. Jurisdictions, whether states or school districts, that are the exception to this state of affairs—that is, jurisdictions that have aligned standards, tests, curricula, instructional materials, and professional development—are privileged sites in which to evaluate the power of a standards-based education *system* to raise achievement. The Pittsburgh Public Schools (PPS) is such a site. In 1992, PPS adopted a strategic plan that called for the district to become a fully standards-based system. The plan called for district policies and practices in support of the core elements of a standards-based system: standards, assessments, accountability, curriculum, and professional development. Between 1992 and 1998, most of the elements of the system were put into place, one by one. The process is incomplete, as we describe shortly. But enough has been done that we can, at this time, evaluate the effects of a nearly complete standards-based system in at least one subject matter—mathematics—in which the Pittsburgh school system has acted energetically.

The PPS mathematics program includes the following components of a standards-based system:

Content and performance standards are described in the district's Mathematics Core Curriculum Framework, along with performance indicators and suggested classroom assessments. The Pittsburgh Core Curriculum Framework for math was developed by teacher committees and guided by the National Council of Teachers of Mathematics and New Standards mathematics standards (National Council of Teachers of Mathematics, 1989; New Standards, 1996).

Standards-based assessments. PPS incorporated the New Standards Mathematics Reference Examination (Harcourt Educational Measurement, 1996-1999) for Grades 4, 8, and 10 into its assessment system in 1996. This was intended as the first step toward eventually replacing norm-referenced assessments with standards-based assessments. In the period under study here (1996-1998) both the New Standards exams and the survey battery of the Iowa Tests of Basic Skills (1993) were administered to elementary school students. Teacher preparation for the New Standards exams was limited in the first year; but beginning with the 1996-1997 school year, all teachers were offered (but not required to participate in) professional development built around the New Standards released tasks and practice exams.

Standards-based instructional materials. Beginning in 1993, PPS adopted NSF-supported *Everyday Mathematics* (University of Chicago School Mathematics Project, 1995) for Grades K-5. This is a program directly mapped to the NCTM Standards and informed by research on children's cognitive development in mathematics. It is well aligned with the District's Core Curriculum for mathematics and the New Standards exams. *Everyday Mathematics* implementation began with the cohort entering kindergarten in 1993-1994; these students were fourth graders in 1997-1998.

Standards-based professional development for teachers and administrators is supported by the Pittsburgh Reform in Mathematics Education project (PRIME), an NSF-supported Local Systemic Change program. This program came into place in the 1996-1997 school year. Designed specifically to develop teachers' capacity to implement *Everyday Mathematics*, PRIME provides in-class support —demonstration lessons, joint planning, coaching—by the demonstration teachers, in addition to summer and after-school professional development workshops.

In this paper, we use data on elementary school mathematics assessments over a three-year period (the school years 1995-1996, 1996-1997, 1997-1998) to explore the effects of the Pittsburgh Public Schools' implementation of elements of a standards-based system. In a concluding section, we consider how the district's very

incomplete accountability system influenced the extent to which the other elements of a standards-based system were implemented.

Method

Variables and Measures

Student achievement. The study used two tests of student achievement in mathematics: the New Standards Mathematics Reference Examination and the Iowa Tests of Basic Skills Survey Battery, Form K.

The New Standards Mathematics Reference Exam (NSRME) is closely aligned to Pittsburgh's Core Curriculum standards and to its officially adopted *Everyday Mathematics* curriculum and functions as an integral part of the Pittsburgh Public Schools' standards-based assessment system. The elementary NSRME is designed to be administered in fourth grade only. The exam consists of three 50-minute sections[2] and contains 20 multiple-choice and 20 performance tasks. It assesses performance in three areas: skills, concepts, and problem solving. *Skill* tasks assess students' use of basic routines and procedures, including computation, measurement, graphing, reading tables, and using tools such as compasses and protractors. *Concept* tasks assess students' use of concepts of number and operation, geometry, measurement, functions and algebra, and statistics and probability to solve problems, represent concepts in multiple ways, and explain those concepts to others. *Problem-solving* tasks assess students' use of concepts and skills to formulate problems, implement solutions, justify conclusions, make generalizations, and use the language of mathematics to explain their reasoning and results. (See Figure 1 for a sample of each type of task.)

New Standards exams are scored by the test publisher. Each performance task is scored according to a task-specific rubric by specially trained individuals who have demonstrated their ability to score performance reliably using the rubric. New Standards scores compare students to established performance standards that say what students should be able to do at different points in their educational career (New Standards, 1996). Standards are set by a national group of teachers and educators. Student performance is reported in five categories:

[2] 50 minutes is the suggested time. Students who are productively working at the end of 50 minutes may continue to work until they have completed the section.

(a) Sample Skills Task

Mowing the Lawn

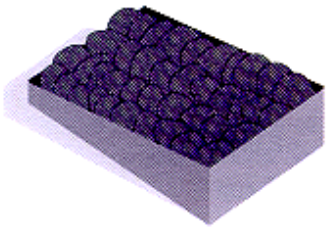
Gerald started mowing the lawn at 3:15 p.m. He finished 1 hour and 40 minutes later.

What time was it when Gerald finished mowing the lawn?

(b) Sample Concepts Task

Spilled Cookies (about 5 minutes)

Harry made cookies for the school bake sale.
He baked 20 chocolate cookies, 12 nutty cookies, and 8 lemon cookies.



He wrapped each cookie separately. He used the same kind of paper for all the cookies, so people couldn't tell one kind from another. Then they spilled out of their box.

Harry put them back in the box without worrying about which was which, and took them to the bake sale.

What is the chance that the first cookie Harry takes out of the box at the bake sale will be chocolate?

Explain how you decided on your answer.

Figure 1. Sample Grade 4 New Standards Reference Examination tasks.

(c) Sample Problem Solving Task

Trail Mix **(about 15 minutes)**

Tran and Jenna are members of a hiking club. They are using this recipe to make trail mix for their club to take on a hike.

TRAIL MIX RECIPE Enough for 1 bag	
Raisins	1 cup
Peanuts	2 cups
Shredded Coconut	$\frac{1}{2}$ cup

Look at the raisins, peanuts and shredded coconut they bought.

Tran and Jenna follow the recipe exactly. They make as many bags of trail mix as possible.

Trail Mix **(continued)**

Answer these questions. Show all of your work or explain how you figured it out.

- How many bags of trail mix can they make?
- Which ingredient do they run out of first?
- How much of the other two ingredients are left over when they are done?

Figure 1. (continued).

- *Achieved the Standard with Honors* denotes performance that is consistently at a higher level than the standard;
- *Achieved the Standard* denotes performance that is consistently at the level of the standard;
- *Nearly Achieved the Standard* denotes performances that show some evidence of being at the level of the standard, but overall the performances do not consistently meet the standard;
- *Below the Standard* denotes performances that show some attempt to respond to the tasks, but the number of successful responses is minimal, and responses are often incomplete; and
- *Little Evidence of Achievement* denotes performances that show almost no attempt to respond, as evidenced by numerous blank answers, entirely unsuccessful answers, and incomplete answers.

The *ITBS Survey Battery* is a 30-minute norm-referenced achievement test. It consists of 33 questions and assesses mathematical concepts (11 questions), data interpretation (6), estimation (8), and routine problem solving (8). Tests are scored by the district, using software developed by the test publisher. Results are reported by percentile rankings. ITBS has been given for many years in Pittsburgh, and many people in Pittsburgh, as elsewhere, consider it a way of insuring that students' basic skills in arithmetic are not neglected as new curricula come into place.

Variations in Implementation

Using districtwide results to measure the effect of the standards-based program assumes that all schools are implementing the program as intended. There was, however, substantial variability from school to school and even classroom to classroom in the extent and quality of implementation of *Everyday Mathematics,* even after the PRIME professional development program was introduced. A more refined picture of the importance of curriculum and professional development can be obtained by examining the effects of these variations on achievement.

In the spring of 1998, the PRIME demonstration teachers who work in each of the Pittsburgh elementary schools were asked to rate first- through fourth-grade teachers on their degree of implementation of *Everyday Mathematics.* These ratings were made separately for each of the two years in which PRIME demonstration teachers were present in the schools (1996-1997 and 1997-1998). Strong implementers were those who (a) used all of the *Everyday Mathematics* components and (b) provided student-centered instruction by giving students opportunities to explore

mathematical ideas, solve problems, and explain their thinking. Weak implementers were either not using *Everyday Mathematics* at all, or were using it so little that the overall instruction in the classroom was hardly distinguishable from traditional mathematics instruction.

Demonstration teachers used the following evidence, obtained from their ongoing work in each school, to determine whether a teacher met the two criteria for strong implementation:

- Students' familarity with activities and procedures specific to the *Everyday Mathematics* program (e.g., MathBoxes, function machines, frames and arrows diagrams, Everything Math Deck). As demonstration teachers conducted demonstration lessons, whether the classroom teacher was implementing *Everyday Mathematics* as intended was readily apparent;
- Opportunities for students to explain their solutions to problems during lessons, and students' comfort in doing so (suggesting that they frequently had such opportunities);
- Students working in groups or with partners when appropriate;
- Classroom appearance—that is, required visual aids (e.g., number line, hundreds chart, weather chart) were displayed, and there was evidence that they were being used as intended; displays of student work showed *Everyday Mathematics* explorations, projects and activities;
- Manipulative materials were accessible and obviously had had prior use;
- Teachers had informed questions about content and instruction; and
- No evidence of the use of other programs.

On the basis of these judgments, 57 teachers in 13 schools were identified as weak implementers, and 54 teachers in 25 schools as strong implementers. Teachers who did not fall clearly into the strong or the weak categories, or about whom the demonstration teachers could not provide information, were excluded from this part of the study. Judgments of teachers were then aggregated to identify Strong and Weak Implementation schools. A school was classified as a Strong Implementation school if there was strong implementation by all Grade 4 teachers in 1997-1998 and all Grade 3 classrooms during 1996-1997. This means that children in Strong Implementation schools would have received at least two years of strong standards-based instruction prior to taking the 1998 New Standards Reference Exam. A school was classified as a Weak Implementation school if all but one or two teachers in the school (in Grades 1-4) were weak implementers. According to these criteria, three schools were identified as Weak Implementers, and eight were identified as Strong

Implementers.[3] The remaining 45 elementary schools in the Pittsburgh system had moderate implementation of *Everyday Mathematics* and were not considered in the degree of implementation study.

Demographic Characteristics of Schools

In order to attribute observed differences in schools' performance to implementation of the program rather than to student characteristics, it is necessary to know the demographic characteristics of the schools in the weak and strong implementation categories. We used the proportion of students eligible for reduced-cost or free lunch as a measure of each school's socioeconomic level. We also described each school in terms of the proportion of its students who were African American. The Pittsburgh Public Schools have virtually no non-English speaking students. We examined the demographic variables for the three Weak Implementation schools and identified the Strong Implementation school that most closely matched each one. The matching Strong Implementation schools were called Similar Strong schools. Table 1 shows the demographic features of the Weak Implementation schools and their corresponding Similar Strong schools. Table 2 shows the demographic characteristics of fourth-grade students in the Weak Implementation schools, Similar Strong schools, and the remaining four Strong

Table 1

Demographic Characteristics of Similar Weak and Strong *Everyday Mathematics* Implementation Schools

Demographic variables	Weak school A	Similar Strong school A	Weak school B	Similar Strong school B	Weak school C	Similar Strong school C
Number of students	421	361	411	230	375	337
% F/R Lunch[a]	88	93	91	81	76	82
% Other parents[b]	82	79	70	64	58	57
% Mobil[c]	15	14	16	14	8	12
% African American	98	99	55	53	45	43

[a] Percentage of students who are eligible to receive free or reduced price lunches.

[b] Percentage of students who do not live with two parents.

[c] Measure of student movement in and out of the schools. The rate is calculated by dividing the total number of student transfers (transfers in plus transfers out) by the total number of different students who attended the school during the school year. (Definition from the PPS School Profiles Report.)

[3] One Strong Implementation school was eliminated from the analyses of student performance due to technical difficulties with the data tape.

Table 2

Demographic Characteristics of the Grade 4 Students in Each Implementation Group

Demographic variables	Weak EM schools	Similar Strong EM schools	Other Strong EM schools
Number of Students	182	118	173
% F/R Lunch[a]	85	83	54
% Other Parents[b]	70	70	50
% Mobile[c]	11	12	5
% African American	65	64	51

Note. EM = *Everyday Mathematics.*

[a] Percentage of students who are eligible to receive free or reduced price lunches.

[b] Percentage of students who do not live with two parents.

[c] Measure of student movement in and out of the schools. The rate is calculated by dividing the total number of student transfers (transfers in plus transfers out) by the total number of different students who attended the school during the school year. (Definition from the PPS School Profiles Report.)

Implementation schools, here termed Other Strong. As can be seen, the Other Strong schools had fewer children on free and reduced lunch, more intact families, less mobility and somewhat fewer African American students.

Prior Academic Achievement of Children in the Schools

The possibility needs to be considered that the Strong Implementation Schools had a more academically able population than the Weak Implementation Schools. We were able to make a crude estimate of the academic ability of students in these groups of schools by looking at the first measure of mathematics performance that the Pittsburgh Schools have on record for their students: end-of-first-grade ITBS scores (from tests administered in spring 1995). Such scores were available for 82% of the students in Strong Implementation schools and for 81% of the students in Weak Implementation schools. The scores of these students on the 1995 ITBS-Math were virtually identical (e.g., 63.1% of students in Strong Implementation schools scored at or above the 50th percentile, 64% in Weak Implementation schools).

Design of the Study

Our study was designed to answer five questions:

Did the standards-based policy produce increases in mathematics achievement? We examined fourth-grade city-wide scores on the New Standards

Reference Exams and the Iowa Tests of Basic Skills over a three-year period, 1996–1998. During the first year for which scores were examined, *Everyday Mathematics* had been officially adopted but was only implemented in kindergarten through second grade. During the second year, the program was implemented through third grade and the comprehensive professional development program (PRIME-LSC) was offered for the first time. During the third year, *Everyday Mathematics* was used through the fourth grade. Because of the District's testing policy focusing on fourth grade, the students studied over the three-year period are from three different cohorts. The only students who might have experienced the *Everyday Mathematics* program throughout their elementary schooling were those tested in 1998. Thus, if the standards-based policy were directly producing increases in mathematics achievement, we would expect to see important rises in scores only in 1998.

How did variations in implementation of the instructional program affect the likelihood of achievement gains? To answer this question, we compared scores on NRSME and ITBS for Weak and Strong Implementation schools for the 1998 school year. We predicted higher scores for Strong Implementation schools.

Were variations in school performance a function of differences in implementation of the curriculum or of overall teacher quality? It is possible that the Strong Implementation schools were staffed by teachers who were generally better mathematics teachers and would have produced higher test performance using any curriculum. To examine this possibility we compared Strong and Weak Implementation schools over the three years, 1996 to 1998. If differences were due to sustained implementation of *Everyday Mathematics* and the accompanying professional development program, we would expect to see better performance in the Strong Implementation schools only in 1998 when the program was fully in place and the cohort of students tested had been in *Everyday Mathematics* throughout their primary school years.

Were variations in performance a function of differences in implementation of the curriculum or in school demographics? We compared students' mathematics scores in 1998 for Weak Implementation schools, Similar Strong schools, and Other Strong schools.

What effects did the standards-based program have on the achievement of African American students, both in absolute terms and in comparison with White

students? To answer this question, we compared African American and White students' scores over the three years of the study. In order to determine whether observed changes were due to the program under study, we focused special attention on comparing the performance of African American students and White students in Weak, Similar Strong, and Other Strong Implementation schools.

Results

Did the Standards-Based Policy Produce Increases in Mathematics Achievement?

New Standards Reference Exam. Figure 2a plots the percentage of students who met or exceeded the standard (i.e., earned grades of Achieved the Standard or Achieved the Standard with Honors) in the three measured areas of Skills, Concepts, and Problem Solving in each of the three years under study. In Skills, there was no significant improvement in scores between 1996 and 1997, but there was a substantial improvement in 1998. This finding matches our prediction exactly. In Concepts and Problem Solving, which started at a very low baseline, there was small but significant improvement even in 1997 and further improvement in 1998.

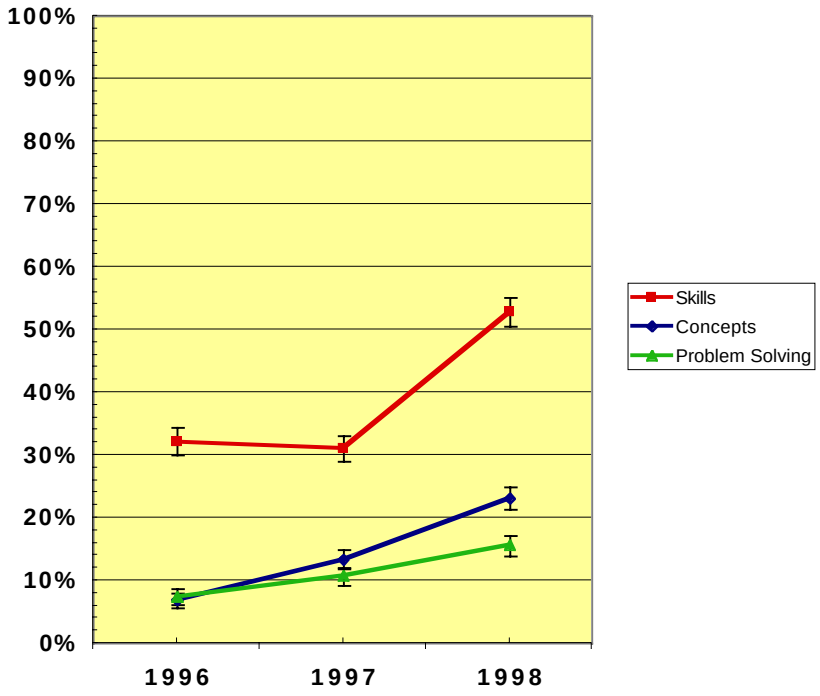


Figure 2a. Districtwide Grade 4 NSMRE performance for 1996, 1997 and 1998. Percentage of students who *achieved the standard.* Error bars denote the 99% confidence interval for each data point.

Looking at the bottom end of the distribution helps to confirm this pattern. Figure 2b plots the percentage of students who were in the two lowest score categories (Below the Standard and Little Evidence of Achievement) in the three measured areas of Skills, Concepts, and Problem Solving in each of the three years under study. In Skills, there was a small but significant increase in the proportion of students in these lowest categories from 1996 to 1997, followed by a large decrease from 1997 to 1998. In Concepts, there was a small but statistically significant decrease from 1996 to 1997, with a dramatic drop from 1997 to 1998. In Problem Solving, there was little change from 1996 to 1997, but in 1998 the proportion of students in these lowest categories dropped significantly.

It is also interesting to note that there was a sharp drop in the proportion of students at the very lowest score level (Little Evidence of Achievement) in Problem Solving (23% in 1996, 27% in 1997, only 7% in 1998). This lowest score level usually means that a student has no ability to even engage with the largely constructed-response tasks and the reasoning called for in them. These are precisely the areas that were neglected in the traditional curriculum but formed a core part of *Everyday Mathematics.*

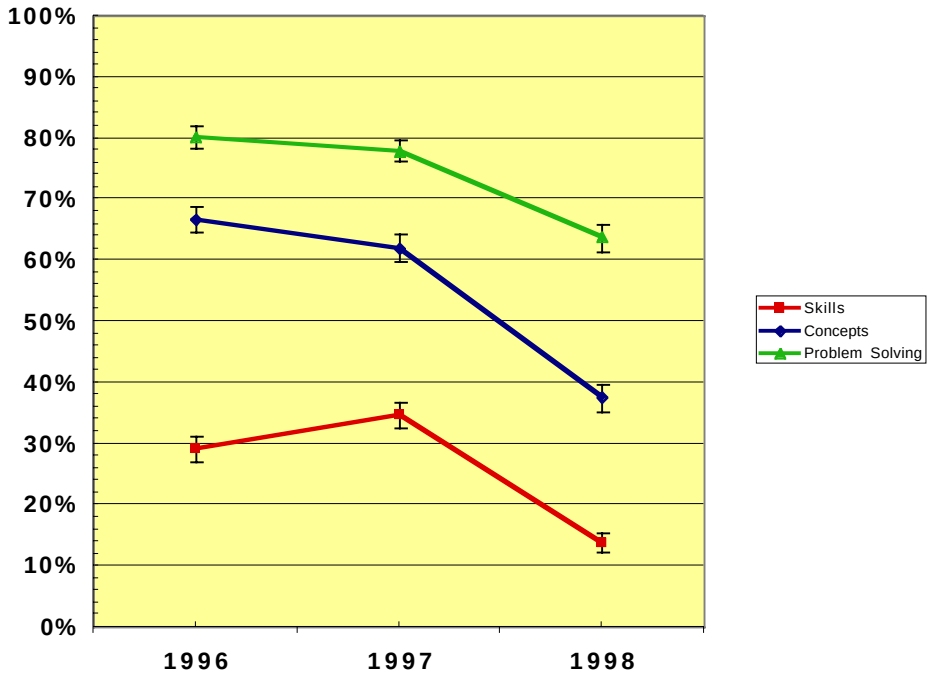


Figure 2b. Districtwide Grade 4 NSMRE performance for 1996, 1997 and 1998. Percentage of students who *scored well below the standard.* Error bars denote the 99% confidence interval for each data point.

Iowa Tests of Basic Skills. Figure 3 shows year-by-year results on the ITBS. The proportion of students at or above the 50th percentile was slightly but significantly (Fisher's Exact Test) higher in 1998 than in 1997; 1998 performance did not differ significantly from that in 1996. The proportion of students at or above the 75th percentile was slightly but significantly higher in 1998 than in both preceding years. In addition, the proportion of students below the 25th percentile dropped slightly but significantly in 1998. There were, overall, very limited changes in ITBS scores. However, there were some small gains, especially in 1998, and it is at least clear that children's gains on New Standards did not come at the expense of more traditional measured skills.

How Did Variations in Implementation of the Instructional Program Affect the Likelihood of Achievement Gains?

Figure 4a compares the proportion of students who met the standard or met it with honors on the 1998 New Standards Exam in Weak and Strong Implementation schools. As is evident, twice as many students met the Skills standard in Strong Implementation schools. The difference due to degree of implementation was even greater for Concepts and Problem Solving. Figure 4b shows the proportion of

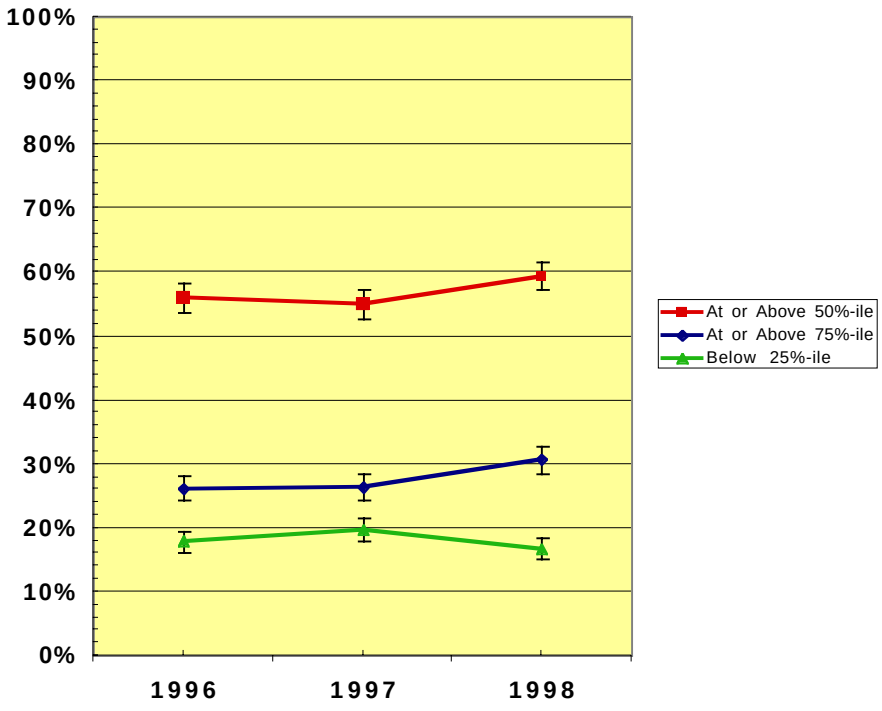


Figure 3. Districtwide Grade 4 ITBS performance for 1996, 1997 and 1998. Error bars denote the 99% confidence interval for each data point.

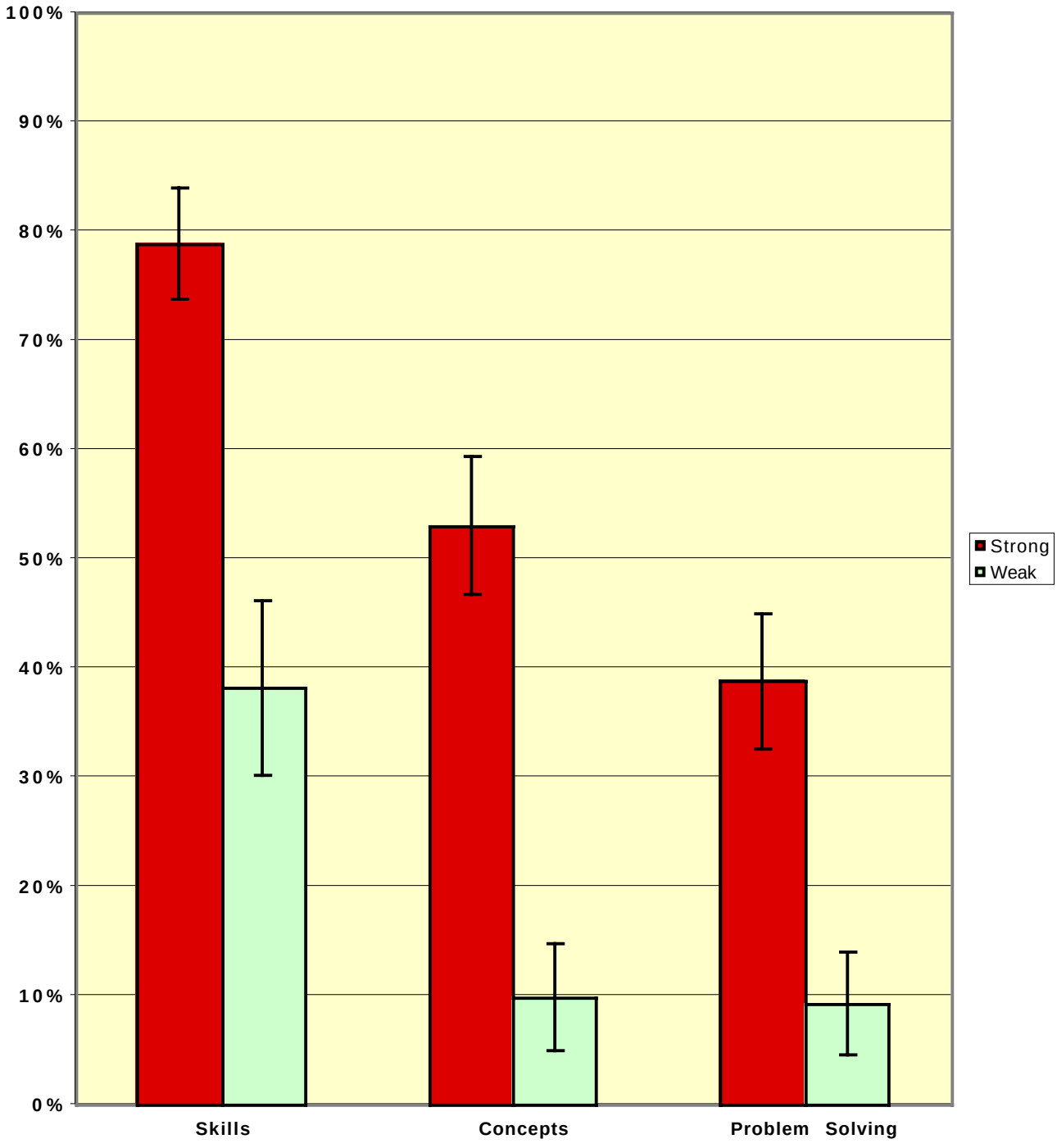


Figure 4a. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation. Percentage of students who *achieved the standard*. Error bars denote the 99% confidence interval for each data point.

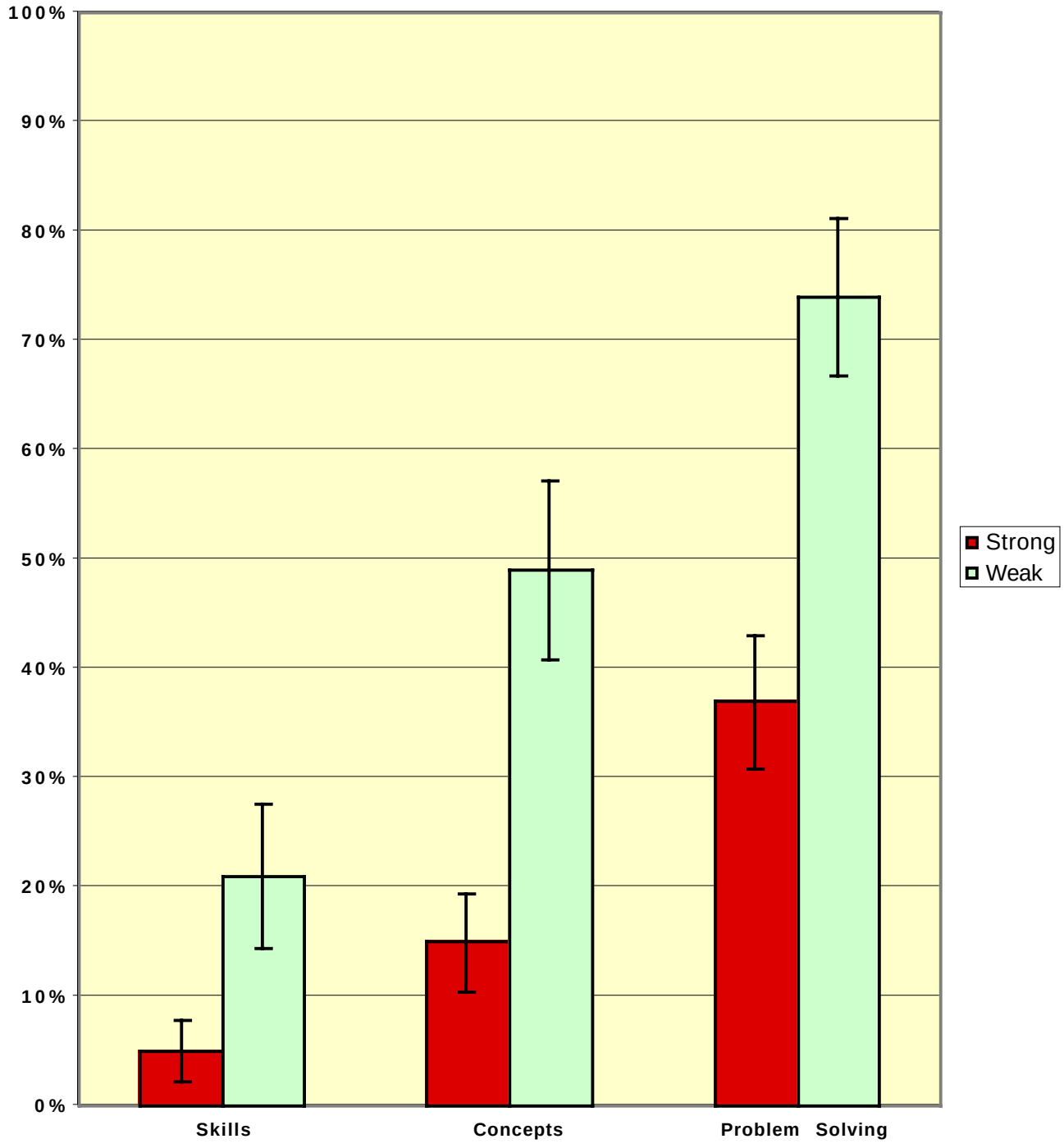


Figure 4b. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation. Percentage of students who *scored well below the standard*. Error bars denote the 99% confidence interval for each data point.

students in the two categories of schools that were well below the standard (i.e., in the two lowest New Standards score categories). The pattern is similar. In Strong Implementation schools, there were virtually no students in these lowest categories on Skills, whereas slightly more than 20% were well below the standard in Weak Implementation schools. The differences between Strong and Weak Implementation schools were even more pronounced for Concepts and Problem Solving.

These strong differences in performance as a result of degree of implementation were echoed in the ITBS scores, as shown in Figure 5. Almost half of

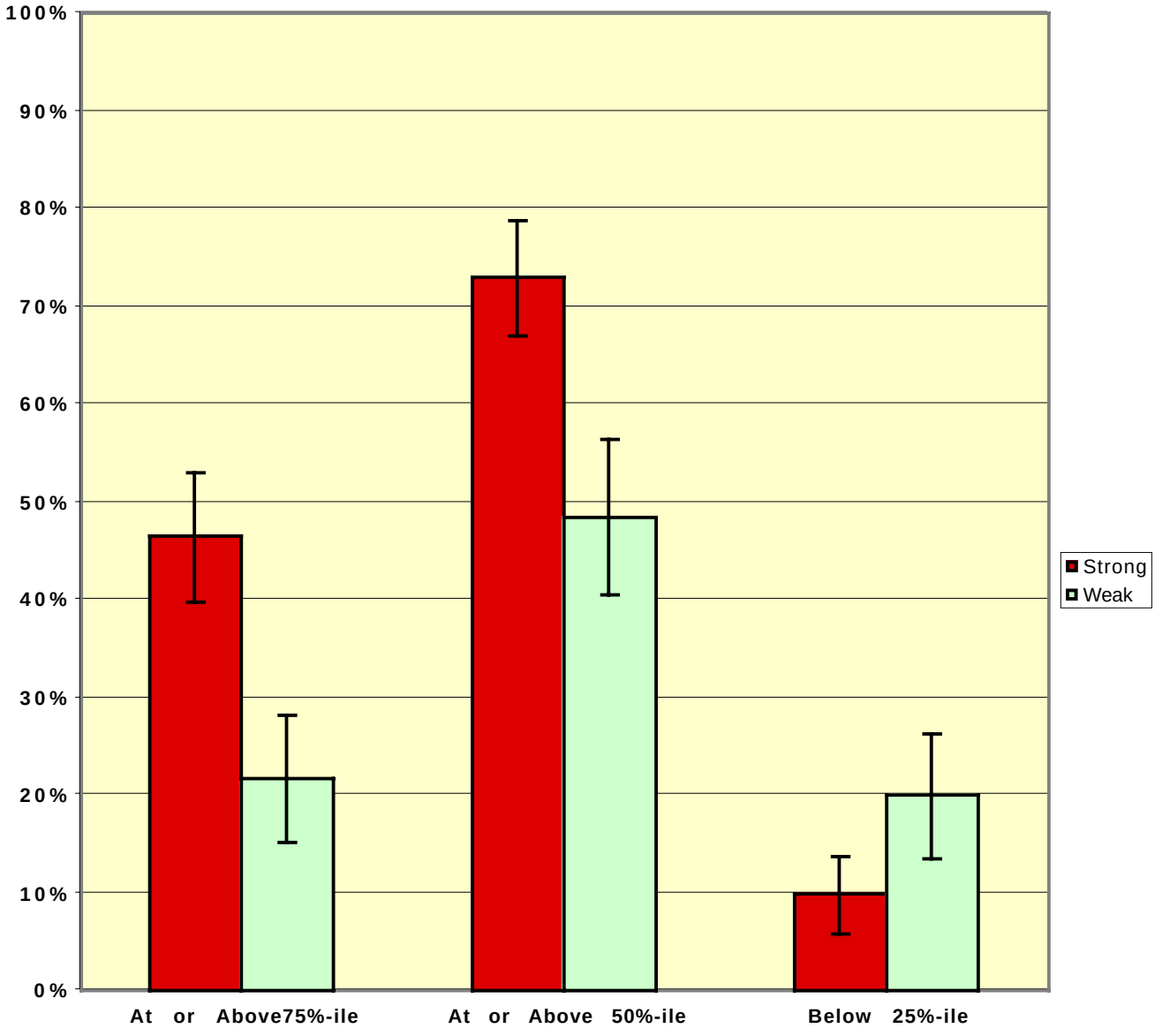


Figure 5. ITBS Grade 4 1998 results by level of *Everyday Mathematics* implementation. Percentage of students scoring at each level. Error bars denote the 99% confidence interval for each data point.

the students in Strong Implementation schools were in the top quartile, over 70% were at or above the traditional "grade level" of the 50th percentile; and only 10% were in the bottom quartile. Thus, a well-implemented curriculum aimed at conceptual understanding and problem solving also produced higher performance on a more traditional test. In Weak Implementation schools, there was a typical distribution for norm-referenced tests, with about 20% of students in the bottom quartile and 20% in the top quartile.

Were Variations in School Performance a Function of Differences in Implementation of *Everyday Mathematics* or of Overall Teacher Quality?

Figures 6a, 6b, and 6c show New Standards scores for three years (1996 through 1998), for Weak and Strong Implementation schools. For Skills (Figure 6a), performance every year was higher in Strong Implementation than in Weak Implementation schools. In Strong Implementation schools, there was a sharp increase in the proportion of students meeting the standard in 1998. In Weak Implementation schools, there was little change across the three years.

For Concepts (Figure 6b), there was no difference in performance between Weak and Strong Implementation schools in 1996. In 1997, Strong Implementation schools showed a small but significant increase in performance, whereas Weak Implementation schools remained static. In 1998, performance in the Strong Implementation schools increased dramatically, whereas performance in the Weak Implementation schools again remained constant. Problem Solving (Figure 6c) scores show a similar pattern of performance, with the exception that performance in the Strong Implementation schools did not increase significantly from 1996 to 1997.

Taken together, these results suggest that teachers in the Strong Implementation schools may have been more effective at teaching skills before the new program than those in the Weak Implementation schools, and slightly more effective at teaching concepts and problem solving. Their students' performance improved only a little in 1997, after a year of experience with New Standards, but substantially in 1998. This finding suggests that although the teachers in High Implementation schools may have been somewhat more skilled, they needed something else—the *Everyday Mathematics* curriculum and the PRIME professional development program, for example—to implement fully the standards.

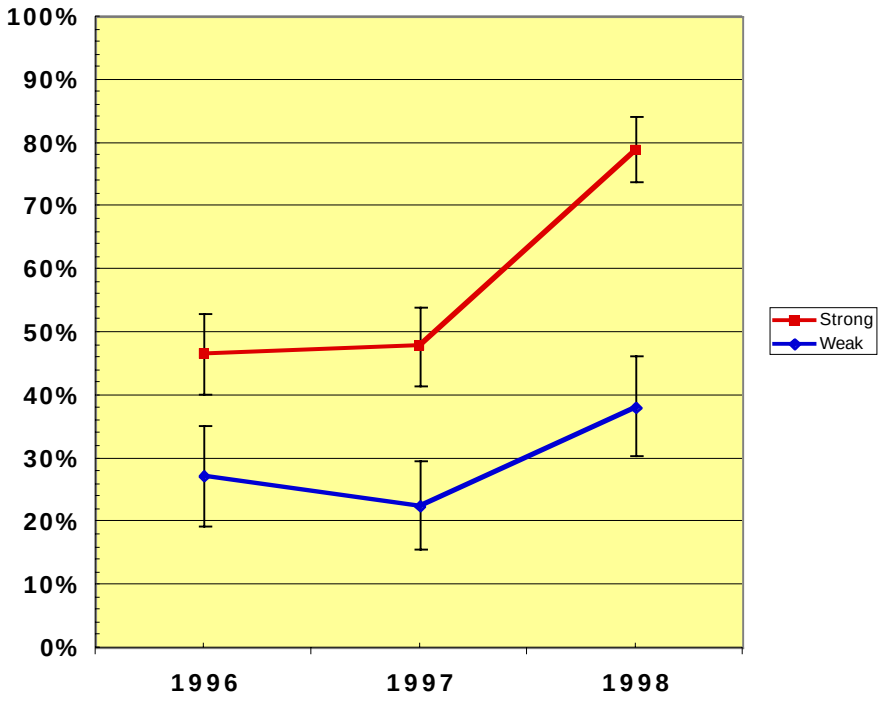


Figure 6a. Districtwide Grade 4 NSMRE performance for 1996, 1997 and 1998 by level of *Everyday Mathematics* implementation. Percentage of students who *achieved the skill standard*. Error bars denote the 99% confidence interval for each data point.

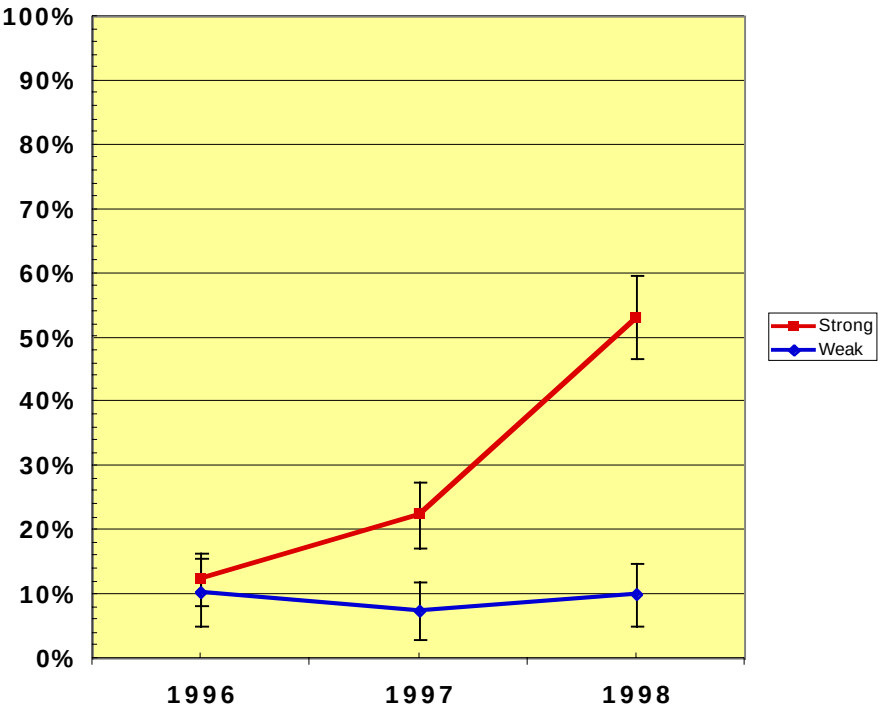


Figure 6b. Districtwide Grade 4 NSMRE performance for 1996, 1997 and 1998 by level of *Everyday Mathematics* implementation. Percentage of students who *achieved the concept standard*. Error bars denote the 99% confidence interval for each data point.

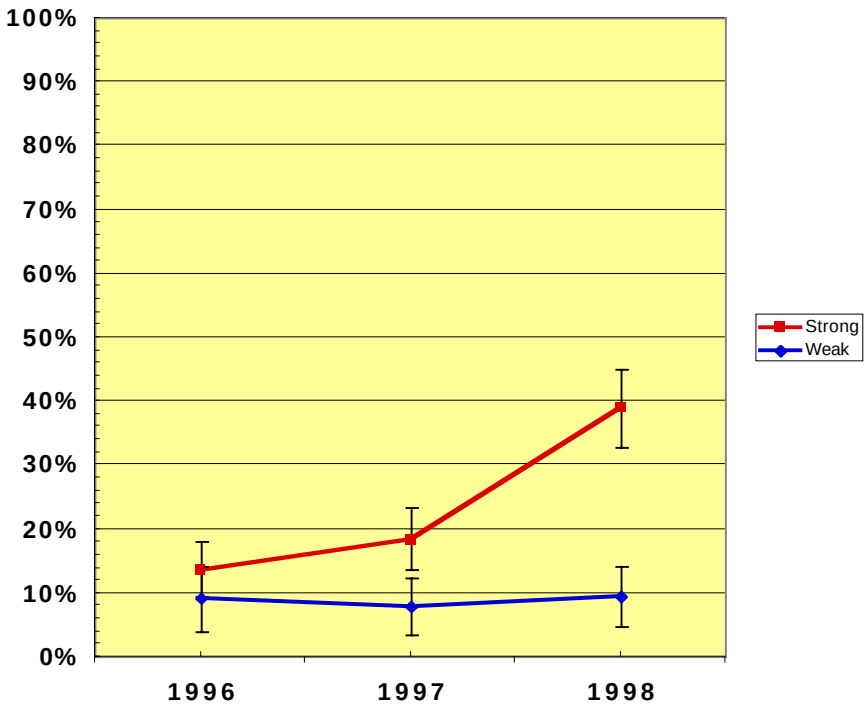


Figure 6c. Districtwide Grade 4 NSMRE performance for 1996, 1997 and 1998 by level of *Everyday Mathematics* implementation. Percentage of students who *achieved the problem solving standard.* Error bars denote the 99% confidence interval for each data point.

Were Variations in School Performance a Function of Differences in Implementation of *Everyday Mathematics* or in School Demographics?

As described earlier, an alternative explanation for the superior performance of Strong Implementation schools is that they had a different demographic profile; that is, fewer minority students and fewer students from low SES households. Figure 7 shows the New Standards performance of students in the Weak Implementation schools compared with students in demographically matched Strong Implementation schools (Similar Strong Schools) and the remaining Strong Implementation schools (Other Strong Schools). As the figure shows, Similar Strong and Other Strong schools showed similar high performance, whereas performance in the Low Implementation schools was dramatically lower than in either of the Strong Implementation groups of schools.

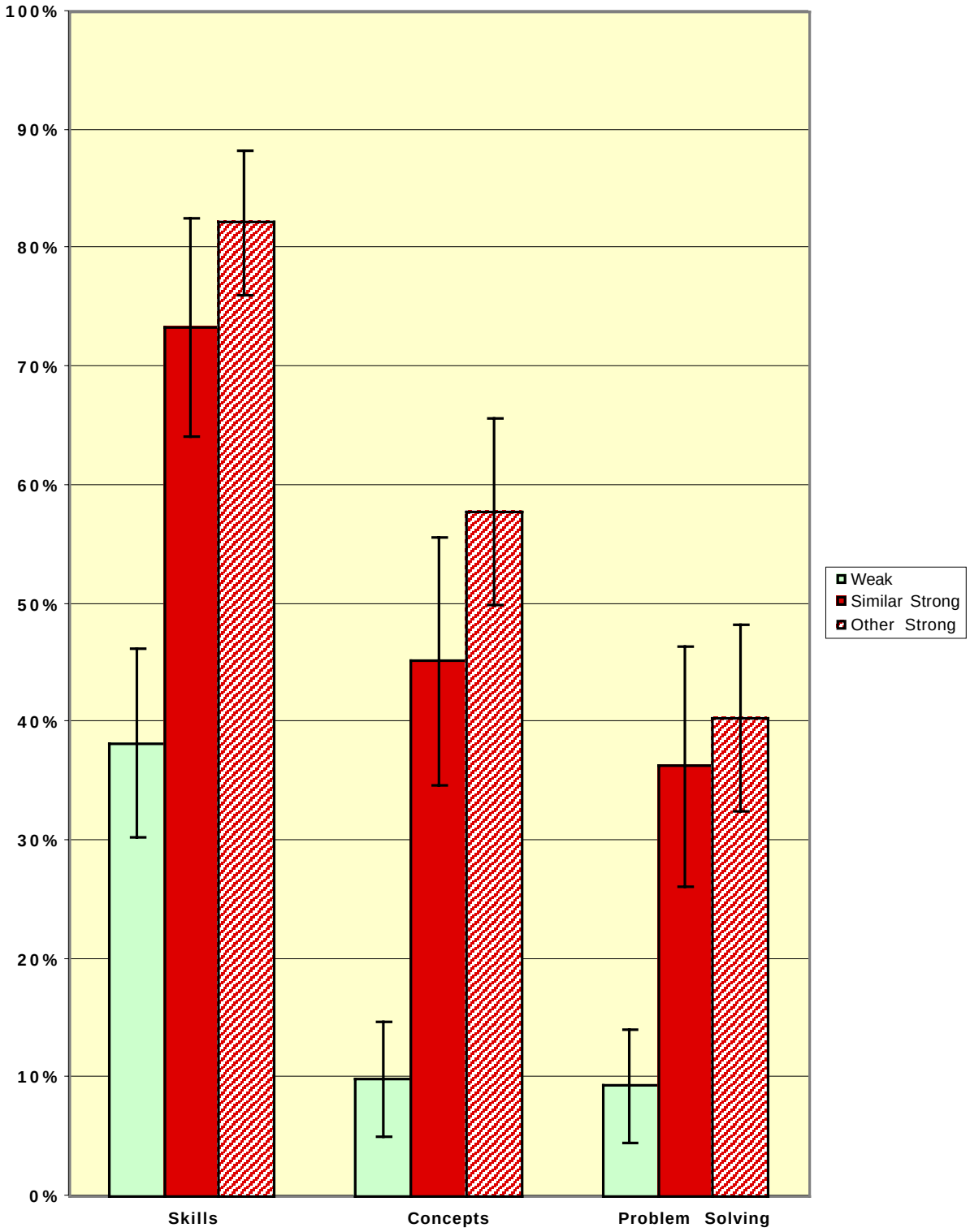


Figure 7. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation and school demographics. Percentage of students who achieved the standard. Error bars denote the 99% confidence interval for each data point.

What Effects Did the Standards-Based Program Have on the Achievement of African American Students—Both in Absolute Terms and in Comparison With White Students?

Figures 8a, 8b, and 8c show the proportion of students who met or exceeded the NSMRE standard broken out by race and school type. On Skills (Figure 8a), African American students in both Similar Strong and Other Strong schools did far better than their peers in Weak Implementation schools. They even performed better than Whites in the Weak Implementation schools. In Similar Strong schools, there was no difference between African American and White students. In the other two categories of schools, differences between the groups of students did not meet our stringent criterion for significance.

Concepts (Figure 8b) and Problem Solving (Figure 8c) show a similar pattern of performance. Both African American and White students in Similar and Other Strong Implementation schools did significantly better than their counterparts in Weak Implementation schools. Differences in performance between White and African American students were statistically significant only in the Other Strong Implementation schools.

Discussion

The findings of this study broadly support the expectations that were laid out in our study design. Taken as a whole, the standards-based policy for mathematics produced an overall rise in mathematics achievement in the district. The gains were sharpest on the New Standards Reference Exam but were present even on the Iowa Tests of Basic Skills Survey Battery. A standards-based education system calls for alignment of standards, assessment, curriculum, and professional development, and one would expect to see the greatest gains on the aligned assessment. It is important for both district and national policy, however, to know that performance on a more traditional measure of math achievement did not suffer (and, in fact, even improved) when a new direction for instruction was introduced. The most dramatic increases in performance were achieved by the 1998 cohort of fourth graders, the first cohort of students to experience the standards-based instructional program from kindergarten through fourth grade. This indicates the important role of a well-aligned instructional program in a standards-based system.

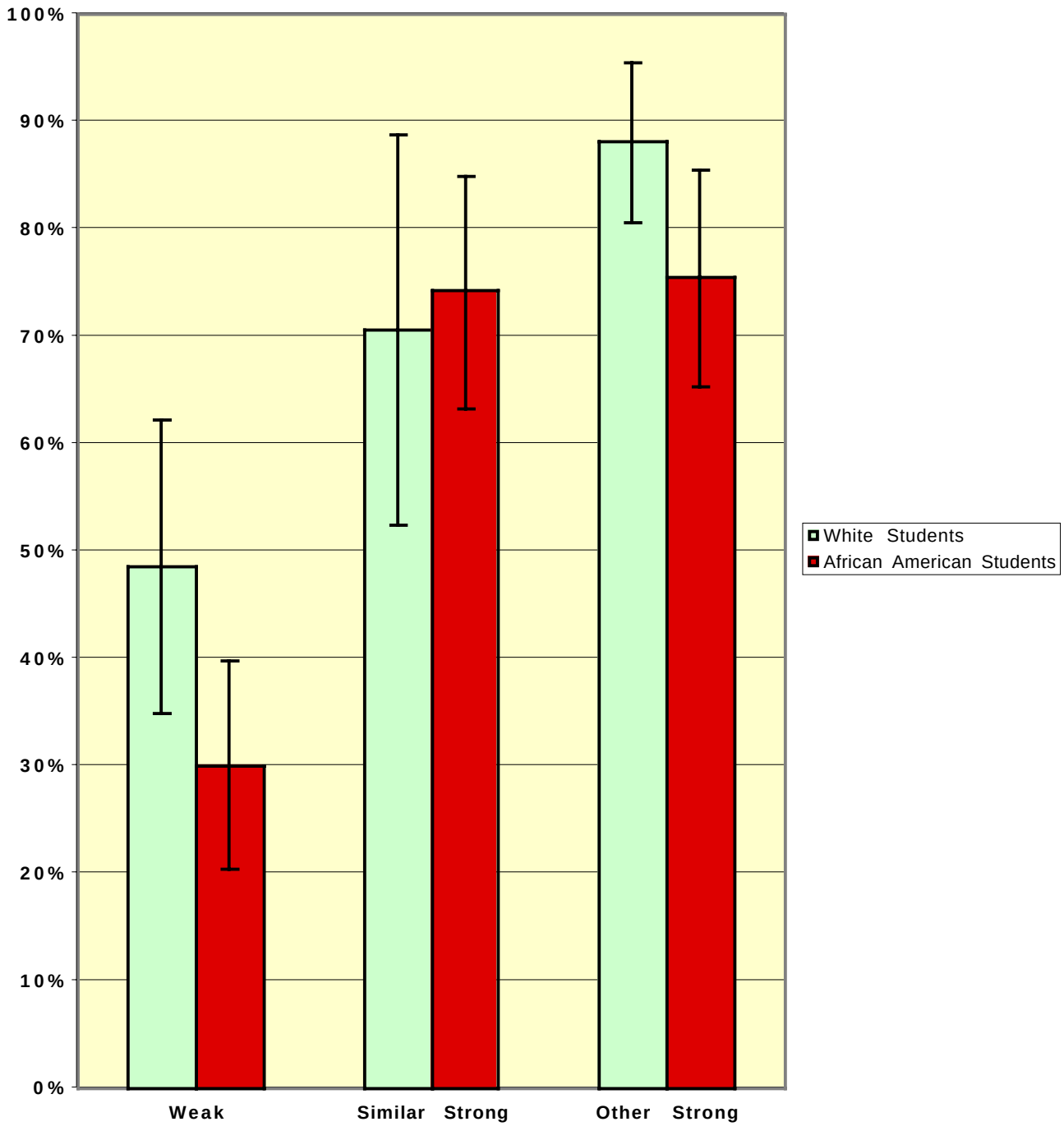


Figure 8a. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation, school demographics and race. Percentage of students who *achieved the skill standard*. Error bars denote the 99% confidence interval for each data point.

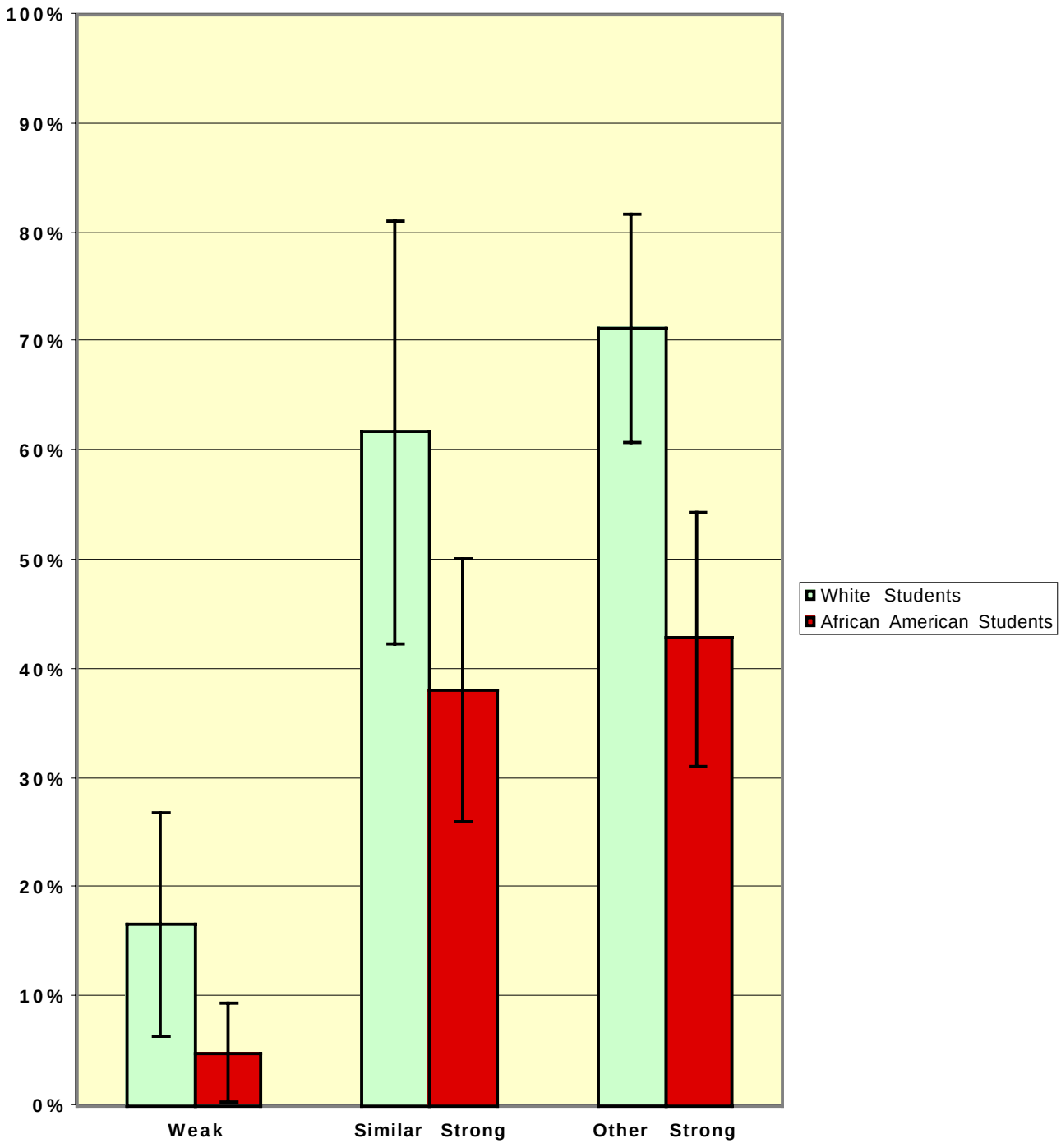


Figure 8b. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation, school demographics and race. Percentage of students who *achieved the concept standard*. Error bars denote the 99% confidence interval for each data point.

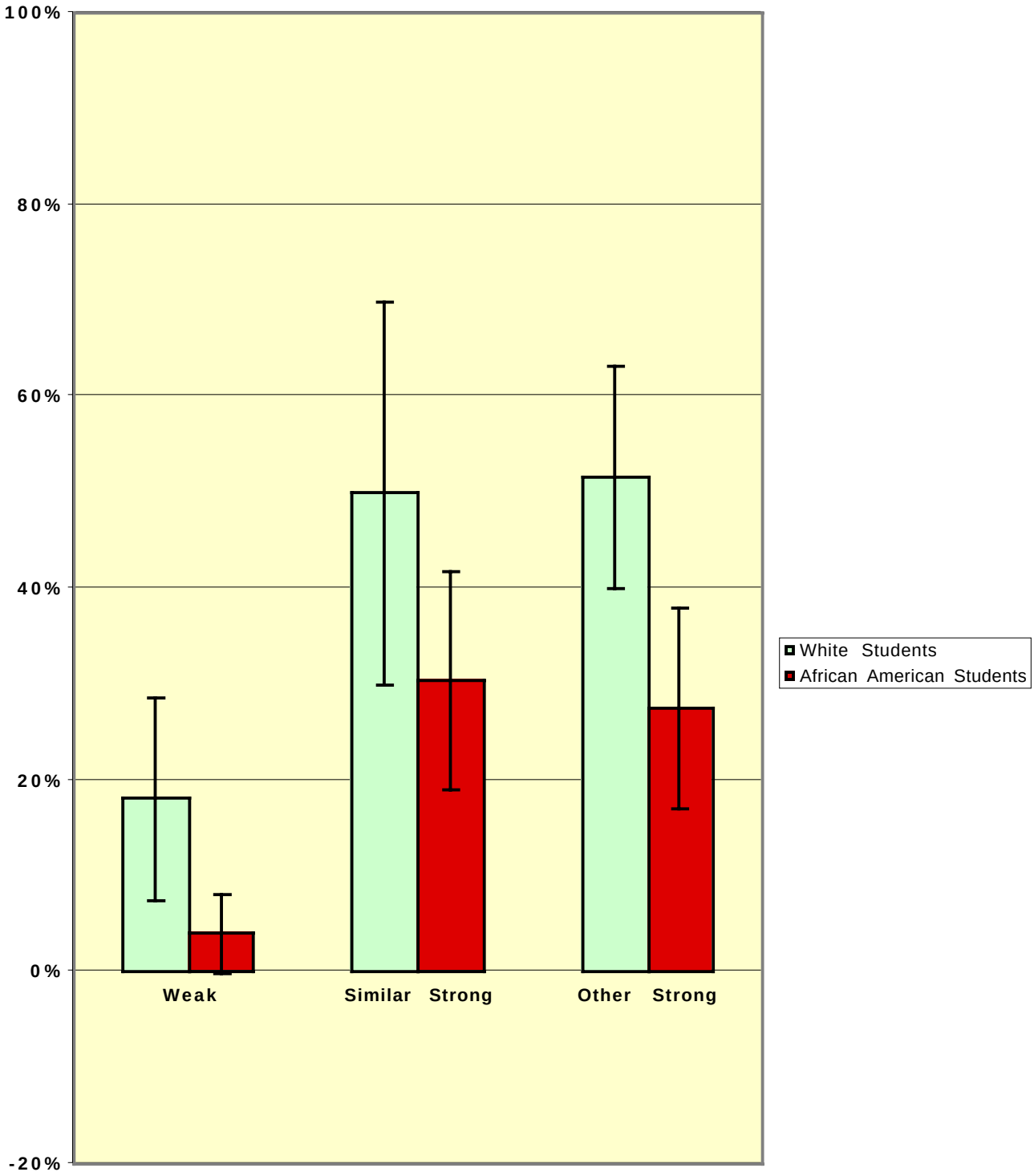


Figure 8c. NSMRE Grade 4 1998 results by level of *Everyday Mathematics* implementation, school demographics and race. Percentage of students who *achieved the problem solving standard*. Error bars denote the 99% confidence interval for each data point.

The pattern of results on the three New Standards standards—skills, concepts, and problem solving—is characteristic of what has been found in other jurisdictions using the New Standards math exams (e.g., Resnick & Harwell, 2000; Yoon & Resnick, 1998). In every year, both before and after the introduction of the curriculum and professional development program, Pittsburgh students did best on skills, next best on concepts, and worst on problem solving. This pattern is not surprising, given the traditional focus of American mathematics teaching on computational skills. Items on traditional tests are often called *problem solving* when they surround a computational task with a "cover story," that is, a simple word problem requiring no problem formulation or explanation. In the New Standards exams, such problems appear on the Skills portion of the test; Concepts and Problem-Solving tasks require substantial understanding, including written explanations for solutions. This is something students in Pittsburgh had had limited experience with prior to the introduction of New Standards and *Everyday Mathematics*. When such tasks became a standard part of the mathematics program, with the adoption of New Standards and *Everyday Mathematics*, scores began to rise.

The especially sharp drop in number of students at the very lowest score level on the Concepts and Problem Solving standards is especially noteworthy. This lowest score level usually means that a student has no ability to even begin to engage with the largely constructed-response tasks and the reasoning called for in them. As teachers throughout the district became more familiar with the expectations of the New Standards exams and with a curriculum that supported concepts and problem solving, very few children remained who could not at least tackle the problem-solving and concept tasks. It is reasonable to infer that a change in teaching was beginning to spread throughout the district.

Nevertheless, there were very large differences among the schools, with Strong Implementation schools showing two to five times more students meeting the standards than Weak Implementation schools on the New Standards Reference Examination. There were highly significant differences between Strong and Weak Implementation schools on the Iowa Test as well. Implementing the instructional program as intended was necessary to get the achievement gains sought.

Our data show that strong implementation of the program, with its associated rises in measured achievement of students, was not due to demographic differences between the Strong and Weak Implementation schools. There were large differences in 1998 achievement between students in Weak Implementation schools and the

demographically matched Similar Strong schools. And there were no differences between the Similar Strong and Other Strong schools. In other words, it was possible to implement the program well in schools with very diverse urban populations of students; and when the program was well implemented, achievement gains were high.

Good implementation of the program also led to especially great improvement in achievement for African American students. On the New Standards Skills standard, the traditional "gap" between White and African American students was essentially closed. On all standards, African American students in Strong Implementation schools performed a great deal better than White students did in Weak Implementation schools.

We do not know for certain to what degree teachers in the Weak Implementation schools might have improved had they participated more energetically in the district's professional development program. Scores in the Weak Implementation schools were somewhat below those of Strong Implementation schools, at least on skills measures, even before the new program and its associated professional development were introduced. Were teachers in these schools somehow less *able* to benefit from the new professional offerings. Put another way, were the Strong Implementation schools simply filled with more "early adapter" teachers, those who take quick advantage of new professional opportunities and often show achievement improvements with whatever programs they attempt to use?

The extraordinary results in the schools that fully implemented Pittsburgh's elementary math program show what is possible in a fully aligned standards-based system. But these very successes also highlight the complex policy environment for efforts to upgrade instruction to meet new standards for academic achievement. We did not set out to study implementation policy and do not have systematic data on how the Pittsburgh Public Schools' accountability system actually functioned. Nevertheless, we can point to some reasons for the school system's having proceeded very cautiously in requiring use of its officially adopted program.

In Pittsburgh, as throughout the country, there exists a pervasive culture of teacher independence. Teachers are expected to individually modify textbook lessons to meet the needs of their students without necessarily consulting with others. This contrasts with standard practice in some other countries, in which joint

lesson study by teachers builds a professional knowledge base that is shared within and across schools (Stigler & Hiebert, 1999). Furthermore, in previous Pittsburgh mathematics program adoptions, modifying the textbook lessons was considered necessary because the adopted textbooks did not provide enough problem-based lessons and manipulative activities. The *Everyday Mathematics* adoption was the first that no longer considered it acceptable for teachers to modify the program substantially or to teach it "their own way"; but many teachers and schools continued to operate under old assumptions.

Everyday Mathematics also presented new challenges to building and central office administrators in monitoring program implementation. Many teachers were mathematically unprepared to teach the curriculum and needed substantial content preparation. *Everyday Mathematics* can also exacerbate weaknesses in classroom management, and many principals may not have had the background to help teachers meet the new management demands. Finally, parents frequently questioned *Everyday Mathematics* and expressed a desire for a more traditional program. Strong Implementation schools had to provide support to parents as well as to teachers. These factors all point to a need for considerable professional development for both administrators and teachers if programs such as *Everyday Mathematics* are to be well implemented across an entire school district.

The situation described here is not unique to the Pittsburgh Public Schools. Similar conditions can be found in many jurisdictions throughout the United States in which practices of site-based management and professional autonomy for teachers have led district leadership to be shy of imposing districtwide programs. Data such as those presented here, showing the power of an aligned system of standards, assessment, curriculum, and professional development, pose a challenge to this practice. If certain programs are demonstrably effective, should schools and teachers have the right, in the name of local autonomy, to continue to use ineffective programs? Might accountability to students and their achievement call for a re-evaluation of some of our practices of local decision making? To say this is not to suggest that matters can be improved merely by officially *requiring* use of an adopted program or of holding schools *accountable* for raising their students' test scores. Official requirements are mediated by the actions of people working in the system, who act in accord with their beliefs and capabilities in the context of official policy. *Everyday Mathematics* called for new professional behaviors by teachers, administrators, and central office staff. Pittsburgh had paper policies and

mechanisms in place that might have been used to monitor implementation of the program. Principals and central office personnel to whom the principals report were often unwilling to confront those who were not fully implementing the program, especially without compelling data about the program's benefits for students. Now that those data are available, Pittsburgh and other school districts in similar situations will have to take on the complex challenges of insuring that their intended policies concerning instruction and professional development become implemented practices.

References

Harcourt Educational Measurement. (1996, 1997, 1998, 1999). *New Standards Mathematics Reference Examination, Forms B-D*. San Antonio, TX: Harcourt, Inc.

Iowa Tests of Basic Skills Summary Battery, Form K. (1993). Itasca, IL: Riverside Publishing Company.

National Commission on Education Standards and Testing. (1992). *Raising standards for American education*. Washington, DC: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

New Standards. (1996). *New Standards Mathematics Reference Examination technical summary* (Vol. 1). Pittsburgh, PA / Washington, DC: Learning Research and Development Center, University of Pittsburgh / National Center for Education and the Economy.

Resnick, L. B., & Harwell, M. (2000). *Professional development and teaching quality in a standards referenced education system* (Final report to CRESST / U.S. Department of Education, Office of Educational Research and Improvement). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Smith, M. S., & O'Day, J. (1990). Systemic school reform. In *Politics of Education Association yearbook* (pp. 233-267). London: Taylor & Francis Ltd.

Stigler, J. W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

University of Chicago School Mathematics Project. (1995). *Everyday mathematics: Teacher's manual and lesson guide*. Evanston, IL: Everyday Learning Corporation.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New Standards Examinations for the California Mathematics Renaissance* (Final report to CRESST / U.S. Department of Education, Office of Educational Research and Improvement). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.