Chapter Six

NAS DESIGNS AND ACADEMIC ACHIEVEMENT

Mark Berends, Sheila Nataraj Kirby, Scott Naftel,
Christopher McKelvey, Sue Stockly, R. J. Briggs,
JoAn Chun, Gina Schuyler, Brian Gill, Jodi Heilbrunn

The overall mission of NAS is to help schools and districts significantly raise the achievement of large numbers of students with whole-school designs and the assistance design teams provide during the implementation process. This chapter provides policymakers and researchers some understanding of the performance progress that NAS made within the partnering jurisdictions during the scale-up phase. This chapter focuses specifically on the following research questions:

- Did NAS schools make gains in test scores relative to all schools in their respective jurisdictions?
- What were the achievement gains across grade levels of individual students in NAS schools compared with non-NAS students?

Before turning to the findings, we must explain what this analysis is and what it is not, so we provide some background on the analysis. We then present our findings using school-level test scores. After that, we discuss the relationship, or lack thereof, between school-level aggregate scores and school-level implementation as we measured it in Chapter Four. We go on to present the findings on student-level test scores in San Antonio and Memphis. Finally, we cover findings from the final set of case studies and others' work before providing a general summary (more details of the methods we used appear in the Appendix).

BACKGROUND OF THE ANALYSIS

A major presumption behind the NAS concept was that each design would be responsible for evaluating its own efficacy. The RFP required evaluation section in the proposal and NAS itself constantly promoted self-evaluation or third-party evaluations for each of the teams. The teams turned out to vary in their ability in this regard and in the energy they spent on it.

NAS requested that we examine the progress teams made toward the goals of improving student performance in the schools undergoing scale-up in the partnering jurisdictions. NAS was not interested in progress made outside of the scale-up sites. Thus, RAND was not asked to evaluate each team's efficacy in improving schools and student outcomes in all of their respective sites. Rather, our work is confined to schools using designs in the partnership districts during the scale-up phase.

While RAND and the Annenberg Advisory Panel recommended NAS develop a set of assessment instruments geared to measure the types of performance the design teams expected, this advice was not taken for several reasons. First, NAS might not have had the resources available for this type of undertaking. Second, the design teams did not agree on a set of assessments, and several did not have assessments in place to examine. Third, and perhaps most important, the partnering districts insisted that the schools be held accountable to the state- or district-mandated assessments. For the most part, NAS and its design teams were concerned that these existing tests were not intended to measure critical-thinking skills or complex reasoning skills on which many of the designs focused (Mitchell, 1996). Even so, NAS and the design teams agreed to partner with these districts and to accept the state- and district-mandated tests as the means of measuring improved performance. It was thought by many at the time that the designs would minimally be able to show progress in these areas, so there should be no concern. Finally, because these were whole-school designs, the schools were an important focus for our analyses. Districts expected average school test scores to improve as a result of implementation of the design.

RAND tracked test score results in partner districts from 1995 to 1998. We focused on evidence based on school- and student-level

achievement. For the *school*-level results, we examined whether NAS schools made gains in reading and mathematics scores relative to all schools in their respective jurisdictions. At the *student*-level, we focused attention on two supportive school districts—San Antonio and Memphis—to understand whether NAS designs were related to student achievement compared with non-NAS schools. The results from San Antonio also enable us to control for other relevant student, classroom, and school characteristics within a multilevel framework.

The performance trends portrayed span only a few years, and several design developers and school reformers emphasize that it takes several years to expect implementation to take hold throughout the school (Sizer, 1992; Hess, 1995; Levin, 1991; Darling-Hammond, 1988, 1995, 1997). In addition, our results clearly show the wide variation in implementation both within schools and among jurisdictions and design teams. Thus, because of this variation in implementation, one should not expect robust performance results across the NAS sites. However, it is important to examine trends in performance to set realistic expectations for meaningful schoolwide improvement.

MONITORING ACADEMIC PROGRESS WITH SCHOOL-LEVEL TEST SCORES

When examining school-level achievement, we analyzed data on trends in mathematics and reading scores for NAS schools and the associated jurisdiction for selected grades in elementary, middle, and high schools, where relevant. Because we were concerned about the variability that particular grade test scores show within a given school, we generally aggregated across NAS schools, using grade enrollment as weights. Thus, the comparisons being made are generally between NAS schools and the district or the state.[1]

---

[1]The comparison we make here between NAS schools and the district averages uses absolute gains. In addition, we also calculated and compared percentage gains in test scores for the NAS schools and the jurisdictions. The results were not substantially different from those presented here. Moreover, although not reported here, we compared the gains in test scores of the individual NAS schools with their past performance to see if the schools made *any* gains over time. Again, the results did not differ from those discussed in this section.

Moreover, it is important to note that some of the designs do not specifically have curriculum and instruction materials per se, and even some design teams that do may not have been implementing that particular design component. This should be kept in mind when examining the results that follow. However, mathematics and reading are central to improving student learning for large numbers of students. These subject area tests are also central to the accountability systems of the jurisdictions in which NAS schools are located. Thus, we focus on these two subject areas.

The fact that NAS schools began implementing at different times makes clear comparisons of gains over time difficult. Wherever possible, we show data for the baseline and baseline plus two years. For some late implementing schools, we show the baseline and baseline plus one-year data. (For more details on these results and the tests used by the various jurisdictions see Berends and Kirby et al., 2001.)

For these results, we relied on the tests administered by the districts as part of their accountability system. While not ideal, these were the tests the jurisdictions, NAS, and the design teams expected to influence during the course of the NAS scale-up strategy. In its initial request for proposals, NAS's intent was for "break the mold" schools. NAS was not interested in incremental changes that led to modest improvement in student achievement compared to conventional classrooms or schools. Rather, the achievement of students was to be measured against "world class standards" for *all* students, not merely for those most likely to succeed. Moreover, design teams were to "be explicit about the student populations they intend to serve and about how they propose to raise achievement levels of 'at risk' students to world class standards" (NASDC, 1991, p. 21).

If such ambitious effects on student achievement occurred, these large test score changes would be reflected in school-level scores. Yet, to fully understand the test score trends of NAS schools three years into scale-up, it is important to keep in mind several issues when examining school-level scores.

First, differences in achievement between schools are not nearly as great as the achievement differences within schools. For the past 30 years, a finding on student achievement that has stood the test of time is that about 15–20 percent of the student differences in

achievement lie *between* schools; most of the achievement differences (80–85 percent) lie *within* schools (Coleman et al., 1966; Jencks et al., 1972; Lee and Bryk, 1989; Gamoran, 1987, 1992). Understanding the differences between schools remains critically important for making changes that maximize the effects of schools on students. However, it is also important to understand the limitations of schools—no matter what the school reform—in explaining the overall differences in student achievement (Jencks et al., 1972).

Second, when examining the grade-level scores over time (e.g., 4th-grade scores between 1995 and 1998), these are based on different cohorts of students taking the tests. These scores are often unstable because some schools have small numbers of students taking the test in any given year, and these scores are more likely to vary from year to year with different students taking the test. Districts and states use such scores in their accountability systems, and over a longer period of time, they provide some indication of a school's performance trends.

Third, while establishing trends in the NAS schools relative to other schools within the same district is informative, it is important to remember the variety of family, school, district, and design team factors that influence these scores. Research on student achievement has consistently found that individual family background variables dominate the effects of schools and teachers (Coleman et al., 1966; Jencks et al., 1972; Gamoran, 1987, 1992), and such effects are not controlled for when describing school-level test scores. More-specific information than districts typically collect or make available is necessary to understand the relative effects of these factors on student achievement.

Fourth, the ways districts report their scores to the public are not always amenable to clear interpretations over time. For example, several districts changed their tests during the scale-up phase, and the tests in some cases have not been equated, so the test scores are not directly comparable over time. Moreover, in some instances, the form in which test score information is reported (for example, median percentile rank) makes it difficult to detect changes in the tails of the distribution. Wherever possible, we have tried to obtain specific test score information at the school level to clarify the interpretations that can be made.

Fifth, the way that we summarize school performance—comparing whether the NAS schools made gains relative to the jurisdiction—may miss some significant achievement effects that could be captured if student-level data were available and comparable across the jurisdictions. That is, our indicator will only reflect large achievement effects of designs. The data provided by the districts do not support more fine-grained analyses to understand smaller, statistically significant effects on student-level achievement scores, particularly for certain groups of students (e.g., low-income or minority students or students with limited English proficiency).

Comparing NAS Schools with District Averages: Setting Expectations

NAS schools were predominantly high-poverty and high-minority, and many faced challenges related to student mobility.[2] It could be argued that comparisons with the district average are unfair to these schools, particularly if they fail to capture smaller, albeit significant, achievement effects.

However, it must be pointed out that NAS and the design teams agreed to be held accountable to district assessments and to improve student learning for substantial numbers of students. Because of these expectations, NAS requested that RAND examine the progress of these NAS schools relative to the district averages to understand whether the NAS expectations of dramatic improvement were met.

Sample of NAS Schools for Performance Trend Analyses

The sample of NAS schools for which we have data on test scores is larger than the sample of schools used for the implementation analysis. Of the 184 schools in the original sample, we have data on 163 schools. Some schools were dropped from the sample because

[2]When examining trends in school performance, it is important to consider the state and district accountability system (Berends and Kirby, 2000; Miller et al., 2000; Koretz and Barron, 1998). For example, different exclusion rules for special population students could result in different rates of achievement growth across jurisdictions and bias outcomes for particular groups of schools. However, the comparisons made here are between NAS schools and the jurisdiction average. Therefore, all the schools are supposed to be subject to similar testing provisions and administration.

they were not implementing: This was true of the Pittsburgh schools and about 12 schools in Dade. Some of our schools were K–2 schools for which there was no testing data available and other schools were missing data on test scores.

Our analysis of performance trends focused on whether NAS schools made gains in test scores relative to their respective jurisdictions.

Overall, the results are mixed (see Table 6.1). Of the 163 schools for which we had data, 81 schools (50 percent) made gains relative to the district in mathematics and 76 schools (47 percent) made gains in reading.

Differences in School Performance by Jurisdiction

Among the four jurisdictions with ten or more implementing NAS schools, Memphis and Kentucky schools appear to be the most successful in terms of improvement in mathematics, while Cincinnati and Washington state do better in reading (Table 6.1).

Differences in School Performance by Design Team

Examining school performance results by jurisdiction inevitably brings up the question: Which design teams appear to be the most successful in improving student test scores? In many ways, this is an unfair question. School performance and implementation vary importantly across jurisdictions. Given:

- the importance of district environments and support in implementation of the designs;
- the uneven implementation of designs across the jurisdictions;
- the uneven distribution of designs across jurisdictions and small sample sizes for some designs;
- the variation in testing regimes; and
- the possible lack of alignment between assessments and design team curriculum, instruction, and goals,

it is difficult to compare "success" rates of various designs in a meaningful and fair fashion. Nonetheless, NAS and the design teams

agreed to be held accountable to district standards, and NAS expected dramatic achievement gains across design teams.

Thus, we present the performance summary results by design to help set expectations for those implementing comprehensive school reforms (see Table 6.1). The results vary across the two subject areas. For example, for the eight AC schools, five made progress relative to the district in mathematics, but only two did so in reading. With the exception of ATLAS and EL schools, about half of the other design team schools made progress relative to the district in mathematics; in reading, fewer than half of AC, CON, and NARE schools made gains relative to the district. RW was the most consistent, with ten out of 21 schools making progress in both reading and mathematics relative to the district. Of the 11 MRSH schools, seven made progress in mathematics and eight in reading.

Once again, we warn that these results need to be interpreted in the context of district environments. Because of the wide variation in implementation and environments that occurs within schools and among jurisdictions, one should not expect robust performance results across the NAS sites after only a couple of years at most. In addition, better and longer-term performance data at the student level are needed in order to make conclusive judgments about designs and their effects on student achievement, controlling for important school, classroom, and student characteristics.

THE LINK BETWEEN IMPLEMENTATION AND PERFORMANCE AT THE SCHOOL LEVEL

One of the goals of the RAND analysis plan is to monitor progress in implementation and performance in NAS schools and to understand the factors that relate to higher implementation and higher performance. Such findings will not only inform New American Schools, but also the CSRD program now under way.

However, as the above section has made abundantly clear, we do not have good, sustained, and coherent measures of school-level achievement scores that are comparable across jurisdictions and across design teams. The summary tables we show above compared gains in NAS schools with changes in the district test scores—any gains—but as we detailed in each section, sometimes the compar-

Table 6.1

NAS Schools Making Gains Relative to Jurisdiction, by Jurisdiction and Design Team, Three Years into Scale-Up

	Number of schools	Number making gains in test scores relative to district
	Jurisdiction	
Math		
Cincinnati	18	9
Dade	11	6
Kentucky	51	30
Memphis	30	16
Philadelphia	19	7
San Antonio	12	4
Washington state	22	9
Reading		
Cincinnati	18	10
Dade	11	5
Kentucky	51	22
Memphis	30	11
Philadelphia	19	11
San Antonio	12	7
Washington state	22	11
	Design Team	
Math		
AC	8	5
AT	24	9
CON	17	10
EL	16	4
MRSH	11	7
NARE	66	36
RW	21	10
Reading		
AC	8	2
AT	24	15
CON	18	6
EL	15	8
MRSH	11	8
NARE	66	27
RW	21	10
	Overall	
Math	163	81
Reading	163	76

isons were across one year, sometimes across two years, and often covered different time periods, where cohorts of schools were involved.

Our data do not show any clear linkage between implementation and performance in NAS schools. This was disappointing and runs counter to conventional wisdom. If the theory of action underlying comprehensive school reform is correct and if these models are implemented in a sustained coherent fashion, then higher implementation should be related to improved outcomes. As Stringfield et al. (1997, p. 43) conclude in *Special Strategies for Educating Disadvantaged Children*, "We know that some programs, well implemented, can make dramatic differences in students' academic achievement." Yet, Stringfield et al. go on to point out the critical challenge in educational reform that has existed in this country for decades:

> [A]fter a third of a century of research on school change, we still have not provided adequate human and fiscal resources, appropriately targeted, to make large-scale program improvements a reliably consistent reality in school serving students placed at risk. (p. 43.)

We offer some hypotheses for the failure to find a link between implementation and performance when examining school-level aggregates. First, despite schools reporting implementation of designs, it remains relatively early for expecting deep implementation that would dramatically affect performance gains. As Sizer (1984, p. 224) points out, "Schools are complicated and traditional institutions, and they easily resist all sorts of well-intentioned efforts at reform." Moreover, as several design developers and school reformers have pointed out, schoolwide change can take more than five years for a school to accomplish meaningful change (Sizer, 1992; Hess, 1995; Levin, 1991; Darling-Hammond, 1988, 1995, 1997).

Some of the design teams emphasize that it takes several years to expect implementation to take hold throughout the school (Bodilly, 1998; Smith et al., 1998). Only with coherent implementation would one expect school test scores to consistently increase throughout the school. Our analysis shows a large number of NAS schools near the midlevel implementation points on scales for the wide array of indi-

cators considered here. Moreover, there is a great deal of variation among teachers within the NAS sites. While there is a range in implementation levels observed in our analysis, it is probable that implementation is not deep enough throughout the schools at this point to raise student scores across grade levels. Over time with more-specific test score information and additional measures of implementation, the empirical link might be observed. This remains an open question.

Second, the nature of our dependent variable—a simple 0/1 variable—does not allow for any gradations in student performance. Had we been able to calculate effect sizes, perhaps we would have seen a link between implementation and performance.

Third, the analysis sample may have failed to find evidence of the link between implementation and student performance, perhaps because of measurement error in our indicators. Although our implementation indicators appear to be credibly constructed and to track well with Bodilly's findings, they may fail to capture important aspects of implementation that are linked to school performance. The great variability that we see within schools in implementation adds to the difficulty in measuring mean implementation levels in a school. The summary measures examined in this study may not have the power to distinguish fully between schools with higher and lower levels of implementation. As we noted in Chapter Four, the majority of the schools' implementation levels were at the midpoints on our scales, and there was a great deal of stability between 1997 and 1999 (i.e., both in mean levels and within-school variance) (see also Kirby, Berends, and Naftel, 2001; Berends and Kirby et al., 2001).

MONITORING ACADEMIC PROGRESS WITH STUDENT-LEVEL TEST SCORES

The ultimate aim of school reform efforts and implementation of NAS designs is to substantially improve student performance. As we pointed out, analysis of grade-level aggregate scores within NAS schools compared with the district are fraught with problems for which it is difficult to control with available data. However, other analyses of student test scores in what were two ostensibly supportive NAS districts—San Antonio and Memphis—reveal that

significantly raising student achievement scores, sustaining them over time, and attributing them to design team activities presents a substantial challenge as well.

Student Achievement in San Antonio

As described in the previous chapter, elementary school students in San Antonio take the TAAS. Given the available data, we conducted two sets of analyses: First, for the entire district we examined the effects of student, teacher, and school characteristics on the 4th-grade TAAS reading and mathematics scores, controlling for prior achievement. Data provided by the San Antonio district and other sources allowed for construction of a data set containing more than 3,800 4th-grade students in about 280 classrooms in all 64 elementary schools in the district. Individual 4th graders' TAAS reading and mathematics scores were regressed against students' prior achievement and student, teacher, classroom, and school characteristics using multilevel models to partition the variation in reading and mathematics achievement into student and classroom components. Second, we analyzed student achievement in a subsample of over 800 students in 63 classrooms for which teachers completed our survey.

The results at the district level provide the context for the subsample. Data gathered from the teacher surveys help inform the district analysis on the impacts of teacher practices and perceptions of student achievement. In addition, these students were administered the Stanford-9 open-ended reading test, making possible an independent measure of student performance without the "high stakes" implications of the TAAS.

After controlling for all of these student, classroom, and school characteristics, we fail to find a significant effect of implementation of NAS designs in San Antonio.[3] This same result came up in estimations of a variety of other model specifications, using other regres-

[3]Because students are nested within classrooms, which are nested within schools, we relied on multilevel modeling techniques to provide more accurate estimates of student-, classroom-, and school-level effects (see Bryk and Raudenbush, 1992; Bryk, Raudenbush, and Congdon, 1996; Singer, 1998; Kreft and De Leeuw, 1998). Further details are available in Berends et al. (2002).

sion techniques such as ordinary least squares, three-level linear models and probit models, where the dependent variables were binary indicators of passing or failing scores.[4] This is not surprising since we are examining effects on spring 1998 scores, and many of the designs had not been in place that long. In addition, implementation was not deep in these schools, given the conflicting reforms that overshadowed implementation of NAS designs in these schools. As such, we would not expect to find effects of implementation on student achievement.

Because instructional conditions varied more between NAS and non-NAS schools during the 1997–1998 school year, we wanted to examine whether such variation in instructional conditions was related to student achievement, controlling for other student, teacher, classroom, and school characteristics. We first examined relationships in all 4th-grade classrooms in the district and then in the sample of classrooms for which RAND gathered additional survey data on classroom instruction and a supplemental reading test (Stanford-9) (see Berends et al., 2002).

We did not find that instructional conditions promoted by reforms such as NAS—including teacher-reported collaboration, quality of professional development, and reform-like instructional practices—were related to student achievement net of other student and classroom conditions.

However, we did find significant effects of principal leadership on the TAAS reading and mathematics scores by 0.15 and 0.21 of a standard deviation gain, respectively. Principal leadership in our analysis was measured by teacher reports about principals who clearly communicated what was expected of teachers, were supportive and encouraging of staff, obtained resources for the school, enforced rules for student conduct, talked with teachers regarding instructional practices, had confidence in the expertise of the teachers, and took a personal interest in the professional development of teachers. Chapter Four described how our previous analyses have shown the importance of principal leadership in implementing the designs

[4]For example, additional analyses of longitudinal student achievement growth models from grades 3–5 do not show significant effects of NAS classrooms and schools compared with non-NAS comparisons.

(Berends and Kirby et al., 2001; Kirby, Berends, and Naftel, 2001). In our San Antonio classroom study, we found a link between principal leadership and student achievement in NAS and non-NAS schools, indicating that leadership is important for academic achievement in general, and to implementation in particular.

Student Achievement in Memphis

Memphis was another supportive district of NAS designs during the scale-up phase. The superintendent provided significant resources toward designs and was committed to NAS's scale-up strategy to widely diffuse the designs within the district. In fact, her leadership during her tenure in Memphis resulted in her being honored as national superintendent of the year.

Memphis has used the Comprehensive Test of Basic Skills, version 4 (CTBS/4) since 1990. This is a commercial, multiple-choice test that measures skills in reading, mathematics, and other subject areas. In the spring of 1998, Memphis adopted the CTBS/5 Complete Battery Plus (Terra Nova). This latter version of the CTBS as tailored to the State of Tennessee is also a multiple-choice test, but concentrates on higher-order thinking skills to a greater extent than the previous CTBS/4. Scores have been equated across the two tests. Produced by CTB/McGraw-Hill, both forms of the test contain items developed specifically for students in Tennessee.

Tennessee has a sophisticated testing and assessment program called the Tennessee Value-Added Assessment System (TVAAS), which enables the tracking of the academic progress of every student in the state in grades 3–8 and beyond (as high school testing is implemented) in science, math, social studies, language arts, and reading (see Sanders and Horn [1994, 1995] for more details on this system and the methodology used to measure student progress). TVAAS reports annually on the gains that students made in each grade and each subject grouped by achievement levels. These reports have information on the three most recent years as well as the three-year average gains. The state monitors all school systems that are not achieving national norm gains; those systems

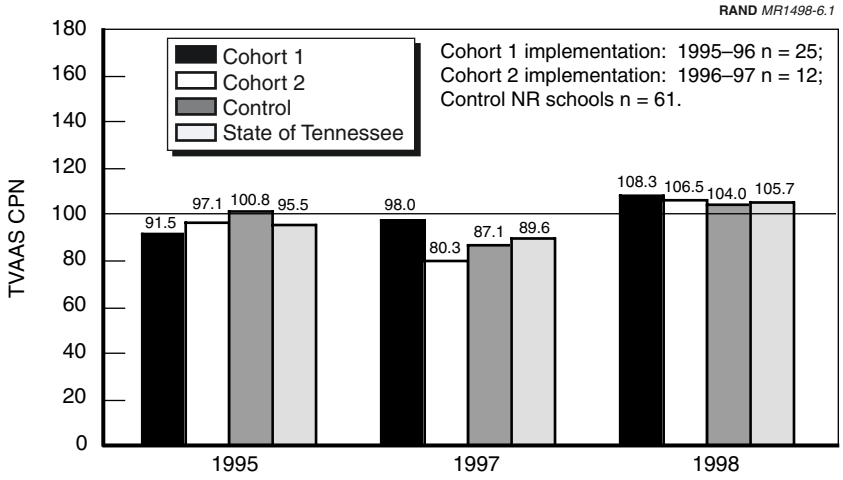
> achieving two or more standard errors below the national norms must show positive progress or risk intervention by the state. Each

> school and system is expected to achieve the national norm gains regardless of whether its scale scores are above or below the national norm. (Sanders and Horn, 1994, p. 302.)

The raw data for TVAAS are the scaled scores from the CTBS/4 and now CTBS/5, which form a part of the Tennessee Comprehensive Assessment Program (TCAP). All students in grades 2–8 are tested yearly; this information is linked to the school and the teacher by subject area and grade. The longitudinal nature of the data allows each student to serve as his or her own "control." TVAAS uses statistical mixed-model methodology to estimate a multivariate, longitudinal model of student achievement and then to aggregate these data to the classroom or the school level. The gain scores of a school's students are estimated and compared with the national norms. Thus, deviations from the national norms can be calculated to see how the school is doing with respect to a national sample of students.

The index of student achievement used in the analyses is the Cumulative Percent of Norm (CPN) mean. This measures the percent of national (expected) gain attained by the school in the reported grades (Bratton, Horn, and Wright, 1996). For example, if a school had a CPN equal to 75 percent in 5th-grade reading, then the average gain of the 5th-grade students in the school was 0.75 of the expected year-to-year gain based on a national sample.

Ross et al. (1998, 1999, 2000, 2001) provide an examination of the relative performance of restructuring elementary schools in Memphis from 1995 to 1998 (see Figure 6.1). They compared gains in the restructuring schools on the TCAP with non-restructured (NR) elementary schools and the state. Their results show that by 1998, both cohort 1 (in year 3 of implementation) and cohort 2 schools (in year 2 of implementation) demonstrated "small, nonsignificant advantages over the NR schools" (Ross et al., 1999, p. 3). An additional important finding is that higher-poverty schools appeared to derive the greatest benefits from these reforms. Their overall conclusion is that although the effects have varied by year and by cohort, restruc-



NOTE: Figure taken from Ross et al. (2001).

Figure 6.1—Memphis City Schools TVAAS Results for All Subjects Cumulative Percent of Norm, Mean Across Grades 3–5

turing shows promise in raising achievement in Memphis elementary schools.[5]

Yet, despite these relatively more-promising results Memphis has decided to drop the designs in favor of more curriculum-specific re-

[5]In our analysis, we found less positive results (see Berends and Kirby et al., 2001) when comparing the NAS-only designs with the district. We worked with Steven Ross of the University of Memphis and William Sanders of the University of Tennessee, who provided the supplementary results in our research. The data we examined for mathematics and reading for elementary schools were somewhat different from Ross et al. (1998, 1999, 2001) for several reasons. First, we compared the NAS schools with the district, and they compared the NAS school designs with "non-restructured" schools between 1995 and 1998. Second, Ross et al. also included some non-NAS schools in their analyses. Third, to be consistent to what we did for other jurisdictions, we compared Memphis NAS schools with the district, using base year of implementation to two years after implementation. Had we used the 1998 results for cohort 1, our results would have looked more similar to Ross et al. Fourth, we also examined secondary schools, where the picture seems somewhat more mixed: It varied by year and the most recent year is the least encouraging, with NAS schools well below the district average.

forms.[6] The former superintendent who brought in the NAS designs and provided significant support for them—about $12 million over six years—took another position. The incoming superintendent announced to the school board that the whole-school reform models would be discontinued in favor of new districtwide curriculum, beginning with a reading program in fall 2001. Similar to the situation in San Antonio, there were concerns about the effectiveness of designs and their ability to teach students more fundamental reading and writing skills. Apparently, recent score results in Memphis did not help in that they were not nearly as positive as in the past year (Ross, S. M., personal communication).

FINDINGS FROM CASE STUDIES

Similar to what we found in our quantitative analyses, the case study work offered some provocative, but inconclusive, information that might lead one to assert that a variety of factors other than design implementation account for the differences in test score gains between the matched pairs of schools that were the focus of the study (Chun, Gill, and Heilbrunn, 2001). These factors include student and family characteristics; stability, experience, and morale of the teaching force; and test preparation programs. Moreover, several factors likely contribute to the absence of a relationship between design implementation and test score results. These include:

- Tests that fail to capture the range of student learning outcomes targeted by NAS designs;
- Pressure on schools to raise test scores immediately and dramatically (which promotes the use of skills-oriented curricula at odds with ambitious, interdisciplinary designs); and
- Low levels of implementation across the board—an absence of truly comprehensive reform.

---

[6]This information is from an article in *The Commercial Appeal* by Aimee Edmonson entitled "Watson Kills All Reform Models for City Schools" (June 19, 2001). See also NAS's response in a press release of June 28, 2001—"New American Schools' Statement on Memphis Superintendent's Decision to Drop Comprehensive School Reform" (http://www.newamericanschools.com/press/062801.phtml).

SUMMARY AND POLICY IMPLICATIONS

Our analysis of performance trends across the set of schools three years into scale-up focused on whether NAS schools made gains in test scores relative to their respective jurisdictions.

- Among the four jurisdictions with ten or more implementing NAS schools, Memphis and Kentucky schools appeared to be the most successful in terms of improvement in mathematics, while Cincinnati and Washington state did better in reading.
- In total, of the 163 schools for which we have data allowing us comparisons in performance relative to the district or state, 81 schools (50 percent) made gains relative to the district in mathematics and 76 schools (47 percent) made gains in reading.

Because of the wide variation in implementation and environments that occurs within schools and among jurisdictions, it may have been too early to expect robust performance results across the NAS sites. However, our implementation analysis shows little increase in level of implementation over time and continuing within-school variation in implementation. Thus, one might expect design adoption to never have any lasting impact on student performance. In addition, better and longer-term performance data are needed in order to make conclusive judgments about designs and their effects on school performance.

The detailed classroom study of San Antonio allowed us to examine whether variation in instructional conditions was related to student achievement, controlling for other student, teacher, classroom, and school characteristics:

- As expected because of the early stages of implementation, elementary students in NAS schools did not significantly differ in their achievement growth compared with students in non-NAS schools.
- More importantly, we did not discover that instructional conditions promoted by reforms such as NAS—including teacher-reported collaboration, quality of professional development, and reform-like instructional practices—were related to student achievement net of other student and classroom conditions.

- However, we did find significant effects of principal leadership on the TAAS reading and mathematics scores.

Evidence from Other Studies

At this point in comprehensive school reform, there is only limited evidence about the effectiveness of design-based models from studies that rely on rigorous comparative evaluation designs. For example, Herman et al. (1999) find only two models were able to provide convincing results in terms of raising student achievement levels. In addition, in evaluations of the Comer's School Development Program, Cook and his colleagues at Northwestern University (see Cook et al., 1999; Cook et al., 1998) found no effect of the model on student achievement in Prince George's County, Maryland, but found small positive effects on students (less than one-tenth of a standard deviation) in Chicago schools. These Cook et al. studies were based on randomized experimental longitudinal designs, and both point to the importance of further longitudinal studies that carefully examine the approaches of design-based assistance providers and the variation in implementation and performance that is likely to occur. Cook et al. (1998) also point to the importance of district-level support and expectations for improving instruction and achievement. Other studies have shown that raising achievement levels in dramatic fashion within urban school districts is a formidable challenge (see Fullan, 2001; Bryk et al., 1998; Orr, 1998).

The evidence reported here suggests variation in implementation and performance and describes a number of factors related to implementation. The evidence suggests that design teams, districts, schools, and teachers have a great deal of work to do to fully implement designs on a broad scale before we can expect to see dramatic, or even significant, improvements in student outcomes. Whether large numbers of schools can implement whole-school designs in a sustainable fashion that can improve student achievement across grade levels remains an open question.