Appendix

METHODOLOGY FOR THE STUDIES ON IMPLEMENTATION AND PERFORMANCE

CASE STUDIES TWO YEARS INTO SCALE-UP (BODILLY, 1998)

Sample

Resources allowed the development of 40 school-level case studies with the intent of performing longitudinal analysis across each school over a two-year period. New American Schools was working with ten jurisdictions, including three state jurisdictions, with specific districts in those states, and seven independent districts.[1] We chose six of these jurisdictions to study in the first year of implementation.[2] We used seven in the second year.[3] That choice was in part determined by the evident progress made in getting the initiative under way in each jurisdiction.

---

[1]This includes selected districts within Kentucky, Maryland, and Washington state, and seven districts: Cincinnati, Dade, Memphis, Pittsburgh, Philadelphia, San Antonio, and San Diego.

[2]In the first year, we chose to study the six jurisdictions that had schools that were beginning implementation that year. These included: Cincinnati, Dade, two districts in Kentucky, Memphis, Pittsburgh, and two districts in Washington state.

[3]In the second year, NAS's relationship with some districts changed; thus, the district sample changed. We added two jurisdictions (Philadelphia and San Antonio) and dropped one (Washington state), making seven jurisdictions in the second year: Cincinnati, Dade, two districts in Kentucky, Memphis, Pittsburgh, Philadelphia, and San Antonio. Five jurisdictions stayed the same over the entire study and two new ones were added in the second year.

In choosing schools to study, we attempted to get at least four schools for each design team, to be able to track differences both among designs and among districts. However, each team does not work in each jurisdiction, and each team is implementing in different numbers of schools. For example, AC had less than 20, the fewest schools, while RW had more than 100 schools implementing at least the *Success for All* portion of the design. Neither were the teams uniformly dispersed throughout all districts. For example, Cincinnati had only CON, EL, and RW schools.

Table A.1 shows the sample for the second year of scale-up. Of the 40 schools we visited in 1997, we had also visited 30 in the previous year. Ten were visited only once. Seven of these ten were visited only once because they were added when we added districts in the second year. Those schools were in San Antonio and Philadelphia.

We attempted to make the choice of schools within a district random. In at least one case, we had little choice but to leave the selection to the district.[4] While not random, our sample was fairly representative of NAS schools in general. The sample included urban and

Table A.1

RAND Sample for Site Visits

	AC	AT	CON	EL	MRSH	NARE	RW	Totals
Cincinnati			2	2			2	6
Dade	1		2		2		2	7
Kentucky						4		4
Memphis	2	2	1	2	2		2	11
Philadelphia		3	1		1			5
Pittsburgh						3		3
San Antonio				2	2			4
Totals	3	5	6	6	7	7	6	40

[4]For example, the state of Florida put a group of Dade County schools on a probation list because of low performance against a set of state indicators. Dade County mandated that all schools on this list adopt the RW design and be off limits to researchers. Thus, this group could not be included in the sample, leaving us with no choice as to which RW schools to include—the only two RW schools not on the state probation list.

rural schools and districts; elementary, middle, and high schools; and schools that were well-resourced and schools that were not.

Data Sources and Collection

We used many sources and types of information:

- Structured interviews (a set of predefined questions, both open- and closed-ended) by telephone about resource usage and in person during field visits. The structured formats varied by type of respondent.
- Observations of activities in schools. These observations were not formal or extensive. We toured each of the schools, sat in on several randomly selected classes, and observed special events at the school scheduled for the day, if they had applications to the design. In several instances we were able to observe critical friends' visits taking place, teacher group meetings, etc.
- Archival data, including documents produced by design teams, schools, and districts describing their efforts; plans by these parties for transformation; and local news releases or newspaper items concerning the local education scene, local political issues, NAS, design teams, or schools using NAS designs.
- Numerical data on each school's enrollment, demographics, test scores, etc.

The major data collection in the field took place during two waves of site visits in spring 1996 and in spring 1997. The latter established the level of implementation of the 40 schools at the end of the second year of scale-up. All interviews probed for the reasons behind differing levels of implementation.

We attempted to ensure that two researchers visited each school for approximately one day. One researcher spent a day at the district collecting information and performing interviews. All interviews had a structured format, with a mix of factual closed-ended questions and open-ended questions. Interviews were translated into condensed formatted sheets that covered specific variables identified to be of interest and coded for later analysis. Specific issues, such as resource usage and the matching process between design teams and schools, were explored using structured phone surveys.

The analysis used a combination of quantitative and qualitative measures. We used the qualitative data to develop a quantitative measure for the dependent variable, the level of implementation.

Measuring the Dependent Variable

The implementation analysis used an audit-like approach to establish the level of implementation in each school. Schools associated with each team were assessed over common areas of schooling we call "elements" (see Table A.2). By *common*, we mean that each design included specific changes to that element from "typical" practice. These common elements were curriculum, instruction, assessments, student assignments, and professional development. But within each element of schooling, the teams varied significantly in what they attempted to accomplish.

Three elements remained, but were not held in common among the teams: staff and organization, community involvement, and standards. That is, not all teams aspired to make significant changes in these areas. Together with the five common ones, these are the eight elements that make up what we refer to as the "school-level component" of the designs. We also tracked progress on these three elements, as applicable.[5]

The specifics of each element for each design team were originally determined by a document review and interview with the design team during the demonstration phase. The elements were sharpened in scale-up by a request from NAS and several districts for design teams to create "benchmarks" of progress for their designs that schools and districts could use to understand where they were

[5]Our analysis of design documents shows that, in fact, the teams have more elements than these eight. Additional elements include governance changes, integrated technology in the classroom, and integrated social services. In scale-up, with the emphasis on developing a supportive environment within the district, these elements became part of NAS's jurisdiction strategy: all of the governance, integrated social services, and technology. We thus still tracked them, but not as part of the school-level designs. Instead, we tracked them as part of the jurisdiction's supportive environment that NAS was to encourage and support.

Table A.2

Elements of Designs

Element	Description
Curriculum	Usually, the knowledge bases and the sequence in which they are covered, whether defined by traditional subject areas or in more-interdisciplinary fashion.
Instruction	The manner in which the student acquires knowledge and the role of the teacher in this process.
Assessments	The means for measuring progress toward standards, either at the school or student level.
Student Grouping	The criteria or basis for assigning students to classes, groups, or programs.
Professional Development	Includes opportunities to develop curriculum and instruction, to develop expertise in using standards, to collaborate with others, and to enter into networks or prolonged discussions with other teachers about the profession. Several teams also planned extensive on-the-job practice, coaching in the classroom, and teaming in individual classrooms, as well as schoolwide forums to change the ways in which teachers deliver curriculum and instruction permanently.
Community Involvement/ Public Engagement	The ways parents, businesses, and others participate in schools and vice versa.
Standards	The range of skills and content areas a student is expected to master to progress through the system and the levels of attainment necessary for schools to be judged effective.
Staff and Organization	The configuration of roles and responsibilities of different staff. Changed organizational structures and incentives encourage teachers to access both staff in-services and professional growth opportunities.

going and when and to determine whether they were making reasonable progress. The benchmarks developed varied significantly from team to team as one would expect; however, all gave descriptions of what teams expected by the final year of a three-year implementation cycle.

We relied on two types of evidence of progress. First, we looked for evidence of implementation in keeping with the benchmarks and expectations provided by the team. Second, we interviewed district and school-level staff to understand their views of the design and how much they had changed their behaviors and to gain descriptions of the level of implementation. We asked how much their jobs had changed so far in relation to where they understood the design to be taking them.

Creating a Scale

The following paragraphs describe the construction of the dependent variable of the analysis—the level of implementation observed.

Level of Implementation

We rated progress in an element using a straightforward scale, as follows:

0 = **Not Implementing.** No evidence of the element.

1 = **Planning.** The school was planning to or preparing to implement.

2 = **Piloting.** The element was being partially implemented with only a small group of teachers or students involved.

3 = **Implementing.** The majority of teachers were implementing the element, and the element was more fully developed in accordance with descriptions by the team.

4 = Fulfilling. The element was evident across the school and was fully developed in accordance with the design teams' descriptions. Signs of institutionalization were evident.[6]

Application and Development of a Summary Dependent Variable

We initially applied these levels of implementation to each element that a design team intended to change in a school.[7] For *each element included in a design*, a score was given based on the observations and interviews conducted at the sites.[8] We then developed an average score for each school to use as a summary variable. First, we summed across the elements of design identified for each design team. For the five common elements, we totaled the values for each element and then divided by five to arrive at a school implementation level.[9] No weighting was attached to particular elements. For assessment of more elements, we summed across those included in the design and divided by the appropriate number (from five to eight). We assigned schools to the above categorizations based on the average score.[10]

---

[6]Implementation analysis often calls this level of implementation *institutionalizing* or *incorporating,* implying a level of stability and permanence. Our research indicates that the transience of the school and district political context often prevents institutionalization. We have thus used *fulfilling* to imply that the elements are present as the design teams intended, but we make no claim as to permanence.

[7]The reader should note that the use of numbers in the above scale does not imply interval-level data. The intervals between these points are not known. For example, a school with an average score of two is not halfway done with implementation. Neither is it twice as far along as a school scoring a one. The leap from planning to piloting might be far less formidable than the leap from implementation to the full vision of the design. In fact, a school scoring a three might take several more years to finish the design fully. The score indicates only what a school has accomplished in the way of implementation, as denoted in the above description.

[8]Reliability between raters was a potential issue in the creation of these scores. Reliability was increased by each rater performing this operation on a sample of schools that they and other raters had visited. The raters then exchanged scores and discussed discrepancies and how to resolve them.

[9]These five elements are curriculum, instruction, assessments, student grouping, and professional development.

[10]In assessing the total score of a school, the following intervals were used: A 0 or less than 0.8 was "not implementing;" a score equal to or greater than 0.8, but less than 1.6, was "piloting," etc.

LONGITUDINAL ANALYSES OF IMPLEMENTATION AND PERFORMANCE (BERENDS AND KIRBY ET AL., 2001; KIRBY, BERENDS, AND NAFTEL, 2001)

The Population of New American Schools for the Longitudinal Evaluation

The original sample of schools consisted of those schools initiating implementation of NAS designs in eight jurisdictions that NAS named as its partners during scale-up in either 1995–96 or 1996–97. These eight jurisdictions include:

- Cincinnati;
- Dade;
- Kentucky;
- Memphis;
- Philadelphia;
- Pittsburgh;
- San Antonio; and
- Washington state.[11]

The choice of these jurisdictions reflected RAND's desire to obtain a sample including all the designs that were participating in the scale-up phase and the judgment that the costs of working in the additional jurisdictions would not yield commensurate benefits. While jurisdictions and their support of the NAS reform will no doubt continue to change over time, these jurisdictions reflected a range of support for implementation—from relatively supportive to no support at all (see Bodilly, 1998).

---

[11]At the time we decided on the longitudinal sample of schools, Maryland and San Diego were not far enough along in their implementation to warrant inclusion in RAND's planned data collection efforts. Since then, several of the design teams report that they are implementing in Maryland and San Diego.

The 1998 Final Analysis Sample

Our aim was to collect data on all the NAS schools that were to be implementing within the partner jurisdictions. NAS believed that as of early fall of 1996, there were 256 schools implementing NAS designs across these eight jurisdictions. However, based on conversations with design teams, jurisdictions, and the schools, the sample was reduced to 184 schools for several reasons:

- There were 51 RW schools in Dade that were low-performing and on the verge of serious sanctions, so the district promised these schools that they would not be burdened with researchers.
- An additional 21 schools declined to participate because they did not want to be burdened with research, were not implementing, or had dropped the design.

Thus, for our surveys of teachers and principals, the target sample was 184 schools (see Table A.3).

Of the 184 schools in our 1997 sample, we completed interviews with 155 principals. Based on our interviews with principals in the spring

Table A.3

1997 Target Sample for RAND's Longitudinal Study of Schools: Principal Interviews and Teacher Surveys

	Design Team							
Jurisdiction	AC	AT	CON	EL	MRSH	NARE	RW	Total
Cincinnati			5	5			6	16
Dade	5		4	1	3		4	17
Kentucky						51		51
Memphis	5	5	5	5	4		9	33
Philadelphia		12	4		2			18
Pittsburgh						12		12
San Antonio				8	5			13
Washington state		8				16		24
Total	10	25	18	19	14	79	19	184

of 1997, most of these schools reported they were indeed implementing a design.[12] Yet, some were not. Figure A.1 shows that 25 of the 155 schools (about 15 percent) reported that they were in an exploratory year or a planning year with implementation expected in the future. About 85 percent (130/155) of the schools for which we had teacher, principal, and district data reported implementing a NAS design to some extent.[13]

Because our interest is in understanding the specific activities that are occurring within the 130 schools that were implementing a NAS design to some extent (the non-white areas of Figure A.1), we limited our analysis sample to these 130 schools.

In the spring of 1998, all 184 schools were once again surveyed. The completed sample size consisted of 142 implementing schools. However, the overlap between the 1997 and 1998 samples was incomplete. For purposes of this analysis, which is partly longitudinal in nature, we limited the analysis sample to schools that met two criteria:

- Schools were implementing in both 1997 and 1998; and
- Schools had complete data (i.e., from teachers and principals) in both years.

Of the 130 schools implementing in 1997 for which we had complete data, seven had either dropped the design or had reverted to planning, and another 17 had missing or incomplete data. Thus, 106 schools met both criteria. Figure A.2 shows the derivation of the sample.

---

[12]The first question we asked principals was about the status of the school's partnership with a NAS design. Principals could respond that they were in an exploratory year (i.e., the school has not committed to a design yet); in a planning year (the school has partnered with a design team and is planning for implementation next school year) in initial implementation for part of the school (i.e., a subset of the staff is implementing); continuing implementation for part of the school; in initial implementation for the whole school (i.e., all or most of the staff are working with the design); or continuing implementation for the whole school.

[13]These were schools that had complete principal data, at least five teachers responding to the teacher surveys, and complete district data.

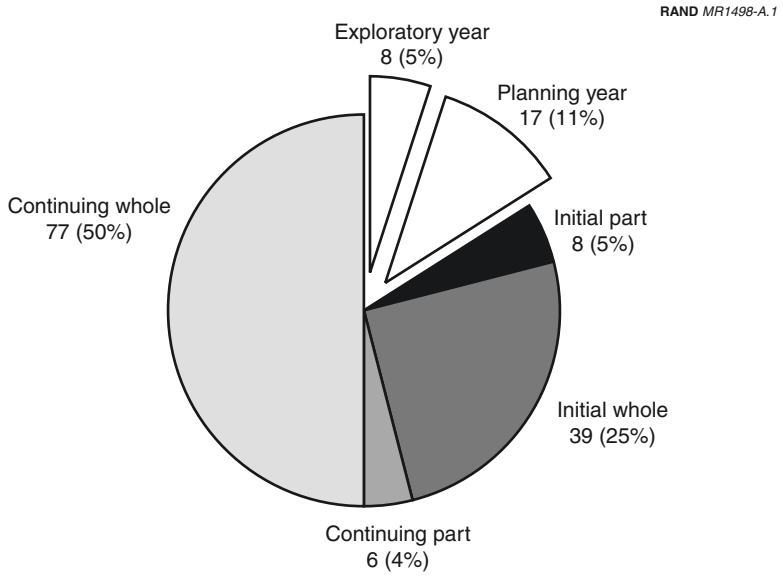
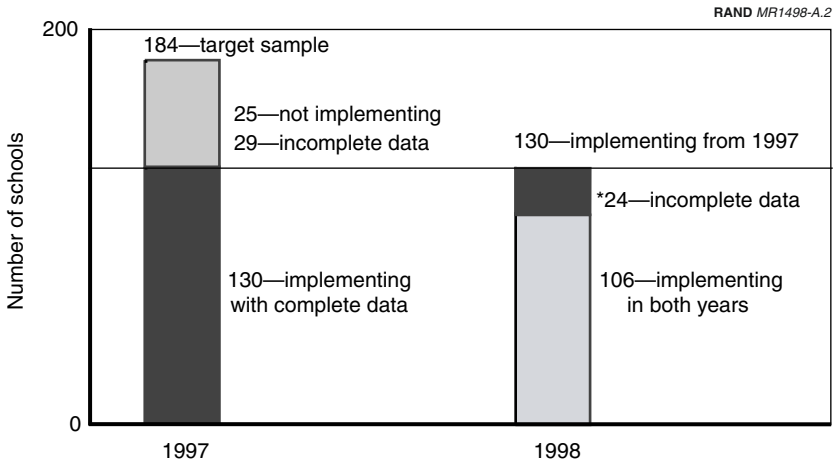


Figure A.1—Principal Reports of Implementation Status, Spring 1997



*6—dropped design, 1—reverted to planning, 17—missing principal or teacher data.

Figure A.2—Derivation of the Sample Analyzed in Berends and Kirby et al. (2001)

Of these 106 schools, there were two schools in Pittsburgh that we later discovered were not implementing and had dropped the design. In fact, throughout RAND's monitoring of the schools in Pittsburgh, there were severe budget crises. RAND's site visits and principal phone interviews consistently revealed that NAS implementation in Pittsburgh was not taking place (also see Bodilly, 1998). As a result, these two schools (and Pittsburgh) were excluded from the analysis; our final sample for the analysis for the implementation study three years after scale-up consisted of 104 schools across seven jurisdictions.

The average school size for the 104 schools was 662 students, but the standard deviation of 434 was quite large. Eighteen percent of the schools had fewer than 400 students. The distribution of the 104 schools across levels revealed that 64 percent were elementary schools, 14 percent middle schools, and 14 percent high schools. Eight percent were mixed levels.

Teachers in our sample were mostly female (84 percent), mostly white (68 percent), and the majority had a master's degree or above (52 percent). More than three-fifths were over 40 years old, mirroring the national teacher profile. Teachers also reported that, on average, they had been in their current school for seven years.

As of spring 1998, 40 percent of the 104 schools reported two years of implementation, 35 percent of the schools reported three years of implementation, and 25 percent of schools reported four years or more of implementation. More than half of the NARE schools reported four or more years of implementation.

The various criteria we used to define the sample all biased the sample to some extent in a positive direction in terms of expected implementation. RAND's sample of NAS sites is drawn initially from a set of NAS schools that expressed interest in implementing designs in districts that had formed a partnership with New American Schools. In addition, we chose schools where principals reported they were implementing the designs either partly or wholly for at least two years in 1998. This was done to ensure some degree of comparability across schools in terms of where they were in implementing designs. But, omitting schools that reported they were not implementing or had just started implementing in 1998 from the sample made our

analysis relatively more likely to find effects of designs on teaching and student achievement, where they existed.

The 1999 Final Analysis Sample for Examining Implementation Trends

Among the 104 schools that formed the longitudinal sample for the three-year scale-up study, we obtained complete data (i.e., from principals and at least five teachers in the school in 1999) on 71 schools. Principals in ten of the 104 schools reported that they had dropped the design but the attrition in the sample was largely due to nonresponse (13 schools were missing principal data as well as some teacher data; ten schools had fewer than five teachers responding to the survey). Thus, the analysis sample for the second study consisted of 71 schools in which principals reported that they were implementing designs in all three years (1997, 1998, and 1999) and which had complete data in all three years.

Table A.4 compares respondents with nonrespondents. In terms of jurisdictions, nonresponse was higher among schools in Washington state, Cincinnati, and Kentucky; in terms of design teams, nonresponse was higher in CON, EL, and NARE schools. Schools that had been implementing for three and five or more years in 1998 were disproportionately represented among the nonrespondents. Nonresponding schools tended to be less poor than responding schools, and to have lower proportions of minority students. However, as measured by the 1998 survey, these schools reported fairly similar levels of overall implementation of NAS designs as the responding schools and the within-school variability in reported implementation was the same. Despite the high attrition and somewhat differing characteristics of the nonrespondents, the patterns of implementation we found were remarkably similar to the findings of the earlier study (Berends and Kirby et al., 2001) based on the 104 schools.

The distribution of the 71 schools in the longitudinal sample by jurisdiction and design team is shown in Table A.5. In the longitudinal sample, a little over one-quarter of the schools were NARE schools, primarily located in Kentucky, while ATLAS, RW, and EL each accounted for 15–18 percent of the sample. AC, CON, and MRSH had

Table A.4

A Comparison of Respondents and Nonrespondents in the 1999 Longitudinal Sample, Based on 1998 Data

Selected Characteristics	Nonrespondents	Respondents
Number of schools		
Cincinnati	6	10
Dade	3	1
Kentucky	7	13
Memphis	5	24
Philadelphia	1	6
San Antonio	2	7
Washington state	9	10
Design team		
AC	1	4
AT	4	13
CON	6	6
EL	6	10
MRSH	0	7
NARE	13	19
RW	3	12
Years implementing in 1998		
2 years	11	31
3 years	13	22
4 years	4	13
5 years	5	5
Percentage		
Percent elementary schools	60.6	66.2
Mean percent students eligible for free/reduced-price lunch	58.5	66.3
Mean percent minority students	52.7	63.0
Total number of schools	33	71

the smallest number of schools in the longitudinal sample. All 71 schools had been implementing for three or more years by 1999. About 44 percent of the sample had been implementing for three years; a little over 30 percent for four years; and the remaining one-fourth of the sample for five years or more (most of these were NARE schools).

Table A.5

Distribution of the 1999 Longitudinal Sample, by Jurisdiction and Design Team

	Design Team							
Jurisdiction	AC	AT	CON	EL	MRSH	NARE	RW	Total
Cincinnati			3	2			5	10
Dade	1							1
Kentucky						13		13
Memphis	3	4	3	4	3		7	24
Philadelphia		5			1			6
San Antonio				4	3			7
Washington state		4				6		10
Total	4	13	6	10	7	19	12	71

Teacher Sample

The sample size of teachers who responded to the survey was approximately 1,700 in 1997, and 1,500 teachers in both 1998 and 1999. The average response rate among teachers in these schools has fallen over time in the 71 schools, from 73 percent in 1997 to 59 percent in 1999. The interquartile range for response rates, representing the middle 50 percent of the distribution, was 41–75 percent. Response rates were generally lower in 1998 compared with 1997, but response rates in 1999 were comparable with those of 1998 in most jurisdictions.

Measuring Implementation Within and Across Designs

Challenges of Constructing Indices to Measure Implementation. Measuring progress in implementation broadly across a wide set of schools in several partnering jurisdictions involved a number of challenges.

First, each design is unique. Attempting to develop a common set of indicators that measures implementation *across* designs is difficult, particularly when design teams adapt their programs to the local needs of the schools (Bodilly, 2001). However, despite their differences, design teams do aim to change some key conditions of schools in common ways, such as school organization, expectations

for student performance, professional development, instructional strategies, and parent involvement.[14] We attempted to draw on these commonalities to guide the construction of an index that could be used to broadly measure "core" implementation across designs.

Second, the difficulties of constructing indices that capture the key components of a design are compounded by the fact that these design components may themselves be evolving (see Bodilly, 2001). For example, design teams may change their implementation strategies because of lessons learned during development and implementation experiences in various sites.

Third, even if one developed measures on which there was general agreement that they fully captured the key facets of designs, the local context introduces a great deal of variability that must be taken into account (Bodilly, 1998; Bodilly and Berends, 1999). For example, while a design may focus on project-based learning over several weeks of the semester, this may be superseded by district-mandated curricula that take priority over significant portions of each school day.

Fourth, because the index is so general, it may be measuring more than just reform implementation.[15] Each of the components is a characteristic of effective schools, so schools may be pursuing these separately as school goals or as part of a district initiative. An increase in any one of these measures may not necessarily mean higher implementation of the model. For example, it may be that the design is helping the school to better attain these goals, or even that the school has been more successful in meeting this goal over time, independent of the model.

Fifth, it is important to note that all the implementation results are based on teachers' responses to surveys. The usefulness of what we can learn and infer from the analyses is heavily dependent on the quality of the data that are obtained from these surveys. In some in-

---

[14]With the recent support of the federal CSRD program, schools need to make sure that their plan covers these areas. If one particular design team or CSRD model does not cover these and several other areas of school improvement, then schools need to adopt more than one design or model (see Kirby et al., in review).

[15]We thank one of our reviewers, Amanda Datnow, for making this point.

stances, what we find has been validated by RAND's early case studies and other research (Bodilly, 1998; Ross et al., 1997; Datnow and Stringfield, 1997; Stringfield and Datnow, 1998), but for some indicators, all we have are teacher-reported survey measures.

Sixth, in the analysis sample of NAS schools that we examined, small sample sizes for some design teams made traditional tests of statistical significance somewhat more difficult to apply. That is, with larger sample sizes, we would have more power to detect differences and effects. Thus, in the school-level descriptive analyses, we focused on what appeared to be educationally substantive differences where appropriate.

Despite these challenges, evaluation remains an important component of any effort to change schools, and it is important to develop and refine sets of indicators that are informative not only for researchers, but for design teams, educators, and policymakers.

Implementation Indices. We developed two implementation indices:

1. A core implementation index that broadly measured implementation of the *major*, shared components of the designs across the sites; and
2. A design team–specific implementation index that measured implementation of both shared and some unique aspects of the designs.

The core implementation index was useful for understanding the progress of the NAS schools during the scale-up phase. The design team–specific index allowed us to measure implementation of each design on components that are unique to, and emphasized by, the design. The shortcoming of this index is that it was not directly comparable across designs, because it varied both in terms of items and number of items included in the index, and thus was not strictly comparable across design teams.

We should reiterate that this design team–specific index was not designed to measure *all* the unique aspects of the designs. Indeed, we could not construct such a measure with the available data, given that this was a broad study of NAS schools, not a detailed case study

of one particular design. As a result, the design team–specific index measures what we consider to be some of the key components of the designs.

Constructing a Core Implementation Index. The core implementation index is a summative scale of teacher responses as to the degree to which the following described their school (on a scale of 1–6, with 1 = does not describe my school, and 6 = clearly describes my school):[16]

- Teachers are continual learners and team members through professional development, common planning, and collaboration;
- Student assessments are explicitly linked to academic standards;
- Teachers develop and monitor student progress with personalized, individualized learning programs;
- Performance expectations are made explicit to students so that they can track their progress over time;
- Student grouping is fluid, multiage, or multiyear; and
- Parents and community members are involved in the educational program.

Teacher responses were averaged across a school to obtain the school mean level of implementation.

The professional life of teachers refers to the roles and relationships in which the teachers participate during the school day. In effect, when referring to restructuring schools, particularly those in poor, urban areas, this involves overhauling the conditions under which teachers work by changing their responsibilities and tasks and by developing a more professional culture in schools (Newmann et al., 1996; Murphy, 1992; Sykes, 1990; Wise, 1989). In contrast to teachers working in isolation without contact with their colleagues (see Louis and Miles, 1990; Lortie, 1970), design teams aim to build a collaborative environment for teachers. Thus, it is important to understand

[16]The alpha reliability of this index was 0.81. The range of correlations for the individual items was 0.21 to 0.57.

the extent to which teachers collaborate and engage in activities together, such as professional development, common planning time, and critiquing each other's instruction.

Each of the designs aims to bring all students to high standards, even though each may differ in the process to attain this goal. To monitor whether designs are making progress toward this end, critical indicators might include the degree to which (a) student assessments are explicitly linked to academic standards, (b) teachers make performance expectations explicit to students, and (c) the curriculum and performance standards are consistent and coherent across grade levels.

Most of the designs are concerned with shaping student experiences within classrooms to further their academic achievement growth. NAS designs embrace alternative instructional strategies that involve different relationships between teachers and students and between students and subject matter. Yet, again, each design differs somewhat in the specific nature of these activities. Conventional classrooms are often characterized as teachers talking *at* students and filling their heads with knowledge, with students responding with the correct answers at appropriate times (see Gamoran et al., 1995; Sizer, 1984; Powell, Farrar, and Cohen, 1985). In contrast, design teams tend to emphasize alternative instructional practices such as students working in small groups, using manipulatives, engaging in student-led discussions, or working on projects that span a long period of time (e.g., a marking period or semester).

The design teams also address a particular set of instructional strategies revolving around student grouping arrangements. How students are grouped for instruction and the effects of this on student achievement are subjects of heated debate among educators and researchers (see Slavin, 1987, 1990; Gamoran and Berends, 1987; Oakes, Gamoran, and Page, 1992; Hallinan, 1994; Oakes, 1994). Yet, most researchers agree that alternatives to inflexible grouping arrangements are worth further exploration. Thus, the NAS designs have experimented with such alternative student groupings. For example, students within an EL or CON design may have the same teacher for a couple of years. RW emphasizes flexible uses of grouping by organizing students according to their achievement levels in reading for part of the day and mixing achievement levels for other

subjects. These groupings are assessed every eight weeks or so to see if students would be better served by being placed in a different group. In short, each of the designs is sensitive to the issue of ability grouping and is working with schools to group students in more-effective ways.

Conventional wisdom suggests that the parent-child relationship and parent involvement in the child's education are critical components of school success. The NAS designs have embraced this issue as well. Several of the designs aim to have individuals or teams within the schools serve as resources to students and families to help integrate the provision of social services to them (e.g., ATLAS and RW). Other designs emphasize students applying their learning in ways that directly benefit the community (e.g., AC, EL, and NARE). Of course, each design desires that parents and community members be involved in positive ways in the educational program.

Table A.6 presents the means and standard deviations of the core implementation index across the 71 schools for 1997, 1998, and 1999.

Constructing the Design Team–Specific Implementation Index. As we mentioned above, designs vary in their focus and core components. As a result, we constructed a design team–specific implementation index that included the six core items of the core implementation index and items that were specific to each design team. Table A.7 lists the specific items included in the specific index constructed for each design team.

Again, the specific measures listed may not have captured all the unique features of the designs. Moreover, the wording of the survey items was more general to broadly compare schooling activities across design teams. Nonetheless, the design team–specific indices created here provide additional information about implementation of some of the unique features of the design teams. Such information was helpful for examining changes over time in the teacher-reported implementation, including changes in the means and variance within and between schools.

Berends and Kirby et al. (2001) reported results for both the core implementation index and the design team implementation index.

Table A.6

Means and Standard Deviations of the Core Implementation Index and Its Components, 1997–1999

	Mean 1997	Mean 1998	Mean 1999	Change, 1997–1999	SD 1997	SD 1998	SD 1999	Change, 1997–1999
Parents and community members are involved in the educational program	3.80	3.85	3.90	0.10	0.79	0.93	0.90	0.11
Student assessments are explicitly linked to academic standards	4.42	4.63	4.79	0.37	0.68	0.64	0.62	–0.06
Teachers develop and monitor student progress with personalized, individualized learning programs	4.01	4.14	4.19	0.18	0.68	0.64	0.62	–0.06
Student grouping is fluid, multiage, or multiyear	3.62	3.79	3.80	0.18	1.25	1.28	1.12	–0.13
Teachers are continual learners and team members through professional development, common planning, and collaboration	4.77	4.87	4.88	0.10	0.60	0.62	0.51	–0.09
Performance expectations are made explicit to students so they can track their progress over time	4.23	4.39	4.43	0.20	0.63	0.53	0.55	–0.08
Core Implementation Index	4.14	4.29	4.32	0.18	0.61	0.57	0.52	–0.08

Table A.7

Survey Items Included in the Design Team–Specific Implementation Index, by Design Team

Survey Items	AC	AT	CON	EL	MRSH	NARE	RW
Core Items							
Parents and community members are involved in the educational program	√	√	√	√	√	√	√
Student assessments are explicitly linked to academic standards	√	√	√	√	√	√	√
Teachers develop and monitor student progress with personalized, individualized learning programs	√	√	√	√	√	√	√
Student grouping is fluid, multiage, or multiyear	√	√	√	√	√	√	√
Teachers are continual learners and team members through professional development, common planning, and collaboration	√	√	√	√	√	√	√
Performance expectations are made explicit to students so they can track their progress over time	√	√	√	√	√	√	√
Design Team–Specific Items							
The scope and sequence of the curriculum is organized into semester- or year-long themes	√						
Students are required on a regular basis to apply their learning in ways that directly benefit the community	√						
Students frequently listen to speakers and go on field trips that specifically relate to the curriculum	√						
This school is part of a K–12 feeder pattern that provides integrated health and social services to improve student learning		√					
Students are required by this school to make formal presentations to exhibit what they have learned before they can progress to the next level		√					

Table A.7—continued

Survey Items	AC	AT	CON	EL	MRSH	NARE	RW
Consistent and coherent curriculum and performance standards have been established across the K–12 feeder patterns		√					
Most teachers in this school meet regularly with teachers in other schools to observe and discuss progress toward design team goals			√				
Technology is an integrated classroom resource			√		√		
Students engage in project-based learning for a significant portion of the school day (i.e., more than one-third of the time)			√				
Technology is used in this school to manage curriculum, instruction, and student progress			√		√		
A majority of teachers in this school stay with the same group of students for more than one year				√			
Students frequently revise their work toward an exemplary final product				√			
There are formal arrangements within this school providing opportunities for teachers to discuss and critique their instruction with each other				√			
This school has the authority to make budget, staffing, and program decisions					√		
Curriculum throughout this school emphasizes preparation for and relevance to the world of work						√	
Students are monitored according to annual performance targets established by the school as a whole						√	
Student assessments are used to reassign students to instructional groups on a frequent and regular basis							√

Table A.7—continued

Survey Items	AC	AT	CON	EL	MRSH	NARE	RW
Students are organized into instructional groups using block scheduling for specific curricular purposes							√
This school has specific activities aimed directly at reducing student absenteeism							√
Students who are not progressing according to expectations are provided with extended days and/or tutors							√
This school has a coordinator, facilitator, or resource specialist assigned on a full- or part-time basis							√
Alpha Reliability Index	0.83	0.80	0.87	0.88	0.90	0.85	0.87

Kirby, Berends, and Naftel (2001) reported results for the core implementation index only because the findings were similar across the two indices.

Measuring Performance in NAS Schools

Monitoring Academic Progress with School-Level Test Scores. As previously stated, because of resource constraints, jurisdictions' hesitancy to have additional testing, and established agreements between NAS and the partner jurisdictions, it was not feasible in RAND's evaluation of NAS to administer a supplemental, common test to the students within the participating schools. Thus, 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 with 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 limits 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 have 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 may 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).

Measure of School Performance. The analyses of performance in NAS schools focused on one main research question—Did NAS schools make gains in test scores relative to all schools in their respective jurisdictions?

To answer this question, we collected 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, data were generally aggregated across NAS schools, using grade enrollment as weights. Thus, we compared NAS schools with the district or the state. However, in a couple of cases, the test score information did not lend itself to aggregation. In these cases, we provided trends for each NAS school in the sample.

The comparison we made between NAS schools and the district averages used 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. Moreover, 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 were not substantially different from those obtained using absolute gains.

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. 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 focused 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 collected data for the baseline and baseline plus two years. For some late implementing schools, we were only able to get baseline and baseline plus one year data. (See Berends and Kirby et al. [2001] for more detail on each of the tests used by the various jurisdictions.)

Earlier, we showed that the NAS schools in this sample were predominantly high poverty and high minority, and many faced challenges related to student mobility.[17] 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

[17]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.

student learning for substantial numbers of students. Because of these expectations, NAS requested that RAND examine the progress of these NAS schools in comparison with the district averages to understand whether NAS's 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 104 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 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 were no testing data available and other schools were missing data on test scores.

CASE STUDIES FIVE YEARS AFTER SCALE-UP

In order to better understand the relationship between implementation and performance, we conducted a case study of matched schools, matched on the basis of design, district, grade span, years of implementation, and implementation level (as measured by our surveys but validated by the design teams). One school was high-performing and the other was not. Although we attempted to get a total of 20 schools, only 13 schools participated in the study (five matched pairs and a triplet): two ATLAS, two CON, five MRSH, and four RW schools. One to two researchers spent a day at each school conducting interviews with principals, groups of teachers, and district officials. We collected data from the design teams about the schools as well as data from the district and the schools themselves on student test scores; demographic and program descriptors; other school programs and interventions; level of implementation; district support of the design; and perceptions about the causes of different levels of performance increase.

METHODOLOGY FOR SAN ANTONIO CLASSROOM STUDY (BERENDS ET AL., 2002)

The San Antonio district has over 90 percent of its students eligible for free/reduced-price lunch; most of the students in the district are either Hispanic (85 percent) or African American (10 percent); and approximately 16 percent of the students in the district are classified as having limited English proficiency. Since 1994, the proportion of San Antonio students failing to earn passing rates on the TAAS in each school year has consistently been the highest or second highest in the county.

It is within this context of high poverty and low student performance that elementary schools in San Antonio began the process of adopting NAS reform models. Of the 64 elementary schools in the district, three schools began implementation during the 1995–1996 school year, nine schools the following year, and 20 schools during the 1997–1998 school year. By the 1998–1999 school year, 39 of 64 elementary schools in the district had adopted NAS designs. Table A.8 lists the number of schools adopting specific designs in each year.

RAND collected data on a sample of 4th-grade teachers and their students during two school years, 1997–1998 and 1998–1999 (see Tables A.9 and A.10). Fourth grade was an advantageous selection for several reasons: most NAS designs were being implemented in elementary schools; the state administered its test to students in the 3rd grade, providing a baseline for test score analysis; and teacher questionnaire items were already developed and tested with 4th-grade teachers. In addition, the school district expressed its preference for a grade four focus.

Table A.8

Elementary Schools Adopting NAS Designs in San Antonio, by Year

	Number of Schools				
	1995–1996	1996–1997	1997–1998	1998–1999	Total
CON			3	1	4
EL	1	2			3
MRSH		5	4	1	10
RW	2	2	13	5	22
NAS Total	3	9	20	7	39

Table A.9

Target Sample of Schools Compared with Final Study Sample, by Type of Data Collection and NAS Design Team

Number of Schools in 1997–1998 School Year					
	Requested to Participate	Returned Teacher Surveys	Returned Principal Surveys	Returned Stanford-9 Testing	Classroom Observations
CON	2	2	2	2	1
EL	2	2	1	1	1
MRSH	4	4	4	4	2
RW	8	8	8	9	1
Non-NAS	10	8	9	10	2
Total	26	24	24	26	7
Number of Schools in 1998–1999 School Year					
	Requested to Participate	Returned Teacher Surveys	Returned Principal Surveys	Returned Stanford-9 Testing	Classroom Observations
CON	2	2	2	2	2
EL	2	2	1	2	2
MRSH	4	4	2	4	2
RW	8	8	7	8	2
Non-NAS	7	7	7	7	2
Total	23	23	19	23	10

Generally, in each school year we were able to gather teacher survey data and supplemental student test scores in reading (Stanford-9), including over 850 students in over 60 classrooms in over 20 schools. Moreover, during the course of this study, we were able to obtain information on all the teachers and students in the district to provide a benchmark for the analyses reported here. In 1997–1998, we were also able to observe and gather classroom artifacts from 12 teachers in NAS and non-NAS schools. In the following year, we gathered such data from 19 teachers. Each of these data collection efforts is described more fully in the sections that follow.

The assistant superintendent's office demonstrated its support for our study by asking principals to announce the study to their staff and to invite all 4th-grade teachers to participate in the study. Once the initial volunteers were reported, RAND attempted to balance the

Table A.10

Target Sample of Teachers Compared with Final Study Sample, by Type of Data Collection and NAS Design Team

Number of Teachers in 1997–1998 School Year			
	Requested to Participate	Returned Surveys & Stanford-9 Testing	Observations
CON	6	6	3
EL	4	2	2
MRSH	12	10	3
RW	26	22	2
Non-NAS	26	23	2
Total	74	63	12
Number of Teachers in 1998–1999 School Year			
	Requested to Participate	Returned Surveys & Stanford-9 Testing	Observations
CON	11	10	4
EL	8	6	5
MRSH	13	11	3
RW	32	27	4
Non-NAS	19	19	3
Total	83	73	19

representation of designs in the sample by approaching schools of underrepresented designs. While the RAND sample of NAS and non-NAS schools cannot be considered random, district staff indicated that the schools selected were typical of elementary schools in the district. Comparisons of demographic and other characteristics for students (i.e., gender, race, limited English proficiency status, special education status, average test scores, and mobility rates) and teachers (i.e., gender, race, highest degree earned, years of teaching experience) indicated no significant differences, on average, between the RAND sample and district populations. Each teacher selected was asked to administer the Stanford-9 to his or her 4th-grade students and to complete a teacher survey. Teacher focus groups were conducted in eight schools during the 1997–1998 school year. A subset of teachers agreed to provide classroom logs, and samples of student work and allowed classroom observations once in the spring of the 1997–1998 school year and three times in the 1998–1999 school year. In addition, principals in the sample schools were asked to complete a telephone interview, during which a survey was completed.

Teacher Data

In the late spring of the 1997–1998 school year, with the help of district staff, we contacted 74 teachers in 26 schools to participate in the study. Three of the schools refused to participate in our study. Of those 74 teachers initially contacted, 63 teachers in 23 schools agreed to participate, returned completed teacher surveys, and their students completed the Standford-9 reading test resulting in an 85 percent response rate for teachers and classes with student achievement scores.

In 1998–1999, we returned to the 23 schools that participated in our study the previous year. Because we wanted to increase our sample of teachers, we supplemented our teacher sample and contacted 83 teachers in these 23 schools. Of those contacted, we received completed teacher surveys and Stanford-9 tests from 73 teachers (88 percent). Between spring 1998 and spring 1999, one of our sampled schools went from having no design in place to adopting RW.

Not all teachers had complete survey data across both years, given that different teachers were included in both years. Thus, for the longitudinal descriptions of NAS and non-NAS classrooms, we tracked indicators for the 40 teachers for whom we had complete data from both the 1997–1998 and 1998–1999 school years (see Table A.11). In addition to these teacher data from RAND surveys, we also obtained information on teachers from the district, such as demographic characteristics (race-ethnicity and gender), years of experience, and highest degree obtained.

The analysis only included the sample of 40 4th-grade teachers who completed surveys in both the spring of 1998 and 1999 and re-

Table A.11

Longitudinal Sample of Teachers in NAS and Non-NAS Schools, 1997–1998 and 1998–1999 School Years

	CON	EL	MRSH	RW	Non-NAS	Totals
Number of Schools	2	2	3	6	7	20
Number of Teachers[a]	4	3	8	11	14	40

[a]Teachers who completed the survey in both spring 1998 and spring 1999 and who were in the same school, same design, and teaching 4th grade in both years.

mained in the same school/design/teaching assignment. This was because our interest lay in examining what changes, if any, occurred during the early stages of implementation in school organization, teachers' professional work lives, and their classroom instruction.

We also compiled survey results from the larger sample (66 teachers in 1998 and 83 in 1999). A comparison of average response rates found few differences between the two samples. A detailed analysis of individual teacher responses found no substantive differences between these larger samples and what we find in the longitudinal teacher sample of 40 teachers.

Because of the small size of the longitudinal sample analyzed, we did not focus much attention on testing the statistically significant differences between NAS and non-NAS teachers. Given the design, most standard statistical tests comparing the 40 NAS and non-NAS teachers in the longitudinal sample would fail to detect many real differences. However, in conjunction with the qualitative data from this study, the NAS and non-NAS comparisons shed light on a variety of factors related to implementing NAS designs in a high-poverty urban district.

Surveys. The teacher survey fielded during the spring 1998 semester and then again in spring 1999 was designed to provide a broad measure of instructional practices in NAS and non-NAS classrooms. Teachers were asked to report on a range of instructional strategies, some of which reflected a focus on basic skills and tended toward more conventional practices, and others of which reflected more reform-like methods. Given that the NAS designs emphasize changes in instructional conditions whether through building basic skills and then higher-order thinking (e.g., RW) or through theme-based projects that last for several weeks (e.g., CON or EL) (see Bodilly, 2001), we would expect the implementation of designs to result in changes in teaching strategies.

General topics covered in the survey include school and classroom characteristics, instructional strategies and materials, skills and assessments emphasized, resources, parent involvement and community relations, impact of design team and reform efforts, professional development, and perceptions and attitudes toward teaching.

Two versions of the survey were fielded in each year, one to 4th-grade teachers in a sample of schools adopting NAS designs, the other to 4th-grade teachers in non-NAS schools. The two forms of the surveys varied only slightly. For instance, three items specifically related to the implementation of NAS designs were not included in the survey received by non-NAS teachers. A few items in other sections also referred specifically to NAS designs. On the non-NAS version, these items were either omitted or had slightly different wording (e.g., whereas NAS teachers were asked about the NAS design being implemented in their school, non-NAS teachers were asked about the school reform efforts in their district). For example, an item on the NAS version that asked if an activity was "specifically oriented toward the design team program activities" was changed to "specifically oriented toward the reform efforts of San Antonio" on the non-NAS version.

These surveys were developed in conjunction with RAND's ongoing case study work (Bodilly, 1998). As part of our overall instrument development, we conducted phone interviews with design team representatives about what specific indicators and events would be observed in design-based classrooms. For the survey development, we also relied on other studies that have examined instruction with surveys (Newmann et al., 1996; Gamoran et al., 1995; Burstein et al., 1995; Porter, 1995; Porter and Smithson, 1995; Porter et al., 1993; see also Mayer, 1999).

Longitudinal Sample of 40 Teachers Compared with Elementary Teachers in District. Overall, it appears that in demographic terms, the longitudinal survey sample of 40 teachers was a fairly representative group of teachers within the school district. There were few differences when comparing teachers in our sample with all 4th-grade teachers in the San Antonio school district (Table A.12). Teachers in this sample and the district as a whole were similar with respect to gender, racial-ethnic characteristics, and average years of experience. Whereas 40 percent of teachers in the district had earned master's degrees, 45 percent of the teachers in the longitudinal sample had attained this level of education.

Table A.12

Teacher Characteristics—Districtwide Versus RAND Survey Sample, 1997–1998 School Year

	District (n = 329)	Survey Sample (n = 40)
Male	11%	8%
With master's degrees	40%	45%
Average years teaching experience	13	13
White	37%	33%
African American	15%	20%
Latino/Latina	47%	47%
Asian American	0.3%	None
Native American	0.3%	None

Observations and Logs of Instructional Activities.[18] In the spring of the 1997–1998 school year, RAND conducted classroom observations of a subsample of 12 teachers from the larger group of 64. These observations consisted of a RAND researcher shadowing a teacher for a day, writing detailed summaries of classroom activities, taking notes on the general description of the classroom and the resources in it, and informally discussing design team activities with the teacher.

School observations first began in the spring of 1998 and continued throughout the 1998–1999 school year. Observations, targeting the 4th-grade level, covered ten different schools. Data were collected in two CON, two EL, two MRSH, two RW, and two non-NAS schools. In the first year of our study, in addition to observations, we aimed to gather more-extensive classroom data through (1) teacher logs of assignments, homework, projects, quizzes/tests/exams, and papers or reports over a five-week period, and (2) illustrative teacher-selected samples of student work related to a major project assigned during the spring semester. Because we could not gain entry into these classrooms until May, right after the administration of TAAS, and because our logs were overly burdensome, the response rate for these 12 teachers was less than desirable. Five of 12 teachers (42 percent) returned completed logs.

[18]Each teacher who participated in this part of the study received a $200 honorarium.

Therefore in the second year, we significantly revamped our data collection methods for observations and logs of instructional activities. Teachers were not asked to submit logs of assignments. Rather, arrangements were made to observe 19 teachers across ten different schools—two CON, two EL, two MRSH, two RW, and two non-NAS schools—on three separate occasions. Moreover, a staff person on site in San Antonio interviewed them at length over the course of one school year. In addition, teachers provided work assignments, lesson plans, and even district memos when appropriate.

Interviews. In the spring of 1998, we conducted focus group interviews with 4th-grade teachers from eight different schools, including schools implementing each of the four NAS designs and some comparison schools. Our aim was to get a representation of teachers within NAS schools to provide information about what activities were undertaken across grade levels. These interviews were conducted to help us better understand design team program characteristics, the nature of instructional strategies, the variety of professional development activities, and the types of available classroom-level resources. Additional information about these schools, professional development activities, and the resources available for design implementation was provided by 45-minute structured interviews with principals.

During the 1998–1999 school year, after each observation, teachers were interviewed about what occurred during the observation as well as about other more-general issues pertaining to design implementation, instructional strategies, professional development, and other matters related to design and district initiatives.

In addition, we conducted interviews of NAS design team leaders, district staff, school instructional leaders, and principals.

Student Data

Data for individual students were obtained mainly through the cooperation of the central office staff, who provided district files on students to RAND for analysis.

Student Achievement. In this study, student achievement was measured in a variety of ways. First, we asked teachers to administer the

Stanford-9 open-ended reading test. We decided to use the Stanford-9 because, as a commercial test that could be nationally normed, it differed somewhat from conventional multiple-choice tests. The Stanford-9 requires students to use an open-ended response format. The test takes about 50 minutes to administer.

In addition, RAND obtained the TAAS mathematics and reading scores for all of the district's 3rd-, 4th-, and 5th-grade students during the time of this study. Our focus was mainly on the 1997–1998 4th-grade cohort. Not only did we track their achievement back to when they were third graders, but we also obtained their scores from the 5th grade to examine achievement growth. Specifically, we analyzed the TAAS mathematics and reading Texas Learning Indices (TLI). These data were linked to teachers and schools in our survey sample. They allowed us to examine achievement across schools and classrooms for the entire district in addition to the RAND sample that included teacher surveys and Stanford-9 tests.

Student Characteristics. Other information available for individual students from district data files included student race-ethnicity, gender, date of birth, poverty status (economically disadvantaged or not), number of weeks the student was in the school during the academic year, limited English proficiency status, and participation in Special Education or Talented and Gifted programs.

Examples of Student Work. The teachers we observed in the 1998–1999 school year were asked to provide examples of students' work. We randomly selected one-quarter of the students in each class every three months. Once a student was selected, his or her name was removed from the class roster. While no criteria were established with regard to what was submitted, we asked teachers to provide examples of typical work assignments that students produced.

We cannot claim that the submitted work was representative of all student assignments made by a given teacher. However, these examples did provide a glimpse of the types of activities assigned by each of the teachers in our sample.