

Using School Reform Models to Improve Reading Achievement: A Longitudinal Study of Direct Instruction and Success For All in an Urban District

Steven M. Ross

*Center for Research in Educational Policy
The University of Memphis*

John A. Nunnery

*Department of Educational Leadership and Counseling
Old Dominion University*

Elizabeth Goldfeder

Aaron McDonald

*Center for Research in Educational Policy
The University of Memphis*

Robert Rachor

Toledo Public Schools

Matthew Hornbeck

Hampstead Hill Elementary School, Baltimore, MD

Steve Fleischman

American Institutes for Research

This research examined the effectiveness in an urban school district of 2 of the most widely used Comprehensive School Reform (CSR) programs—Direct Instruction (DI), implemented in 9 district elementary schools, and Success for All (SFA), implemented in 2 elementary schools. In examining impacts on student achievement and school change outcomes (e.g., teacher buy-in, school climate), a mixed-method re-

search design was employed, encompassing both quantitative and qualitative analyses. Student achievement results on the reading sections of the Ohio Proficiency Test and the Stanford Achievement Test showed that both DI and SFA schools performed comparably to other district schools after statistically adjusting for school and student variables. Qualitative measures indicated generally positive support for both models by teachers, principals, and parents. However, in the case of DI, findings indicated weaknesses in implementation due largely to uncertainties involving school versus district roles and inadequate training. Results are discussed with regard to the influences of contextual and implementation variables on judging CSR model effectiveness in general and for specific schools and districts.

With the passage in 2002 of the federal No Child Left Behind legislation, increased focus has developed for identifying "proven" practices that demonstrably raise student achievement. Federal initiatives, such as the Comprehensive School Reform (CSR) program (U.S. Department of Education, 1999) and funding formulas under NCLB, were intended to facilitate restructuring efforts to replace traditional, unsuccessful strategies with scientifically based practices that raise student achievement. Initiated in 1999–2000, Comprehensive School Reform Demonstration (CSR) funded more than 1,800 schools nationally with a minimum of \$50,000 each year for 3 years (Hatch, 2001). To date, there are over 380 models that have been adopted under CSR support (Desimone, 2002). Most of these models attempt to implement whole-school restructuring by reforming curriculum, instruction, and organizational structures (Datnow & Stringfield, 2000). Among the best-known and most frequently used externally developed designs are Success for All (SFA; Slavin & Madden, 2001), Accelerated Schools (Hopfenberg & Levin, 1993), the School Development Program (Haynes, 1998), and Direct Instruction (DI; Englemann, Becker, Carnine, & Gersten, 1988). These and other widely disseminated CSR models (Borman, Hewes, Overman, & Brown, 2003) were developed to be systematically replicated so as to produce comparable implementations and benefits for multiple schools across the country.

Despite the long history and recent proliferation of CSR designs, there is surprisingly limited rigorous, scientific, or independent evidence on their effectiveness in either implementation quality or, most critically, raising student achievement. According to Borman and colleagues (2003), prior to their recent meta-analysis of 29 popular models, there have been only five major practitioner-oriented reviews or catalogs (see Herman, 1999; Northwest Regional Educational Laboratory, 1998, 2000; Slavin & Fashola, 1998; Wang, Haertel, & Walberg, 1997). Taken as a whole, these reviews show the quantity of rigorous research performed on CSR models to be extremely limited. Thus, little evidence exists from which to make educational decisions about model selection. In Herman's (1999) frequently cited review, for example, only three out of 24 models reviewed—DI, High Schools That Work, and SFA—were judged as having "strong evidence" of positive effects. Notably, for eight of the models (33%), not enough research evidence was available to make a judgment.

In their meta-analysis, Borman and colleagues (2003) found an overall achievement advantage ranging from one-tenth to one-seventh of a standard deviation for the CSR models examined. Models classified as having the strongest evidence of yielding benefits were DI, SFA, and the School Development Program, with effect sizes in the +0.15 to +0.21 range. Discouragingly, 17 out of the 29 models were classified as having insufficient statistically reliable or generalizable results to judge effects.

Even where seemingly rigorous studies have been performed, the determination of educational program effects in school settings may be influenced by many extraneous variables (Berliner, 2002). One such variable is potential bias due to design factors that might favor the experimental over the control program or due to the involvement of developers in researching their own models (e.g., Pogrow, 1998, 2000). A second factor is evidence becoming dated as a result of a model undergoing changes over time or being implemented in schools affected by different national or local policies than existed in the past.¹ A third factor is the many contextual variables that influence how a program or model is perceived by school staff, integrated with administrative structures and other initiatives, implemented by individual teachers, and sustained over time (Cuban, 1993; Datnow & Stringfield, 2000; Fullan, 2000). In view of these considerations, it is not surprising that a given CSR model can have positive effects at one school but fail to succeed at a similar school in the same geographic area and school district (see Berends, Chun, Schuyler, Stockley, & Briggs, 2002; Ross, Sanders, Wright, Stringfield, Wang, & Alberg, 2001).

Complementing educators' interests in identifying what works in schools is the current national demand, spurred by the reauthorization of the Elementary and Secondary Education Act (U.S. Congress, 2002), for increased rigor in educational research (see Feuer, Towne, & Shavelson, 2002). This study was designed to address the need to expand the evidence base from rigorous research on CSR models by examining the longitudinal educational impacts in an urban school district of two of the most widely used models—DI (e.g., Englemann, 1969) and SFA (Slavin & Madden, 2001). An assumption supporting the rationale for this and similar field studies was that, due to their complexity and susceptibility to intervening variables, educational programs do not have fixed effects on student achievement that will necessarily be duplicated from site to site (Berliner, 2002; Ross et al., 2001). Multi-site replicated studies (Slavin & Madden, 1993) thus add to the literature base on overall program effects, while identifying the particular contexts and variables associated with higher or lower success.

A second, related assumption was that a mixed-method research design (Onwuegbuzie & Teddlie, 2003), which combines rigorous quantitative analyses

¹For example, research on Direct Instruction examining implementations of DISTAR in the 1970s (e.g., Kaufman, 1972, 1973, 1974) would seem less valid for informing practices today than would recent studies conducted in CSR contexts (e.g., Mac Iver, Kemper, & Stringfield, 2003).

with qualitative inquiry to describe influential processes and contextual factors (Feuer et al., 2002), has strong advantages over using either a quantitative or qualitative approach alone. Variables such as school climate, teacher support, quality of professional development, adequacy of resources, political events, participant characteristics, and administrative policies typically help to explain why the program was more or less difficult to implement in the target context. Such variables also help to examine the viability of theoretical assumptions for why given reforms may work to raise achievement in certain situations but not in others. Datnow, Hubbard, and Mehan (2002), for example, viewed school reform as a co-constructed process, shaped by dynamics that occur in diverse contexts, such as the classroom, school, district, design team, state, and federal levels. Thus, the reform model that worked well when first developed and implemented by the developers may have quite different impacts after being funneled through the interactive influences of the multiple stakeholders at each new site. Alternatively, Desimone (2002) and Porter (1994) interpreted the success of CSR in terms of model components (i.e., specificity), compatibility with existing reform efforts (consistency), political and control factors in the district (authority and power), and the sufficiency of the time available to invoke change (stability). According to both of these frameworks, the success of the reforms examined in this study should be shaped by interactions identified between pervasive district influences and idiosyncratic model-specific and school variables.

A third assumption was that acquiring scientifically valid evidence about educational programs or models not only depends on the rigor of the research design employed but also on the availability of appropriate data from schools. An especially apt structure for conducting school-based research can be created through a partnership between individual school districts, which are interested in obtaining research results on the program of interest, and external research groups, which are interested in conducting such studies while offering the expertise and objectivity of third-party researchers (Ross, 2000). This type of collaboration, involving Toledo Public Schools (TPS) and this research team, formed the basis for this district-wide impact evaluation of DI and SFA.

STUDY CONTEXT AND REFORM MODELS

During a 4-year period prior to this study, the district attempted to revitalize its lowest performing schools by making DI and SFA available to nine urban elementary schools serving largely disadvantaged student populations. A brief review of each model is provided in the sections below.

Direct Instruction (DI)

DI is grounded in Sigfried Englemann's theory that learning can be accelerated and improved if teacher-delivered presentations are clear, fast-paced, and highly

interactive for learners (Englemann, 1969; Englemann, Becker, Carnine, & Gersten, 1988). The initial DI model was published by Science Research Associates under the trade name Direct Instruction System for Teaching Reading and Arithmetic (DISTAR), and focused on reading, language, and math. Presently, the model has expanded to other subjects (e.g., science and handwriting), and includes prescribed curricula and teaching materials, and extensive professional development to teachers and principals. Instructional components include scripted lessons, cross-grade and between-class ability grouping, and frequent assessment of mastery.² DI is currently being used in over 300 schools nationally.

Research on DI has generally been supportive (see reviews by Adams & Englemann, 1996; Borman, 2002; Herman, 1999) of its effects on student achievement. In Borman and colleagues' (2003) meta-analysis of CSR research, DI was one of only three out of 29 models classified as demonstrating the strongest level of effectiveness based on the number of studies, generalizability of the findings, and the magnitude of the effect size from suitably rigorous studies. Across studies, DI was associated with effect sizes of $d = +0.21$ for the overall sample, and $d = +0.15$ for comparison and third-party studies only. The vast majority of the studies, however, were conducted in the 1980s and earlier. More recent outcomes have been mixed, with statistically significant advantages found for DI over control schools in Houston (Carlson & Francis, 2002) and Baltimore (Mac Iver, Kemper, & Stringfield, 2003) but no effects in an additional Baltimore study of overall (not continuously enrolled only) student populations (Butler, 2003).

Success for All (SFA)

SFA was developed in the late 1980s by Robert Slavin and his associates at Johns Hopkins University (Slavin & Madden, 2001). The overall goal is to enable every child in participating schools to read at grade level by the end of Grade 3. SFA emphasizes strategies for early intervention and prevention of reading failure. Key components of SFA consist of: (a) a research-based reading program comprised of Reading Roots in Grades K-1 and Reading Wings in higher grades; (b) a strong emphasis on developing both phonemic awareness and comprehension skills; (c) individual tutoring by certified teachers for students most in need; (d) regrouping of students so that ability-grouped multi-age classes are established for a daily 90-min reading block; (e) a family support team to bolster attendance and parent involvement; (f) a full-time facilitator; and (g) extensive, ongoing professional development.

Research evidence on SFA is considered strong relative to other CSR models based on the large number of studies and the consistency of the findings (Herman, 1999). The typical SFA study involves matching SFA schools to demographically

²For more information, see <http://www.nwrel.org/scpd/catalog/>.

similar control schools in the same district (Slavin & Madden, 2001). In Borman and colleagues' (2003) meta-analysis, SFA was classified along with DI and the School Development Program (Haynes, 1998) in the select "strongest evidence of effectiveness" category. Results showed a statistically significant overall effect size for SFA of $d = +0.18$, and a significant but lower effect size of $d = +0.11$ for third-party comparison studies only. Critics of CSR have called for more studies to be conducted by third-party researchers who are independent of the programs being evaluated (Pogrow, 1998, 2000). Among the most recent SFA studies is a third-party investigation conducted in Louisville, Kentucky (Muñoz & Dossett, 2004; Muñoz, Dossett, & Judy-Gullans, 2003). Using a matched-control design with three SFA schools and three control schools, the researchers found statistically significant positive effects for SFA schools on CTBS reading scores, attendance, and reduction of suspensions.

RESEARCH PURPOSE AND QUESTIONS

The primary purpose of this research was to provide a rigorous, third-party study of the success of reform efforts in impacting positive school change and student achievement in the nine schools implementing DI and SFA. Specific research questions were:

1. What were program (DI and SFA) impacts on student achievement outcomes over time?
2. What were program impacts on school climate?
3. What school and district variables appeared to influence the effectiveness of the programs and the schools' ability to implement the programs effectively?

METHOD

Sampling

Eight schools participated in the portion of the study focusing on student achievement, including three DI schools that began implementation in 1997–1998, three DI schools that began implementation in 1999–2000, and two SFA schools that began implementation in 1999–2000. For the qualitative study, an additional DI school, in only its first year of implementation, participated. The DI schools served a predominantly African American, high-poverty student population (see Table 1). The SFA schools had substantially lower poverty rates (around 40%), with SFA-1 serving a diverse population (42% African American, 36% White, and 22% other ethnicities) and SFA-2 serving a predominately White population.

The comparison sample for the student achievement analysis consisted of all other district elementary schools ($n = 36$), the scores for which were adjusted for school and student characteristics. For the qualitative analyses, experts in the school district, in collaboration with the researchers, selected matched Control schools for DI and SFA based on prior achievement, socioeconomic status (SES) factors, and ethnicity. There were six DI Control schools and two SFA Control schools.

Data Sources and Instrumentation

DI, SFA, and Control programs were all assessed using the instruments described below, with the exceptions of the Control schools not having (a) targeted observations, and (b) facilitator interviews.

Achievement analyses. Both student-level and school-level analyses were conducted using a two-level hierarchical linear model (HLM). Average program effects for DI and SFA on (a) Stanford 9 Total Reading in Grade 2 and (b) reading scores on the Ohio Proficiency Test in Grades 4 and 6 were estimated. Inferential tests, adjusting for prior achievement and poverty, were run to determine whether students attending schools using either program scored significantly higher than students attending other district schools. Similarly, the relationship between pretest and posttest scores within each school (i.e., the pretest-posttest regression slope) was estimated and tested to determine variations as a function of program type. Average effects and effects on pretest-posttest slopes were estimated and analyzed using a two-level hierarchical linear model for each grade level (2, 4, and 6).

TABLE 1
Participant School Demographics, 2000–2001 School Year

Program/School	Enrollment	Free Lunch (%)	African American (%)	White (%)
1997–1998 DI schools				
DI-1	475	92.6	94.7	2.9
DI-2	369	88.4	95.9	2.4
DI-3	394	63.1	91.4	6.1
1999–2000 DI schools				
DI-4	495	95.9	97.2	1.0
DI-5	363	51.6	97.2	1.4
DI-6	262	70.7	96.9	0.4
1999–2000 SFA schools				
SFA-1	302	40.1	42.4	36.4
SFA-2	863	45.6	10.7	75.9
Overall District Mdns	416	71.9	29.8	53.1

Note. The school year in the Program/School column represents the first year of comprehensive school reform model implementation. DI = Direct Instruction; SFA = Success for All; Mdns = Median scores.

for 3 school years (1998–1999, 1999–2000, and 2000–2001). Further description of the analytical design is provided in the Results section.

School Climate Inventory (SCI). The main purpose of the School Climate Inventory (SCI) is to assess teacher perceptions of impacts of reform initiatives in relation to seven dimensions logically and empirically linked with factors associated with effective school organizational climates (Butler & Alberg, 1991). The inventory contains 49 items, with 7 items comprising each scale. Responses are scored through the use of Likert-type ratings ranging from 1 (*strong disagreement*) to 5 (*strong agreement*). Each scale yields scores ranging from 1 to 5, with higher scores being more positive. Additional items solicit basic demographic information.

Face validity of the school climate items and logical ordering of the items by dimensions were established during the development of the inventory (Butler & Alberg, 1991). Subsequent analysis of responses, including a concurrent validity study by Sterbinsky (2001), collected through administration of the inventory in a variety of school sites, substantiated validity of the items and scales. Dimension descriptions and current internal reliability coefficients on the seven dimensions of the inventory, obtained using Cronbach's alpha, are as follows: *Order*—the extent to which the environment is ordered and appropriate student behaviors are present ($\alpha = .84$); *Leadership*—the extent to which the administration provides instructional leadership ($\alpha = .83$); *Environment*—the extent to which positive learning environments exist ($\alpha = .81$); *Involvement*—the extent to which parents and the community are involved in the school ($\alpha = .76$); *Instruction*—the extent to which the instructional program is well developed and implemented ($\alpha = .75$); *Expectations*—the extent to which students are expected to learn and be responsible ($\alpha = .73$); and *Collaboration*—the extent to which the administration, faculty, and students cooperate and participate in problem solving ($\alpha = .74$).

Reading Teacher Survey (RTS). All teachers who taught reading classes at each DI, SFA, and Control school were asked to complete the RTS, which contains 20 items to which teachers respond using a five-point Likert-type scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The survey was adapted from the Comprehensive School Reform Teacher Questionnaire (CSRTQ), a validated 28-item instrument (Sterbinsky, 2001) used to assess teacher reactions to overall school restructuring efforts using numerous CSR models (including DI and SFA). Adaptation of the selected CSRTQ items for this RTS identified the specific reading program in the case of DI and SFA, but referred generically to the "reading program" for the Control schools. Among the areas assessed were professional development, impacts on students, changes in teaching, support for the program, effects on technology use, and involvement of parents and the community. Kuder-Richardson Formula-20 (Kuder & Richardson, 1937). Reliability coefficients were +0.82, +0.80, and +0.86 for the three respective versions (DI, SFA, Control), thereby showing relatively high internal consistency.

Targeted observations. Independent, out-of-state consultants with expertise in the implementation of DI and SFA conducted site visits at the schools implementing the two models. Because the observers for both models had previously performed implementation checks as trainers or coaches, they employed checklists and rating scales adapted from the formal model-specific evaluation materials. For example, the SFA data collection form consisted of five open-ended sections concerning impressions of (a) school climate, (b) staff interactions, (c) staff-student interactions, (d) school organization, and (e) attractiveness and adequacy of facilities. Rating scales for "extent" and "quality" were then used to assess specific program components (e.g., Early Learning, Reading Roots, etc.). A final open-ended section prompted the expert to describe implementation strengths and weaknesses. Comparable program-specific assessments were used in the DI site visits.

Interviews and focus groups. Interviews were conducted by the external research team with (a) key stakeholders (district leadership, union leadership, and the president of the school board), (b) principals of DI, SFA, and Control schools, (c) building-level union representatives, and (d) DI and SFA school facilitators. Teacher focus groups of about 1 hr in length were held to provide background information about schools' usage of their selected programs. The principal at each of the nine schools participated in a 1-hr, on-site interview concerning his or her experiences with and reactions to the design implementation and the associated outcomes for the school, students, faculty, parents, and community. The Control school principals also participated in a similar structured 1-hr interview concerning their school's board-adopted program.

Procedure

The SCI and RTS were administered by a district staff member at school faculty meetings early in the spring semester. Confidentiality of responses was maintained by having teacher leaders directly mail the completed survey (not including any identifying information) to the researchers. All interviews were conducted by the researchers in early spring. Qualitative analyses, guided by Miles and Huberman's (1994) analysis model, were performed on open-ended survey and interview responses. The procedure consisted of transcribing ideas and concepts, deriving patterns and concepts, identifying themes, and revision and refinement based on member checking and interrater review. Once a final set of themes was established for each question, the data were examined to derive frequencies of occurrence for the respondent groups. Using this procedure, the closed-ended responses were supplemented by determining for each question the major themes, and the manner and frequency that the themes were experienced by individuals. Student achievement data files were prepared by the district researchers but analyzed by the external team.

RESULTS

Student Achievement

Two-level hierarchical linear model specification. Average program effects for DI and SFA programs on Stanford 9 Total Reading in Grade 2 and reading scores on the Ohio Proficiency Test in Grades 4 and 6 (see Table 2) were estimated. The HLM inferential tests adjusting for school pretest means and poverty (percentage free or reduced-price lunch) were run to determine whether students attending schools using either program scored significantly higher than students attending other district elementary schools. Similarly, the relationship between pretest and posttest scores within each school (i.e., the pretest-posttest regression slope) was estimated and tested to determine whether this relationship varied as a function of program type. Smaller slopes indicate a more equitable distribution of achievement outcomes across levels of pretest. Average effects and effects on pretest-posttest slopes were estimated and tested using a two-level hierarchical linear model for each grade level (2, 4, and 6) for 3 school years (1998–1999, 1999–2000, and 2000–2001).

Level 1 (within schools). A regression of posttest scores on pretest scores was generated for each district school, with the pretest variable centered on the grand pretest mean for all schools. These analyses yielded two outcome variables for each school: (a) an adjusted posttest mean, and (b) a pretest-posttest regression slope. The adjustment procedure rendered program effect comparisons equivalent across schools, despite variation in the average pretest scores.

Level 2 (between schools). Three school-level variables were used to estimate reading program effects on adjusted means and pretest-posttest slopes: (a) percentage of students receiving free or reduced-price lunch, (b) DI program, and (c) SFA program. Coefficients estimated for DI and SFA represent the difference between the average adjusted means (or slopes) for schools implementing these programs and the expected adjusted mean given the level of poverty in the school. Thus, inferential statistical tests on the effectiveness of these two

TABLE 2
Pretest-Posttest Measures for Evaluation of DI and SFA Programs

Cohort (Grade)	Pretest	Posttest
Second	1 st grade Stanford 9 Total Reading	2 nd grade Stanford 9 Total Reading
Fourth	3 rd grade Off-grade Proficiency Test ^{a,b}	4 th grade Ohio Proficiency Test ^b
Sixth	5 th grade Off-grade Proficiency Test ^{a,b}	6 th grade Ohio Proficiency Test ^b

Note. DI = Direct Instruction; SFA = Success For All.

^aRiverside Off-Grade Proficiency Test. ^bReading scale score.

programs control for both differences in average pretest scores and differences in school-level poverty rates. A positive coefficient associated with a program indicates that students attending program schools scored higher than would be expected, given their pretest scores and the poverty level of the school.

Program effect size estimates. An effect size estimate was computed for each program for each year and grade level by dividing the coefficient associated with the program derived from the hierarchical linear model by the pooled within-schools standard deviation of the posttest in the baseline year (1998–1999). The effect size estimate (*ES*) represents the difference between the actual and expected adjusted means for a program in standard deviation units, or z-scores.

DI Program Effects on Adjusted Means

Schools implementing the DI program had significantly lower posttest means than expected in 2001 fourth grade ($t = -2.37$, $df = 40$, $p < .05$) and 1999 sixth grade ($t = -2.30$, $df = 40$, $p < .05$; see Table 3). Otherwise, there were no significant differences between DI schools and other district schools after adjusting for poverty rates and beginning levels of student achievement. In second grade, the overall DI effect size estimates declined from $ES = +0.12$ in 1999 to $ES = -0.17$ in 2001. These effect size estimates were small and not significantly different from zero. In fourth grade, DI effect size estimates were progressively larger and negative, from $ES = -0.09$ in 1999, to $ES = -0.14$ in 2000, to $ES = -0.25$ in 2001. This trend was reversed for sixth-grade DI effects, which improved from $ES = -0.31$ in 1999 to $ES = -0.18$ in 2001. Generally, DI effect size estimates were negative and small to moderately large in size.

SFA Program Effects on Adjusted Means

No significant differences were associated with the SFA program at any grade level or year (see Table 3). Effect size estimates for SFA were near zero for all grade levels and years, with the exception of Grade 2 in the baseline (preprogram) year, 1999 ($ES = -0.19$), and Grade 4 in the second implementation year, 2001 ($ES = -0.20$).

Effects on Pretest-Posttest Slopes

Neither the DI nor the SFA program schools had significantly different pretest-posttest slopes than expected for district schools with similar levels of poverty for any grade level or year considered (see Table 4). This suggests that achievement gains in schools implementing either of the programs are distributed in a manner similar to that in other district schools. Furthermore, school poverty rate was not significantly related to the slopes for any of the comparisons made. Thus, while slopes varied from school to school, there was no apparent relationship be-

TABLE 3
Program Effects on Adjusted Posttest Means

Grade/School Year	DI		SFA	
	Coefficient	t-ratio	Coefficient	t-ratio
Second				
98-99	1.37	0.68	-2.08	-0.47
99-00	-0.71	-0.34	-0.04	-0.01
00-01	-1.91	-0.83	-0.10	-0.03
Fourth				
98-99	-1.91	-0.85	1.65	0.36
99-00	-2.93	-1.10	-1.28	-0.29
00-01	-5.37	-2.37*	-4.30	-1.13
Sixth				
98-99	-7.97	-2.30*	1.27	0.18
99-00	-5.65	-1.74	-2.33	-0.36
00-01	-4.52	-1.68	0.70	0.13

Note. SFA programs did not begin implementation until the 1999-2000 school year. DI = Direct Instruction; SFA = Success for All. *Significant at $p < .05$, $df = 40$.

tween poverty level of the district schools and the distribution of achievement outcomes across different levels of prior achievement. For example, students with high prior achievement in high-poverty and low-poverty schools were likely to demonstrate similar success in current achievement.

School Effects

Effect size estimates were computed at each grade level (2, 4, and 6) for all Program schools from 1998-1999 to 2000-2001. All posttest means were first adjusted for pretest scores. School effect sizes were ascertained by subtracting the expected adjusted posttest mean for a school based on poverty rate from the school's obtained adjusted posttest mean. This quantity was then divided by the pooled within-groups standard deviation of the posttest in the baseline year (1998-1999) to derive an effect size estimate. These descriptive patterns are summarized below by program and cohort.

1997-1998 DI school effects. Three schools (DI-1, DI-2, and DI-3) began implementation during the 1997-1998 school year.

- In second grade, DI-3 exhibited moderately large and positive effects from the baseline year, 1999 ($ES = +0.14$), to 2001 ($ES = +0.46$). DI-1 students, however, consistently achieved at a normative level for the district (0.00 to +0.13). At DI-2, effect sizes were moderately large and negative (-0.26 to -0.38).

TABLE 4
Program Effects on Pretest-Posttest Slopes

Grade/School Year	DI		SFA	
	Coefficient	t-ratio	Coefficient	t-ratio
Second				
1999	-0.08	-0.96	-0.66	-0.45
2000	-0.14	-1.49	0.00	0.02
2001	-0.11	-1.47	0.11	1.05
Fourth				
1999	0.04	0.33	-0.25	-0.90
2000	-0.14	-0.74	0.04	0.13
2001	-0.02	-0.15	-0.04	-0.15
Sixth				
1999	-0.07	-0.66	-0.29	-1.21
2000	0.14	1.18	-0.22	-0.92
2001	-0.01	-0.08	-0.16	-0.55

Note. SFA programs did not begin implementation until the 1999-2000 school year. DI = Direct Instruction; SFA = Success for All.

- Effects in fourth grade were small and neither consistently positive or negative for these schools.
- Sixth-grade effects at DI-1 and DI-3 were moderately large and negative in 1999 ($ES = -0.30$ and -0.25), but improved to normative levels by 2001.

1999-2000 DI school effects. DI-4, DI-5, and DI-6 began implementation in 1999-2000.

- In second grade achievement at DI-6 declined precipitously from the baseline year, 1999, ($ES = +0.38$), to 2001 ($ES = -0.85$). In contrast, achievement effects improved substantially at DI-4 during the same period (from +0.13 to +0.37). At DI-5, achievement scores were substantially below the district average ($ES = -0.39$) in 1999, but normative in 2001 ($ES = -0.12$).
- In fourth grade, all three schools were slightly below the district average gain in 1999 ($ES = -0.08$ to -0.16). DI-4 and DI-6 exhibited a modest improvement in 2000 but moderate decline in 2001.
- In sixth grade, DI-4 exhibited a consistent pattern of improvement from 1999 ($ES = -0.35$) to 2001 ($ES = +0.07$). DI-5 showed a large decline in achievement from 1999 ($ES = -0.09$) to 2000 ($ES = -0.54$), but improved substantially in 2001 ($ES = -0.12$). Results at DI-6 were comparable to the district average in 1999 and 2000, but declined modestly in 2001 ($ES = -0.23$).

1999-2000 SFA school effects. Two schools, SFA-1 and SFA-2, began implementation during the 1999-2000 school year. Because SFA-1 was a rela-

tively new school, 1999 data were not available for Grades 2, 4, and 6. Effects for all grades and years at the SFA schools were almost equal to the district average, ranging from a low of $ES = -0.13$ for Grade 4 in 2001 at SFA-2, to a high of $ES = +0.12$ for Grade 2 in 2001 at SFA-1.

SCHOOL CLIMATE

Data analysis for the SCI involved comparing the Program (DI and SFA) schools to the Control schools on the seven climate dimensions, using one-way multivariate analyses of variance (MANOVA).

DI

A summary of the dimension means and standard deviations for the DI and Control schools is provided in Table 5. Because the DI results were fairly consistent across schools, aggregate findings are presented in the table. Although the means in general were slightly lower than those for SCI national norms, none of the dimension means, except for Order (both DI and Control M 's = 2.49), indicates overly negative school climate outcomes for either group. The MANOVA comparing DI ($n = 184$) and Control ($n = 168$) responses was not significant, $F(7, 247) = 1.27, p = .264$. The largest difference, in fact, was only .09 points, favoring the DI schools on Involvement.

SFA

Unlike the situation for DI, the two SFA schools demonstrated clearly contrasting climate outcomes. Accordingly, Table 6 provides a summary of the descrip-

tive results by school. MANOVA results significantly favored SFA-1 over the Control schools, $F(7, 38) = 6.61, p < .001$. Univariate tests yielded differences, all supporting SFA-1, on five out of the seven climate dimensions (see Table 6). Effect sizes were extremely high, ranging from +1.05 (Expectations) to +1.80 (Environment). In contrast, SFA-2 was inferior to the Control schools, $F(7, 64) = 9.74, p < .001$, showing significant deficits in univariate tests on six dimensions. Effect sizes were large and negative, ranging from -0.77 (Expectations) to -1.75 (Collaboration).

READING TEACHER SURVEY

DI

Frequency data for the DI sample revealed the following items as receiving 80% or higher agreement or strong agreement by teachers: understanding the program (88%), receiving adequate training (81%), and receiving support from the school's external facilitator or other staff (83%). For the Control sample, no item received higher than 80% agreement. At the other extreme, both groups strongly disagreed (over 50%) that parents or community members were more involved because of the program.

A one-way MANOVA comparing DI and Control school means on the survey yielded a highly significant effect, $F(20, 207) = 10.27, p < .001$. Follow-up univariate analyses, as summarized in Table 7, were significant on 12 of the 20 items, all favoring DI. As shown in the table, nearly all of the effects were at least moderate in size, approximating or exceeding $ES = +0.50$.

TABLE 5
Descriptive Results on the School Climate Inventory:
DI ($n = 7$) Versus Control Schools ($n = 6$)

Dimensions	National Norms		DI (teacher $n = 184$)		Control (teacher $n = 168$)	
	M	SD	M	SD	M	SD
Collaboration	3.74	0.59	3.12	0.56	3.18	0.65
Environment	3.79	0.70	3.16	0.69	3.16	0.80
Expectations	3.85	0.62	3.32	0.69	3.36	0.63
Instruction	4.02	0.46	3.59	0.63	3.65	0.53
Involvement	3.77	0.61	3.31	0.67	3.22	0.57
Leadership	3.92	0.68	3.39	0.77	3.39	0.82
Order	3.24	0.81	2.49	0.91	2.49	0.66

Note. 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree; DI = Direct Instruction.

TABLE 6
Descriptive Results on the School Climate Inventory:
SFA ($n = 2$) Versus Control Schools ($n = 2$)

Dimensions	National Norms		SFA-1 (teacher $n = 21$)		SFA-2 (teacher $n = 46$)		Control (teacher $n = 45$)	
	M	SD	M	SD	M	SD	M	SD
Collaboration	3.74	0.59	4.35*	0.31	2.30*	0.56	3.39	0.68
Environment	3.79	0.70	4.44*	0.42	2.52*	0.64	3.32	0.68
Expectations	3.85	0.62	4.27*	0.48	3.21*	0.52	3.65	0.62
Instruction	4.02	0.46	4.53*	0.34	3.77	0.46	3.80	0.57
Involvement	3.77	0.61	4.26*	0.39	3.04*	0.59	3.53	0.57
Leadership	3.92	0.68	4.10	0.68	2.59*	0.77	3.54	0.79
Order	3.24	0.81	3.56	0.79	2.35*	0.68	3.32	0.74

Note. 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree; SFA = Success for All. *Significantly different from the Control schools at $p < .01$.

TABLE 7
Summary of Follow-up DI vs. Control Group Comparisons on
the Reading Teacher Survey

Items	DI (teacher n = 126)		Control (teacher n = 129)		F (1, 226)	p	ES
	M	SD	M	SD			
1. I have a thorough understanding of this school's reading program.	4.26	0.73	3.86	0.95	14.78	0.000	0.47
2. I have received adequate initial and ongoing professional development/training for implementation of my school's reading program.	4.09	0.93	3.38	1.07	29.39	0.000	0.71
3. Professional development provided by external trainers, model developers, and/or designers has been valuable	4.02	0.86	3.33	0.96	36.13	0.000	0.75
4. Guidance and support provided by our school's external facilitator, support team, or other resource personnel have helped our school implement its reading program.	4.13	0.78	2.87	1.01	117.45	0.000	1.40
5. Teacher are given sufficient planning time to implement our reading program.	3.66	1.14	2.99	1.08	22.51	0.000	0.60
6. Materials (books and other resources) needed to implement our reading program are readily available.	3.94	1.11	3.24	1.23	17.06	0.000	0.62
7. Our school has sufficient faculty and staff to fully implement its reading program.	3.90	1.01	3.61	0.90	5.94	0.016	0.30
8. Because of our program, technological resources have become more available.	2.55	0.99	2.73	0.93	1.78	0.183	-0.19
9. Our reading program has changed classroom learning activities a great deal.	3.87	0.90	3.07	0.90	37.75	0.000	0.89
10. Student achievement has been positively impacted by our reading program.	3.83	1.12	3.09	0.89	28.58	0.000	0.73

(continued)

TABLE 7 (Continued)

Items	DI (teacher n = 126)		Control (teacher n = 129)		F (1, 226)	p	ES
	M	SD	M	SD			
11. Children in this school are more enthusiastic about learning because of our reading program.	3.26	1.26	2.88	0.86	5.78	0.017	0.36
12. Because of our reading program, parents are more involved in the educational program of this school.	2.34	1.06	2.22	0.85	0.57	0.451	0.11
13. Community support for this school has increased since our reading program has been implemented.	2.50	1.07	2.38	0.89	0.98	0.323	0.13
14. Students have higher standards for their own work because of our schools' reading program.	2.85	1.24	2.57	0.86	3.38	0.067	0.26
15. Teachers are more involved in decision making at this school than they were before we implemented our reading program.	2.71	0.87	2.67	0.91	0.02	0.897	0.04
16. Our reading program adequately addresses the requirements of children with special needs.	3.46	1.15	2.85	1.10	14.01	0.000	0.54
17. Because of our reading program, teachers in this school spend more time working together to develop curriculum and plan instruction.	2.62	1.08	2.49	0.97	0.43	0.512	0.13
18. Teachers in this school are generally supportive of our reading program.	3.54	1.07	3.29	0.84	3.29	0.071	0.26
19. The elements of our reading program are effectively integrated to help us meet school improvement goals.	3.44	1.02	3.28	0.95	1.07	0.30	0.16
20. This school has a plan for evaluating all components of our reading program.	3.20	1.02	2.72	0.95	10.96	0.001	0.48

Note. 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree; DI = Direct Instruction; ES = size estimate.

SFA

Because the two SFA schools responded similarly on the RTS, they were combined for this analysis. Descriptive results indicated agreement by 80% or more of the SFA teachers that: they had a thorough understanding of the program (96%), had adequate initial and ongoing professional development (89%), received valuable external professional development (88%), and changed classroom learning activities a great deal (83%). In contrast, at least 80% of Control teachers agreed with only one item—that they had a thorough understanding of the reading program (91%). For the Control group only, there was majority disagreement on items concerning parent involvement (59%), community involvement (62%), teacher participation in decision making (56%), and teachers spending more time working together (65%).

The one-way MANOVA comparing the SFA and Control means was highly significant, $F(20,58) = 4.25, p < .001$. Follow-up univariate ANOVAs yielded significant program differences on 13 of the 20 items, all favoring SFA (see Table 8). The effect sizes for those comparisons were all moderate to large in magnitude (ranging from +0.44 to +1.12).

TARGETED OBSERVATIONS

DI Targeted Observations

A team of experienced DI implementers visited the seven DI schools for 1 day per school. A summary of their qualitative report is as follows (for a detailed description and actual report, see Ross et al., 2003). In general, the observations revealed inconsistent and incomplete levels of implementation of the DI components considered critical for accelerating student progress to reach grade-level performance. There was some anecdotal evidence indicating positive changes in most schools, such as a reduction in the number of children who were nonreaders, better student behavior, and increasing instructional groups in the lower grades performing at or above grade performance levels.

The primary recommendations indicated the need for: (a) greater emphasis on accelerating student performance in kindergarten and first grade, (b) longer reading instructional time to accelerate and expand student reading proficiency, (c) additional daily reading instruction for children who are functioning below grade level, (d) structured reading in a wide variety of materials, (e) greater emphasis on implementation of the DI language curricula, (f) more intensive professional development to enable all teachers to reach high levels of proficiency in teaching DI, (g) training and support for building principals to take a more active role in supporting implementation of the DI model, and (h) more coordinated district management of DI funding and school usage to ensure consistency and quality in implementation.

TABLE 8
Summary of Follow-up SFA vs. Control Group Comparisons on the
Reading Teacher Survey

Items	SFA (teacher n = 53)		Control (teacher n = 34)		F (1, 77)	p	ES
	M	SD	M	SD			
1. I have a thorough understanding of this school's reading program.	4.55	0.57	4.26	0.96	3.31	0.073	0.39
2. I have received adequate initial and ongoing professional development/training for implementation of my school's reading program.	4.40	0.86	3.97	1.09	3.99	0.049	0.45
3. Professional development provided by external trainers, model developers, and/or designers has been valuable.	4.25	0.81	3.50	1.11	16.90	0.000	0.80
4. Guidance and support provided by our school's external facilitator, support team, or other resource personnel have helped our school implement its reading program.	4.18	0.91	3.06	1.13	25.54	0.000	1.12
5. Teachers are given sufficient planning time to implement our reading program.	3.17	1.27	2.88	1.32	0.91	0.342	0.22
6. Materials (books and other resources) needed to implement our reading program are readily available.	4.09	1.02	3.59	1.13	4.15	0.045	0.47
7. Our school has sufficient faculty and staff to fully implement its reading program.	3.17	1.33	3.74	0.96	3.33	0.072	-0.47
8. Because of our program, technological resources have become more available.	2.63	1.01	2.91	0.97	1.25	0.267	-0.28
9. Our reading program has changed classroom learning activities a great deal.	4.13	0.88	3.15	1.02	24.22	0.000	1.04
10. Student achievement has been positively impacted by our reading program.	3.81	1.16	3.12	1.20	8.19	0.005	0.58
11. Children in this school are more enthusiastic about learning because of our reading program.	3.30	1.20	3.09	1.11	2.21	0.141	0.18

(continued)

TABLE 8 (Continued)

Items	SFA (teacher n = 53)		Control (teacher n = 34)		F (1, 77)	p	ES
	M	SD	M	SD			
12. Because of our reading program, parents are more involved in the educational program of this school.	3.21	1.19	2.47	0.93	8.01	0.006	0.67
13. Community support for this school has increased since our reading program has been implemented.	2.94	1.13	2.35	0.81	5.74	0.019	0.58
14. Students have higher standards for their own work because of our schools' reading program.	3.12	1.17	2.91	1.07	0.76	0.387	0.18
15. Teachers are more involved in decision making at this school than they were before we implemented our reading program.	3.02	1.19	2.50	0.90	5.96	0.017	0.48
16. Our reading program adequately addresses the requirements of children with special needs.	3.47	1.38	2.76	1.10	7.63	0.007	0.55
17. Because of our reading program, teachers in this school spend more time working together to develop curriculum and plan instruction.	3.00	1.27	2.35	1.01	7.17	0.009	0.55
18. Teachers in this school are generally supportive of our reading program.	3.28	1.32	3.15	1.13	0.18	0.672	0.10
19. The elements of our reading program are effectively integrated to help us meet school improvement goals.	3.75	1.06	3.27	1.13	5.82	0.018	0.44
20. This school has a plan for evaluating all components of our reading program.	3.58	1.08	2.65	1.18	17.60	0.000	0.83

Note. 1 = Strongly Disagree; 2 = Disagree; 3 = Neutral; 4 = Agree; 5 = Strongly Agree; SFA = Success for All; ES = Size estimate.

SFA Targeted Observations

An expert observer visited the two SFA schools for 1 day each. All available classrooms during the SFA reading blocks and segments of 20-min tutoring sessions were observed at each site. In addition, the prior year (2000–2001) Implementation Reports composed by external trainers from the SFA Foundation were reviewed to corroborate information gathered directly through observation.

Overall, the district SFA implementation was characterized as “uneven” (for a detailed description, see Ross et al., 2003). Although observations revealed that implementation of the SFA program was generally above average in curricular areas and both schools provided the requisite 90 min of reading instruction each day, concerns and recommendations focused primarily on weaknesses in (a) usage of SFA’s writing components, (b) the family support implementation, and (c) ongoing teacher training to compensate for high staff turnover.

Contrasts between the SFA schools were also observed. SFA-1 had attractive facilities, served only 300 students, had very positive school climate (as corroborated by the SCI results reported above), limited class size to only 22 students, and attracted a relatively stable student population. On the negative side, SFA-1 was lacking a full-time facilitator at the time of this study. In contrast, SFA-2 was in poor physical condition, served almost 900 students, and had fairly poor school climate. Positive aspects of SFA-2 were its employment of two full-time facilitators and commitment to SFA as the central school focus.

QUALITATIVE SYNTHESIS

The following sections present a synthesis of qualitative findings obtained from the DI and SFA principal interviews, the building representative interviews, and the teacher focus groups. Categories were derived from the coding analyses described previously. In the interests of brevity, Control school interviews are not presented and only major Program group findings are highlighted.³

DI

Most effective elements. The different respondent groups consistently identified several advantages of DI, including the emphasis on phonics, the consistency of methods across classes, the 90 min of reading, and the overall reading work completed in the lower grades.

³Readers interested in more detailed findings should see Ross and colleagues (2003).

Least effective elements. Two of the most common concerns about DI were the lack of emphasis on comprehension skills and the excessive time demands (e.g., paperwork, pacing, and scheduling). Other concerns included perceived deficiencies of silent, sustained reading, alignment with standards, quality and diverse reading material, and supplemental materials needed to implement the entire DI program.

Implementation of reading program. Four of the six DI principals conveyed that implementation was proceeding well. The most commonly cited obstacle, according to three principals and three teacher focus groups, was high student mobility. Several building representatives and teacher focus groups felt that teachers, especially in the intermediate grades, were not using the model the way the developers intended.

Teacher support. Most principals and building representatives indicated good teacher support for DI. One principal reported that "There does tend to be some resistance by new teachers, who have not been trained in DI and who feel that DI is too hard and requires a lot of grading." Two teacher focus groups were positive, but four indicated mixed support for the program, especially in higher grades.

Classroom level changes. According to principals, classroom changes included: very structured classes, stronger teacher emphasis on reading, more focused reading instruction, vocabulary being a part of each lesson, more books in classrooms, attempts at reduced seatwork, and student placement according to reading level. Some respondents felt that usage of team-based and cooperative approaches had increased. Teacher focus groups most strongly conveyed that DI materials are better targeted, teaching is easier, and classrooms are more structured, consistent, and disciplined.

Impact on student achievement. The DI principals and teachers generally perceived both reading and writing to have improved. Building representatives, however, saw comprehension skills as still weak in the intermediate grades. Whereas most respondents felt that student discipline in reading classes had improved, there were mixed opinions regarding DI effects on student enthusiasm for reading.

Professional development. Only two principals felt that the DI training provided to teachers was effective. The remainder conveyed that more extensive training was needed, with special provisions for new and transfer teachers. Three teacher focus groups specifically noted that there is little follow-through after initial training and that more mentoring, observing, and critiquing are needed.

Community support. In most of the DI schools, principals reported strong community support for DI. For example, one principal conveyed that parents love DI and are ready to advocate for keeping DI, adding, "There is no way that parents will let DI be taken away from their school." Teachers and building representatives reflected the mixed view that although parent and community support for keeping DI was high, parents were only minimally involved with the school and monitoring students' work at home.

SFA

Most and least effective elements. According to the two SFA principals, the most effective elements of the SFA model are the 90 min of reading time and the ability grouping for reading. The teacher focus group and building representatives reported that the most effective elements are the consistency from grade to grade and the phonics and tutoring offered to the primary grades.

The least effective elements, according to the principals, are the nonfiction segment and time constraints. To the building representatives and teacher focus groups, the least effective elements of the SFA model were the scripted lessons, the lack of alignment to state standards, and the writing component.

Resources needed. With regard to resource needs, respondents identified additional funding to pay for the program, smaller class sizes, additional facilitators, more books to implement the program, and more space.

Program implementation and teacher support. Respondents at SFA-1 were consistently positive about the implementation and teacher support. The SFA-2 responses were less supportive but still mostly positive. An obstacle to full implementation identified multiple times in the interviews was the high teacher turnover rate.

Classroom-level changes. The most salient changes identified were smaller reading groups, attention to individual needs, increased cooperative and team-based learning, and improved accommodation and inclusion of special needs children.

Impact on students. Respondents from both schools believed that SFA improved student ability and motivation to read, although this effect was not necessarily reflected in standardized test score gains. Another benefit was stronger student inter-relationships and fewer discipline problems.

Impact on teachers. The principals believed that SFA increased teacher collegiality and interactions. Teachers, however, were equivocal, feeling that

even though the program seemed to improve student learning, it also increased time demands and stress.

Professional development. In general, professional development for SFA was described positively by most respondents. There was some criticism concerning the adequacy of initial training for first-year and transfer teachers, and follow-up training for veteran staff.

DISCUSSION

In this section, we examine and synthesize results relative to each of the major research questions that guided the study.

What Were Program Impacts on Student Achievement Outcomes?

Student achievement at the DI and SFA schools was generally at levels that would be expected of district schools serving similar populations. It is important to consider that at the time of this analysis, the two SFA schools and three of the six DI schools were in only their 2nd year of implementation, a period that many experts on school reform would consider too brief to produce strong gains (Fullan, 2000; Levin, 1993). Still, given that DI and SFA are structured programs rather than philosophical frameworks or complex CSR models, initial impacts on student performance could reasonably be expected.

The results overall and for each model further failed to reveal significantly different pretest-posttest slopes than expected for district schools with similar levels of poverty for any grade level or year considered. Although both DI and SFA are directly oriented to foster growth in at-risk populations (Englemann et al., 1988; Herman, 1999; Slavin & Madden, 2001), the results failed to show differential benefits for low or high achievers compared to students receiving the district model.

The individual school achievement test results for both models were mixed, showing some gains and some deficits relative to the district. Two of the DI schools and one SFA school emerged as showing generally positive trends, whereas two DI schools and the other SFA school showed a mostly negative pattern. In the sections below, we draw on the results from this study to explore possible relationships between school process and context variables (e.g., climate and teacher support) and student achievement outcomes.

What Were Program Impacts on School Climate?

The school climate results showed comparable scores for DI and Control schools. Whereas school climate was slightly below national norms for both groups of schools (especially in the Order dimension), these results do not reflect serious problems that would negatively impact implementation of a program or student achievement.

SCI results for SFA schools were relatively low at SFA-2 but high at SFA-1. Recent research reveals school climate to be a significant predictor of schools' success in raising achievement over time (Bobbett, Ellett, Teddlie, Olivier, & Ruggett, 2002; Bryk & Schneider, 2002). As use of SFA continues at the two schools, it should be interesting to determine if these climate patterns continue and relate to differential student achievement trajectories.

What School and District Variables Appeared to Influence the Effectiveness of the Programs?

Implementation. Although the expert DI observers noted positive outcomes such as students becoming more interested and involved in reading, they raised several specific concerns perceived as limiting the potential to raise student achievement at the program schools (Question # 1). Among the more salient needs identified included: (a) accelerating student performance in kindergarten and first grade, (b) increasing reading time per day beyond the 90-min block, (c) providing supplementary (after-school) tutoring for low-performing children, (d) providing structured reading for a wide variety of materials, (e) placing greater emphasis on the DI language programs, (f) providing more intensive professional development for teachers, (g) selecting school facilitators who have sufficient DI experience and expertise, (h) involving principals to a greater extent in leading DI implementations, and (i) establishing more focused district-level leadership in the implementation of DI. It is noteworthy but not surprising that prior research on DI has shown usage of model-related teaching strategies to relate to better student performance (Carlson & Francis, 2002).

Interpretations of the above impressions should be qualified due to the limited duration of the observations. Still, many of the experts' impressions of incomplete implementation were triangulated by views conveyed from other sources (e.g., model facilitators, teachers, and district leaders). Supplementary interviews with district leaders further expressed concerns that teachers often opted not to participate in summer or other training where not required to by contract. As a result, these individuals were not fully trained. A second theme was that the district was reevaluating its ability and willingness to provide substantial funding and other resources to support implementation, viewing these areas as needing to become increasingly the responsibility of the individual schools.

The SFA expert viewed the implementations at both schools as “above average,” although still in developing stages after less than 2 years of model adoption. Teacher and principal reactions were also generally positive about their schools’ ability to implement the SFA model components effectively. Implementation challenges center on professional development and training issues, including the lack of a full-time facilitator for most of the school year at one school, an absence of any formal SFA training for tutors, and inadequate classroom space. Despite the challenges, staff support at both sites, but especially at SFA-1, which showed a more positive achievement pattern, was regarded as sufficiently high to produce adequate to above-average implementation.

Stakeholder reactions. Across the multiple data sources (teacher survey and various stakeholder interviews), the models were viewed as having beneficial impacts on students’ skills and interests in reading. However, when we compare the teacher survey results to elementary school norms from the original CSRTQ survey (see Sterbinsky & Ross, 2003), the DI and SFA means tend to be comparable or only slightly higher. Based on prior research on school reform, teacher and principal support seems a necessary but not sufficient condition for effective program implementation and increases in student achievement to take place (Desimone, 2002; Muncey & McQuillan, 1996; Smith et al., 1997). Judging by the very low Control school means on the teacher survey, the DI and SFA teachers by comparison appear appreciative of the additional resources, structure, and professional development received.

Relationship of school variables to student achievement. A fundamental question in school reform research concerns the extent to which improvements in school culture and program implementation are associated with gains in student achievement (e.g., Bryk & Schneider, 2002; Datnow & Stringfield, 2001). Quantitative analysis of such relationships was precluded in this study by the relatively small number of program schools in the DI and especially SFA subsamples. However, for exploratory purposes, we addressed the question through a qualitative examination of the data.

First, we classified the program schools according to their longitudinal achievement patterns.⁴ Most of the schools performed comparably to district-wide norms after adjusting for school and student variables. However, four DI schools, two showing positive and two showing negative trends, were distinguishable from the others. Next, we examined school data reflecting program implementation, school climate, and teacher support for the reforms. Consistent with the reform literature, the two high-achieving schools (DI-1 and DI-2) had average to above-average

⁴A detailed reporting of the methodology and results is provided in Ross and colleagues (2002). A summary of main results is presented here in the interests of brevity.

school climate and above-average implementation. Furthermore, one of the low-achieving schools (DI-3) had relatively negative school climate, markedly so on the Order dimension. Program implementation at that school was also impeded by high student mobility and lack of faculty support for the model. Contrary to these patterns, however, school DI-1 (high-achieving) had very low teacher support for the program, whereas DI-4 (low-achieving) had both positive school climate and high teacher support.

Both SFA schools performed comparably to the district norms, but the longitudinal achievement pattern was more positive for SFA-1 than for SFA-2. This apparent trend directly mirrored school status variables, which indicated somewhat stronger program implementation and noticeably more positive school climate at SFA-1 than SFA-2 (see Table 6).

These results overall suggest that positive indicators of school reform status (e.g., favorable climate, high teacher support, strong program implementation) are generally predictive of greater success in raising achievement. But school reform is a complex process involving interactions between many events, policies, and stakeholder groups (Datnow et al., 2002). Thus, as exemplified by the anomalous patterns for DI-1 and DI-4, predictions made from isolated status variables are far from certain.

CONCLUSIONS

As indicated in the introduction to this article, Desimone (2002) employed the policy attributes theory developed by Andrew Porter and his colleagues (Porter, 1994; Porter et al., 1988) as an explanatory framework for successful policy implementation. Following Desimone’s approach, we conclude by interpreting this findings relative to the same five theory components: specificity, consistency, authority, power, and stability.

Specificity incorporates clarity and concreteness in strategies, materials, information, and monitoring. Relative to other popularly used reform designs, both DI and SFA are relatively strong in this domain (Herman, 1999). Both, for example, have prescribed curricula, defined strategies of teaching and adaptation to individual differences, and extensive and structured professional development programs. Nonetheless, some weaknesses noted here in the specificity of the actual model implementations (e.g., lack of a facilitator in SFA-1; inadequate professional development for new DI teachers) would have likely limited program effectiveness.

Consistency denotes smooth integration with other reform efforts at the school, district, and state levels. In the case of SFA, there were some concerns expressed about the recency of its curriculum alignment with state standards. There were also isolated comments by teachers and building representatives regarding the need for stronger coordination of both DI and SFA with district policies and goals. However, compared to other systemic reform contexts (e.g., see

Berends, Kirby, Naftel, & McKelvey, 2001), the present DI and SFA implementations appeared to provide reasonable consistency levels.

The *authority* dimension entails the interactive roles of teachers, the principal, and the district in the model selection and implementation. A related dimension, *power*, concerns the degree to which these and other stakeholders exercise control over decision making and policies. In the case of DI, but less so for the more newly adopted SFA, these two domains seem to have been highly influential in affecting the quality of model implementation. Interviews with various stakeholder groups revealed how DI had originally been strongly advocated by neighborhood groups supporting schools that served predominately disadvantaged, African American students. Some informants recalled teacher and principal buy-in to be "mostly positive" (if not universal) for the first three schools, and "from accepting to positive" for the later-implementing group of three schools. Especially in the case of the latter cohort, orchestrated community pressure was believed to have alienated some teachers who consequently felt forced to come aboard. But as often occurs with reforms (Berends et al., 2002; Ross, 2001), the level of buy-in at all schools was perceived to have eroded over time with changes in school staff and leadership.

A second perceived major source of authority and power influence was the local teacher union, which had a strong district role and presence. Specifically, union policies protected teachers' freedom to choose whether to participate in professional development activities occurring outside of regular school hours (e.g., summer or weekend training). In general, the teacher groups most in need of model-specific professional development were novice and transfer teachers who were not employed at their present schools when the models were adopted. But also for the latter reason, these groups were least likely to be motivated to participate in optional training opportunities.

The school district was a third source of authority and power. Most representatives of the different stakeholder groups (teachers, principals, community members, school board, and teacher union) agreed that school district leaders had been supportive of the reform models during the first few years of implementation. School district leaders further conveyed an open and hopeful attitude that the models would ultimately prove successful in raising student achievement. However, patience was rapidly wearing thin in the absence (in the case of DI) of tangible results. Even though the district had invested extensive financial and professional development resources in the models, it was increasingly viewing implementation costs and demands as the responsibilities of the individual schools.

The combination of these major sources of authority (i.e., the strong community investment, teacher union policies, and district support), in the authors' opinion, served to weaken some schools' ownership of the models and accountability for their success. To the extent that schools perceive external entities to be the own-

ers, the quality of implementation and sustainability are likely to be negatively impacted (Berends et al., 2002; Ross, 2001). Some informants, however, believed that a somewhat different power-authority perception developed at other schools. At these schools, there was relatively strong ownership of the model but the belief that the district was responsible for providing the resources, advocacy, and technical assistance needed to ensure success. Unlike the systemic reforms efforts of San Antonio (Berends et al., 2002) and Memphis (Ross, 2001), the district role here was more piecemeal (several schools at a time) and oriented to shifting responsibility increasingly over time to the individual schools.

The fifth dimension in Porter's framework is *stability*. For effective change to occur, there must be sufficient time for the reform strategies to be mastered by teachers and integrated with existing school structures. In the case of the two SFA schools and three of the DI schools, the 2 years of implementation completed at the time of this study may have been too limited to produce strong gains in achievement (see Fullan, 2000). For SFA-2, the noticeably low school climate scores, implementation concerns, and somewhat negative early achievement patterns may portend ultimate failure, unless positive changes occur. For the three DI schools that were completing their 4th year, the time window for showing success, especially for the multiple district stakeholder groups, is rapidly closing. More immediate indications of implementation improvement and student achievement gains will be expected for support to continue.

Today, in the era of NCLB, there is greater explicit demand than ever before for educational research to identify for practitioners and policy makers the programs and models "that work." Looking at the mixed findings that occur when complex models are employed in diverse real-world contexts (Berliner, 2002; Slavin 2002), the latter goal seems unlikely to be attained via a simple box score or cumulative effect size statistic. For example, detractors of CSR (or of the specific models examined) may be inclined to invoke the present results to bolster a negative view of effectiveness. The model developers, on the other hand, could justifiably argue that positive outcomes require full implementation (that was lacking here). A third view might be a consumer-oriented definition of model effectiveness as a combination of both ease and cost of implementation (process) and likely student achievement (outcomes). Each field experiment thus constitutes an additional trial yielding context-specific results. In this case, the implication is that schools' ability and willingness to implement DI fully were hampered by factors limiting the effective acquisition of resources (e.g., technical assistance, curricular materials, and funding), school ownership, and teachers' involvement in professional development. SFA failed to show achievement gains and had some minor implementation weaknesses, but had been adopted for only 2 years. Based on these results, the presence or absence of similar conditions in other urban districts should therefore relate to these models' success.

REFERENCES

- Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle, WA: Educational Achievement Systems.
- Berends, M., Chun, J., Schuyler, G., Stockley, S., & Briggs, R. J. (2002). *Challenges of conflicting school reforms: Effects of New American Schools in a high-poverty district*. Santa Monica, CA: RAND.
- Berends, M., Kirby, S. N., Naftel, S., & McKelvey, C. (2001). *Implementation and performance in New American Schools: Three years into scale-up*. Santa Monica, CA: RAND.
- Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher*, 31(8), 18–20.
- Bobbett, J. L., Ellett, C. D., Teddlie, C., Olivier, D., & Ruggett, J. (2002, April). *School culture and school effectiveness in demonstrably effective and ineffective schools*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Borman, G. D. (2002). *Experiments for educational evaluation and improvement*. Unpublished manuscript, University of Wisconsin-Madison.
- Borman, G. D., Hewes, G., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research*, 73(2), 125–230.
- Bryk, A., & Schneider, B. (2002). *Trust in school: A core resource for improvement*. New York: Russell Sage Foundation.
- Butler, P. A. (2003). Achievement outcomes in Baltimore City Schools. *Journal of Education for Students Placed At Risk*, 8, 33–60.
- Butler, E. D., & Alberg, M. J. (1991). *The Tennessee School Climate Inventory: Resource Manual*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.
- Carlson, C. D., & Francis, D. J. (2002). Increasing the reading achievement of at-risk students through Direct Instruction: Evaluation of the Rodeo for Teacher Excellence (RITE). *Journal of Education for Students Placed At Risk*, 7, 141–166.
- Cuban, L. (1993). *How teachers taught: Constancy and change in American classrooms, 1890–1980* (2nd ed.). New York: Teachers College Press.
- Datnow, A., Hubbard, L., & Mehan, H. (2002). *Extending educational reform: From one school to many*. New York: Routledge-Falmer.
- Datnow, A., & Stringfield, S. (2000). Working together for reliable school reform. *Journal of Education for Students Placed At Risk*, 5, 183–204.
- Desimone, L. (2002). How can comprehensive school reform models be successfully implemented? *Review of Educational Research*, 72, 433–480.
- Engelmann, S. (1969). *Preventing failure in the primary grades*. Chicago: Science Research Associates.
- Englemann, S., Becker, W. C., Carnine, D., & Gersten, R. (1988). The Direct Instruction Follow Through Model: Design and outcomes. *Education and Treatment of Children*, 11, 303–317.
- Feuer, M. J., Towne, L., & Shavelson, R. J. (2002). Scientific culture and educational research. *Educational Researcher*, 31(8), 4–14.
- Fullan, M. (2000). The return of large scale reform. *Journal of Educational Change*, 1, 5–28.
- Hatch, T. (2001). What does it take to "Go to Scale"? Reflections on the promise and the perils of comprehensive school reform. *Journal of Education for Students Placed At Risk*, 5, 339–354.
- Haynes, N. (1998). Overview of the Comer School Development Program. *Journal of Education for Students Placed At Risk*, 3, 3–11.
- Herman, R. (1999). *An educators' guide to schoolwide reform*. Arlington, VA: Educational Research Service.
- Hopfenberg, W. S., & Levin, H. M. (1993). *The Accelerated Schools resource guide*. San Francisco: Jossey-Bass.
- Kaufman, M. (1972). *The effect of the DISTAR Instructional System: An evaluation of the 1971–1972 Title I program* (Report No. PS006218). Winthrop, MA. (Eric Document Reproduction Service No. ED 070525)
- Kaufman, M. (1973). *The effect of the DISTAR Instructional System: An evaluation of the 1972–1973 Title I program* (Report No. PS007945). Winthrop, MA. (Eric Document Reproduction Service No. ED 110171)
- Kaufman, M. (1974). *The effect of the DISTAR Instructional System: An evaluation of the 1973–1974 Title I program* (Report No. PS007944). Winthrop, MA. (Eric Document Reproduction Service No. ED 110170)
- Kuder, G. F., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151–160.
- Levin, H. (1993). *Learning from Accelerated Schools*. Unpublished paper, Stanford University, Palo Alto, CA.
- Mac Iver, M. A., Kemper, E., & Stringfield, S. (2003). *The Baltimore curriculum project: Fourth-year report*. Report No. 62. Baltimore, MD: Johns Hopkins University, Center for Research on the Education of Students Placed At Risk. Retrieved on February 11, 2004 from <http://www.csos.jhu.edu/crespar/techReports/Report62.pdf>.
- Miles, M. B., & Huberman, A. M. (Eds.). (1994). *An expanded sourcebook: Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.
- Muncey, D. E., & McQuillan, P. J. (1996). *Reform and resistance in schools and classrooms: An ethnographic view of the Coalition of Essential Schools*. New Haven, CT: Yale University Press.
- Muñoz, M. A., & Dossett, D. H. (2004). Educating Students Placed At Risk: Evaluating the Impact of Success for All in Urban Settings. *Journal of Education for Students Placed At Risk*, 9, 261–277.
- Muñoz, M. A., Dossett, D., & Judy-Gullans, K. (2003, April). *Targeting at-risk students: Evaluating the impact of Success for All in urban settings*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Northwest Regional Education Laboratory. (1998). *Catalog of school reform models: First edition*. Portland, OR: Author.
- Northwest Regional Education Laboratory. (2000). *Catalog of school reform models: Second edition*. Portland, OR: Author.
- Onwuegbuzie, A., & Teddlie, C. (2003). A framework for analyzing data in mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 351–383). Thousand Oaks, CA: Sage.
- Pogrow, S. (1998). What is an exemplary program, and why should anyone care? A reaction to Slavin and Klein. *Educational Researcher*, 27(1), 22–29.
- Pogrow, S. (2000). Success for All does not produce success for students. *Phi Delta Kappan*, 82, 67–81.
- Porter, A. C. (1994). National standards and school improvement in the 1990s: Issues and promise. *American Journal of Education*, 102, 421–449.
- Porter, A. C., Floden, R., Freeman, D., Schmidt, W., & Schwillie, J. (1988). Content determinants in elementary school mathematics. In D. Grouws & Thomas Cooney (Eds.), *Perspectives on research on effective mathematics teaching* (pp. 343–356). Reston, VA: National Council of Teachers of Mathematics.
- Ross, S. M. (2000). *Telling the story of comprehensive school reform: Using third-party evaluators in school districts*. Washington, DC: New American Schools.
- Ross, S. M. (2001). *Creating critical mass for restructuring: What we can learn from Memphis*. Charleston, WV: AEL Policy Briefs.
- Ross, S. M., Nunnery, J. A., Goldfeder, E., McDonald, A. J., Rachor, R., Hornbeck, M., et al. (2002). *Using School Reform Models to Improve Reading Achievement: A Longitudinal Study of Direct Instruction and Success For All in an Urban District*. Memphis, TN: Center for Research in Educational Policy, The University of Memphis.
- Ross, S. M., Nunnery, J., Goldfeder, E., McDonald, A. J., Rachor, R., Hornbeck, M., et al. (2003). *Progress and options regarding the implementation of Direct Instruction and Success For All in Toledo Public Schools: Final Report*. Memphis, TN: The University of Memphis, Center for Research in Educational Policy.

- Ross, S. M., Sanders, W. L., Wright, S. P., Stringfield, S., Wang W., & Alberg, M. (2001). Two- and three-year achievement results from the Memphis restructuring initiative. *School Effectiveness and School Improvement, 12*, 323-346.
- Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.
- Slavin, R. E., & Fashola, O. S. (1998). *Show me the evidence?* Thousand Oaks, CA: Corwin.
- Slavin, R. E., & Madden, N. A. (1993, April). *Multi-site replicated experiments: An application to Success For All*. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.
- Slavin, R., & Madden, N. (2001). *One million children: Success for All*. Thousand Oaks, CA: Corwin.
- Smith, L., Maxwell, S., Lowther, D., Hacker, D., Bol, L., & Nunnery, J. (1997). Activities in schools and programs experiencing the most, and least, early implementation success. *School Effectiveness and School Improvement, 8*, 125-150.
- Sterbinsky, A. (2001). *Comprehensive School Reform Teacher Questionnaire: Criterion Related Validity Study*. Technical Report. Memphis, TN: Center for Research in Educational Policy, The University of Memphis.
- Sterbinsky, A., & Ross, S. M. (2003). *Summary of CSRTQ Reliability Studies*. Technical Report. Memphis, TN: Center for Research in Educational Policy, The University of Memphis.
- U.S. Department of Education. (1999). *Guidance on the comprehensive school reform program*. Washington, DC: Author.
- Wang, M. C., Haertel, G. D., & Walberg, H. (1997). *What do we know? Widely implemented school improvement programs*. Philadelphia: Temple University Center for Research in Human Development and Education.