

Student Gains in a Privately Managed Network of Charter Schools Using Direct Instruction

**Richard Cross is the former Director of Assessment of Advantage Schools, Theodor Rebarber the former Chief Education Officer, and Steven Wilson the former Chairman and CEO. The authors contributed equally to the manuscript and are listed alphabetically. Correspondence should be directed to Steven Wilson at StevenFWilson@attbi.com. The authors appreciate the assistance of Professor Daniel Mueller of Indiana University and Professor Michael Welker of Franciscan University of Steubenville with an earlier version of this manuscript.*

Abstract: A private education management company, Advantage Schools, was established in 1996 to open and operate public charter schools under contract to local boards serving students primarily from urban, economically disadvantaged families. Advantage's objective was to create new schools where students would learn at an accelerated rate, learn to read in kindergarten, attain national norms in the elementary grades, and ultimately undertake college level work in the last 2 years of high school. The approach had four key components: charter public schools as the organizational form; a rigorously academic school design; a school culture centered on student achievement; and the discipline of private management. A research-based school design was developed centering on Direct Instruction in the elementary grades; an

extended school day and year; a structured behavior system that stressed positive reinforcement over admonishment; and a leadership, staffing, and supervisory model that emphasized accountability for student achievement. Direct Instruction titles, including *Reading Mastery* (Engelmann, S., & Bruner, 1995), *Reasoning and Writing* (Engelmann, S., Arbogast, Seitz Davis, Grossen, & Silbert, 1993), and *Connecting Math Concepts* (Engelmann, S., Carnine, D., Engelmann, O., & Kelly, 1994), formed the core of the Advantage elementary curriculum.

The results from this evaluation of academic achievement during the 1999–2000 school year suggest the potential of the model. Over 7,600 students, predominately from low-income urban families, were educated in the new schools during the study period. Data from 5,874 students in Grades k–7 who were tested in both the fall and the spring of that year were included in this evaluation. Aggregated across subject areas and grade levels, these students gained an average of 3.6 NCE (effect size 0.19) on the SAT-9 and 10.1 NCE (effect size 0.52) on the WRMT-R. These results suggest that the four-part approach may offer an effective and replicable approach to urban schooling that can be brought to scale with existing financial resources.

Introduction

A private company, Advantage Schools, was formed in 1996 to provide new educational choices to urban families. There were already instances of individual charter schools in inner

city communities serving a few hundred students that outperformed the surrounding traditional public schools, sometimes dramatically. Other charters disappointed, performing no better or worse. The challenge therefore remained to develop interventions in urban schooling that could be brought to scale to serve thousands of students, and at current spending levels. The company's goal was to create a network of high quality public schools that enable students—regardless of socioeconomic background—to reach high levels of academic achievement.

Working under contract to local charter school boards, Advantage opened 20 public schools primarily in urban areas in nine states and the District of Columbia. The first two schools opened in the fall of 1997 in Rocky Mount, North Carolina and Phoenix, Arizona. In the following 3 years, additional schools were opened in Charlotte; Chicago; Detroit; Jersey City and Newark; Philadelphia; Dallas; Houston; Midland; San Antonio; Washington; Worcester and Malden, Massachusetts; Kalamazoo and Benton Harbor, Michigan; Fulton County, Georgia; and Albany, New York. Some schools elected over time to self-manage or to contract with another management company. During the school year of this study, 1999–2000, Advantage operated 14 schools for the entire academic year.

The School Design

Overview of Advantage Schools

Most of the new schools opened as elementary schools with Grades k–5 and approximately 550 students, and then expanded by a grade each year as students were promoted, many with the intention of eventually extending through the 12th grade. The oldest school, in Phoenix, Arizona, extended through the seventh grade at the time of this evaluation. In the long term, Advantage sought to demonstrate that most students, regardless of background, who began in the program from the

early grades could be rigorously prepared to attend four-year colleges upon high school graduation, with a significant number having attained the prestigious International Baccalaureate standing (International Baccalaureate Organization, 2001).

The founders of the organization combined four elements—the governance structure of the charter school (including parental choice); a distinctive school culture centered on academic achievement; the focus and accountability of a private company overseeing school implementation; and a powerful, research-proven school design (developed by Chief Education Officer Theodor Rebarber)—to elevate student achievement levels. Direct Instruction was used in all schools for instruction in the core subjects of reading, language, and math, and the program benefited from the expertise of many educators with Direct Instruction expertise, including Vice President Kathleen Madigan.

Creating high performing new schools requires sustained engagement with all of the elements of effective schools. The founders believed the greater autonomy and flexibility afforded public charter schools, compared to traditional district schools, would permit a rigorous school design to be implemented with fidelity and sustained over time. As both parents and teachers would have chosen their new school, the first as customers and the second as employees, a greater commitment to fulfilling the school's academic mission might be engendered. Equally important, newly created schools would not inherit the culture of any existing public school (which frequently emphasizes compliance with external rules over the academic results of its clients) and a new culture focused on staff accountability and high expectations for student achievement could be created.

This culture was to be sustained and enhanced by the assured, steady leadership of the schools' directors; accordingly, consider-

able attention was given by Advantage to recruiting directors and to determining candidates' fit with the client board, community, and the Advantage model. In addition to a broad range of entrepreneurial, academic management, and leadership skills, the successful school director would be a passionate exponent of a liberal education for all students and foster a culture among all members of the school community that expects and assumes the academic success of every child.

If Advantage was to be held accountable to its client school boards for student achievement, it in turn must have the necessary authority to achieve these results. Under its management contracts to local school boards, the company had broad responsibility for launching the new school and managing its day-to-day operations in fulfillment of the school's charter.

Advantage typically identified and secured the site, developed plans and obtained permits to build the school or convert an existing structure, secured the capital for construction, and oversaw its completion. Advantage recruited the school's executive leadership team (the school director, the assistant director for instruction, the behavior intervention specialist, and the business manager) and faculty, publicized the opening of the new program, and enrolled the student body. Once open, Advantage oversaw the ongoing operations, reporting to the school's board.

While the governance benefits of charter schools were important, the founders believed that policymakers had overstated the gains that these benefits would alone bestow, while understating the importance of a research-proven curriculum and sound instruction. Some academic gains might result merely from the staff engagement and sense of mission that the charter structure frequently engenders. But more important was the unusual autonomy afforded charters, including from the policies and dictates of the central district office. That autonomy might facilitate the precise implementation of a comprehensive

school design, including high academic standards, a highly structured curriculum, and a uniform behavior system.

The founders believed that all students should benefit from a liberal education—a rigorous academic education that provides sound preparation for a four-year baccalaureate program, the type of education traditionally reserved for the nation's elite—regardless of their backgrounds or society's assumption of their likely future vocations. A liberal education today would combine the classical focus on language and mathematics with the modern disciplines of science, history, and literature. The systematic study of these subjects would convey important knowledge and skills, cultivate aesthetic imagination, and teach students to think critically and reflectively about the world in which they live.

Elementary School Program

Advantage's school design was grounded in empirical research on effective instruction (Adams & Engelmann, S., 1996) and effective school characteristics (Purkey & Smith, 1983). Properly implemented at the client schools, the design was intended to produce strong annual gains in student achievement that are replicable across sites and student populations. The core principles of the design were detailed curriculum and lesson plans; intensive professional development that is tightly aligned with the curriculum; frequent assessment of student mastery; and ongoing collaborative problem-solving, based on student data, to diagnose and intervene with learning problems before students fall behind in lesson mastery.

Key components of the elementary design were the Direct Instruction programs in the core subjects of reading, language, and math; world language instruction starting in the second grade (generally Spanish); art and music instruction; a character education program; a Code of Civility for the entire school community; a behavior system that emphasized positive

reinforcement over admonishment; a pervasive culture centering on expectations of high achievement; a newly renovated facility; an unusual level of investment in pre and in-service professional development and academic oversight; an extended school day and school year; and school uniforms for all students.

In the elementary grades, Advantage used the Direct Instruction curriculum to teach all children reading, writing, and math. The major Direct Instruction programs used included *Language for Learning* (Engelmann S., & Osborn, 1998), *DISTAR Language* (Engelmann S., & Osborn, 1997), *Reading Mastery* (Engelmann, S., & Bruner, 1995), *DISTAR Arithmetic* (Engelmann, S., 1997), *Connecting Math Concepts* (Engelmann S., Carnine, D. Engelmann, O., & Kelly, 1994), *Reasoning and Writing* (Engelmann, S., Arbogast, Seitz Davis, Grossen, & Silbert, 1993), *Spelling Mastery* (Dixon, Engelmann, S., Bauer, Steely, & Wells, 1990), *Expressive Writing* (Engelmann, S., & Silbert, 1985), and all *Corrective* programs, including *Decoding* (Engelmann, S., Johnson, & Carnine, L. 1988), and *Comprehension* (Engelmann, S., Haddox, Osborn, & Hanner, 1998). Direct Instruction has been credited with a high level of effectiveness with all levels of students, but particularly with students from underprivileged backgrounds (Adams & Engelmann, S., 1996).

These programs provided every Advantage elementary teacher with Direct Instruction lessons, honed over years of study with actual students. Each Direct Instruction lesson offers polished and effective instructional strategies for teaching key concepts and skills. When correctly implemented, students are grouped according to their present knowledge of the subject and engaged in a fast-paced, interactive learning dialogue. Direct Instruction is designed to ensure that all students in a working group grasp the concept being taught and are able to verbalize or write the response that demonstrates this mastery; only then do they move on to other topics. Because students

experience frequent academic success—some for the first time—Direct Instruction can build student self-esteem and enjoyment of learning. Properly implemented, children make steady academic progress, take pride in their accomplishments, and acknowledge praise as something they have earned (Adams & Engelmann, S., 1996).

In three schools, Advantage had begun to implement one or both of two additional programs, Junior Great Books (Junior Great Books Foundation) and Accelerated Reader (Renaissance Learning). The Junior Great Books program, which begins in kindergarten, introduces students to an array of classic stories, fairy tales, fables, and legends. Relying heavily on Socratic discussion, the elementary program prepares students for seminar-style courses that would be used increasingly at the middle and high school levels. Junior Great Books allows students to apply and sharpen their reading skills, while learning to discuss their interpretations of literature. The Accelerated Reader program was chosen to enhance the Direct Instruction reading curriculum by encouraging students to read at home for an hour every day. Students' reading is monitored with simple computer-based tests that provide some assurance that each book was, in fact, read. In addition to the core subjects of reading, writing, and mathematics, Advantage elementary students also studied history, science, foreign language, music, and art, with the goal of mastery of a broad array of important knowledge and skills.

Closely related to the character education program was the strong sense of discipline Advantage wanted to see in force at every school. The Code of Civility, a blueprint for appropriate conduct, describes how these concepts play themselves out in positive behavior, consequences for misbehavior, and interventions to prevent misbehavior from reoccurring. Advantage saw behavior as an instructional challenge that is best taught in an environment that relies on explicit teaching of correct behavior and consistent application of positive

reinforcement and encouragement when such behavior is demonstrated. The goal was to maintain the most effective environment's ratio of positive to corrective feedback to students of approximately four to one—far different from the predominantly negative feedback that most disadvantaged students receive. Parents, teachers, and school leaders explicitly agreed to support and adhere to the Code.

Professional Development, Support, and Accountability

In the Advantage model, greater resources were directed at ongoing professional development and curriculum implementation than in most public schools. First, leadership team members and teaching staff participated in a 2-week on-site intensive training program in the school's curriculum and behavior system, led by Direct Instruction trainers. In addition, the school leadership teams came together each summer for a national leadership conference to learn new skills, share best practices with their colleagues, and plan for the new school year.

Professional development was not limited to intensive summer meetings. The Advantage organization included multiple layers of professionals who were responsible for ongoing staff development and problem solving. The academic operations of every two schools were supported by a curriculum implementation specialist (CIS) recruited nationally from experts in Direct Instruction implementation. The CISs were coordinated by the vice president for instruction, who was charged with ensuring the fidelity of the school design's implementation at each location. Each school included two full-time staff members who trained teachers and monitored the implementation of the academic program: the assistant director for instruction (ADI) and the behavior intervention specialist (BIS). The BIS provided professional development in classroom management techniques as well as assistance to

teachers with particularly challenging students. The ADI provided training in instructional methods, was responsible for the development of the instructional staff, and oversaw the overall academic implementation, including scheduling of instruction, grouping, and placement of students. Neither the BIS nor the ADI had regular teaching responsibilities. In addition, the teachers at each grade level met weekly to review lesson progress and mastery data. The head teacher for each grade met weekly with all of the other lead teachers, the ADI, School Director (principal), and the CIS.

The corporate office attempted to hold every level of the organization accountable for student achievement and for their efforts to constantly improve teaching and learning. Consistent with this culture, both the ADI and the CIS made use of an in-class coaching model. They visited classrooms, observed instruction, and, when necessary, intervened to provide immediate feedback or instructional modeling to the teacher. In addition, the CIS reviewed weekly school records on student grouping, scheduling, the implementation of the behavior management program, and Code of Civility.

Because Direct Instruction has been studied with thousands of students (Adams & Engelmann, S., 1996), the schools benefited from expectations of the rate of progress in the curriculum by which to assess their own performance. Each week, the CIS inspected data on lesson progress and periodic mastery tests from every instructional group and compared these results with targets for the groups. CISs could rapidly intervene in instructional groups that were moving slower than expected and assist staff with pacing, delivery, and classroom management. Because the CISs were familiar with the individual student make-up of each group, they could monitor lesson progress and mastery for all students and determine which students would benefit from reassignment to another instructional group with a different pace, either to firm up their mastery or to encounter new material more rapidly.

Evaluation Methods

Advantage evaluated its client schools by several measures. As every school was a school of choice, enrollment levels and reenrollment rates each fall revealed parents' confidence in their schools. Surveys of parents added other information on parent satisfaction. Weekly lesson progress and periodic mastery assessments embedded in the curriculum tracked the rate and depth of learning. State assessments provided a basis for comparison of student performance to expectations in each state. But only national standardized tests provided a consistent and objective basis for assessing the academic performance of a school by comparing it to national norms. These nationally normed standardized tests provide the results reported in this evaluation.

However, even with standardized tests, evaluation of an educational innovation such as Advantage Schools is fraught with stubborn methodological problems. Ideally, students' progress would be compared to that of a randomly assigned control group that was similar in every way except that it did not receive the Advantage education. Short of this, it would be desirable to compare Advantage Schools' results to a control group that, while not randomly assigned, was similar in (a) socioeconomic characteristics, (b) academic characteristics, and (c) family support for education. However, identification, recruitment, and assessment of such a group was well beyond the means of this evaluation. Instead, we compare the performance of Advantage school students to national test norms. This allows us to draw conclusions about the rate of progress of these students relative to the rate of progress of students who contributed to the national norms. In this evaluation, we address the question of whether Advantage students experienced achievement gains that were less than, comparable to, or superior to their peers nationally.

Participants

Twice during the 1999–2000 academic year, students in all grades took standardized achievement tests. The pretest included 7,379 students, and the posttest included 7,209 students. Since the principal question is about academic gain, only students with valid pretests and posttests are included in the reporting of results. Of the 7,687 students who were enrolled in Grades k–7 in Advantage schools at some time during the academic year, 5,874 had test results that could be matched for both pretest and posttest. Results from these matched students form the data set used in this evaluation. The discrepancy between the number of matches and enrolled students is explained by students who (a) enrolled in the schools after the middle of September when the fall test was given, (b) left the school before the posttest in late April, or (c) were absent during one or both rounds of testing. This discrepancy is substantial, but most of the Advantage schools were in areas where one typically observes high student mobility.

Students with no English proficiency were not tested. Unfortunately, the number of students excluded for this reason is not known. Students judged to have some, but limited, English proficiency were included in the assessment.

Student and family characteristics were not considered in the process of enrollment in Advantage schools. All applicants were admitted except where there were more applications than capacity; in these cases, admission was based on lottery. Demographic characteristics of Advantage schools were generally similar to central city schools, with a high percentage of students who were eligible for the Federal Free or Reduced-Price Lunch (FRL) program. Table 1 lists the spring enrollment of each school, estimated during March and April, and the percentage of students qualifying for the FRL program. We believe that, due to under-reporting by parents, the figures reported in Table 1 significant-

ly understate the actual percentage of economically disadvantaged families the schools served.

Advantage served a higher percentage of economically disadvantaged families than do American schools on average. Seventy-one percent of students system-wide were qualified for FRL. In comparison, the Department of Education (National Center for Educational Statistics, 2000) reports that in 1993–1994, 33% of students nationwide participated in the FRL program and in inner cities elementary schools, 52% participated. The comparison is not perfect because (a) both sets of data are based on parents' willingness to report low income, (b) *qualifying* for the program (basis for Advantage data) and actually *participating* in it (National data) are somewhat different, and (c) Advantage data are from 1999–2000 and the national data are from 1993–1994. Nonetheless, it appears to be safe to conclude that the Advantage schools served a high proportion of economically disadvantaged families.

At the time of admission, Advantage students were assigned to a grade in the standard manner, according to their chronological age or their prior placement. The grade level to which the student was assigned at the time of testing was recorded as the grade level used for test administration and scoring. Thus, none of the evaluation results reflect out-of-level testing. For schools reporting information on gender (over 90% reported this information), the ratio of girls to boys was 49:51.

Measures and Test Administration

Each fall and spring, Advantage students at all grade levels took subtests from the Stanford Achievement Test-Ninth Edition (SAT-9; Harcourt Brace, 1996) and students in Grades k–2 also took the Woodcock Reading Mastery Tests-Revised (WRMT-R; Woodcock, 1998). Pretests (fall) for most students were administered during the period from September 15 through October 15, 1999 and posttests

(spring) were given between March 16 and April 15, 2000. For approximately 30% (300) of the students in the Phoenix school, SAT-9 results from the previous spring (1999) were used as pretests. Results from all students who had pretest and posttest results on SAT-9 or the WRMT-R were analyzed.

Stanford Achievement Test–Ninth Edition (SAT-9). Students in Grades k–7 took subtests of the SAT-9 Form S or Form T (Harcourt Brace, 1996). All students took the Mathematics subtest. Students in Grades k–2 took the Listening subtest and those in Grades 3–7 took the Language subtest. Students in k–2 did not take an SAT-9 reading test as their reading skills were assessed with the WRMT-R (see below). Students in 3–7 completed the Reading subtest of the SAT-9.

Table 1

Percentage of Students Qualifying for Free or Reduced-Price Lunch (FRL), by School

School	Enrollment (<i>N</i>)	% FRL
Albany, NY	380	67
Chicago, IL	698	94
Charlotte, NC	422	53
Dallas, TX	333	89
Houston, TX	447	74
Jersey City, NJ	492	76
Kalamazoo, MI	582	81
Midland, TX	601	61
Newark, NJ	482	77
Philadelphia, PA	541	53
Phoenix, AZ	1,031	85
San Antonio, TX	652	71
Washington, DC	451	77
Worcester, MA	575	26
Advantage Total	7,687	71

The SAT-9 was designed and developed to measure skills, knowledge, and understanding important to growth across the curriculum in the nation's public and private schools. The SAT-9 reflects over 70 years of test development and research on measuring achievement and critical thinking skills in reading, mathematics, language arts, listening, social studies, and science. The scope and sequence of test content were developed following the review of national and state curricula and curriculum standards, current textbook series and instructional materials, and educational research. During test development, all items were reviewed and statistically checked for possible gender, ethnic, and cultural bias prior to the publication of the final form of the tests (Harcourt Brace, 1996). Data from a nationally representative sample of public and private schools were collected in 1995 and used to form the national norms. Standardization and equating of scores across test forms occurred in the spring and again in the fall of 1995. KR-20 internal consistency estimates for all tests, levels, and forms range from .81 to .96. SAT-9 Forms S and T are equivalent forms and adjacent test levels were successfully equated. Alternate forms equivalence is documented with correlation coefficients in the SAT-9 tests ranging from .79 to .93, with a median equivalent forms coefficient of .88. Corresponding raw score test means and variance estimates differed by no more than two points. Median correlations for consecutive levels of the test (e.g., third grade and fourth grade) were .90 for Reading, .86 for Mathematics, .85 for Language, and .79 for Listening (Harcourt Brace, 1996). The SAT-9 provides norms for fall and spring administration. This allows computation of growth across a school year in comparison to the norm group (Harcourt Brace, 1996).

Woodcock Reading Mastery

Tests-Revised (WRMT-R). Students in Grades k-2 took the WRMT-R Form G (Woodcock, 1998). The WRMT-R is one of the most respected standardized reading

tests. Students in this study were administered three of the tests: Word Identification, Word Attack, and Passage Comprehension. The WRMT-R is widely used in educational program evaluation and research. We used the WRMT-R NU-Normative Update (Woodcock, 1998) for deriving norm-based scores. These norms are based on a stratified random sample of students in public and private schools in the United States who were tested in 1995 and 1996. The WRMT-R correlate highly with the Iowa Test of Basic Skills (.78 to .83), the Wide Range Achievement Test (.86 to .88), and other reading achievement tests (Woodcock, 1998). The WRMT-R provides monthly norms so growth across a period of months can be understood in comparison to the norm group.

Test Administration. Each school received SAT-9 testing materials, which included test booklets, answer documents, and instructions for test administration from the publisher or local district. Publisher-supplied practice materials for Grades k-2 were distributed to students a few days before the test to familiarize those who may never have taken a standardized multiple choice test with testing procedures. Teachers were instructed to follow all testing procedures exactly as specified in the administration manual. Typically, the classroom teacher administered the test to a classroom of students over 3 consecutive days, in the morning, for periods not exceeding 120 min. For students who were absent during the regular classroom administration, special small group testing was made available. Testing group sizes varied from 5 to 30 students, except for students requiring special accommodation. Sixty-one students (0.85% of the posttests) were provided testing accommodations as required by their Individual Education Plans. Accommodations typically included extending the duration of the test, or testing in smaller groups. Students with limited English proficiency were tested in English without any accommodations.

Consistent with the publisher's standard procedures, teachers administered the tests and returned them directly to the testing service or local district for scoring. For each administration of the test, the publisher provided a data file including item responses and scores for each student on each subtest to Advantage's corporate office. Pretest and posttest data for each student were hand-matched based on a combination of student name, sex, grade, and date of birth.

School administrators entered WRMT-R raw scores into *ASSIST for the WRMT-R* (AGS Software, 1998) to derive standard scores. Data files containing student name, grade, and scores for each subtest were generated by *ASSIST* and were sent to an independent testing firm, Data Analysis and Testing Associates of Concord, Massachusetts, for compilation. Data for this article were taken from the testing firm's data files.

Results

All computations in these analyses used the NCE score, which is a norm-based standard

score resulting from the division of the normal curve into 99 equal units. Like the percentile score, the NCE describes each student's relative rank within the normative sample. If a student makes progress across time that is typical of that seen in the norm group, the NCE score would remain constant. If the student makes faster progress than is typical, the NCE would increase. However, unlike percentile, NCE scores are equal interval and therefore may correctly be compared across different parts of the scoring range. The NCE score is well suited to estimating changes in performance over time as a change or gain score. The NCE score is also useful for averaging results across groups and across tests. The NCE scale has a mean of 50 and a standard deviation of 21.06.

Table 2 summarizes the pre and posttest results on the SAT-9 subtests. This table gives the grade levels of students who completed each subtest (second column), the number of scores (third column), pretest and posttest means and standard deviations (fourth and fifth columns), and several descriptions of change in performance from pretest to posttest. The simplest description of change is simply gain score (sixth column). In order to

Table 2
SAT-9 Results, All Grades and All Subtests, in NCE

Subject	Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
Reading	3-7	2,437	36.1 (19.1)	39.5 (19.9)	3.4 (11.0)	0.17	15.23	< .0001
Math	k-7	5,631	35.5 (19.1)	38.4 (19.4)	2.8 (13.3)	0.15	15.90	< .0001
Language	3-7	2,246	36.8 (19.5)	42.2 (19.4)	5.3 (12.8)	0.27	19.72	< .0001
Listening	k-2	2,227	38.7 (19.4)	43.4 (19.4)	4.7 (15.9)	0.24	13.87	< .0001
Composite		5,874	36.5 (19.4)	40.2 (19.4)	3.6 (13.2)	0.19	21.00	< .0001

put the size of the gain in context, we report effect sizes of the change (seventh column). These are the gain scores divided by the pooled standard deviation. The effect size describes the change in standard deviation units. The scores were also subjected to tests of statistical significance. Test results were entered into a matched pairs procedure that compares means using a two-tailed paired *t* test. These tests results in *t*-values (eighth column) and *p*-values (ninth column) are shown in the table. The *p*-values give the probability of obtaining a gain score of this size due to chance alone. Small *p*-values are interpreted to mean that the results are not likely due to chance alone. We used the Dunn-Sidak method to adjust the *p*-value so that the set of tests reported on each table would have a family-wise alpha level of no more than 0.10. Small *p*-values, however, should not be interpreted to mean that the results are large enough to be educationally important. To make this important determination, we must examine the size of the gain scores and effect

sizes and judge whether this size of change is meaningful in an educational context.

The results reported in Table 2 indicate that Advantage students were performing below average when they entered the school year. Pretest scores ranged between 35 and 40th NCE (24th to 31st percentiles). Posttest scores were also below average, though they were higher; they ranged between the 38th and 43rd NCE (29th to 37th percentile). Averaged across grades, Advantage students made gains on all subtests of the SAT-9. The largest gain was in 5.3 NCE in Language (Grades 3–7), the smallest was in 2.8 NCE in Mathematics (Grades k–7) and the average across all tests was 3.6 NCE (effect size 0.19). The effect sizes suggest that these gains should be considered to be small to moderate in size. The *p*-values indicate that all of these differences are statistically significant, that is, results of this size are unlikely due to chance alone.

Table 3
SAT-9 Reading Subtest Results, by Grade, in NCE

Reading							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
3	534	40.1 (17.5)	41.2 (19.0)	1.1 (11.5)	0.06	2.22	0.026*
4	754	33.2 (19.9)	37.0 (19.3)	3.8 (11.0)	0.20	9.51	< .0001
5	729	34.9 (21.0)	39.6 (19.1)	4.7 (10.6)	0.23	11.86	< .0001
6	334	38.8 (20.2)	42.0 (20.6)	3.2 (10.5)	0.16	5.62	< .0001
7	86	37.1 (18.3)	40.5 (17.8)	3.4 (9.3)	0.19	3.36	0.001
Composite	2,437	36.1 (19.9)	39.5 (19.5)	3.4 (11.0)	0.17	15.23	< .0001

* Not significant with the Dunn-Sidak correction.

Table 3 reports the results of the analysis of SAT reading subtest for each Grade 3 through 7. Each grade level was clearly performing below the SAT average on both the pretest and posttest. The grades ranged from NCE 33 to 40 on the pretest and from 37 to 42 on the posttest. Each grade showed a gain from pretest to posttest that was statistically significant. The gain for third-grade students was very small (1.1 NCE; effect size 0.06). At grade levels 4 through 7, gains were between 3.2 and 4.7 NCE, which could be considered to be small to moderate.

Results for the mathematics subtest are shown in Table 4. These results are quite variable across grades. Results for kindergartners had a

different pattern than any of the other grades. Kindergartners had the highest pretest scores, the highest posttest scores and the largest gain (6.5 NCE, effect size 0.32). In contrast, though first graders entered the year with the second highest pretest score, they showed a slight decline from pretest to posttest (-0.5 NCE, effect size -0.03). Grades 3 through 6 showed more consistent results with gains of 3 to 4 NCE and effect sizes 0.15 to .023. Overall, math subtest results do not lend themselves to simple interpretation and the composite gain of 2.8 NCE could be misleading because individual grade levels diverge so greatly.

Language subtest results, reported in Table 5, indicate gains at every grade level. With the

Table 4
SAT-9 Math Subtest Results, by Grade, in NCE

Math							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
k	893	41.2 (20.7)	47.7 (20.3)	6.5 (17.1)	0.32	11.39	< .0001
1	942	40.0 (18.1)	39.5 (19.1)	0.5 (14.8)	-0.03	-1.02	0.30
2	939	33.7 (18.0)	34.5 (18.7)	0.8 (12.1)	0.04	2.03	0.04*
3	829	32.6 (18.5)	35.9 (19.2)	3.3 (12.6)	0.17	7.52	< .0001
4	824	32.6 (18.9)	35.9 (18.2)	3.3 (11.1)	0.18	8.54	< .0001
5	781	32.3 (18.5)	36.4 (18.2)	4.1 (10.6)	0.23	10.80	< .0001
6	338	35.3 (19.1)	38.3 (19.1)	3.0 (10.4)	0.15	5.20	< .0001
7	85	35.3 (15.5)	36.3 (14.2)	1.0 (8.2)	0.06	1.08	0.28
Composite	5,631	35.5 (19.1)	38.4 (19.4)	2.8 (13.3)	0.15	15.90	< .0001

* Not significant with the Dunn-Sidak correction.

Table 5
SAT-9 Language Subtest Results, by Grade, in NCE

Language							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
3	660	34.9 (18.6)	42.1 (17.9)	7.2 (14.1)	0.40	13.32	< .0001
4	616	37.1 (17.9)	41.3 (18.9)	4.2 (11.9)	0.23	8.84	< .0001
5	598	35.9 (20.9)	41.4 (20.8)	5.5 (12.5)	0.26	10.70	< .0001
6	289	41.8 (20.9)	44.2 (20.4)	2.4 (12.0)	0.11	3.33	< .0009
7	83	38.9 (19.2)	46.8 (19.9)	7.9 (12.0)	0.40	5.97	< .0001
Composite	2,246	36.8 (19.5)	42.2 (19.4)	5.3 (12.8)	0.27	19.72	< .0001

exception of sixth grade, all gains were in the range of 4.2 to 7.9 NCE (effect sizes 0.26 to 0.40). The sixth grade showed a much smaller gain of 2.4 NCE. Interestingly, the two largest

gains were at the lowest and highest levels tested—third and seventh grade. Both of these showed substantial gains in excess of 7.0 NCE and effect sizes of 0.40.

Table 6
SAT-9 Listening Subtest Results, by Grade, in NCE

Listening							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
k	842	37.1 (19.4)	47.0 (20.2)	9.9 (16.0)	0.50	17.90	< .0001
1	717	41.7 (17.9)	43.3 (17.7)	1.6 (14.8)	0.09	2.87	0.004
2	668	37.6 (20.6)	39.0 (19.3)	1.4 (15.0)	0.07	2.39	0.02*
Composite	2,227	38.7 (19.4)	43.4 (19.4)	4.7 (15.9)	0.24	13.87	< .0001

* Not significant with the Dunn-Sidak correction.

Table 7
WRMT-R Results, by Subtest, in NCE

WRMT-R							
Test	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
Word ID	2191	47.4 (23.0)	56.6 (21.5)	9.2 (20.5)	0.41	21.06	< .0001
Word Attack	2180	48.5 (21.2)	64.9 (21.4)	16.3 (22.1)	0.77	34.42	< .0001
Passage Comp	2202	47.0 (24.8)	52.3 (22.0)	5.3 (23.2)	0.23	10.69	< .0001
Composite	2,215	47.9 (19.8)	58.0 (19.9)	10.1 (16.7)	0.52	58.80	< .0001

Results of the Listening subtest given to students in kindergarten through second grade are shown in Table 6. On this subtest, kindergartners demonstrated a large gain of 9.9 NCE (0.5 effect size). At the first and second grade levels, however, students showed near zero changes (effect sizes less than 0.1). The average gain of 4.7 is not very meaningful because of the great differences among grade levels.

Table 7 shows results for the WRMT-R scales. The data in this table represent averages across all grade levels. On the composite score, the 2,215 k–2 students achieved a large gain of 10.3 NCE points with an effect size of 0.52. Clear gains are apparent on all of the subtests. The greatest gain was seen in Word Attack with a 16.3 point NCE gain (effect size 0.77), the second largest in Word Identification with a gain of 9.2 NCE (effect size 0.41). The smallest gain was in Passage Comprehension with 5.3 points (effect size 0.23).

Results for each grade level on the Word Identification test of the WRMT-R are given in Table 8. Kindergartners scored a very large gain on the subtest (20 NCE and 1.01 effect size), first graders showed a moderate positive effect of 5.9 NCE (0.25 effect size) and the

second graders made very little gain (1.4 NCE and 0.07 effect size).

Word Attack results are summarized in Table 9. On this subtest, kindergartners evidenced a massive gain of 30.5 NCE and a very large effect size of 2.04. First and second graders also registered substantial gains of 7.2 NCE (effect size 0.31) and 11.2 NCE (effect size 0.50).

Table 10 reports the results for each grade level on the Passage Comprehension subtest of the WRMT-R. On this comprehension subtest, kindergartners made very little change (1.5 NCE and 0.06 effect size) while first and second graders showed more substantial changes. First graders gained 7.9 NCE (effect size 0.35) and second graders improved by 6.5 NCE (effect size 0.31).

Summary of Achievement Test Results

During the 1999–2000 academic year, Advantage students posted gains on the SAT-9 and WRMT-R that were administered in the fall and then again in the spring. In such normed tests, students who remain at the same percentile rank (or NCE) across the tests' administrations are

Table 8
WRMT-R Results, Word ID Subtest, in NCE

Word ID							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
k	742	46.8 (20.1)	66.8 (19.4)	20.0 (22.3)	1.01	24.44	< .0001
1	734	47.8 (26.7)	53.7 (20.0)	5.9 (18.1)	0.25	8.79	< .0001
2	715	47.5 (21.8)	48.9 20.9	1.4 (15.4)	0.07	2.23	0.02*
Composite	2,191	47.4 (23.0)	56.6 (21.5)	9.2 (20.5)	0.41	21.06	< .0001

* Not significant with the Dunn-Sidak correction.

learning at a pace comparable with peers at their level nationwide, while students who gain in percentile rank (or NCE) are learning at an accelerated pace. At each test administration, the norm group is composed of a representative sample of students nationally in the same grade taking the test at the same time. Summary gains

of Advantage students in 1999–2000 are shown in Table 11. Students in kindergarten through second grade gained 18.8 percentile rank points on the nationally normed WRMT-R. Students in these early grades also posted substantial gains in Listening of 8.1 percentile rank points on the SAT-9 and moderate gains in Math of 3.6 per-

Table 9
WRMT-R Results, Word Attack Subtest, in NCE

Word Attack							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
k	729	44.9 (13.3)	74.5 (16.6)	30.5 (17.9)	2.04	-26.01*	< .0001
1	735	51.6 (24.2)	58.8 (22.2)	7.2 (22.1)	0.31	8.79	< .0001
2	716	49.1 (23.7)	60.4 (20.8)	11.2 (18.6)	0.50	16.10	< .0001
Composite	2,180	48.5 (21.2)	64.9 (21.4)	16.3 (22.1)	0.77	34.40	< .0001

* Wilcoxon sign rank *z*.

Table 10
WRMT-R Results, Passage Comprehension, in NCE

Passage Comprehension							
Grade	<i>N</i>	Pretest	Posttest	Gain	Effect	<i>t</i>	<i>p</i>
k	748	57.2 (25.6)	58.8 (22.2)	1.5 (30.1)	0.06	1.41	0.16
1	737	40.3 (23.7)	48.2 (22.1)	7.9 (21.2)	0.35	10.10	< .0001
2	717	43.1 (21.6)	49.6 (19.9)	6.5 (15.1)	0.31	11.52	< .0001
Composite	2,202	47.0 (24.8)	52.3 (22.0)	5.3 (23.2)	0.23	10.69	< .0001

centile rank points. In the higher grades, students also gained against their peers nationally over the course of the school year. Students in Grades 3 through 7 gained on average 6.2 percentile rank points across all subjects on the SAT-9. Across all grades and subjects tested, students gained 9.1 points in National Percentile Rank during the school year.

Discussion

In the number and demographics of the students it served in the period of study, the system of schools managed by Advantage Schools resembled that of a small urban school dis-

trict. Within this system of schools, students learned at an accelerated rate in the 1999–2000 school year. The two reading tests utilized to evaluate the effectiveness of Advantage Schools resulted in somewhat different patterns of results. The WRMT-R, used with k–2nd-grade students, showed a robust average gain of 10.1 NCE (effect size = 0.52); whereas the SAT-9, administered to 3rd–7th-grade students, indicated a modest gain of 3.4 NCE (effect size = 0.17). This discrepancy is especially noteworthy because it reveals several important cautions for interpreting the results of this evaluation. On a broad level, it indicates Advantage did not have a single simple effect on “reading.”

Table 11
Summary Gains, All Schools, WRMT-R and SAT-9, in NPRs

	Reading k–2	Reading 3–7	Math k–7	Language 3–7	Listening k–2	Average k–7
Fall	46.0	25.4	24.6	26.5	29.6	29.8
Spring	64.8	30.9	29.0	35.5	37.7	38.9
Difference	18.8	5.4	4.4	8.9	8.1	9.1
<i>N</i>	2,215	2,437	5,631	2,246	2,227	5,874

Rather, the effect of Advantage was more complex. The effect appears to be very large for younger students on the tasks of the WRMT-R and more subtle for older students on the tasks of the SAT-9. Three factors might account for the discrepancy: (a) the tasks presented in the two tests, (b) the grade level of the students, and (c) differences in the amount of instructional time devoted to reading at the various grades. We administered three subtests of the WRMT-R; two of these subtests focus on word reading skills (i.e., Word Attack and Word Identification) and only one focuses on comprehension (i.e., Passage Comprehension). The SAT-9 reading score is more dependent on comprehension tasks. The two very large gains seen on the WRMT (effect sizes of 1.01 and 2.04) were for kindergartners on the two-word reading tests. Thus, the difference in test content could account for some of the differences. However, this does not explain all of the differences as students in first and second grade showed substantially larger gains on the WRMT-R Passage Comprehension subtest (0.35 and 0.31) than did the older students on the SAT-9 Reading subtest (0.06 to 0.23).

A second major factor that might be supposed to explain the discrepancy in reading results is the grade level of the students. This might appear to be particularly plausible because the proportion of a student's school career that was spent at an Advantage school is related to their grade level. Most of the Advantage students were in their first year of attendance at an Advantage school during the evaluation year. Thus, kindergartners had all of their school experience in an Advantage school, most of the first graders had one half their schooling in Advantage, most of the third graders had only one third, and so on up to the seventh graders who had only one seventh of their experience at Advantage. Kindergartners might be considered to show the most pure effects of Advantage while seventh graders would be considered to reflect an attenuated effect. This explanation would fit the pattern of the discrepancy

between WRMT-R and SAT-9 results. However, it is not supported by the patterns within each test. For example, this explanation would suggest that we should see a gradual decrease in effect size on the SAT-9 reading test from third grade through seventh grade. This is not the case. In fact, third grade registered the smallest effect size on SAT-9 reading (0.06 compared to the other grades between 0.16 and 0.23). The results on the WRMT-R Passage Comprehension subtest further contradict this explanation; kindergartners have the smallest effects (0.06 compared to 0.35 and 0.31).

A third factor that may have contributed to larger gains in kindergarten through second grade is the amount of classroom time allocated to reading. Advantage policy provided for at least nine periods each week of reading in Grades k–2; at least five periods of reading for students in Grades 3 and above who were “on grade level” (about 30% to 50% of the students), and at least seven periods for students who were “below grade level” (50% to 70% of the students). This could contribute to larger gains in reading achievement seen in students from the lower grades.

Students were overwhelmingly from economically disadvantaged families as indicated by the percentage of students eligible for free- and reduced-lunch, and their achievement level on entry to the Advantage schools was generally low as measured by their pretest scores. Therefore, outcomes do not appear to be a result of Advantage schools serving students who would be expected to score higher than the national average based on their socioeconomic background or their initial skill levels.

One must be cautious in interpreting these results. They cover only one academic year. It is not known if students would enhance or even maintain their ranking over time as they were promoted through the grades. It also is not known if the quality of the program's implementation would be sustained over multiple years by school staff.

The number of students, for this kind of study, is substantial. With this large number of students, it is unlikely that the results were caused by the specific personal characteristics of individuals or the random fluctuations of scores that would be expected anytime students are tested and retested.

This is a multisite evaluation. The results reflect 14 implementations of the model. The overall effects on student achievement suggest that the results are not due to a single exceptional administrator or other characteristics unique to a small number of sites. As all sites posted achievement gains (although one school's average gain was nonsignificant), the factors causing these gains appear to have been replicated across the sites.

Although the Advantage intervention *as a whole* had positive effects on achievement, it is not possible to determine which of the specific components of the Advantage program were or were not important in producing this effect. We speculate, however, that rigorous Direct Instruction program implementation, the alternate setting of charter schools, and the focus that private management brought to the implementation contributed to the results.

Reflections on the challenge of large scale reform

The difficulty of markedly elevating achievement within the structures of traditional urban school districts suggests that alternative approaches like that developed and implemented by Advantage should be pursued. Within many traditional urban districts, even under the leadership of a determined superintendent, standardized test scores remain stubbornly unchanged: Each year, some grades move slightly up, but others move down. It is possible that traditional large urban school districts impose too many obstacles to fundamental reform. New settings, such as networks of

semiautonomous schools within or outside the jurisdiction of existing school districts, similar to that developed by Advantage, might provide a setting more hospitable to reform and to the successful implementation of Direct Instruction programs.

Among elementary curricula, Direct Instruction is perhaps unrivalled in its academic efficacy. But it relies on precise implementation and is widely misunderstood and therefore easily mischaracterized by its opponents. Its more visible features—choral delivery, teacher-led instruction, homogeneous grouping—arouse hostility from educators schooled in the progressive orthodoxy. This may explain why it is not more broadly deployed.

Within the culture of traditional districts, it is difficult to establish a school culture with a broadly shared commitment to such a distinctive model. To cite several challenges, teachers are generally assigned to schools, regardless of their instructional convictions, rather than choosing to work with a particular program. But the more distinctive and unorthodox the model, the more essential is staff subscription to it. Moreover, each year seniority-based “bumping” undermines both the teamwork and skill development essential to program implementation, as teachers newly trained in the model are reassigned to other schools and teachers not trained in the model take their places. Further, administrators often accede to requests from staff and parents to “modify” the program. Paradoxically, they do so out of desire to sustain the commitment of these parties to the new program, in an educational culture that prizes teacher invention, not the mastery of a proven educational protocol like Direct Instruction. Individually, these adjustments may have little impact, and the requested changes often seem intuitively reasonable. But intuition is of little value to administering Direct Instruction implementation. In aggregate, the adjustments erode, in succession, the fidelity of implementation, the academic gains achieved, and ultimately staff and institutional

commitment to the program. It is then almost impossible to rebuild that commitment. Inadequate investment in professional development and the absence of firm-handed curricular oversight—both common implementation problems—greatly increase this risk.

Beyond the implementation of a highly engineered program like Direct Instruction, other reform elements also pose challenges to traditional urban districts. A longer school day and year, a large investment in professional development, the creation of schools of choice for both staff and students, rewards for teachers who perform the best and the termination of those who chronically fail to meet standards, attractive and renovated school facilities, a researched-based behavior system (including a focus on praise rather than admonishment)—all of these reforms may in fact be essential to addressing chronic urban under-achievement, but all are also difficult to realize in many school systems.

The four-part model—charter or other semiautonomous public schools as the organizational form; a rigorously academic school design based on Direct Instruction; a distinctive school culture centered on student achievement, transmitted to all school constituents by a powerful school leader; and the discipline of private management—may provide an alternative approach that is replicable and can be brought to scale. First, charter, “pilot,” and other forms of semi-autonomous new public schools (whether under the jurisdiction of the district or an outside chartering authority) provide administrators charged with raising achievement levels with the requisite authority to allocate scarce financial resources and make personnel decisions. The new schools bring together like-minded staff members who have at least an initial willingness to implement unorthodox instructional and behavioral models. Parents bring a level of enthusiasm from having selected the school for their children. This setting permits the second component, a rigorous school design based on Direct Instruction, to be implemented uncom-

promisingly. Third, the identification and appointment of school leaders who can articulate—unabashedly, at every occasion, and to every audience—the school’s audacious expectations for its students’ academic achievement, as well as oversee day-to-day implementation—is critical. There is no substitute for such a leader who can establish and sustain a school culture that prizes student achievement and constantly works to advance it. Lastly, private management may be freer to administer the schools to maximize results rather than to secure and demonstrate compliance.

Admittedly, this model faces challenges of its own. Charter schools still face great difficulty in securing adequate facilities and, unlike traditional schools, must pay occupancy costs out of scarce operating funds. The involvement of private management organizations, especially if organized as for-profit companies, arouses, at the very least, suspicion—even when they bring necessary capital. Advantage Schools initially stumbled in duplicating services already operated at scale by the district, such as meals, transportation, personnel and financial management, and regulatory compliance functions. In turn, these difficulties eroded the commitment of key stakeholders (parents, staff, client boards, state regulators, and investors) and detracted from the significant gains students were making.

The Advantage experience may suggest an approach to raising achievement, at scale, at existing spending levels, and with sustainable levels of staff commitment. The approach warrants further study at scale. The implementation of Direct Instruction-based school designs in newly created, semiautonomous schools may create settings where the many elements of effective schools can be consistently and repeatedly realized. If, in such new public schools, urban students from economically disadvantaged families can learn at an accelerated rate from the earliest grades, they may rise in time to perform on a par with their peers nationally.

References

- Adams, G. L., & Engelmann, S. (1996). *Research on Direct Instruction: 25 years beyond DISTAR*. Seattle, WA: Educational Assessment Systems.
- Advantage Schools. (2001). *Annual Report on School Performance*. Boston, MA: Advantage Schools, Inc.
- Advantage Schools. (2001). *Annual Report on School Performance: Technical Report*. Boston, MA: Advantage Schools, Inc.
- American Guidance Service. (1998). *ASSIST: Automated system for scoring and interpreting standardized tests, Manual for the WRMT-R*. Circle Pines, MN: American Guidance Service.
- Dixon, R., Engelmann, S., Bauer, M. M., Steely, D., & Wells, T. (1990). *Spelling mastery*. Chicago: Science Research Associates.
- Engelmann, S. (1997). *DISTAR arithmetic*. Worthington, OH: SRA Macmillan/McGraw-Hill.
- Engelmann, S., Arbogast, A. B., Seitz Davis, K. L., Grossen, B., & Silbert, J. (1993). *Reasoning and writing*. Chicago: Science Research Associates.
- Engelmann, S., & Bruner, E. C. (1995). *Reading mastery I*. Worthington, OH: SRA Macmillan/McGraw-Hill.
- Engelmann, S., Carnine, D., Engelmann, O., & Kelly, B. (1994). *Connecting math concepts*. Chicago: Science Research Associates.
- Engelmann, S., Johnson, G., & Carnine, L. (1988). *Corrective reading: decoding A*. Chicago: Science Research Associates.
- Engelmann, S., Haddox, P., Osborn, J., & Hanner, S. (1998). *Corrective reading: comprehension A*. Chicago: Science Research Associates.
- Engelmann, S., & Osborn, J. (1997). *DISTAR language*. Worthington, OH: SRA Macmillan/McGraw-Hill.
- Engelmann, S., & Osborn, J. (1998). *Language for learning*. Columbus, OH: SRA/McGraw-Hill.
- Engelmann, S., & Silbert, J. (1985). *Expressive writing I*. Chicago: Science Research Associates.
- Harcourt Brace and Co. (1996). *Stanford Achievement Test Series, Ninth Edition, Technical Data Report*. San Antonio, TX: Author
- International Baccalaureate Organization. (2001). The Diploma Programme. Retrieved October 30, 2001, from http://www.ibo.org/ibo2/en/programmes/prg_dip.cfm
- Kaufman, A. S. (1990). *Practice Effects. Encyclopedia of Human Intelligence*. (Vol. 2). New York: MacMillan.
- National Center for Educational Statistics. (2000). *NAEP 1999 Long-Term Trend Summary Data Tables*. <http://nces.ed.gov/naep3/tables/Ltt1999/>
- Purkey, S. C., & Smith, M. S. (1983). Effective schools: A review. *The Elementary Schools Journal*, 83(4), 427–452.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests-Revised, Normative Update, Form G & H, Examiner's Manual*. Circle Pines, MN: American Guidance Service.
- U.S. Department of Education, *Digest of Education Statistics 2000*, National Center for Education Statistics, <http://nces.ed.gov/pubs2001/digest/dt372.html>