CHAPTER 10

What Did Students Learn?: 1982–1994

Elizabeth G. Cohen, Julie A. Bianchini, Ruth Cossey, Nicole C. Holthuis, Christopher C. Morphew, and Jennifer A. Whitcomb

From the beginning of the program, educators considering adoption of complex instruction, funding agencies, researchers in cooperative learning, and even sociologists have wanted to know the bottom line: What did students learn? Were the learning/achievement gains statistically significant? Did students in complex instruction learn more than comparable students in conventional classrooms? To answer these questions, the authors of this chapter describe achievement data collected by the program with a variety of tests, both standardized and content-referenced.

The first part of the chapter summarizes and interprets the results of repeated evaluations of the effects of complex instruction, using the *Finding Out/Descubrimiento* (De Avila & Duncan, 1982b) curriculum in Grades 2–5. At the elementary level, evaluators used the California Test of Basic Skills (1982), and an *FO/D* content-referenced test.

The next section of the chapter reviews and interprets achievement results for the middle school. The middle school data include results of content-referenced tests for social studies, mathematics, and Human Biology. Table 10.1 contains summary information on tests, grades, and numbers of students and classrooms that will assist the reader in following the discussion.

Certain patterns emerge from a review of results of these evaluations. In the concluding section of the chapter, we synthesize what we have learned from these results. Some of these generalizations are substantive and have to do with the conditions under which the most impressive gains in achievement occur. Others foreshadow the methodological discussion of Chapter 11, which offers a critique of achievement testing. Finally, we discuss important products of CI, such as intellectual and social skills, that these achievement tests have not measured.

Table 10.1: Summary of Test Data

Year	Test Type	Subscale or Subject Matter	Grade	Number of Classrooms	Number of Students	Tables/ Figures
Achievement Tests for Elementary School						
1982–83	CTBS	Math Concepts, Math Applications, Computation	2, 3, 4	12 *FO/D* 3 comparison	102	Tab.10.2 Fig.10.2
1983–84	CTBS	Science Math C & A Computation	2, 3, 4, 5	17	334 230 252	Tab. 10.2
1987–88	CTBS	Total Math	3, 4	4	65	Fig.10.1
1989–90	Content-Referenced Multiple Choice	Science	2, 3, 4	10	202	Fig.10.3
Achievement Tests for Middle School						
1991–92	Content-Referenced Multiple Choice	Social Studies	7, 8	26 CI 9 comparison	84–382	Tab.10.3 Tab. 10.4
1992–93	Content-Referenced Multiple Choice	Social Studies	7, 8	25 CI 2 comparison	669	Tab. 10.5
1992–93	Items from QCAI Rubric-scored	Mathematics	7, 8	14	272	Figs.10.4 & 10.5
1992–93	Content-Referenced Rubric-Scored	Human Biology	8	10	260	Tab.10.6
1993–94	Content-Referenced Rubric-Scored	Human Biology	6	3	80	Tab.10.6

ACHIEVEMENT RESULTS IN ELEMENTARY SCHOOLS

We review three sets of data from standardized tests: 1982–83, 1983–84, and 1987–88. The data set for 1982–83 included three comparison classrooms where *FO/D* was not used. For the *FO/D* content-referenced test, there is one set of data from 1989–90 (see Table 10.1).

Nature of Schools and Classrooms

Audiences hearing about CI often assume that the program works in university laboratory schools or at least in the wealthy suburbs near Stanford University. Actually, the elementary schools represented in this chapter were from the San Jose Unified School District, from working-class suburbs of the Bay Area, and from a rural district near Fresno. Classrooms varied from largely segregated Latino or Southeast Asian to heterogeneous classrooms with a mix of middle-class Anglos and working-class Latinos. Many of the participating students were from non-English-speaking backgrounds and were experiencing difficulties in basic skills. Educators implemented *FO/D* in an attempt to improve basic skills while at the same time addressing the need for development in conceptual aspects of science and mathematics.

Standardized Tests

Three subscales of the CTBS math tests are relevant to the curriculum activities of *FO/D*: Math Concepts, Math Application, and Computation. In addition, in 1983–84, the school district chose to administer the science portion of the CTBS test to all elementary classrooms, including those where *FO/D* had been implemented. Results for *FO/D* students were examined, even though the test did not include physics and chemistry, the major scientific content of the *FO/D* materials.

Pre-Post Test Gains: NCE Analysis. Using data from the CTBS standardized achievement test, we first present a comparison of student performance in the fall of the school year with performance in the spring. De Avila, in an unpublished proposal, examined the gains statistically by testing for significance of the difference between average fall and spring scores for the 1982–83 and 1983–84 school years. This analysis employs Normal Curve Equivalents (NCEs). At the time of this statistical analysis, NCEs were the preferred statistic for evaluation of programs such as Title I. NCEs are normalized standard scores with a mean of 50 and a standard deviation of 21.06 (Linn, 1979). These statistics permit standardization across a variety of forms of the normed test administered at different grade levels and in different years. Improvement in NCEs between fall and spring means that students gained

more than the nationally normed population. If they gained the same amount as the normed population, their score would stay the same. Thus, any increase in average NCE means that students are gaining more than is to be expected.

Table 10.2 presents pre- and posttest scores for the math computation, concepts, and application subscales, total math scores, and science scores, where available. The average NCE for the 1983–84 sample starts off considerably higher than that for the 1982–83 sample. The higher pretest scores for 1983–84 were probably due to the inclusion of one magnet school with a significant number of middle-class students. Also included was one school with gifted bilingual classes.

Students made statistically significant gains from fall to spring in mathematics in both years and in science in 1983–84, when the science test was

Table 10.2: Pre–Post Math and Science CTBS Test Scores: Normal Curve Equivalents for 1982–83 and 1983–84, Grades 2–5, for Classrooms Using FO/D

	1982–83		1983–84	
Test	Pretest	Posttest	Pretest	Posttest
Computation				
Mean	24.92	35.5	44.88	54.6
SD	(9.86)	(7.04)	(16.71)	(24.71)
t	13.47*	(*n* = 102)	11.16*	(*n* = 252)
Math Concepts				
Mean	29.93	36.64	47.47	52.61
SD	(7.74)	(6.13)	(16.98)	(15.87)
t	8.53*	(*n* = 102)	6.11*	(*n* = 241)
Math Applications[a]				
Mean			46.52	51.78
SD			(18.74)	(15.97)
t			5.33*	(*n* = 230)
Total Math				
Mean	37.08	57.04	43.47	53.47
SD	(18.56)	(20.87)	(15.11)	(15.28)
t	8.36*	(*n* = 102)	15.24*	(*n* = 329)
Science[b]				
Mean			44.67	51.32
SD			(18.82)	(17.79)
t			8.17*	(*n* = 334)

* $p < .001$
[a] Concepts and Applications Subscales combined for 1982–83
[b] Data not available for 1982–83

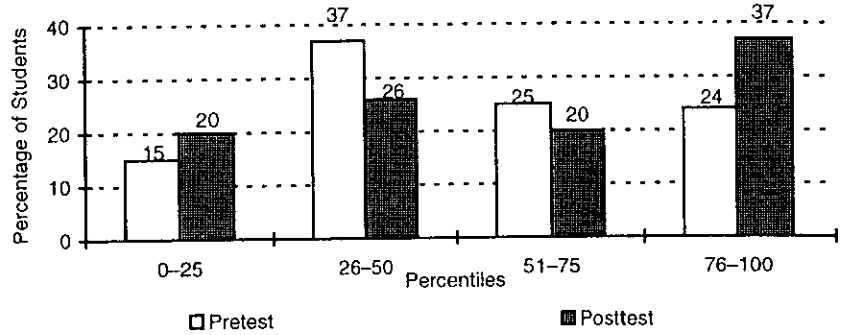
administered. The gains in all subscales and in the math total battery were markedly higher than would have been expected according to national norms. The gains in Math Computation were even greater than the gains in Concepts and Application. Gifted bilingual students and the students from the magnet school showed excellent gains, along with those whose pretest scores were lower.

Gains in 1987–88: A Shift in the Distribution. These data come from a suburb of Fresno, California, where Professor Teresa Perez of Fresno State University had worked with several schools in implementing CI. A sample of students on whom there were spring scores in 1987 were tested once more, in Spring 1988, after a year of *FO/D*. Data are available on 89 students in 2 second- and 2 third-grade classrooms from two schools in 1987. Of these youngsters, 65 experienced *FO/D* in third and fourth grades during the following school year.

For this set of data, the analysis compared the percentage of students falling into each quartile of the percentile distribution on the CTBS. Figure 10.1 shows these percentages for the students before and after experience with CI for the total math score. There were similar results on the CTBS reading scales. Examination of the bar chart shows the dramatic increase in the percentage of students in the top quartile (from 24% to 37%) after the year of experience with CI. In the previous spring, this sample was much more likely to be in the second quartile than anywhere else in the distribution. It was also the case that there were 5% more students in the lowest quartile in Spring 1988 than in the previous spring.

Use of "untreated" comparison groups. De Avila (Cohen & De Avila, 1983), using standardized achievement test data from 1982–83, compared

Figure 10.1: Percentage of Students per Quartile, Pre vs. Post *FO/D* for Math, 1987–88

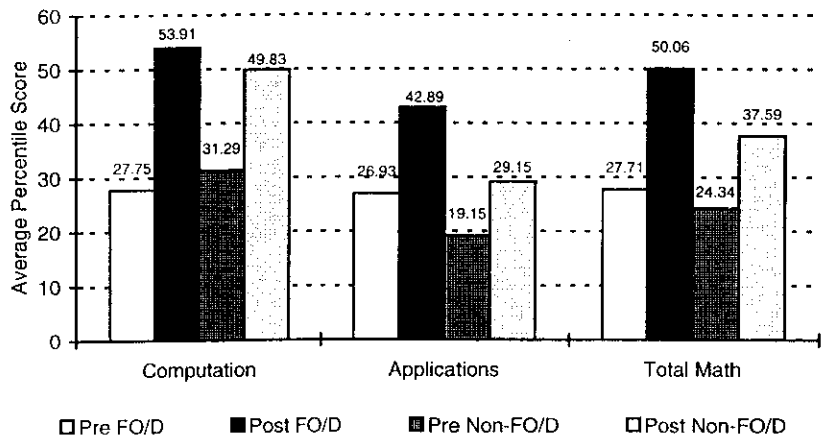


the 12 classrooms that experienced *FO/D* with students from three classrooms (n = 41) that did not. All these classrooms were members of the Bilingual Consortium of San Jose, a federally funded project that offered extensive staff development to teachers of all member classrooms. The teachers in the *FO/D* classrooms received staff development from program staff members in addition to the regular offerings of the Consortium. Figure 10.2 presents the average percentiles for pre- and posttests for *FO/D* students and for the comparison students (Cohen & De Avila, 1983).

The percentiles are more directly interpretable than the normal curve equivalents in Table 10.2. For Total Math and Computation, the *FO/D* students moved from a very low standing in terms of the nationally normed population (around twenty-fifth percentile) to grade level, or fiftieth percentile. Clearly, the comparison students also gained relative to the nationally normed population. De Avila used t-tests to assess the difference in gain scores between the *FO/D* and comparison students. The *FO/D* students showed significantly greater gains in Math Computation ($t = -1.7$, $p < .05$) and in the Total Math scores ($t = -2.55$, $p < .01$) than the comparison students. Although the *FO/D* students also gained more on average than the comparison students on the Concepts and Application subscales, the difference in gain scores was not statistically significant. Cohen and De Avila (1983) pointed out that it was difficult to show significance with only 41 comparison students.

In Figure 10.2, one can easily see the differences between pre- and posttest scores for the two groups on each of the subscales and on total test

Figure 10.2: *FO/D* vs. Non-*FO/D*, Pre–Post Test Scores in CTBS Math, 1982–83: Grades 2–4 (N = 104)



scores. The difference in the height of the pretest bar and the posttest bar displays graphically the larger gains of the *FO/D* students. One can also see how both groups moved upward in comparison to the national norms from fall to spring (Cohen & De Avila, 1983).

A Content-Referenced Test for *FO/D*

The *FO/D* content-referenced test contains items that reflect the vocabulary, concepts, and applications of the *FO/D* curriculum. The first version of this test was created and administered in the pilot year 1979–80. Ten years later, in 1989–90, the test was revised with the help of classroom teachers and administered once more as a pre- and a posttest during the school year.

The revised test consists of 100 items, 65 items covering science and 35 on mathematics. As a result of the revision, there were items for all 17 units of the *FO/D* curriculum, although teachers were not able to cover all these units in one academic year.

There were three types of items: concepts and vocabulary, simple applications, and complex applications. The term *concepts and vocabulary* refers to science and mathematics vocabulary, including concrete items such as candle or washer, and abstract concepts such as circumference or fulcrum. *Simple applications* refer to the recognition of a concept embedded in its context, either in a question or in a fill-in-the-blank item. For example:

> Which of the following is a liquid?
> _____Dough _____ Milk _____ Salt ____ Measuring cup
> Liters are useful for measuring _____.
> ____Bugs _____ Milk ____ Your height ____ Your weight

Complex applications refer to concepts that involve abstractions that are outside the students' everyday life experience or that use more than one kind of abstraction. For example:

> Rachel put some white powder into a cup. Then she added vinegar. She noticed bubbles in the cup. What did Rachel observe?
> Acid _____ Base _____ Solution _____ Reaction

The test sample was 202 students in 10 classrooms from Grades 1–4 in Redwood City and Milpitas, California. The sample included classrooms where only a handful of students were reading at grade level. Eleven non-English-proficient students and 87 limited-English-proficient students took the test. Teachers administered the test orally in English and in Spanish. The students could choose whether to take the test in English or in Spanish. The students

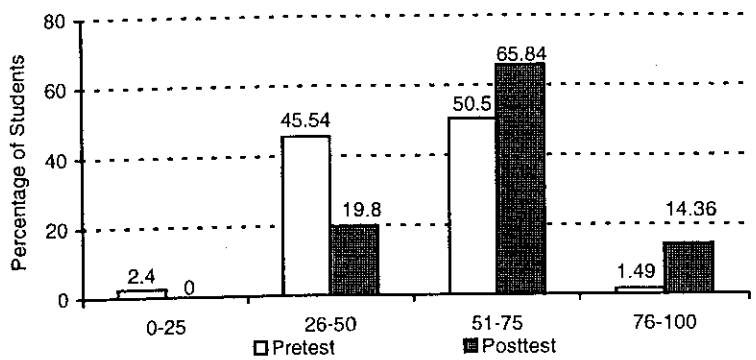
circled responses, often choosing between pictures. The test did not require that students be able to read or write at grade level. In the case of other language minority students, the teacher used the English version, but an instructional aide who spoke the student's native language often helped. In the lower grades, most teachers took several days to administer the pretest. The posttests were administered during May 1989.

Students gained an average of 11 points out of the 100 points on the test. The average pretest score was 50.3. On the posttest, over 80% of the students answered more than half the items correctly. Figure 10.3 illustrates student growth on the test by dividing the percentage of correct items into quartiles. The percentage of students who fell into each of the quartiles in the pre- and posttest can be directly compared. Only a few students fell into the lowest quartile on the pretest and none did on the posttest. On the pretest, 45.54% of the students fell into the second quartile, while in the posttest, only 19.80% were in this category. Half of the sample scored in the third quartile at the time of the pretest, but this category increased to 65.84% on the posttest. Most impressive, however, was the increase in the proportion of students in the top category from 1.49% to 14.36%.

With respect to the different subscales, the highest gain, 6.4 items out of 50, or 12.8%, was achieved on the Concepts and Vocabulary subscale. The lowest gains were in the Complex Applications subscale, which is, by definition, the most challenging. Out of 15 items on this subscale, the average gain was only 1.3 items. There were about equal gains on the items drawn from science and mathematics.

According to the final report on this evaluation (Cohen & Lotan, 1990), when data were analyzed according to groups divided on the basis of English

Figure 10.3: Percentage of Students per Quartile Pre vs. Post *FO/D* Content-Referenced Test, Grades 2–4, 1989–90



language proficiency (non-English proficient; limited English proficient; and fully English proficient), all groups made considerable gains, but the limited-English-proficient group showed the highest gains. With respect to gender, although girls scored somewhat lower than boys on the pretest, the gains of the former were somewhat higher than those of the latter. As a result, there was no difference between boys and girls on the posttest.

Interpretation of Elementary Achievement Results

Initially, De Avila viewed the two CTBS subscales of Math Concepts and Application as the most theoretically relevant to the curriculum (Cohen & De Avila, 1983). The children repeatedly solved word problems on the *FO/D* worksheets. Unlike the decontextualized problems in the typical arithmetic lesson, these problems were a natural adjunct to their group activities: How many liters do you think this will hold? How many liters did it hold? How far off were you? The children were not only gaining experience with word problems, but were using computation to answer questions about their own activities.

De Avila and Cohen were initially surprised by the strength of the results in computation. They viewed computation as relatively routine learning in comparison to the problems included in the Math Concepts and Application subscales. Since *FO/D* stressed the development of thinking skills, they did not expect it to have a dramatic effect on computation items. However, there were strong results in computation in the earliest evaluation (1979–80), and the gains in computation were even more dramatic in the analyses of 1982–83 and 1983–84. In retrospect, De Avila and Cohen realized that there were multiple opportunities to practice computation in connection with the worksheets. Even more important, students in the second and third grades could understand the underlying concepts of arithmetic in ways that the ordinary classroom drills did not provide. Instead of the drills of arithmetic lessons, students carried out arithmetic operations to solve real problems in which they manipulated materials, discussed solutions, and made estimates. Conversations with classroom teachers revealed that they did not ordinarily teach arithmetic with emphasis on concepts. It was probably the combination of the drills in the regular classwork and the more conceptual approach to arithmetic built into the worksheets, that regularly produced dramatic gains in the Computation subscale.

The significant gains on the science test in 1983–84 were also surprising. The science content of *FO/D* was physics and chemistry and was not directly related to the science content of the tests. Nonetheless, the general emphasis on scientific reasoning, analysis, and construction of hypotheses may have assisted the students in taking this test.

The overall results of achievement testing at the elementary school showed strong gains according to the standardized tests and the content-referenced test. From an absolute perspective, over 80% of the students answered more than half the items on the content-referenced test correctly. This is an impressive accomplishment by any criterion. It means that a considerable proportion of the students, many of whom belong to linguistic minorities, improved their general knowledge of mathematics and science. The gains among the students with limited English proficiency were also remarkable. Observers often noted these students making full use of activity cards in English and Spanish. Having individual reports and the content-referenced test, as well as the activity cards, in both languages represented a great advantage for the limited-English students with a Spanish-speaking background.

Differences by Grade Levels. From the earliest studies, the youngest children made the most impressive gains. This was always surprising to the teachers of the second graders, who felt that the program was too difficult and was stretching the children to their utmost. Both in 1982–83 and in 1983–84, the analysis of normal curve equivalent scores on the CTBS shows a consistent pattern of the highest gains for the second grade. The lowest gains were for Grade 5, which was included in the 1983–84 sample. For example, in the total math battery, the average gain for second graders was 20.86 points in 1982–83 and 13.20 in the following year. In contrast, the gains for Grade 4 were 12.95 in the first year and 12.24 in the second year. Fifth graders in the second year gained only 7.29 points.

A similar pattern was found in the *FO/D* content-referenced test. Whereas the first and second graders showed an average gain of 13.6 and 13.0 points, respectively, the fifth graders gained only 6.9 points. The comparison of gain scores according to grade showed a systematic decrease with each grade.

These grade differences are probably a joint function of characteristics of the curriculum and of the achievement measures. With respect to the curriculum, *FO/D* was designed for the early elementary years. Program staff advised fifth-grade teachers to use supplementary materials and to demand more writing of the students because the curriculum was developmentally appropriate for younger children. Because so many of the fifth-grade students were functioning far below grade level, some teachers felt that the curriculum was just right for their students. However, without enrichment, *FO/D* did not present as many opportunities for growth to fifth graders as to second graders. The curriculum was enormously stimulating for second graders, and they had so much more to learn. With respect to measurement, the content-referenced test contained simple vocabulary like the word *candle* that was no challenge for fifth graders, and as a consequence they received relatively high pretest scores.

In the case of the standardized achievement tests, CTBS Computation scales were directly relevant to the regular mathematics curriculum in the second and third grades and to the kinds of computations required by the worksheets. In contrast, fifth-grade CTBS tests in mathematics did not center on simple computation and included mathematics that was not represented in the worksheets. Thus, the measurement was a much better match to the curriculum in the early elementary grades.

ACHIEVEMENT RESULTS IN MIDDLE SCHOOLS

In this section of the chapter we present the achievement results for social studies, mathematics, and science in the middle school. There are two data sets for social studies from 1991–92 and from 1992–93, each including results from comparison classrooms that did not work with multiple-ability curricula. For mathematics, we report on results for 1992–93, and for science for 1992–93 and 1993–94.

Social Studies

We have 2 years of social studies test results in classrooms that used curricula especially developed for CI and designed to fit within the California curriculum framework (California State Department of Education, 1988). Researchers constructed content-referenced tests designed to reflect material that students were supposed to cover, according to the California curriculum framework. Teachers administered these tests before and after the multiple-ability curricular units in 1991–92 and 1992–93. In both years there were comparison classrooms in which teachers who did not use CI also administered the pre- and posttests.

Test Construction and Administration. The multiple-choice tests for the seventh and eighth grades in social studies had two major sections: factual information and higher-order thinking. The pre- and posttests were identical. The seventh-grade test had 50 items, 33 factual and 17 higher-order thinking items.[1] It covered topics on Feudal Japan, the Crusades, the Mayan culture, and the Reformation. The higher-order thinking items were analogies, using simple language but requiring very abstract thinking. A sample item follows:

> The way the Muslims felt after the Crusaders captured Jerusalem was like the way you would feel after

a. winning the lottery.
b. not getting invited to a party.
c. catching a cold.
d. having your home robbed of all its valuables.

The eighth-grade test had 40 items, 30 factual and 10 higher-order thinking items.[2] It covered materials on the following topics: Manifest Destiny, the Civil War, and the Rise of the Industrial Era. These topics/eras are all covered in the textbook used in the classrooms and in the materials from the Teachers' Curriculum Institute program[3] that the teachers utilized with multiple-ability groupwork tasks.

For each of the higher-order thinking items in the 1991–92 administration, some students were asked to explain why they chose the answer they did. Analysis of student responses led to modifications of the test for 1992–93. In an attempt to reduce the impact of reading skill on test outcome, teachers read test items out loud in the second year.

In the first year, students took the pretest before any units had been implemented and the posttest after all units for the year had been completed. Conditions of administration were changed in the second year so that teachers were instructed to administer a pre- and a posttest directly before and after the relevant unit. This strategy was designed to avoid the poor motivation of some students to do well on a test administered in June that had no connection with their grades in social studies. All but one teacher in a CI classroom and one teacher in a comparison classroom complied with this directive.

Results for 1991–92. Pre- and posttests were administered in 26 seventh- and eighth-grade social studies classrooms from 5 schools and in 9 comparison classrooms, all from the same school. In an attempt to control on school effects, comparison classrooms were selected from a school where CI was also being implemented. Teachers of comparison classrooms received no preparation in CI. There were 5 comparison classrooms at the seventh-grade level taught by two teachers, and 4 comparison classrooms at the eighth-grade level taught by two teachers.

Not all the teachers taught all the units for which there were test items. Therefore, the size of the sample varies by unit. In analysis of the data, test items were divided into subtests according to the unit to which they referred. Only subtests on units that were covered were scored for a given class. The total number of students taking each subtest varied from 84 students who studied the Crusades, to 382 students who studied the Civil War.

Table 10.3 presents the average pretest scores and posttest scores for CI classrooms and for comparison classrooms. The table is divided according to

Table 10.3: Average Pretest and Posttest Scores for Middle School Social Studies by Unit: For Seventh- and Eighth-Grade Complex Instruction vs. Comparison Classrooms, 1991–92

		Average Scores					
		Pretest		Posttest		*n*	
Grade	Unit	CI	Comp.	CI	Comp.	CI	Comp.
7	Feudal Japan	3.39	3.85	4.99	4.9	118	99
7	Crusades	4.94	—	7.33	—	84	—
7	Maya	4.59	—	6.89	—	298	—
7	Reformation	4.73	—	7.31	—	300	—
8	Manifest Destiny	5.51	5.82	6.8	6.13	305	38
8	Civil War	8.69	8.87	10.62	10.99	306	76

the unit and the grade to which the specific test items apply. There is only one unit on which comparisons could be made at the seventh-grade level and two at the eighth-grade level. The average pretest scores for CI classrooms are a little lower than those in comparison classrooms, while the posttest scores are somewhat higher in CI classrooms than in comparison classrooms in two of the units. This is also the case for the size of the gain scores.

Statistical analysis of these data showed a consistent effect on posttest scores of the individual's sixth-grade reading score, for each of the units, even when the pretest score was controlled. Thus, there is a clear effect of lack of reading skills on the scores on this test.

On the items requiring analogies between central concepts and other settings, the average gain scores of the CI seventh- and eighth-grade classrooms are significantly higher than those in the comparison classrooms ($t = 2.366$, $p < .05$). The lower gains in the comparison classrooms were not due to a ceiling effect, because their pretest scores allowed ample room for improvement. Table 10.4 examines the effects of being in a CI classroom on posttest scores on items requiring higher-order thinking. The regression analysis controls for the effects of differing pretest scores and differences in reading scores. The table shows significant favorable effects of being in a CI

Table 10.4: Regression of Social Studies Posttest Score on Pretest Score, Complex Instruction vs. Comparison Classrooms, and Reading Score: For Higher-Order Thinking Items in 1991–92, Seventh and Eighth Grade

Predictors	*B*	Beta	*p*
Seventh Grade (n = 356)			
Constant	.214	.000	.000
Pretest Score	.271	.265	.000
CI vs. Comparison	.087	.201	.000
Reading Score	.002	.300	.000
R^2 = .226			
Eighth Grade (n = 344)			
Constant	.121	.000	.000
Pretest Score	.201	.200	.000
CI vs. Comparison	.081	.147	.001
Reading Score	.004	.449	.000
R^2 = .324			

classroom for both seventh and eighth graders. There were no effects of being in a CI classroom for items requiring factual recall.

Results for 1992–93. Achievement data were collected for 11 seventh-grade and 14 eighth-grade classrooms. Among the seventh-grade classes were a number from one school where a combination of factors made implementation very difficult. These included a minimum of support from the school administration, teachers who had severe disciplinary problems, and a school with a long history of low expectations for student performance and problems with deviant student behavior. There were only two comparison classrooms; they were both seventh grade and came from the problematic school site.

Different teachers taught different numbers of units. In order to standardize for this variation in number of items, Table 10.5 presents "batting averages," or the average percentage of correct items for pre- and posttests on higher-order thinking and factual items for seventh and eighth graders separately. The results for the two comparison classrooms appear separately under the seventh grade heading in the table. The seventh graders showed statistically significant gains on higher-order thinking skills, moving from an average of 37 to 49% correct (t = 3.31, p < .05). The gains for the eighth graders were larger, moving from 40 to 55% (t = 9.78, p <. 001). On factual items, the seventh graders showed only a 5.6% gain, while the eighth graders gained 19.6%.

Table 10.5: Average Percentage Correct Answers for Seventh and Eighth Grade Social Studies Tests on Higher-Order Thinking and Factual Items: Complex Instruction Classrooms and Two Seventh Grade Comparison Classrooms, 1992–93

		Higher-Order Thinking Items		Factual Items	
		Average % Correct	*n*	Average % Correct	*n*
Seventh Grade					
Complex Instruction	Pretest	37.1	265	38.6	266
	Posttest	48.8	237	44.2	237
Comparison Classroom A	Pretest	36.9	25	31.4	26
	Posttest	47.5	22	38.5	22
Comparison Classroom B	Pretest	32.9	28	38.87	29
	Posttest	91.4	18	75.7	18
Eighth Grade					
Complex Instruction	Pretest	40.3	344	38.4	348
	Posttest	55.4	304	58.0	304

Of the two comparison classrooms, Classroom A did a little less well than CI seventh graders on higher-order thinking skills, and a little better on factual items. Classroom B had scores higher than anything ever seen with these tests: a batting average of 91.4% on the posttest items that required higher-order thinking skills and 75.7% on the factual items. The pretest scores were similar to those of the CI seventh graders, if not a little weaker, so the improvement scores were very large.

Multivariate analysis revealed two additional factors that affected individual achievement. One was the percentage of students with below grade level skills in reading in the classroom. The more of these students there were, the lower were the batting averages as well as the gains. The second factor was the number of units the teacher had taught and tested. The more units students had experienced, the better they tested.

Interpretation of Social Studies Results. The results for 1991–92 show the strong effects of CI on the ability of the students to answer the items requiring analogies, clearly an example of higher-order thinking. Being in a CI classroom had a significantly favorable effect on these items but no effect on factual items. Many of the activities in the social studies units created for CI required students to draw analogies between historical events and current events. For example, when students studied political cartoons of the

Reformation period, they drew analogous political cartoons for current events illustrating the theme of challenging of authority of institutions. They composed a song on current events analogous to a Crusader song they heard and analyzed. It is very encouraging that a set of lively activities can help students to think abstractly—students for whom this kind of test question is frequently difficult.

The purpose of the CI activities is to develop concepts rather than to increase factual knowledge. The latter goal does not require such elaborate curricula, although it is significant that teaching in this way does not impair students' gains in factual knowledge.

In the second year, the results for the seventh graders, although statistically significant, were puzzling. The absolute gains were very small in comparison to those of the eighth graders. In addition, one of the comparison classrooms did better than any of the CI classes. How can one account for these results? Questionnaire data from these seventh-grade teachers revealed the fact that for most units, students experienced only one multiple-ability activity per unit. They rarely had the chance to grasp the central concepts through experiencing multiple activities. These seventh-grade teachers spent far less time per unit than teachers in the previous year or the eighth-grade teachers of the second year. They did not take time to prepare the students for the historical period with readings, direct instruction, or lecture/discussion before they moved into the group activities. In contrast, the eighth grade teachers had a wealth of supporting materials from the Teachers' Curriculum Institute and spent much more time per unit both with supporting activities and with CI. Moreover, the eighth-grade teachers' topics were more familiar; teachers had a better background for teaching the Civil War than a seventh-grade unit such as Feudal Japan.

In order to assess the effects of these implementation problems on achievement in social studies, we regressed posttest scores in social studies on pretest scores and two measures of implementation. The Rigor Index was the first of these measures. It contained a measure of the frequency with which teachers took the time to rotate groups of students among activities; a measure of the frequency with which students finished their individual reports; whether these reports were completed during class time; and the proportion of reports on which the teacher provided feedback to the students. This information came from a teacher questionnaire administered at the end of the first year of implementation. In addition to the Rigor Index, we used the average percentage of students who were disengaged according to staff observation with the Whole Class Instrument. Both the Rigor Index and the percentage disengaged had powerful effects on posttest scores, holding constant pretest scores. Thus, the lack of time and management problems were clearly major barriers to achievement in 1992–93.

The results for the second comparison classroom were very strange. Careful examination of the figures shows that as many as 10 students who took the pretest were missing on the posttest. This dropoff is not characteristic of any other classroom in the sample. Inquiries revealed that this teacher was a newcomer to the school, had transferred from a high-achieving school, and was disappointed to be teaching in such a problematic school. This teacher had a reputation for drilling students. Moreover, he was one of the teachers who gave all the unit posttests together at the end of the school year. We have no way of knowing what happened in this classroom, but in any case this comparison classroom raised as many questions as it answered.

Mathematics

Ruth Cossey developed the mathematics curricula for CI and designed the evaluation. In the 1992–93 school year, she conducted an evaluation of what students learned in mathematics classrooms using CI. This chapter reports results with one of the assessment instruments she used. (For information on other instruments, see Cossey, 1997).

Because there was variation in the particular units used in different classrooms, Cossey sought an assessment instrument that would capitalize on the commonalities of the math programs. All teachers agreed to provide a mathematics program that emphasized problem solving, reasoning, and communication. All teachers taught a statistics unit that was either developed by CI staff or enhanced in consultation with CI staff, and all teachers treated geometric concepts such as area and perimeter in their curricula.

QUASAR Cognitive Assessment Instrument. Cossey selected the assessment tool Quantitative Understanding: Amplifying Student Achievement and Reasoning (QUASAR). QUASAR (Silver & Lane, 1995) is a national middle school mathematics program, launched in 1989 to demonstrate the feasibility of implementing mathematics programs that promote thinking and reasoning skills in schools located in economically disadvantaged communities. To help monitor the adequacy of the new program, the project developed the Quasar Cognitive Assessment Instrument (QCAI) (Lane, Liu, Stone, & Ankenmann, 1993). QCAI seemed to match the general programmatic goals of the teachers, even though it did not match the specific content of any teacher's curriculum.[4]

From this 36-item instrument, Cossey selected 18 open-ended tasks for CI classrooms. There were two different forms of the test, each using a different set of nine items. The tasks were "open" either in the solution paths possible, the answers, or both. In these tasks students are asked to construct rather than select correct responses. In some of the items students are asked

to show their work; some ask them to explain their answers. The emphasis on divergent thinking and mathematics communication was consistent with principles of CI curricula in mathematics. The items assessed such abilities as estimating the area of an irregular shape or recognizing the underlying mathematical structure of a number pattern. Relevant to the work on statistics was a task assessing students' understanding of the concept of average in which they are required to interpret information presented in a bar graph.

Use of an instrument designed for another mathematics program to evaluate CI had particular strengths and weaknesses. Among the instrument's strong points were meticulous attention to making the test items friendly and accessible to inner-city youth through extensive field testing and external equity panel reviews. Another strength was its alignment with the general instructional goals of the National Council of Teachers of Mathematics. Still a third strength was its attempt to uncover divergent thinking.

Finally, QUASAR agreed to provide focused holistic scoring of the responses on each task by an ethnically diverse, highly trained cadre of scorers with an interrater reliability of at least 90%. QUASAR developed a general rubric incorporating three components: communication, strategic knowledge, and mathematical conceptual and procedural knowledge. Criteria representing the three overlapping components were identified for each of five score levels (0 to 4). For example, under the heading "mathematical knowledge," a score level of 4 requires an understanding of the problem's mathematical concepts and principles, the use of appropriate terminology and notations, and algorithms that are executed completely and correctly. A score of 4 under the heading "strategic knowledge" requires identification of all the important elements of the problem and an appropriate and systematic strategy for solving the problem as well as clear evidence of a solution process that is complete and systematic. To obtain a score of 4, the student must also communicate clearly and unambiguously and use supporting arguments that are sound and complete. The other levels of scoring specify less satisfactory levels of performance on these components. Using criteria specified at each level of the general rubric, QUASAR staff developed specific rubrics for each task. External reviewers examined the specific rubrics along with sample student work scored at each of the five levels (Lane & Parke, 1992).

There were also certain weaknesses in the choice of this instrument as an evaluation tool for CI. As mentioned earlier, QCAI does not represent a tight curricular match for any of the CI classrooms. Moreover, given that only one of the teachers had a strong mathematical background, the 5-month span between pre- and posttest administrations of the measure may not have allowed teachers sufficient time to demonstrate an increased ability to provide a program with a radically new emphasis on mathematical thinking and communication. Still another weakness is that QCAI does not provide a direct

measure of students' sustained mathematical performance. Students had less than 5 minutes to respond to each task. (The time was equivalent to that allowed students in the QUASAR project to complete the same instrument.) The requirement of quick responses may have prevented some students from showing the full range of their mathematical problem-solving and communication ability.

Administration. The QCAI was administered in a 45 minute time period in 14 classrooms of eight middle grade mathematics teachers in the fall and spring of the 1992–93 school year. As described above, there were two forms of the test; for each form, there were two different versions of the booklet, with the same items arranged in different order. As an aid to Spanish-speaking students, each task was presented in both Spanish and English on facing pages. The distribution of the two test forms was in random order in the fall in each class. In the spring, each student present in class was given a booklet with preprinted name, teacher, and class identification to ensure that each would receive a different form in the spring. Cossey employed this strategy to avoid practice effects of taking the same test twice. There were 272 students who took both fall and spring assessments; 113 students from Cohort 1 took Form 1 in the fall and Form 2 in the spring; 159 students from Cohort 2 took Form 2 in the fall and Form 1 in the spring.

Scoring. Under QUASAR's supervision, personnel at the Learning Resource Development Center at the University of Pittsburgh scored the tests holistically. Raters used the specific rubrics developed on a large sample of student responses. Initially, two raters read each student's response. If the two initial raters disagreed by one point, the final score was the average of the two ratings. If the two initial raters disagreed by more than one point, the response of the senior adjudicating rater became the final score. Additionally, as part of an interrater reliability check, a senior rater scored every tenth response. Cossey examined the incidence of disagreement for a subsample of students and found that raters disagreed by more than one point on fewer than 5% of the responses.

Results. A student could not earn a score of 2.5 or better on a question without a reasonable level of mathematical communication. Initial examination of the percentage of items with scores equal to or greater than 2.5 showed that the two forms of the test were not equivalent. Cohort 1 had 31.09% items with score of 2.5 or better on the pretest, whereas Cohort 2 had 40.04% of the items with these scores. The nonequivalence of the two forms made it impossible to talk about gains. Adjusting for differences in the forms did not clarify the data analysis.

The most conservative solution to this problem of analysis was to consider the assessment as one of the program in general, rather than an assessment of individuals. (This is similar to QUASAR's treatment of data from this instrument.) Thus, the analysis is of test forms separately, recognizing that the people who had posttest scores were not the same people who had pretest scores on a given form.

Cossey and Holthuis examined the shift in the sample from low scores to medium or high scores on the posttest. For this purpose, an average score was calculated for each student. These averages were then classified on the pre- and the posttest according to three categories: low (0–1.5); medium (1.52–2.5); and high (2.52–4.0). Students in the low category were those who were not able to communicate their strategies, a dimension required for higher scores. Figures 10.4 and 10.5 show the distribution of scores on the pretest and posttest by test form. These two bar charts indicate that a greater percentage of students were able to achieve a high average score on the posttest than on the pretest, regardless of form. Conversely, fewer students scored low on the posttest than on the pretest. On form 1, 39.9% of the students fell in the medium category on the posttest as compared with 31.9% on the pretest. The equivalent figures for form 2 were 42.1% on the posttest and 39.6% on the pretest. On this form, the easier of the two, the sharpest rise in percentage of students was in the high category (from 22.6% to 35.1%). While these results are encouraging, the grand mean of average scores on both posttests falls within the medium range.

There were no effects of grade on individual pretest or posttest scores either for the percentage of points earned on the whole test or for the percentage of items with scores over 2.5. There was, however, a positive correlation between CTBS reading score and the percentage of questions scored

Figure 10.4: Comparison of Distribution of Total Scores on Pretest and Posttest, QCAL: Test Form 1, Mathematics, Grades 7 & 8, 1992–93 (N = 159)

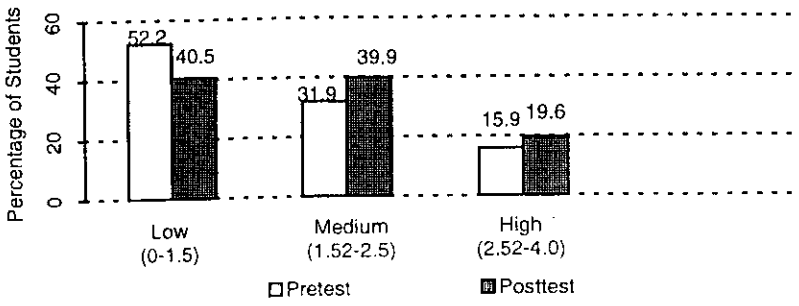
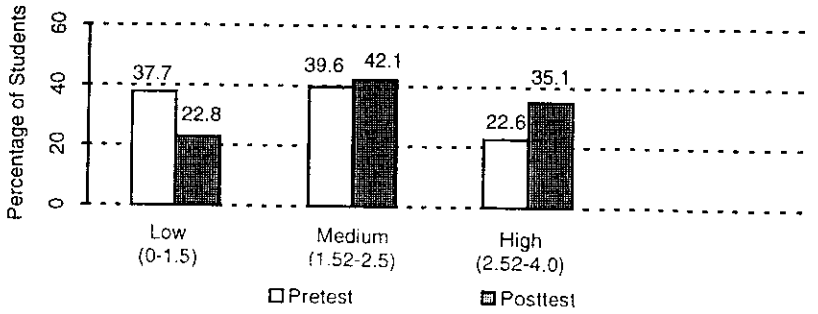


Figure 10.5: Comparison of Distribution of Total Scores on Pretest and Posttest: Test Form 2, Mathematics, Grades 7 & 8, 1992–93 (N = 159)



high on the pretest (r = .554, p < .001) and on the posttest (r = .516, p < .001). There were no gender differences on the pre- and posttest scores.

Interpretation of Results in Mathematics. The analysis documented a modest shift among these students in the direction of improved mathematical communication. Even if the test forms had been equivalent, it was more desirable to examine shifts in overall categories than the average size of individual gain scores. Individuals can show changes in scores from pre- to posttest that represent a change from no understanding to a very weak understanding of what was required. Although the absolute size of the gain score may look impressive, this person is not reaching the level of performance desired. This person's gain is indistinguishable from that of a student who gained the same number of points, but actually moved from an average to an excellent understanding and ability to communicate. The shift to higher categories is especially important because scores in these categories were only given to students who showed some skill in mathematical communication.

Of all the multiple-ability curricula, the program in mathematics was the most challenging from the teachers' perspective. Because it was developed in line with the most recent reforms in mathematics, teachers were quite unfamiliar with the goals and objectives, such as improved mathematical communication. Moreover, there was no supporting textbook for the units, nor were there specially developed supporting materials. Thus, beyond the group activities and what staff developers offered, teachers had to draw on their own understanding of the mathematics involved. When students arrived at novel solutions to the open-ended problems, it was up to the teacher to provide feedback as to whether they had gone astray entirely or had a

solution that reflected the central concepts, such as proportionality. Few teachers had the mathematical background to do this. Thus, modest results for many of the classrooms are not too surprising.

Science

Julie Bianchini conducted an extensive evaluation of CI classrooms using the Human Biology Middle Grades Life Science Curriculum (Hum Bio) (Stanford University Middle Grades Life Science Curriculum Project, Field Test Version, 1994). The goal of this curriculum is to challenge students to learn key science concepts, to apply scientific information to real-world situations, and to practice decision-making skills. The text, laboratory activities, and multiple-ability group activities integrate the natural and social sciences (Heller & Kiely, 1997). As a result of a collaboration between the Program for CI and the Human Biology Project, Bianchini and Nicole Holthuis created multiple-ability activities suitable for CI to accompany selected units of the curriculum. They designed these activities to accompany textbook modules and laboratory activities that were in the process of development.

Bianchini and her team conducted an evaluation study of 13 middle school science classrooms over the course of 2 years. During the 1992–93 school year, approximately 260 sixth- and eighth-grade students in 10 classrooms participated, as did 80 sixth-grade students in three classrooms during 1993–94.

Nature of the Tests. Bianchini constructed content- referenced tests for four of the six Hum Bio units: Circulation, Respiration, Digestion, and Systems. The purpose of the tests was to assess students' factual knowledge, conceptual understanding, and ability to apply and synthesize scientific information. In constructing tests for particular units, Bianchini tried to reflect the multiple-ability nature of CI group tasks within the constraints of a paper-and-pencil format. Many of the questions contained diagrams or illustrations in an attempt to make the test less reading-dependent and easier to understand. For example, in one question regarding the concept of systems, students were given a drawing of an ecological system, including a forest, stream, factory, and polluted factory waste water. They were asked to identify the system's components, to offer three consequences of the system, and to predict the effects of changing one component. Other questions asked students to represent their knowledge with pictures instead of words. Many of the questions were open-ended, requiring students to construct their own short answers.

Administration and Scoring. Teachers administered tests on a unit-by-unit basis: They gave a pretest prior to implementation and a posttest upon unit completion. During the 1992–93 school year, sixth- and eighth-grade students completed one to three CI unit tests: Systems, Circulation 1, and/or Digestion. The following school year, sixth-grade students completed two CI unit tests: Circulation 2 (a revised Circulation 1 test) and Respiration.

Researchers scored these unit tests in 2–3 hour blocks over the course of a year. The rubrics included both criterion- and norm-referenced guidelines. For example, for the Systems unit test, the researchers established a point scale and a set of general criteria for each of the six questions. Then they used a sample of students' pre- and posttests to construct more specific guidelines for those questions requiring open-ended responses. They made adjustments to the range of acceptable answers and/or the point scale to better fit the kind and quality of student work.

There were one to seven rounds of reliability completed for each open-ended question. The researchers began by selecting a set of six to ten pretests and posttests. After each researcher had given initial scores to the set, the group discussed answers and reached consensus on final scores. Then reliability for each researcher was calculated by dividing the number of initial scores that matched final scores by the total number of scores. If the reliability score or calculation was low, researchers pulled a different set of tests and completed another round. The group did not move to scoring until after they felt very comfortable with the question. For several questions, researchers were unable to achieve individual reliability. For these items, the questions were scored in pairs or as an entire group. The average percentage agreement among the scorers was 80%. On average, researchers spent 75 hours per 100 tests.

Results of Science Tests. The five tests were not of equal value (a total of 56 points was possible on Systems, 43 on Respiration, 51 on Digestion, 93 on Circulation 1, and 83 on Circulation 2). To make the tests comparable, percentage totals were calculated for the pretest, posttest, and gain scores by dividing the test scores by the total number of points possible. Table 10.6 provides a summary of the percentage pretest, posttest, and gain scores for each test.

T-tests indicated that the posttest scores were significantly higher than the pretest scores for each of the five tests ($p < .001$ for each). The average scores on the Systems pretest (36.0%) and posttest (60.9%) were the highest of the five tests. Learning gains were greatest for Systems (24.9%) and lowest for Digestion (7.3%).

Because a major objective of CI is the development of higher-order concepts and processes, it is important to analyze higher-order questions separately. Bianchini, Holthuis, and Nielsen (1995) categorized higher-order questions as those that asked students to apply, analyze, and/or synthesize

Table 10.6: Percentage Correct for Pretests, Posttests, and Percentage Gain Scores: Science Testing with Five Tests, Grades 6 and 8, 1992–93 and 1993–94

Test	*n*	% Pretest	%Posttest	% Gain
Systems	206	36.0	60.9	24.9
		(16.2)[a]	(18.9)	(17.4)
Respiration	65	27.4	42.2	14.8
		(13.3)	(19.1)	(11.8)
Digestion	172	35.4	42.7	7.3
		(16.2)	(16.8)	(10.8)
Circulation 1	135	21.6	35.6	14.0
		(10.2)	(14.4)	(10.5)
Circulation 2	69	17.3	36.1	18.8
		(8.2)	(18.1)	(12.8)

[a] *SD* in parentheses

scientific knowledge. When these researchers analyzed the percentage of pretest, posttest, and gain scores on the base of the total number of higher-order questions per test, they found similar trends to those for the overall scores. The posttest scores on the higher-order questions were significantly higher than the pretest scores for each of the five tests, as determined by *t*-tests. Gains on higher-order thinking items were largest on Systems (students scored 21.8% better on the posttest) and smallest on Digestion (5.6%).

The inclusion of diagrams and pictures in many of the test questions and the requirement for drawings and diagrams in answers were strategies intended to make the tests more accessible. However, students did not consistently score higher on the pictorial questions. A qualitative analysis (Bianchini et al., 1995) suggested that in some instances pictorial questions constrained or confused student responses. In other instances, students clearly benefited from the acceptance of drawings as answers; they were better able to convey what they knew through an illustration than through words. Despite these and other attempts to make tests more accessible to all students, reading scores were significantly correlated with pre- and posttest scores on each of the five unit tests. Moreover, reading scores were significantly correlated with percentage gain scores on Respiration, Circulation 1, and Circulation 2.

Given the widely discussed gender gap in science achievement, it was particularly important to compare the scores of boys and girls. Because some tests were given to eighth graders, it was possible to see whether there was

evidence of a gender gap beginning among the older students. Among the sixth graders, although gain scores for boys and girls did not differ significantly, girls scored significantly higher than boys on some of the pre- and postests (Circulation 1 and Digestion). The evidence from the eighth grade is mixed. Girls did significantly better than boys on the Systems pretest, but boys made significantly greater gains than girls on that test. On the digestion test, girls once again had significantly higher pretest scores, but there was no difference between the boys and girls in the percentage gain.

Interpretation of Science Results. The evaluation documented significant gains on all of these assessment instruments. Of course, without standard for comparison, it is impossible to know if the students gained more than they would have without the use of CI. Even from the perspective of absolute scores, it is difficult to assess whether the scores signal a reasonable level of understanding of the topics covered. The only a priori criterion of what students were supposed to gain was a better understanding of the central concept of each unit of multiple-ability activities. There were no a priori criteria of what additional skills, factual knowledge, and concepts students were supposed to learn about circulation, digestion, and respiration.

Most of these students had little background in science. Although teachers had the draft of the new textbook for the Middle Grades Life Science Curriculum as well as laboratory exercises suitable for many of the topics (Stanford University Middle Grades Life Science Curriculum Project, Field Test Version, 1994), they varied as to how much use they made of these materials. Thus, some teachers relied heavily on multiple-ability group activities to do the bulk of the teaching, something the activities were not designed to do. This was not helpful to students with a minimal background in science, who therefore came to the group activities with a minimum of understanding and orientation. In addition, some students lacked skills required by the test, such as the ability to draw a well-labeled diagram.

Some of these units, like Circulation, were much larger than others, such as Systems. Circulation had more than 15 multiple-ability activities available to the teacher. Not all of these activities were used, nor was all the other available material from the unit presented. As a result, it was probably the case that some students were tested on material they had never studied or experienced.

The use of open-ended assessment requiring scoring with rubrics proved to be an expensive and time-consuming strategy. Nevertheless, the answers provided by students were a rich source of information about the strategies and limitations of the curriculum, their understanding of science, and particularly their misconceptions and gaps in background.

SUMMARY AND IMPLICATIONS

In this chapter we have chronicled years of achievement testing with standardized tests and with content-referenced tests for elementary students and for middle school students. Sometimes we have examined the absolute gains from pretest to posttest, and sometimes we have made comparisons between CI classrooms and other classrooms that did not use these strategies but covered similar curricular materials. Overall, students in CI classrooms showed significant gains from pre- to posttests and in comparison to students in other classrooms. The learning gains are both in the areas of factual knowledge and in higher-order thinking skills. The younger children show impressive gains on standardized achievement tests in math computation and in math concepts and application. They also significantly increased their understanding of science concepts and vocabulary and their ability to apply these ideas. At the middle school level, we have demonstrated gains in knowledge of science, social studies, and mathematics. Only in social studies do we have reasonable data from comparison classrooms, permitting the conclusion that CI resulted in greater improvement in higher-order thinking skills than alternative educational treatments.

We have been frank about the limitations of any one of these evaluations. Many lack comparison groups. Others used tests that were less than ideal. Still others revealed problems in test administration. Yet, across all these studies, there is real strength in the consistency of significant learning gains for CI students across varying content, grade levels, individual achievement levels, levels of English language proficiency, and different schools. One could select only the most successful of these evaluations for presentation, but we believe that it is the *array* of findings over a long time under varying conditions that is truly impressive.

Lessons Learned

These results illustrate some important underlying lessons to be learned from a study of achievement data. The first lesson is that the strength of the gains in test scores appears to depend on the match of the test to the curriculum. For example, the gains in middle school mathematics were not very great, but the test was by no means an exact match to the curriculum. Similarly, if we examine the outstanding gains on the Systems unit in the science curriculum, we can see that this success was partly a product of a brief, self-contained unit where the test items faithfully reflected the nature of the group activities and the central concept of the unit. This was less true of the other units where items referred to a wider body of knowledge and did not always reflect group activities. In middle school social studies, where the test questions reflected

the analogies students had to construct in the course of activities, there was a clear superiority in higher-order thinking for students who had experienced these activities in comparison to students who had not.

The second lesson was a sad but important generalization: There will not be impressive learning gains if CI is not implemented properly and/or if there are severe problems of classroom management (see Chapter 3). The second year of testing in social studies revealed weak gains for the seventh-grade classrooms, gains that were smaller than those of the eighth grade classrooms. Analysis showed that the smaller achievement gains were a product of failure to rotate students between activities, failure to give adequate time to the curriculum, and high rates of student disengagement in some of the classrooms. It is significant that these failures took place in a school that did not give teachers adequate support. It would appear that CI requires a modicum of organizational and classroom "health" for successful learning outcomes.

In addition, teachers need considerable support from staff developers and from department chairs if they are to spend more time on particular curricular themes or topics. If teachers feel pressured to cover curriculum by their departments or by their own desire to get through the textbook, they are likely to sacrifice rotation in CI. This means that each student participates in only one activity exemplifying the central concept, a serious dilution of the organizing principle of the curriculum. This particular problem was characteristic of middle schools and was not present at the elementary school level.

A third lesson to be learned from the results is the importance, beginning at the level of middle school, of supporting materials to prepare students before they begin the group activities. Learning outcomes for the middle school typically require more than small group activities. The understanding and knowledge that is tested cannot all be achieved in a group activity.

In the case of the new mathematics curriculum, there was no textbook and a lack of supporting material. Some of the modest learning gains probably should be attributed to these problems. In the case of some of the larger science units, such as Circulation, the teachers varied in their use of textual material and often omitted laboratory activities. Qualitative analysis showed that students had many misconceptions and lacked a basic understanding of the scientific method. They needed something more than the group activities in order to do well on the test.

An excellent example of the need for basic instruction in addition to CI was observable in the responses of students when asked why they answered some items the way they did on the seventh-grade Crusader unit. The responses made it painfully clear that they didn't know what was meant by the term *Muslim*. Teachers (and curriculum developers) had assumed this

word was understood; thus, they never included necessary direct instruction and explanation. This lacuna caused students considerable difficulty on the test.

It is necessary for the teacher to combine CI with other teaching activities. Materials from the Teachers' Curriculum Institute include excellent slides for slide lectures and discussions, and exercises in other social studies skills designed for pairs of students. All of these activities are intended to accompany textbooks. The results in achievement for the eighth graders using these materials are markedly superior to the results for the seventh graders who did not have the rich, supporting materials and activities.

The final lesson has to do with the relationship of reading skills to the test scores. Despite efforts to revise the tests to make them less dependent on reading skills, there was a continuing relationship between an individual's reading score and gains on these tests. We have frequently found these relationships between gain scores and reading scores, even when the test items are read out loud in English and Spanish. Such a correlation does not tell us whether the problem lies in an inability to read and comprehend the test, or with a failure to learn in response to instruction and interaction.

In some analyses, holding the individual's reading score constant, there were additional effects of the percentage of students in the class with low reading scores. Teachers in such classes may be less demanding of their students, or there may be a lack of resources within many of the small groups to interpret the activity cards and the individual reports. For optimal results with CI, it is necessary to seek out truly heterogeneous classrooms where there are some students in each group who represent grade level reading skills.

Unmeasured Results

We would like to close this chapter with the issue of what we are and are not measuring with the instruments described. CI always involves multiple-ability curricula as well as particular instructional strategies. It is not possible to generalize about what students have learned as a result of CI without dealing with the particular content of these curricula. Given the range of conventional achievement measures we have used, we have only measured learning objectives that relate to the content of the curricula.

There are other intellectual and social skills that are not measured by these tests. Students in CI learn to deal with uncertain problems and to use resources among fellow students rather than to depend on adult authority; they also learn how to plan and carry out projects as a cooperative group. These are the kinds of learning outcomes that we have not measured and do not even know how to measure. Yet, in some sense, these are among

the important learning outcomes that teachers and students discuss when they talk about CI.

We have confined ourselves in this chapter to paper-and-pencil instruments administered at the individual level. Yet students in this approach do their learning in groups; and their learning is of a particularly active variety. Clearly, performance assessments of individuals and groups have a better chance of reflecting the outcome of a classroom experience that features active group learning. Performance assessments might have the additional advantage of overcoming the persistent relationship we have documented between reading scores and measures of what students have learned.

The task of assessing what students have learned, although not a standard sociological problem, has raised many new questions relevant to the work of sociologists of education. Sociologists use data from achievement tests as a dependent variable, but they do not often reflect on the nature of these instruments. Chapter 11 considers limitations and potential of paper-and-pencil assessment instruments from a sociological perspective.

NOTES

1. Jennifer Whitcomb and Elizabeth Cohen took primary responsibility for constructing the seventh-grade test. Many of the factual items were adapted from published tests of the textbook *Across the Centuries* (Armento, Nash, Salter, & Wixson, 1991a).

2. Bert Bower and Elizabeth Cohen took primary responsibility for constructing the eighth-grade test. All content was available in the text, *A More Perfect Union* Armento, Nash, Salter, & Wixson, 1991b). Seven teachers reviewed a prototype of this test. Based on their comments, Bower and Cohen made changes.

3. The Teachers' Curriculum Institute, under the leadership of Bert Bow produces commercially available social studies materials that contain slide lectures, activities for pairs, and multiple-ability activities for four- and five-person groups.

4. QUASAR uses different but overlapping versions for seventh and eighth graders. Cossey used the same form for both grades. Technical information about reliability and validity studies of QCAI is provided in Lane, Stone, Ankenmann, and Liu (1994); Lane, Liu, Stone, and Ankenmann (1993); and Stone, Ankenmann, Lane, and Liu (1993).