Data Collection

This study used five primary sources of data from Suburban High School. First, it analyzed results of an Algebra Achievement test completed by eleventh graders in one Traditional cohort and two Reform cohorts. Second, it used student scores from a sixth grade test administered by the Educational Records Bureau as a covariate. Third, it analyzed transcripts from an automated data base containing information from the spring of 1991 through the spring of 2001. Fourth, it analyzed documents provided by the school, including course syllabi and annual school profiles. Fifth, it analyzed information provided in conversations with key informants at the school.

Algebra Achievement Test

This study used a 3-part Algebra Achievement test designed by the Core-Plus Mathematics Project. Part 1 emphasized the type of contextualized problem solving that is typical of Core-Plus, IMP, and other reform curricula. Part 2 emphasized problems typical of traditional mathematics curricula: context-free symbolic manipulations that call for transformation of algebraic expressions and solutions of equations and systems. Items in Part 2 were adapted from released ACT examinations and from items that commonly appeared on college placement tests. Part 3 required collaborative work on a single extensive open-ended problem and was designed to be completed by students in pairs. The Algebra Achievement test was intended to be administered at the end of Grade 11 and focuses on algebra topics that are generally completed by that time.

The Algebra Achievement test designed by Core-Plus has several advantages. Like the IMP, Core-Plus is a curriculum developed under a National Science Foundation grant to implement the NCTM *Standards* at the high school level. The Algebra Achievement test was designed specifically to fulfill the purpose of the proposed study: to compare the effects of a *Standards* -based curriculum to those of more conventional curricula. Since this study compares learning under the IMP curriculum to learning under a more traditional curriculum, it is important to use a test that is fair to both. The Algebra

Achievement test accomplishes this, by measuring both the kind of problem solving and applications emphasized by the NCTM *Standards*, as well as more traditional mathematics skills. Further, there is no chance that the test was unconsciously "tailored" to favor either the IMP or the Traditional curriculum, as this test was not designed by the researchers in this study or by anyone involved with either curriculum.

In order to sample a wide variety of problems, the Core-Plus researchers designed four parallel forms for Part 1, two parallel forms for Part 2, and three parallel forms for part 3. They administered the test via matrix sampling; that is, each student was randomly given one form for each of the three parts of the test. However, matrix sampling was not feasible at Suburban High School, given both the smaller sample size and the desire of Suburban High School teachers to maintain a simple testing program so results could be easily explained to the community. Therefore, this study used one form for each part of the test, selected by teachers at Suburban High: Part 1, form C; Part 2, form A; and Part 3, form A. As noted by the test authors (Huntley, et al., 2000), scores across forms of this test tend to be consistent, so the decision to use only one form was expected have little negative impact on the validity of results at Suburban High. The three parts of the Algebra Achievement test used in this study are contained in Appendix A.

Testing in the spring of 1997: Traditional cohort. In the spring of 1997, Suburban High School students in the Traditional cohort were in eleventh grade and nearly all of them were enrolled in mathematics. They completed the three parts of the Algebra Achievement test in mathematics class during two days in May 1997. On the first day of testing, individual students completed Part 1 of testing. On the second day of testing, individual students completed Part 2. Then, students within classrooms chose partners and together these pairs completed Part 3 of the test. Suburban High School mathematics teachers conducted the 1997 testing and archived the results so it would be possible in later years to compare the achievement of students who had studied under the

new curriculum and schedule to that of the 1997 eleventh graders, who had studied under a traditional curriculum and schedule. In 1997, 89.9% of eligible eleventh graders participated in at least one day of testing. Some of the students who missed the test were unable to participate because of school-scheduled extracurricular activities, and others did not participate due to absence.

Pilot Testing in the spring of 1999. During two days in May 1999, teachers at Suburban High School administered the Core-Plus Algebra test to eleventh graders school-wide. A pilot study compared results of this assessment to those of the May 1997 assessment. Lessons learned from the pilot study indicated that a number of steps needed to be taken to ensure that future testing conditions would be as close as possible to what they had been in 1997. Specifically, in 1999 many students were administered the test in settings that did not resemble a mathematics class, proctored by a non-mathematics teacher who did not create a serious atmosphere. Often, calculators were not available when they should have been. These problems were corrected in the spring of 2000, when the testing to be used for this proposed study was conducted.

Testing in the spring of 2000: First Reform cohort. Suburban High School students in the First Reform cohort completed the three parts of the Core-Plus Algebra test during two days in May 2000, when they were in eleventh grade. Because of the semestered block schedule, many eleventh graders were not enrolled in mathematics during this spring semester. Therefore, for the one hour needed each day for test administration, eleventh graders moved to a mathematics classroom or other classroom proctored by a mathematics teachers—or, in a few cases, by a science teacher. Since all students were enrolled in English during the second semester of eleventh grade, the classroom to which students reported was determined by their English class. In 2000, 90.4% of eligible eleventh graders participated in at least one day of testing. As before,

some of the students who missed the test were unable to participate because of school-scheduled extracurricular activities, and others did not participate due to absence.

Observers reported that the atmosphere and testing conditions in 2000 were very similar to what they had been in 1997. However, discussions after the testing raised concern about the way students were assigned to pairs during the second day of test administration. As in 1997, individual students completed Part 1 on the first day of testing and Part 2 at the beginning of the second day of testing. Then, students within classrooms chose partners and together these pairs completed Part 3 of the test. However, in 2000 students were tested within English class groupings, so it was likely that many pairs consisted of students who had completed differing levels of mathematics. This contrasted with the situation in 1997, when students were tested within a mathematics class, and so automatically paired with another student who had completed the same level of mathematics. Testing conditions in 2001 were adjusted to correct this potential problem.

Testing in the spring of 2001: Second Reform cohort. Suburban High School students in the Second Reform cohort completed the three parts of the Core-Plus Algebra test during two days in May 2001, when they were in eleventh grade. Testing conditions were the same as in the spring of 2000, with two exceptions.

First, students were given class credit for showing up at the test. This change was intended to increase the participation rate, and may have been marginally successful in doing so. In 2001 91.4% of eligible juniors participated in at least one day of testing.

Second, when administering Part 3, teachers requested that when choosing partners, students select someone whose most recent mathematics course was the same

level as their own. This was intended to make testing conditions more similar to what they had been in 1997.

Scoring procedures. The Algebra Achievement test contained open-ended questions that needed be scored using a rubric. For this study, a number of changes were made to the rubric used by the original designers of the test, so that the rubric would be easier to use validly and reliably. The most important change was the selection of anchor papers and practice papers, keyed to each possible score for each item in Parts 1, 2, and 3. In almost all cases, anchor papers and practice papers were selected from actual student responses to earlier administrations of the assessment that had been conducted by Core-Plus researchers. In the few instances where no student paper exemplified a particular response covered by the rubric, this researcher developed an appropriate "anchor paper". Procedures for training scorers for this study were based on professional standards used for the National Assessment of Educational Progress (NAEP), as described by Bourgeacq, et al. (1997). Appendix B contains the rubric used for scoring, and appendices C and D contain the anchor and practice papers used in training.

Because Part 3 is the most difficult section to score, each student submission of Part 3 was reviewed by three independent raters. The raters used a scale of 0 to 4. Two independent raters scored each question on Part 1 and Part 2 of the test. In cases of disagreement, raters reached consensus by discussion and persuasion, not voting. Part 1 and Part 3 were scored by an expert panel of college mathematics professors and retired high school mathematics teachers. Because Part 2 was relatively easier to grade, it was scored by two undergraduate mathematics majors.

Scoring was accomplished at two separate times. In the winter of 1999-2000, all tests from the Traditional cohort and the Pilot cohort were scored, as part of the pilot study. Then, in the summer and fall of 2001, Part 2 and Part 3 of the tests from 1997 were re-scored, and Part 1, Part 2, and Part 3 of the tests completed in 2000 or 2001 were scored. The 1997 Part 2 tests were re-scored because it proved impossible to get the same individuals who had scored Part 2 tests for the pilot study to complete the scoring in 2001, and it was deemed important to have the same raters for tests completed by students in the Traditional cohort and tests completed by students in the Reform cohorts.

Scorers who had graded Part 1 and Part 3 of the test in the pilot study were available to complete scoring in 2001. Before beginning to score the new tests, the scorers completed a "drift test" by re-scoring 20 tests that they had scored during the pilot study. The 20 tests to be re-scored were mixed in with 20 new tests, so that the scorers were more or less blind as to whether they were re-scoring an old test or scoring a new test.

The drift test found that there may have been a systematic difference between original scores and re-scores for Part 3. The consensus score was higher on the re-score than on the original score for 3 of the 20 tests, and the same on the re-score as on the original score for remaining 17 of the 20 tests. With the re-score being higher 15% of the time, it was possible that scorers had drifted towards scoring more leniently. The first attempted to correct this was by re-training. However, a second drift test found that the Part 3 scorers had overcompensated, consistently scoring more harshly on the re-score. Therefore, it was decided to mix the 1997, 2000, and 2001 Part 3 tests together, and have the raters score tests from all three cohorts at once. This new set of scores was used for

all analyses in this study; scores on Part 2 and Part 3 that had been computed during the pilot study were discarded.

In contrast to results for Part 3, results from the drift test for Part 1 indicated that there was no systematic difference between scores that had been assigned in the pilot study and scores assigned in the drift test: A 95% confidence-interval showed that the difference between the original score and re-score was probably between -.06 standard deviations and +.05 standard deviations. Therefore, it was decided to utilize scores for Part 1 of the 1997 tests that had been computed during the pilot study, and scores for Part 1 of the 2000 and 2001 tests that were computed during 2001.

Agreement between the raters who scored Parts 1 and 2 was remarkably high. Recall that one pair of raters scored Part 1, questions 1 and 2; one pair of raters scored Part 1, question 3; two pairs of raters scored Part 1, question 4, and one pair of raters scored Part 2. Among all these pairs of raters, the correlation between a student's score assigned by the first rater and that same student's score assigned by the second rater ranged from a low of .991 for the "least agreeing" pair of raters to a high of .998 for the "most agreeing" pair of raters. This high agreement indicates that procedures for scoring Parts 1 and 2 erred on the side of caution. The combination of rubrics, anchor items, practice papers and training yielded scoring of very high reliability.

For Parts 1 and 2, each pair of raters scored between four and nineteen separate items; the correlation between the raters was based on the mean of all the items that pair scored. In contrast, raters for Part 3 scored only one item; moreover, the item they scored was particularly involved and difficult to score. The relative difficulty of getting a reliable score for Part 3 was the reason for using the consensus among three scorers in

order to determine students' scores. Given the difficulty of the task, agreement among the raters was reasonably high. The correlation between scores assigned by Rater 1 and those assigned by Rater 2 was .892; the correlation between scores assigned by Rater 2 and those assigned by Rater 3 was .849; the correlation between scores assigned by Rater 1 and those assigned by Rater 3 was .904. The correlation with the consensus score was: for Rater 1, .959; for Rater 2, .909; for Rater 3, .909.

Student Test Scores for Grade 6

Since the early 1980s, the school district where Suburban High School is located has conducted yearly testing using an exam designed by the Educational Records Bureau. Most years, the testing was conducted for all grades from 3 through 10. Since 1996, testing reports have been available on computer disks. The school system has kept an archive containing hard copies of student scores prior to that time. As explained below, Grade 6 test scores from the Traditional and from the two Reform cohorts were used as a covariate in this study.

Transcripts

Suburban High School maintains student transcripts on an automated database. This study used the data base to examine complete transcripts for students in the graduating classes of 1995 through 2001, plus incomplete transcripts available as of spring, 2001 for graduating classes of 2002, 2003, and 2004.

Documents

This study examined the following documents:

1. Syllabi for courses taken by students in the Traditional cohort and in the Experimental cohort;

2. "Complementary Materials" designed as supplementary mathematics resources by Suburban High School teachers;
3. Yearly *School Profiles* published the district in which Suburban High School resides;
4. Yearly *Testing Reports* published the district in which Suburban High School resides.

The Complementary Materials contain page references for readings and problems in the traditional Algebra and Geometry texts that complement topics covered in the IMP modules. Although the Suburban High School did not develop the Complementary Materials until the third year they were utilizing IMP, today every student has access to these traditional textbooks, and their teachers use these Complementary Materials to devise supplemental assignments. The yearly *School Profiles* describe student achievement the preceding year on various measures including participation rate and grades in Advanced Placement exams. *School Profiles* from 1995 through 2001 were available. The yearly *Testing Reports* describe results of Grade 3-10 testing using a test published by the Educational Records Bureau.

Key Informants

This study was completed in close collaboration with two key informants:

1. Mrs. Sullivan, the former mathematics department chair at Suburban High School, who was responsible for implementing the IMP curriculum, and
2. One of the mathematics teachers who first taught IMP at Suburban High School. She spent a year on sabbatical working as an IMP trainer with teachers at Suburban and other high schools, and has since returned to her

teaching position at Suburban High School.

Other information was provided by the current mathematics department chair at Suburban High School.