Limitations: Threats to Internal Validity

Threats to Internal Validity Attributable to Non-Random Group Assignment.

In a "true experiment" subjects are randomly assigned to treatment and control groups. If assignment is nonrandom, then the research design is referred to as a "quasi-experiment," and the two groups investigated are commonly referred to as the "treatment" group and the "comparison" group (Krathwohl, 1993).

By this definition, the current study was a quasi-experiment. This increases the risk that there may have been pre-existing differences between students in the Traditional cohort and students in the Reform cohorts at the time they entered ninth grade. The two groups were chosen to be comparable in many ways. They lived in the same community. Many attended the same middle school and used the same middle school mathematics curriculum. They attended the same high school. Due to low staff turnover, they had many of the same teachers both in middle school and high school. Figuratively and in many cases literally, students in the Reform cohorts were the younger brothers and sisters of students in the Traditional cohort.

However, the possibility that some of the results reported here may have been due to a different experience in middle school mathematics class cannot be completely discounted. Both the Traditional cohort and the Reform cohorts used the *Visual Mathematics* curriculum (Foreman & Bennett, 1991) in middle school. However, the Traditional cohort was the first group at the school to use that curriculum. When an innovative program like *Visual Mathematics* is adopted, there is often an "implementation dip" resulting in lower achievement for the first few years (Busick & Inos, 1992; Fullan & Miles, 1992). The Traditional and Reform cohorts had similar scores on mathematics ability in sixth grade, but as Table 2 (see Chapter 3) shows the Traditional cohort had lower scores in seventh grade. As has been discussed in Chapter 3 of this study, this was likely due to a change in the test. Nonetheless, given the possible "implementation dip" the lower scores may have reflected a true difference in what was learned during middle school.

"History" could have affected results in other unknown ways. For example, the size of the junior class increased from 227 in 1996-97 to 279 in 2000-01. Perhaps there were other demographic changes that could have affected test results.

Other Threats to Internal Validity

There were other factors that could have affected achievement and were impossible to control. For example, the individuals teaching both calculus and pre-calculus changed in 1999-00, coincident with the arrival of the first largely-IMP group in the calculus class. Further, Suburban High School adopted a new Calculus textbook, *Calculus: Graphical, Numerical, and Algebraic* (Finney, et al., 1999). The Calculus teacher believed the new text was more in line with the evolving reforms in the AP Calculus curriculum. Together, the change in staff and text provide a possible alternate explanation for some of the success in the AP Calculus program. Various other staffing changes could have affected results on the Algebra Achievement test.

Every effort was made to keep the Algebra Achievement test a low stakes assessment. Teachers at Suburban High school did not review it carefully, and were unlikely to have identified specific topics on the test that they should teach. Nonetheless, the test was administered repeatedly over several years, so the possibility that some teachers unconsciously modified their instruction to accommodate material on the test cannot be completely discounted.

While every effort was made to make testing conditions as comparable as possible for the three administrations of the test, once the block schedule was implemented it was not possible to replicate precisely the testing conditions that had occurred in 1997. Changes made in test administration due to lessons learned during the pilot study partly addressed this problem, but could not overcome it completely. For example, in 1997, students completed the Algebra Achievement test during their mathematics class. In 2000 and 2001, not all students were enrolled in mathematics at the time of testing, so they were administered the test in English classrooms, proctored by mathematics teachers. This meant that pairs completing Test 3 in the Reform cohorts were often from different mathematics classes, whereas in 1997 they had been from the same mathematics class. Due to fears that this difference in pairing might affect results, teachers in 2001 requested that students within each English class pair with someone who had been in their mathematics class. While there is no guarantee that all students complied with this request, many were reported to have done so. Concern that results on Test 3 may have

been due to the way students were assigned to pairs is mitigated by the fact that in 2001, when students were generally paired with partners from the same mathematics class, scores on Test 3 were even higher than in 2000 when no attempt had been made to replicate 1997 testing conditions by matching pairs within a mathematics class.

It should also be noted that only about 90% of eligible eleventh graders completed the Algebra Achievement test each year. Non-participants included some students who were committed to extra-curricular activities that conflicted with test administration times and students who were absent either because of illness or because they elected to skip the test. Had all students taken the test, results might have been different.

As described in Chapter 4, the decision to test students at the end of eleventh grade may have overestimated the achievement of students currently enrolled in mathematics class. A delayed test conducted at the beginning of twelfth grade may have provided a more realistic measure of how much of the algebra students learned by the end of eleventh grade was remembered and available for use in subsequent courses. Thus, the Algebra Achievement test as conducted may have biased results against students using the semestered block schedule, since a portion of them (28.7%) were not enrolled in mathematics class at the time of testing.

Finally, this study measured the overall effects of a semestered block schedule and the IMP curriculum on algebra learning throughout Grades 9-11. Under the Traditional schedule, approximately one-third of the students at Suburban High School enrolled in some form of second-year algebra in twelfth grade. The majority enrolled in Algebra 3/Trigonometry, a course mainly designed to review material they had previously studied, but some students were studying second-year algebra for the first time. Meanwhile, approximately one-third of the students in the Reform cohorts did not complete the core IMP sequence until twelfth grade, and thus had not yet had a full opportunity to learn the algebra content that was tested. Had the testing been conducted at the end of Grade 12, both groups would surely have known more algebra than they did at the end of Grade 11, and the results would have been different in unpredictable ways. It is particularly regrettable that most of the students in the Reform cohorts who were in Academic Assisted classes did not complete the core IMP sequence until twelfth grade, subsequent to the time of testing. Under the new program, Academic Assisted students were offered exposure to more sophisticated mathematics than had previously been the case, and it would have been desirable to measure the full effects of the program on this particular group.