

ABSTRACT

Title of dissertation: THE JOINT IMPACT OF BLOCK SCHEDULING AND
A STANDARDS-BASED CURRICULUM ON HIGH
SCHOOL ALGEBRA ACHIEVEMENT AND
MATHEMATICS COURSE TAKING

Steven Lee Kramer, Doctor of Philosophy, 2002

Dissertation directed by: Associate Professor Patricia F. Campbell
Department of Curriculum and Instruction

This study investigated mathematics learning at a suburban United States high school that simultaneously adopted a semestered (4 x 4) block schedule and the Interactive Mathematics Program (IMP), a curriculum designed to implement the National Council of Teachers of Mathematics (NCTM) *Standards*. Previous research has often found that at sites where a block schedule was adopted without changes to mathematics curriculum and instruction, mathematics achievement has declined. In contrast, when a semestered block schedule and the IMP curriculum were implemented jointly, with extra time allocated to planning and staff development, the two innovations were followed by improvements in student mathematics achievement.

At the end of Grade 11, compared to an earlier cohort of students who had used a traditional schedule and curriculum, students using the block schedule and IMP were better able to formulate algebraic models, interpret graphs and tables, solve algebra

problems presented in context, and work in pairs to solve an extended, open-ended, applied algebra problem. Students who had used a traditional schedule and curriculum were better able to perform symbolic procedures presented in standard format.

Over 4 years of high school, students using the block schedule and IMP on average spent 67 more hours enrolled in mathematics courses than had earlier groups of students who used a traditional schedule and curriculum. Of students using the block schedule and IMP, 58% completed four core courses from the IMP curriculum and went on to enroll in at least one additional advanced mathematics class. In contrast, only 48% of students using a traditional curriculum and schedule completed three core courses (algebra 1, geometry, and algebra 2) and went on to enroll in at least one additional advanced mathematics class.

Under the semestered block schedule about 15% of the student body enrolled in a newly offered Advanced Placement (AP) statistics course, and half of those enrolled completed the AP statistics exam for college credit. AP Calculus enrollment remained about constant after the school adopted the block schedule and IMP, but a larger percentage of AP Calculus students completed the more demanding BC course, and Advanced Placement exam grades improved.

THE JOINT IMPACT OF BLOCK SCHEDULING AND A STANDARDS-
BASED CURRICULUM ON HIGH SCHOOL ALGEBRA
ACHIEVEMENT AND MATHEMATICS COURSE TAKING

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2002

Advisory Committee:

Associate Professor Patricia F. Campbell, Chair/Advisor
Professor James T. Fey
Associate Professor Anna O. Graeber
Associate Professor Kenneth Berg
Associate Professor Paul J. Smith

©Copyright by
Steven Lee Kramer
2002

Acknowledgements

The research conducted for this study was supported in part by a grant from the National Science Foundation. The idea for the research project came from “Mrs. Sullivan” and the dedicated mathematics teachers at “Suburban High School,” and it was they who did the truly hard work of pioneering an exciting mathematics program. They deserve far more credit than I for whatever positive fruits this research project may bear. My doctoral advisor, Dr. Patricia Campbell and the regional director for IMP training in the Suburban High School area have provided critical support throughout this project. Lastly my wife, Risa, and my daughters Helen, Abby, and Frances have supported me with love throughout this endeavor.

Table of Contents

Acknowledgements	ii
Table of Contents	iii
List of Tables	xii
List of Figures.....	xv
Chapter 1: Rationale.....	1
Statement of the Problem.....	1
Standards-based Curricula	1
Block Scheduling.....	2
Possible Synergies Between a Standards-based Curriculum and a Semestered Block Schedule.....	7
Research Questions.....	8
Primary Research Questions	9
Secondary Research Questions	10
Overview of Method	11
Research Opportunity at Suburban High School	11
Testing the Traditional cohort.....	11
Testing the Pilot cohort.	12
Testing the first reform cohort.	12

Testing the second reform cohort.	13
Data Analyzed.....	13
Analysis of Algebra Achievement tests.	13
Transcript analysis.	14
Document analysis.	14
Key informants.	14
Definition of Key Terms	14
Chapter 2: Review of the Literature	17
Semestered (4x4) Block Scheduling.....	18
Effects of Semestered Block Scheduling on Mathematics Instruction.....	19
Reduced effectiveness of lecturing.	19
Decreased breadth and increased depth of coverage.	20
Need to adjust the mathematics curriculum.	21
Advanced placement classes.	22
Impact of student absences.	23
Retention of learning after a gap in sequential instruction.	23
Effects of Semestered Block Scheduling on Mathematics Achievement.....	26
Student grades.	26
Mathematics test scores: Canadian studies.	27
Mathematics test scores: United States studies.	30
Interaction of Semestered Block Schedule with Curriculum.....	34
Interaction of Semestered Block Schedule with Teaching Methods	35

The Interactive Mathematics Program (IMP)	38
IMP Synergies With a Semestered Block Schedule	40
IMP Connections to Learning Theory	42
Critiques of the IMP Curriculum	43
Wu’s detailed review and critique of the IMP curriculum.	43
Critiques of constructivism and situated cognition.....	46
Studies of Achievement under IMP and other Reform Mathematics Curricula.....	53
Studies of achievement under IMP’s sister curricula.	54
Studies of achievement under IMP.....	64
Chapter 3: Method	71
Participating Students	72
Participating Teachers	73
Professional Development	73
Planning Time.....	74
Traditional Schedule	74
Traditional Curriculum	75
Reform Cohorts	76
The Semestered Block Schedule.....	76
The IMP Curriculum.....	76
Schedule and Curriculum: Summary	81

Data Collection	83
Algebra Achievement Test	83
Testing in the spring of 1997: Traditional cohort.....	84
Pilot Testing in the spring of 1999.....	85
Testing in the spring of 2000: First Reform cohort.....	85
Testing in the spring of 2001: Second Reform cohort.....	86
Scoring procedures.....	87
Student Test Scores for Grade 6	90
Transcripts.....	90
Documents	90
Key Informants	91
Data Analysis: Eleventh Grade Algebra Tests.....	92
Covariate: Sixth Grade Test Scores	92
Dependent Variables.....	96
Test reliability.....	99
Statistical Methodology: Rules for Establishing Confidence Intervals.....	99
Statistical Models Used.....	100
Supplemental Analyses of Specific Skills	101
Data Analysis: Student Transcripts	102
Groups to be Compared	104
Statistical Methodology	105
Data Analysis: Documents and Key Informants.....	107

Summary of Method.....	107
Chapter 4: Results.....	109
Algebra Achievement at the End of Eleventh Grade	109
Results from Test 1: Understanding and Solving Algebra Problems presented in Context.....	110
Test 1 results without controlling for prior ability.....	111
Test 1 results after controlling for prior ability.....	113
Results from Test 2: Algebraic Symbol Manipulation	115
Test 2 results without controlling for prior ability.....	116
Test 2 results after controlling for prior ability.	117
Results from Test 3: Pairs of Students Solving an Extended Applied Algebra Problem	119
Test 3 results without controlling for prior ability.....	120
Test 3 results after controlling for prior ability.....	123
Eliminating two alternate explanations for Test 3 results.....	128
Test 3: Analysis of interaction for pairs with two Grade 6 ability estimates.....	130
Test 3: Opposite interaction pattern for pairs with one Grade 6 ability estimate.	Error! Bookmark not defined.
Test 3 interactions: A possible interpretation.	Error! Bookmark not defined.
Specific Effects of Traditional Curriculum/Schedule and IMP Curriculum/Block Schedule.....	Error! Bookmark not defined.
Evaluating expressions.....	Error! Bookmark not defined.

Simplifying expressions.....	Error! Bookmark not defined.
Solving linear equations.....	Error! Bookmark not defined.
Solving quadratic equations.....	Error! Bookmark not defined.
Representational fluency: translating the graph of a line to an equation.....	Error! Bookmark not defined.
Representational fluency: Non-procedural tasks.	Error! Bookmark not defined.
Formulation and interpretation of mathematical models.....	Error! Bookmark not defined.
Summary of detailed differences between groups.....	Error! Bookmark not defined.
Effects of Measuring Achievement the end of Grade 11 Instead of at Other Possible Measurement Times.....	Error! Bookmark not defined.
Possible underestimate of algebra achievement for those students not currently enrolled in mathematics.	Error! Bookmark not defined.
Unmeasured effects of twelfth-grade coursework. ...	Error! Bookmark not defined.
Transcript Analysis.....	Error! Bookmark not defined.
Results from Analysis of Complete High School Transcripts.....	Error! Bookmark not defined.
Comparison of total hours enrolled in mathematics class.	Error! Bookmark not defined.
Comparison of hours enrolled in advanced mathematics classes. .	Error! Bookmark not defined.
Advanced Placement Courses.....	Error! Bookmark not defined.
AP Course Enrollment.....	Error! Bookmark not defined.

AP Exam Grades.....	Error! Bookmark not defined.
Chapter 5: Summary and Conclusions	Error! Bookmark not defined.
Summary of Results on the Algebra Achievement Test ...	Error! Bookmark not defined.
Summary of Results from the Transcript Analysis	Error! Bookmark not defined.
Summary of Results Regarding Advanced Placement Courses	Error! Bookmark not defined.
Interaction of the Semestered Block Schedule and IMP Curriculum	Error! Bookmark not defined.
Limitations: Threats to Internal Validity	Error! Bookmark not defined.
Threats to Internal Validity Attributable to Non-Random Group Assignment.	Error! Bookmark not defined.
Other Threats to Internal Validity.....	Error! Bookmark not defined.
Some Strengths of the Research Design	Error! Bookmark not defined.
The Nature of the Implementation at Suburban High School	Error! Bookmark not defined.
Interaction of Community and Schedule	Error! Bookmark not defined.
Planning Time and Professional Development.....	Error! Bookmark not defined.
Modifications to the IMP Curriculum.....	Error! Bookmark not defined.
Visual Mathematics	Error! Bookmark not defined.
A Strong Grass Roots Leader	Error! Bookmark not defined.

Unusually Strong Mathematics Teachers**Error! Bookmark not defined.**

Collegial Mathematics Department**Error! Bookmark not defined.**

Time to Prepare for the Change**Error! Bookmark not defined.**

Elements of Community Support.....**Error! Bookmark not defined.**

Use of Multiple Measures

.....**Error! Bookmark not defined.**

Synergy of IMP and the Semestered Block Schedule

.....**Error! Bookmark not defined.**

Future Research

.....**Error! Bookmark not defined.**

References.....**Error! Bookmark not defined.**

Appendix A: Core-Plus Algebra Test Forms Used In This Study**Error! Bookmark not defined.**

Part 1: Understanding and Solving Algebra Problems in Context.....**Error! Bookmark not defined.**

Part 2: Symbol Manipulation**Error! Bookmark not defined.**

Part 3: Extended, Applied Algebra Problem**Error! Bookmark not defined.**

Appendix B: Scoring Rubrics	Error! Bookmark not defined.
Guiding Principles for Scoring Exams.....	Error! Bookmark not defined.
Rubric for Part 1.....	Error! Bookmark not defined.
Rubric for Part 2.....	Error! Bookmark not defined.
Rubric for Part 3.....	Error! Bookmark not defined.
Appendix C: Sample Anchor Papers	Error! Bookmark not defined.4
Sample Anchor Papers for Part 1, Problem 3.3.....	Error! Bookmark not defined.
Sample Anchor Papers for Part 2, Question 10	Error! Bookmark not defined.3
Sample Anchor Papers for Part 3.....	369
Appendix D: Sample Practice Papers	Error! Bookmark not defined.2
Key for Practice Set A, Part 1, Question 3	383
Practice Set A for Part 1, Question 3.3	Error! Bookmark not defined.
Key for Practice Set A, Part 2.....	Error! Bookmark not defined.
Practice Set A for Part 2, Question 10	Error! Bookmark not defined.
Key for Practice Set A, Part 3.....	Error! Bookmark not defined.
Five Sample Student Papers from Practice Set A, Part 3	Error! Bookmark not defined.

Appendix E: Detailed Description of IMP Modules

.....Error! Bookmark not defined.07

Appendix F: Methodology for Rasch Analysis and Multiple ImputationError!

Bookmark not defined.1

Multiple Imputation.....Error! Bookmark not defined.

The Rasch Model.....Error! Bookmark not defined.

References Used in Appendix FError! Bookmark not defined.

List of Tables

Table 1. Sequence of IMP modules completed by each ability group at Suburban High School	78
Table 2. Implementation of Semestered Block Schedule and IMP Curriculum at Suburban High School	82
Table 3. Yearly Median National Percentile Rank on Educational Records Bureau CTP Quantitative Ability Test.....	93
Table 4. Difference Between Groups on Understanding and Solving Algebra Problems in Context.....	112
Table 5. Difference Between Groups on Understanding and Solving Algebra Problems in Context After Controlling for Grade 6 Test Scores	114
Table 6. Difference Between Groups on Algebraic Symbol Manipulation.....	116
Table 7. Difference Between Groups on Algebraic Symbol Manipulation After Controlling for Grade 6 Test Scores	118
Table 8. Difference Between Groups on Test 3, an Applied Algebra Task	121
Table 9. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability and Interaction (265 pairs for whom some ability estimate is available)...	125
Table 10. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability (265 pairs for whom some ability estimate is available)	126
Table 11. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability and Interaction (157 pairs for whom most accurate ability estimate is available).....	130
Table 12. Scores on Evaluating Expressions	Error! Bookmark not defined.

Table 13. Scores on Simplifying Expressions with Integer Coefficients **Error!**

Bookmark not defined.

Table 14. Scores on Simplifying Fractional Expressions .**Error! Bookmark not defined.**

Table 15. Scores on Solving Linear Equations, without Context or Access to Calculators
.....**Error! Bookmark not defined.**

Table 16. Scores on Solving Linear Equations in Context **Error! Bookmark not defined.**

Table 17. Scores on Solving Quadratic Equations Presented without Context or Use of
Graphing Calculators**Error! Bookmark not defined.**

Table 18. Percent of Students Showing a Good Understanding of Problems Involving
Quadratic Equations Presented in Context, with Access to Graphing Calculators
.....**Error! Bookmark not defined.**

Table 19. Achievement in Solving Quadratic Equations, Reported by Highest Level
Algebra Course Student Enrolled in Through 11th Grade..... **Error! Bookmark not defined.**

Table 20. Scores on Translating the Graph of a Line into an Equation.**Error! Bookmark not defined.**

Table 21. Opportunity to Learn: Scores on Translating the Graph of a Line into an
Equation for Reform Cohort Students who Completed Relevant Modules..... **Error!
Bookmark not defined.**

Table 22. Scores on Interpreting and Translating Line Graphs **Error! Bookmark not defined.**

- Table 23. Scores on Translating an Exponential Function **Error! Bookmark not defined.**
- Table 24. Ability to Translate an Exponential Function to an Equation **Error! Bookmark not defined.**
- Table 25. Difference Between Groups on Mathematical Formulation.. **Error! Bookmark not defined.**
- Table 26. Difference Between Groups on Interpreting Algebraic Models..... **Error! Bookmark not defined.**
- Table 27. Average Number of Hours Each Student Was Enrolled in Mathematics Class, Traditional vs. Reform Groups **Error! Bookmark not defined.**
- Table 28. Total Hours Enrolled in High School Mathematics for Traditional vs. Reform Groups..... **Error! Bookmark not defined.**
- Table 29. Difference Between Groups on Mean Hours Enrolled in Advanced Mathematics Classes..... **Error! Bookmark not defined.**
- Table 30. Total Hours Enrolled in Advanced Mathematics Courses for Traditional vs. Reform Groups..... **Error! Bookmark not defined.**
- Table 31. Advanced Placement Mathematics Course Enrollment by School Year and Grade Level..... **Error! Bookmark not defined.**
- Table 32. Advanced Placement Exam Grades by School Year and AP Mathematics Course **Error! Bookmark not defined.**
- Table 33. Advanced Placement Calculus BC Exam Grades by School Year..... **Error! Bookmark not defined.**

List of Figures

Figure 1. Sample IMP small group activity.....	41
Figure 2. Percent scoring at each level on Test 3, an applied algebra task.....	123
Figure 3. Low ability pairs: Percent scoring at each level on Test 3, applied algebra task.	Error! Bookmark not defined.
Figure 4. High ability pairs: Percent scoring at each level on Test 3, applied algebra task.....	Error! Bookmark not defined.
Figure 5. Four groups: Percent scoring at each level on Test 3, applied algebra task.	Error! Bookmark not defined.
Figure 6. Evaluating expressions.....	Error! Bookmark not defined.
Figure 7. Simplifying expressions with integer coefficients.....	Error! Bookmark not defined.
Figure 8. Simplifying expressions involving ratios.....	Error! Bookmark not defined.
Figure 9. Linear equations or inequalities without context or access to graphing calculators.....	Error! Bookmark not defined.
Figure 10. Linear equations in context.....	Error! Bookmark not defined.
Figure 11. Quadratic equations without context or access to a graphing calculator.	Error! Bookmark not defined.
Figure 12. Quadratic equations in context.....	Error! Bookmark not defined.
Figure 13. Translating the graph of a line into an equation.....	Error! Bookmark not defined.
Figure 14. Translating graphs (non-procedural).....	Error! Bookmark not defined.
Figure 15. Exponential function in tabular form.....	Error! Bookmark not defined.

- Figure 16. Formulating an algebraic model.....**Error! Bookmark not defined.**
- Figure 17. Interpreting an algebraic model.....**Error! Bookmark not defined.**
- Figure 18. Total Hours Enrolled in Mathematics Class....**Error! Bookmark not defined.**
- Figure 19. Total hours enrolled in advanced mathematics courses.**Error! Bookmark not defined.**

Chapter 1: Rationale

Statement of the Problem

This study investigated the joint effects of two reforms that are beginning to have widespread impact on mathematics instruction in high schools throughout the United States and Canada. The first reform is adoption of a problem-centered mathematics curriculum initially designed to conform with the National Council of Teachers Mathematics' (NCTM's) *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). The second reform is adoption of a semestered block schedule.

Standards-based Curricula

In 1989, NCTM promulgated the *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). The *Standards* called for mathematics classes to include less traditional teacher exposition and student practice, while increasing the amount of project work, group and individual assignments, and discussions between teacher and students and among students (NCTM, 1989, p. 10). The *Standards* reemphasized and strongly endorsed earlier statements by the NCTM that “problem solving must be the focus of school mathematics” (NCTM, 1980, p. 2; NCTM, 1989, p. 6). Later *Standards* documents (NCTM, 1991, 1993, 2000) provided a vision that was consistent with a problem-centered approach to mathematics instruction.

During the 1990s, researchers developed several new high school curricula specifically designed to conform with the vision described in the *Standards*. There has been considerable controversy about the worth of these new curricula (“An Open Letter to United States Secretary of Education, Richard Riley”, 1999). Research to date has

generally shown that these new curricula tend to have positive effects on student achievement (Koedinger, Anderson, Hadley, & Mark, 1997; Huntley, Rasmussen, Villarubi, Sangtong, & Fey, 2000; Merlino, Wolff, & Tolbert, 2000; Webb, in press). Nonetheless, the evidence is not conclusive, as these studies have generally been small in scale, with the number of schools in each study ranging from one to six. Further, in all cases teachers using the new curricula have been volunteers, and in many of the studies the students have been volunteers as well. More research is needed, to determine whether earlier results can be replicated and to extend study of reform mathematics to situations where results cannot be attributed to a volunteer effect.

Block Scheduling

Over the last decade, block schedules have become increasingly popular in high schools throughout the United States and Canada. For example, in North Carolina, the percent of high schools using a block schedule grew from 1.6% in 1992-3, to 35% in 1994-5, to 64.8% in 1996-7, to 73.6% in 1997-8 (North Carolina Department of Public Instruction, 1999). Under a traditional high school schedule, each student typically enrolls in seven courses at a time. Each course meets daily for about 40-50 minutes over a 180-day academic year. In contrast, under the most common type of block schedule, the 4x4 or “semestered” plan, each student typically completes four courses in each semester. Each course meets daily for about 80-90 minutes over a 90-day semester.

Block scheduling has become popular among administrators for a number of reasons. School atmosphere tends to be calmer, and many schools have reported fewer discipline problems under a block schedule (Carroll, 1994a; Hackman, 1995; Meadows, 1995; Reid, 1995a, 1995b; Sessoms, 1995). With fewer classes per day, there are fewer

transitions, so students can spend more time in class instead of walking through hallways between classes. Moreover, some surveys indicate that both students and parents tend to prefer a semestered block schedule to a traditional schedule (Stevens, 1976).

Principals and teachers have reported that low-achieving students find it easier to focus on four classes at a time, instead of the usual seven (Alam & Seick, 1994; Averett, 1997; Carroll, 1994a; Reid, 1995a, 1995b). Also, under a semestered schedule, students have more opportunity to retake courses that they have failed (Reid, 1995a, 1995b; Stevens, 1976; Williams, 1985). Perhaps for these reasons, case studies indicate that failure rates usually decrease after a school switches to a semestered block schedule (Hottenstein & Malatestsa, 1993; Governor Thomas Johnson High School, 1995; Hackman, 1995; Reid, 1995a, 1995b; Schoenstein, 1995b).

Finally, in response to concerns about student achievement stimulated by publications such as *A Nation at Risk* (National Commission on Excellence in Education, 1983), many states and school districts increased the number of credits required for graduation (Porter, 1998). A block schedule makes it easier for principals to meet these requirements, as more courses per year are offered.

Despite the advantages cited above, there is reason to be concerned that changing to a block schedule can actually decrease student achievement, especially in mathematics. Block schedules have been used widely in parts of Canada since the 1970s, well before they became popular in the United States. A number of large-scale studies have investigated achievement effects of block scheduling in Canada (Raphael, Wahlstrom, & McLean, 1986; Marshall, Taylor, Bateson, & Brigden, 1995; Wild, 1998). They have

consistently reported that mathematics achievement under a semestered block schedule is lower than under a traditional schedule.

In contrast, anecdotal reports (Kramer, 1996), as well as some achievement data from North Carolina (Averett, 1994; North Carolina Department of Public Instruction, 1997) indicate that mathematics achievement may improve if a semestered block schedule is adopted at a school that implements instructional changes consistent with those called for by the NCTM *Standards* (NCTM, 1989). However, this possibility has never been systematically studied.

It is important to investigate how a semestered block schedule affects mathematics achievement at schools that are making substantial efforts to implement the reform called for in the NCTM *Standards* (NCTM 1989, 1991, 1993, 2000). As described below, there are three additional reasons to conduct further investigations of how block scheduling affects mathematics achievement.

First, research to date on block scheduling has often been confounded by a possible volunteer effect. Just as studies of *Standards*-based curricula compared achievement of teachers and students who volunteered to use the new curriculum to that of teachers and students who did not volunteer, so most studies of block scheduling have compared achievement at schools that voluntarily adopted a block schedule to achievement at schools that voluntarily maintained a traditional schedule. Data from North Carolina (North Carolina Department of Public Instruction, 1997) indicates that, *prior to the change*, schools that switched to a block schedule had systematically lower achievement than did schools that did not choose to switch. The volunteer effect may account for some of the discrepancies between results reported by the North Carolina

studies, which controlled for prior school-level achievement, and the Canadian studies, which did not do so.

Another reason to conduct further research about the affects of semestered block schedules on mathematics achievement is the need for better measures than have been used in past studies. Previous studies have used only a single measure of achievement, either a standardized test (Raphael, Wahlstrom, & McLean, 1986; Marshall, Taylor, Bateson, & Brigden, 1995) or a standard end-of-course exam (North Carolina Department of Public Instruction, 1997; Wild, 1998). Research comparing student learning under traditional mathematics instruction to student learning under *Standards*-based mathematics instruction has found that the results can vary, depending on what type of test is used to measure achievement (Wood & Sellers, 1996; Huntley, et al., 2000). Tests emphasizing procedures and symbol manipulation tend to favor traditional instruction, whereas tests emphasizing problem solving in context tend to favor reform instruction. It is important to conduct research that not only investigates the joint impact of a block schedule and a curriculum designed to implement the NCTM *Standards*, but also uses a variety of measures designed to test the strengths both of the traditional schedule/curriculum and of the block schedule/*Standards*-based curriculum.

Finally, it is important to investigate how a block schedule affects mathematics achievement at a site where course syllabi have been adjusted to fit the schedule, by spreading mathematics content over a larger number of courses. In the United States, students generally take a larger number of courses per year under a block schedule, with fewer total minutes allocated to each course. For example, a school might change from a traditional schedule offering each student six or seven courses per year to a semestered

block schedule offering eight courses per year (four each semester). Even though the total instructional time available is increased because students are spending fewer minutes walking through hallways between classes, each individual course has fewer allocated minutes of instruction than would an analogous course under a traditional schedule. Thus, one would expect it to be difficult for teachers to cover as much content per course under a semestered block schedule as they would under a traditional schedule. In fact, teachers have reported difficulty with content coverage under a block schedule (Usiskin, 1995; Sturgis, 1995). It should be noted that both Usiskin and Sturgis reported teacher experiences under an alternating-day block schedule, an arrangement in which students take eight classes at a time, each class meeting every other day. However, an alternating-day block schedule affects allocated classroom time in precisely the same way as does a semestered block schedule, so problems in content coverage would likely be the same for both types of block schedule.

Thus, under a block schedule less content can be covered per course than under a traditional schedule. However, more courses are available per year under a block schedule than under a traditional schedule. One solution may be to change syllabi when moving to a block schedule, so that the same content is covered over a larger number of courses (Harter, 1994; Kramer, 1996). Unfortunately, in most cases research on achievement effects of block scheduling has been conducted at sites where this approach has not been implemented. The one exception was a study conducted by Zhang (2000) for the North Carolina Department of Public Instruction. His study compared 214 schools that had adopted a 4x4 block schedule between 1995 and 1997 to 68 schools that had maintained a traditional schedule. He reported that, after controlling for race, socio-

economic status, and 1993-94 school level scores, students in the block scheduled schools outscored those in traditional schools on a state-mandated End-of-Course test in Algebra 1. Zhang attributed the improvement in part to the fact that some students in the 4x4 schools completed the End-of-Course test after taking “regular” algebra in a block semester, while others completed the test after taking Algebra IA and Algebra IB over two semesters.

Zhang’s study investigated algebra achievement at the time students had completed a particular curriculum, delivered either in the “regular” algebra format, or the Algebra IA/Algebra IB format. It did not control for the possible costs to other mathematics learning of spreading Algebra I over two courses. For example, did fewer students in the schools using 4x4 schedules complete Algebra 2, so that by the end of high school they knew less algebra than did students in the traditionally scheduled schools? The current study addresses this possibility in the case of one particular school, by testing all students at the end of eleventh grade. Thus, instead of asking how much algebra students learned in one particular course, it asks how much algebra students learned as a result of their high school program through Grade 11, taken as a whole.

Possible Synergies Between a Standards-based Curriculum and a Semestered Block Schedule

There is reason to suspect that a *Standards*-based curriculum implemented together with a semestered block schedule might have a more positive effect on student mathematics achievement than would either reform by itself. It has been conjectured that lecturing is less effective under a block schedule than under a standard schedule (Kramer, 1996). In contrast, the kinds of group work, in-depth investigations, and problem solving

emphasized by *Standards*-based curricula may actually be easier to accomplish in a longer time block than in the shorter periods offered in a traditional schedule (Averett, 1994; Meadows, 1995, Sturgis, 1995).

Thus, adopting a *Standards*-based curriculum may help compensate for a weakness of mathematics instruction under a block schedule (the relatively poorer results of lecture), while allowing the advantages of the schedule (more total instructional time, better school atmosphere) to positively impact student learning. The semestered block schedule may facilitate successful implementation of *Standards*-based practices such as increased group work, in-depth investigations, and problem-centered instruction.

Research Questions

This study compares the mathematics achievement of students using the reform-based *Interactive Mathematics Program* (IMP) curriculum within a semestered block schedule to the mathematics achievement of students learning a traditional curriculum within a traditional schedule. To narrow the scope of the problem, the primary research questions focus on students' knowledge of algebra. As noted by Huntley, et al. (2000), algebra has been at the heart of high school mathematics for many years, and high student achievement in algebra is generally seen as the hallmark of preparedness for advanced mathematical and scientific studies.

Student achievement data was collected at the end of Grade 11. It was assumed at the time of data collection that most students would have completed their algebra work by the end of eleventh grade.

The high school where this study was conducted used between class ability grouping. Under the block schedule/IMP curriculum the nature of this ability grouping

changed: under the new program, lower-level and higher-level classes used the same text, whereas previously they had used different texts. Also, the new curriculum affected the decision rules for assigning students to academic “levels” for mathematics instruction. Thus, this study’s analyses examine both the main effects of schedule/curriculum on student achievement, and the interaction between schedule/curriculum and prior student ability.

Primary Research Questions

- i. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to solve algebraic symbol manipulation problems? Do the results of this comparison differ depending on students’ prior ability?
- ii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to interpret and solve challenging algebra problems presented in context? Do the results of this comparison differ depending on students’ prior ability?
- iii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to collaboratively solve and communicate their solution to a complex open-ended algebra problem? Do the results of this comparison depend on students’ prior ability?

Secondary Research Questions

A semestered block schedule affords students the opportunity to take more mathematics classes during high school than they would under a traditional schedule. The curriculum being investigated by this study attempts to take advantage of this situation by covering fewer content objectives in each mathematics course, while expecting students to complete more mathematics courses during high school. However, the fact that students have the opportunity to complete additional mathematics courses does not necessarily mean that they will avail themselves of this opportunity. Actual student course-taking patterns are important to study both in their own right and because they will inform results discovered by the algebra achievement testing. The next two research questions address the issue of student course-taking.

- iv. How did students enrolled in a reform-based curriculum and a semestered block schedule differ from students enrolled in a traditional curriculum and traditional schedule in the number of registered mathematics class hours by the end of Grade 12?
- v. How did students enrolled in a reform-based curriculum and a semestered block schedule differ from students enrolled in a traditional curriculum and traditional schedule in participation in advanced courses, as measured by the number of registered hours in advanced mathematics classes by the end of Grade 12, by the number of students enrolling in Advanced Placement courses, and by scores on Advanced Placement tests?

Previous authors have pointed out that, while a semestered block schedule might provide an opportunity for students to spend more time studying mathematics,

administrative policies can either enhance or impede this opportunity (Harter, 1994; Kramer, 1996). Such policies are essentially part of the “treatment,” just as much as the mathematics curriculum and the format of the schedule itself. In addition to administrative policies, other aspects of the school and community could influence the way a semestered block schedule and the IMP curriculum affect student mathematics achievement and student enrollment in mathematics courses. If other schools are to learn from the experience at Suburban High School¹, special aspects of the school and community need to be described. The sixth research question addresses these issues.

- vi. According to administrators and faculty who assumed critical responsibilities for implementing the shift to a semestered block schedule and reform-based mathematics curriculum, what school administrative policies and what unique aspects of the school and community affected mathematics course enrollment and mathematics achievement under the new schedule and curriculum?

Overview of Method

Research Opportunity at Suburban High School

This research project is part of an ongoing research effort that was begun during the 1996-97 school year by teachers at Suburban High School. Suburban is the only high school in a small school district located in a middle class suburb of a large U.S. city. During the 1996-97 school year, Suburban high school initiated a semestered block schedule for all ninth graders. In that same year, ninth graders were enrolled in IMP, as a phase-in of this new reform-based curriculum. Note that during this first year at

¹ “Suburban High School” is a pseudonym

Suburban High School, only ninth graders used IMP and the semestered block schedule. These students and their successors continued in this schedule/curriculum.

Testing the Traditional cohort. In order to assess their understanding of algebra, in the spring of 1997 all eleventh graders at Suburban High School completed an algebra test designed by the Core-Plus Mathematics Project. Core-Plus, like IMP, is a high-school curriculum designed to implement reform-based mathematics. Their algebra test was specifically designed to compare the effects of a *Standards*-based curriculum to

those of more conventional curricula. The test is organized into three parts. Part 1 emphasizes the ability to understand and solve algebra problems presented in context, as is typically emphasized by Core-Plus, IMP, and other reform curricula. Part 2 emphasizes 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 requires collaborative work on a single extensive open-ended applied problem and is completed by students working in pairs. Items from Part 1, Part 2, and Part 3 of the Core-Plus assessment define the Algebra Achievement test for this study.

The eleventh graders tested in the spring of 1997 had used a traditional schedule and traditional curriculum throughout Grades nine through eleven. They form a “Traditional” cohort to be contrasted with later cohorts of students at Suburban High School who were taught using a semestered block schedule and the IMP curriculum.

Testing the Pilot cohort. As the ninth graders who had piloted the semestered block schedule and the IMP curriculum during the 1996-97 school year advanced through high school, they continued using the new schedule and curriculum. In the spring of 1999, when they were eleventh graders, this “Pilot” cohort completed the same Algebra Achievement test that had earlier been completed by eleventh graders in the 1997 Traditional cohort. A pilot study used this test data to compare how well students in the two cohorts (Pilot versus Traditional) understood algebra.

Unlike later cohorts of students at Suburban High School who used IMP exclusively, the Pilot students reflected a mixture of curricula: Honors students, who had begun studying traditional algebra while in eighth grade, continued using the traditional curriculum, while all other regular education students used the IMP curriculum.

Testing the first reform cohort. In the spring of 2000, eleventh graders at Suburban High School again completed the Algebra Achievement test. Students in this “First Reform” cohort had used a semestered block schedule and the IMP curriculum throughout high school. Since teachers already had experience using the IMP curriculum and block schedule with the pilot cohort, it is likely that by the time students in the First

Reform cohort were exposed to the new program their teachers had gotten past the “implementation dip” often experienced by major school reforms (Busick & Inos, 1992; Fullan & Miles, 1992). Also, lessons learned from the pilot study made it possible to ensure testing conditions for the experimental cohort were similar to those used for the Traditional cohort. For these reasons, it is reasonable to assume that a comparison of test scores of students in the First Reform cohort to those of students in the Traditional cohort will provide a fair indication of how a semestered block schedule, implemented jointly with the IMP curriculum, affected algebra achievement at Suburban High School.

Testing the second reform cohort. In the spring of 2001, eleventh graders at

Suburban High School once again completed the Algebra Achievement test. Students in this “Second Reform” cohort again used a semestered block schedule and the IMP curriculum throughout high school. This study combines scores from the First and Second Reform cohorts to estimate achievement of students who used a semestered block schedule and the IMP curriculum, and to compare their achievement to that of students in the Traditional cohort.

Data Analyzed

In order to answer the research questions, this study analyzed six types of data:

1. Scores from the Algebra Achievement test;
2. Mathematics achievement tests administered prior to high school, used as a covariate to compare relative prior ability of students using the Traditional and Reform programs.
3. Student transcripts;
4. Course syllabi, textbooks, and other documents used to teach mathematics at the school;
5. School system documents, including annual testing reports and annual school profiles;
6. Qualitative data from key informants at the Suburban High School.

Analysis of Algebra Achievement tests. This study compared tests taken by the

Traditional cohort of eleventh graders at Suburban High School who had completed a traditional mathematics curriculum within a traditional seven-period per day schedule, to those taken by two Reform cohorts of students who attended the same school three and four years later, completing a reform-based mathematics curriculum within a semestered block schedule.

Transcript analysis. This study analyzed student transcripts to see whether mathematics course taking changed under the new schedule and curriculum. It compared Traditional to Reform students on three measures: total number of hours registered in mathematics courses, number of hours registered in advanced mathematics, and number of students enrolled in Advanced Placement courses.

Document analysis. This study reviewed annual “school profiles” published by the Suburban High School to determine the number of students each year taking Advanced Placement exams administered by the College Board, as well as to determine student grades on the Advanced Placement exams they took.

This study also analyzed course syllabi published by the school. Combined with the transcript analysis, these documents provided insight regarding students’ opportunity to learn the key topics evaluated in the Algebra Achievement test.

Key informants. If other educators are to learn from Suburban High School’s experience, it is important to understand not only the nature of the new schedule, curriculum, teaching methods, and syllabi, but also how course taking changed and why. Therefore, this study examined the administrative or counseling policies that influenced course-taking decisions. Interviews with key informants, combined with the transcript analysis, provided data to answer these questions.

Definition of Key Terms

Advanced Mathematics Classes: The term “Advanced Mathematics Classes” refers to classes for which material traditionally covered in Algebra 1, Geometry, and Algebra 2 is a prerequisite.

Alternating-day or A/B block schedule: A semestered block schedule contrasts with an A/B block schedule, in which students take 8 classes at a time, each class running 80-90 minutes every other day.

Constructivism: A constructivist theory of knowledge is characterized by two basic principles: (a) learners actively construct knowledge through interaction with their surroundings and experiences, and (b) learners interpret these occurrences based on

existing knowledge and their rendering of the experienced observations and actions (Noddings, 1990).

Implementation Dip: A period commonly seen in successful change initiatives, where individuals have given up old practices, but not yet mastered new and potentially more effective practices that they have adopted. During this period, student performance may go down, only to be followed by later improvements (Fullan & Miles, 1992; Busick & Inos, 1992).

IMP: The Integrated Mathematics Program (IMP) curriculum authored by Fendel, Resek, Alper, and Fraser (1997) is published by Key Curriculum Press. IMP was one of five reform-based high school mathematics curricula whose development was funded by the National Science Foundation. The IMP curriculum is built around complex, open-ended problems. It emphasizes in-depth understanding of mathematical concepts and techniques. IMP promotes students' active role in the classroom, working together in teams, talking with each other about mathematics, and making oral and written presentations about challenging problems.

NCTM Standards: Originally, the National Council of Teachers of Mathematics published three *Standards* documents. Only the first of these, the Curriculum and Evaluation Standards for School Mathematics (NCTM, 1989) was available at the time Fendel, Resek, Alper, and Fraser began developing IMP. Unless otherwise specified, this is the document referred to by the term *NCTM Standards*. (Note: the other *Standards* documents (NCTM, 1991, 1993) are compatible with the Curriculum and Evaluation Standards and with the general vision of IMP. So are the Principles and Standards for

School Mathematics (NCTM, 2000), recently published to update the original *Standards* documents.)

Quasi-Experiment: A research design in which treatment and comparison groups are formed by some means other than random assignment (Krathwohl, 1993).

Retention Interval: The gap in time between when a student studied mathematics content, and when the student was tested on knowledge of that content.

Semestered or 4x4 Block Schedule: Under a “semestered block schedule,” also called a “4x4 block schedule,” high school students take four classes at a time, and each class lasts one semester. In general, such classes run about 80-90 minutes a day.

Situated Cognition: A “situated cognition” theory of knowledge is characterized by the principle that the activity in which knowledge is developed and deployed is not separable from learning and cognition. Rather, it is an integral part of what is learned (Brown, Collins, & Duguid, 1989).

Chapter 2: Review of the Literature

The idea for conducting an investigation into the joint effects on mathematics achievement of a semestered block scheduling and an NCTM *Standards*-based mathematics curriculum began with a review of the literature on block scheduling that was published in the *Mathematics Teacher* in December, 1996 (Kramer, 1996). Based on both the literature reviewed and on interviews with teachers, that article noted that lecturing appears to work less well under a block schedule than under a traditional schedule. As a consequence, researchers generally recommended including a larger percentage of participatory activities in each block-scheduled class period. The article concluded with the following observation:

If block scheduling were implemented with adequate planning time and staff development, and with administrative policies that maintained the number of classroom hours allocated to mathematics over a student's high school career, it is quite possible that achievement would be higher than it had been under a traditional schedule. To date, unfortunately, such an implementation has not been studied.

In response to the article in *Mathematics Teacher*, Mrs. Sullivan², the mathematics supervisor at Suburban High School contacted the current author. She indicated that her school was in process of adopting a semestered block schedule and was implementing all of the recommendations contained in the *Mathematics Teacher* article. Further, they were simultaneously adopting IMP, a problem-centered curriculum designed to implement the NCTM *Curriculum and Evaluation Standards* (NCTM, 1989). She believed that IMP was particularly well suited to a block schedule and was likely to further enhance student achievement. She invited the author to help her investigate the

² "Mrs. Sullivan" is a pseudonym

achievement effects of the new schedule and curriculum at Suburban High School. Thus the current study was born.

This Chapter summarizes and updates the original block scheduling literature review published in 1996. It also reviews the literature on IMP and on the achievement effects of other high school curricula designed to implement the NCTM *Standards*.

Semestered (4x4) Block Scheduling

Block scheduling is not a new phenomenon. It has been widely used in British Columbia, Ontario, and Alberta since the 1970s. In the United States block schedules became increasingly popular throughout the 1990s, and today have spread to high schools in all regions of the nation.

This study focused on one of the two most common types of block schedule, namely, the semestered or 4x4 block schedule. A semestered block schedule typically consists of four courses meeting 80-90 minutes daily for 90 days. The other common type of block schedule is the alternating-day, or A/B block schedule. An alternating day block schedule typically consists of 8 courses meeting 80-90 minutes every other day for the entire 180-day school year. Besides these two most common types of block schedules, other forms of block scheduling have been tried in various places. The most important are “interdisciplinary” block schedules (Sigurdson, 1982) and “quarter” plans (a more intensive extension of semestered schedules, in which students usually take two academic courses at a time, with each course meeting about 150-180 minutes daily for one quarter of the school year.) Canady and Rettig (1995) provide an extensive discussion of various forms of block scheduling.

Administrators are often attracted to block schedules for a number of reasons. There is evidence that student discipline improves under all major forms of block scheduling (see, e.g., Carroll, 1994; Hackman, 1995; Hillcrest High School, 1995; Meadows, 1995; Reid, 1995a; and Sessoms, 1995), as do student attitudes towards school (Averett, 1994, Kramer 1997a, 1997b, Meadows, 1995, Stevens, 1976). Semestered and other intensive forms of block scheduling (e.g., the “quarter” plan) appear to lead to reduced dropout rates (Kramer, 1997a; Sharman, 1990). Finally, administrators often gain flexibility in scheduling by having each student take eight courses per year under a block schedule, versus six or seven courses per year under a traditional schedule.

Effects of Semestered Block Scheduling on Mathematics Instruction

Reduced effectiveness of lecturing. In general, the literature reviewed indicated that teaching by lecture alone works less well in a longer time block (Howard High School, 1994; King, Clements, Enns, Lockerbie, & Warren, 1975; King, Warren, Moore, Bryans, & Pirie, 1978; Meadows, 1995; O'Neil, 1995; Reid, 1995a; Sturgis, 1995). Not surprisingly, students seem to find it difficult to sit through a class that consists essentially of two lectures conducted in sequence. Instead, researchers generally recommended including a larger percentage of participatory activities in each block-scheduled class period.

The literature available, however, has two weaknesses. First, the conclusion has not yet been confirmed by student performance data. All of the literature that recommended placing additional emphasis on non-lecture teaching modes based its conclusions on surveys of administrators, teachers, and/or students.

Second, the literature indicates that lecturing is less effective under a block schedule for all subjects in general; it does not address mathematics in particular. In fact, there is some evidence that mathematics teachers may be less likely to change their teaching methods under a block schedule than are teachers in other departments. Reid (1995a) interviewed five principals of schools in British Columbia that had switched to an intensive (four quarters, two courses per quarter) block schedule. He found that mathematics teachers in these schools had a harder time changing their teaching methods than did teachers from other departments.

The need to adopt new teaching modes is reflected in a comment that was made repeatedly by mathematics teachers interviewed by Kramer (1996): Experienced teachers said they "felt like first year teachers" after switching to a block schedule. Apparently, pedagogical methods that teachers had learned from experience in traditional classrooms did not translate successfully into block scheduled classrooms.

Thus, it is reasonable to draw the tentative conclusion, based on the opinions of teachers, administrators, and students, that teachers in mathematics need to reduce their amount of lecturing in order to maintain student interest under a block schedule. Furthermore, mathematics teachers are likely to find such a change more difficult than are teachers in other subject areas, and thus may have a particularly strong need for staff development and extra planning time to assist them in making the change (Brophy, 1978;

King, et al., 1978; Watts & Castle, 1992; Averett, 1994; Meadows, 1995; Reid, 1995a; Canady & Rettig, 1995; Salvaterra & Adams, 1995; Kramer, 1997a).

Decreased breadth and increased depth of coverage. A literature review by

King et al. (1975) reported that both mathematics and French teachers experienced particular difficulty covering the equivalent of two classes of material during a double length period. A follow-up study made detailed observations of classrooms in six schools with semestered block schedules. Comparing mathematics classrooms in these schools to ones they had observed operating under a traditional schedule, they noted that under a block schedule mathematics teachers frequently used up more instructional time to cover the same content (King, Warren, Bryans, & Pirie, 1978).

More recently, Usiskin (1995) reported the opinions of several teachers who had taught under both traditional and alternating-day block schedules using textbooks developed by the University of Chicago School Mathematics Project. All agreed that under the alternating-day block schedule, not as much content was covered. However, it is possible that Usiskin's results can be explained by the fact that there are fewer allocated hours per course. Sturgis (1995) reported that some teachers at a school in Maine were covering less content after switching to an alternating-day block schedule, but that this was balanced by an opportunity to go into more depth. While the data collected by both Usiskin and Sturgis related to alternating day block schedules, it is likely that their observations apply to semestered block schedules, which share the characteristics of more hours per class period combined with fewer allocated hours per course typical of an alternating day schedule.

A number of surveys support the observation that teachers perceive the longer time blocks as providing an opportunity to teach concepts in greater depth. This is true for both semestered and alternating-day block schedules. A survey of teachers at 21 North Carolina schools that were in their first or second year of implementing a semestered block schedule reported that over 70% of those surveyed felt that the block schedule had a "moderate positive effect" or "strong positive effect" on each of the following: 1) problem solving ability, 2) higher order thinking, 3) in-depth knowledge, and 4) retention of subject matter (Averett, 1994). A survey of teachers at four Maryland high schools in their first or second year of implementing a semestered block schedule obtained similar results (Meadows, 1995). Other studies have reported increased depth of

coverage by teachers at schools using an alternating-day block schedule (Sessoms, 1995; Sturgis, 1995).

Need to adjust the mathematics curriculum. A number of researchers have

noted that faculties need to adjust the curriculum under a block schedule (King, et al., 1975; Harter, 1994; Sessoms, 1995). Administrators often move to a block schedule in order to enable students to take a larger number of courses each year, while allocating fewer hours per course. For example, a school that had previously offered students seven 45-minute periods daily might move to a schedule that offered students four 80-minute periods daily. Under a semestered block schedule, each course would then meet for half the year. Students would be replacing two 45-minute periods with one 80-minute period, but would be taking one extra course yearly.

Under this type of schedule, students often enroll in a larger number of core courses (Cameron, 1995), and in particular in a larger number of mathematics classes (Edwards, 1995). One study reported that students enrolling in additional courses were often of two types: students who had failed a class and were retaking it, and top students who were taking two mathematics courses a year (Kramer, 1996).

Harter (1994) noted this kind of schedule allows schools to offer students more time to take mathematics, not less. He emphasized that students unlikely to succeed within existing time constraints could benefit from two-term core mathematics courses, that honors programs could offer two semesters of challenging mathematics yearly to interested students, and that all students could benefit from courses in statistics and data analysis.

However, administrative constraints can make it impossible for schools to take advantage of this opportunity. Harter (1994) reported that principals in North Carolina's block scheduled schools were often allotted faculty slots within distinct classifications (i.e., so many mathematics teachers, so many science teachers, so many art teachers, etc.) and that these allotments could result in an imbalance between student needs and faculty available to teach particular courses. According to Harter, this resulted in many schools on the four-period day (i.e., with semestered block schedules) encouraging students to take elective courses outside the core disciplines, while discouraging or even denying some students access to two courses per year in the same academic discipline. He concluded that the potential advantages of a block schedule could be realized fully only if high school principals were given more flexibility in assigning teaching positions. Kramer (1996) interviewed eight teachers from schools having success with block scheduling. Seven reported adjusting the mathematics curriculum to accommodate the new schedule. Curriculum changes mentioned by teachers in the sample included the following:

- creating a 2-part algebra class for “lower” mathematics students;
- replacing the normal Algebra I/Algebra II sequence with a sequence of three (shorter) Algebra courses;
- modifying Geometry and Algebra I courses to eliminate topics taught in Algebra II;
- splitting a combined Algebra II/Trigonometry class into two separate classes;
- adding new courses (e.g., in statistics) for students who finish up the regular sequence.

A good example of the potential importance of reworking curriculum while moving to a block schedule is seen in the following description, sent in response to a letter to the editor of Mathematics Teacher:

Our school is in its third year of using the block schedule where students complete one course in 18 weeks. The only problem we had was with Algebra. All of the other courses were easily adapted to the 85 minute class, but algebra in 18 weeks is just too fast. We met with such failure the first year that our administration readily went along with changing algebra to a full year class (36 weeks). Algebra is the foundation of all of our other classes and the students need to have a solid foundation before we can expect them to succeed in following courses. The full year classes in algebra allow plenty of time to do the exploration and hands on activities that help the students get a better understanding. In the 18 week algebra, it was practically impossible to do any activities--we felt like we were flying thru the material and losing lots of students in the process. There's not a lot of algebra that can be dropped (chopped, condensed, combined, etc.) without adversely affecting future courses.

Since changing to 36 weeks of 85 minutes of algebra each day, we have had a much better success rate and the geometry and algebra II teachers have noticed a difference. (Kramer, 1996, pp 760-761).

Advanced placement classes. Advanced placement (AP) classes present special challenges to schools switching to block schedules. One teacher summarized the situation as follows:

AP exams are given in May only. Students who take the AP course first semester are rusty, students who take the AP course second semester won't cover enough by May (Kramer, 1996, p. 761).

In some cases, schools using block schedules offer Calculus and other AP courses as double-length courses that run the entire year (Edwards, 1995; Governor Thomas Johnson High School, 1995) or three quarters of the year (Schoenstein, 1995), with the last quarter perhaps offering a special topic, like a probability and statistics class. One article described a block scheduled school that switched its AP courses back to standard 45-minute classes. For example, AP English and AP social studies ran in back-to-back 45-minute classes for the entire school year. The authors noted, however, that several teachers thought this was a step backwards (Salvaterra & Adams, 1995).

The College Board (1998) reported that in 1998 four instructional schedules were widely used for AP courses:

1. traditional schedule of 30- to 60-minute sessions each day during the school year
2. schedule of 61- to 90-minute sessions every day during the school year
3. semesterized fall block course
4. semesterized spring block course

Schools using semestered block schedules have the option of offering Calculus and other AP courses in any one of the latter three formats: as double-length courses running all year, in the fall, or in the spring. The College Board compared achievement on the 1997 AP Calculus AB exam for each of these four types of schedule, using a Bonferroni-adjusted significance level of .08 for each of the six possible pair-wise comparisons. They found that, after adjusting for prior ability as measured by scores on the mathematics portion of the PSAT/NSMQT administered in 1995 or 1996, student AP scores could be predicted by how much time they had spent studying the topic. Students in double-length courses running all year scored significantly higher than students in year-long 30- to 60-minute courses, who in turn scored significantly higher than students enrolled in AP Calculus AB for just the spring semester or just the fall semester. Scores of students enrolled in spring-semester or fall-semester courses were not significantly different from each other.

Impact of student absences. A student who misses a day under a block schedule misses nearly twice as much lesson time. Thus, teachers have indicated that absences are more disruptive to student learning under both semestered and alternating-day block schedules than they are under traditional schedules. A majority of North Carolina teachers responding to Averett's (1994) survey indicated that, under a semestered block schedule, their students had difficulty in recovering from absences. This, along with difficulty in accommodating transfer students, was one of the two major weak points they noted.

Usiskin (1995) reported similar opinions among teachers using materials from the University of Chicago School Mathematics Project in an alternating-day block schedule. Further, Sturgis (1995) reported that an alternating-day block schedule made it more difficult for teachers to ensure students made up missed homework after an absence.

Retention of learning after a gap in sequential instruction. Under a semestered block schedule, students often enroll in mathematics during only one of the two yearly semesters. Parents and teachers have often expressed the concern that mathematics achievement might be harmed because of students' inability to retain information when the gap between one mathematics course and the next one could be more than one year (Lindsay, 2002).

Joseph Carroll, one of the earliest and best-known proponents of block scheduling, provided data relevant to this question. He reported the results of nine tests that compared achievement of students with a gap between learning and test taking of three months to that of students with a gap of six months; as well as nine tests that compared achievement of students with a gap between learning and test taking of three months to that of students with a gap of nine months (Carroll, 1994, p. 11). Although Carroll claimed that levels of retention were comparable between students with longer and shorter gaps, his data appear to show otherwise. Two of the nine tests that compared a three-month gap to a six-month gap showed statistically significant higher scores after the shorter gap; four of the nine tests that compared a three-month gap to a nine-month gap showed significantly higher scores after the shorter gap. Differences on the remaining thirteen tests were not statistically significant. Carroll also reported six tests that compared achievement of students with a gap between learning and test taking of six months to that of students with a gap of nine months or twelve months, and found no significant differences on any of the six tests. It should be noted that Carroll did not provide details on these tests, so it is unclear what academic subjects displayed the effects, or even what *p*-value was used to judge “significance.”

The potential impact of a gap in sequential instruction was addressed in a 3-year longitudinal study involving 214 students in London, Ontario who completed their ninth-grade mathematics course in 1972. Of these students, 107 studied tenth-grade mathematics in all-year schools, 63 studied tenth-grade mathematics in the first semester (fall of 1972), and 44 studied tenth-grade mathematics in the second semester (spring of 1973). At the end of their ninth-grade year (i.e., June, 1972), all students were given a 28-item test, consisting of a 10-item Numerical Skills subtest and an 18-item Algebraic Skills subtest. The three groups scored nearly identically on both subtests.

Each group was administered the same test at the beginning of their tenth-grade mathematics course. Thus, the 44 second semester students had a longer gap (summer-plus-fall) before beginning instruction than did students in the other two groups (summer only). Although there were no differences among the three groups on the Basic Skills subtest, the second semester group (i.e., the group with the longer time gap) scored lower than the other two groups on the Algebraic Skills subtest.

The test was administered again at the end of tenth-grade instruction: January, 1973 for the first-semester group, and June of 1973 for the all-year group and the second-semester group. By the end of tenth-grade instruction, the second semester group had caught up with the other two, so that there were again no differences in test scores on either subtest. Finally, all three groups were administered the test in fall of 1973 (i.e., the beginning of eleventh grade), and all three maintained their scores, with the groups receiving nearly identical results on both subtests.

Thus, when tested, students with an extra semester time gap did have more difficulty recalling recently learned concepts, but they recovered quickly during the subsequent mathematics course. Over the longer term, there were no negative effects caused by the gap (Smythe, Stennett, & Rachar, 1974; Stennett & Rachar, 1973). Shockey (1997) conducted a similar study at two suburban United States high schools using semestered block schedules and produced similar results. Participants in her study were 172 students at the two schools who were enrolled in pre-calculus in the spring of 1997. There were three groups of students at each school. One group ($n=53$) had a

“retention interval” of 0 months. These students had completed Algebra 2 during the first semester of the 1996-97 school year, beginning the pre-calculus course within a week of completing Algebra 2. A second group ($n=55$) had a retention interval of 8 months. These students had completed Algebra 2 during the spring of 1996. The third group of students ($n=64$) had a retention interval of 12 months. These students had completed Algebra 2 during the fall semester of the 1995-96 school year.

Students completed district mandated end-of-course tests in Algebra 2 three times: once at the end of their Algebra 2 course, once at the beginning of the pre-calculus course, and once after approximately four weeks during which the pre-calculus teacher reviewed Algebra 2 concepts. In addition, at the end of the pre-calculus course students completed a district mandated end-of-course test in pre-calculus. The Algebra 2 test contained two parts: a 37-item multiple choice test, and a rubric scored performance-based assessment. For the administration at the end of Algebra 2, only scores for the multiple choice portion of the end-of-course assessment were available for analysis, and there was no statistically significant difference on that part of the test. When tested pre-review and post-review during the pre-calculus course, there was no significant difference among the groups on the performance assessment portion of the test.

On the multiple choice portion of the pre-review test taken at the beginning of pre-calculus the students with a retention gap of 0 months significantly outscored the students with a retention gap of 8 or 12 months. At the end of approximately four weeks of review, there was no significant difference between the multiple choice scores of students with 0 months retention gap and those with 8 months' retention gap. Students with 12 months retention gap had partially caught up, but still scored significantly lower than did students with 0 months retention gap. However, when tested on pre-calculus concepts at the end of the pre-calculus course, there was no significant difference among the groups, nor even a trend towards higher scores for the students with 0 months retention gap.

In general, Shockey's (1997) results confirm those reported earlier in London, Ontario (Smythe, Stennett, & Rachar, 1974; Stennett & Rachar, 1973): A retention gap caused students to score more poorly on a test of the immediately preceding course, but had no negative effect on students' ability to learn material in a subsequent course. However, one aspect of Shockey's (1997) data raises a possible concern. She observed a total of four teachers, two in each of two high schools, teaching pre-calculus in a semestered block schedule. One of the teachers spent 15 days reviewing Algebra 2 concepts, one spent 20 days, one spent 21 days, and one spent 22 days. On average, more than 20% of the 90-day pre-calculus course was spent on review.

Would so much review have been necessary had students been enrolled in a regular schedule, so that the retention gap before beginning pre-calculus was 3 months for everyone, instead of ranging between 0 and 12 months? Shockey's data leave this question unanswered. Surveys and interviews have provided mixed messages in response to this same question. Students and teachers at six Ontario schools with semestered block schedules indicated on a questionnaire that students encountered difficulty in returning to a subject after a break of a semester (King, et al., 1978). In contrast, other researchers provided anecdotal evidence that teachers could discern very little difference in the amount of review needed by students who had a retention gap of three months and that needed by students with a longer retention gap (Canady and Rettig, 1995; Kramer, 1997b). In short, this question is still unanswered, and further research will be needed to address it.

Effects of Semestered Block Scheduling on Mathematics Achievement

Student grades. Many studies have analyzed the academic impact of block scheduling by comparing student grades under a block schedule with grades under a traditional schedule. Most have reported that grades under a block schedule are higher (Carroll, 1994; King, et al., 1975; Pulaski County High School, 1995). One case study reported that mathematics grades, in particular, had improved (Reid, 1994).

However, Wild (1998) reported an opposite trend among British Columbia schools in 1995-96 and 1996-97, with students in semestered block schedules tending to have lower teacher-assigned grades than either students in schools with traditional all-year schedules or students in schools using "quarter" plans with even longer time blocks. There were several differences between Wild's study and the others, any of which may account for the difference. He compared grades across schools, rather than reporting change of grades within schools; his sample size was larger; and his study was conducted in British Columbia in the mid-1990s.

Whatever the effect of a semestered block schedule on student grades, it is not certain that improved grades reflect increased learning. Three studies have presented data indicating that improved grades under a block schedule may be the result of grade inflation, and thus not a valid measure of academic achievement (Gore, 1995; King, et al., 1975; Wild, 1998).

One study conducted by Cobb, Abate, and Baker (1999) did manage to look at grade point averages while avoiding the problem of grade inflation under a block schedule. They did so by looking at the grade point averages of 150 tenth and eleventh graders in a traditionally scheduled high school who had previously attended a junior high school using a semestered block schedule. They found that this group's grade point averages in tenth and eleventh grade were significantly higher than those of a matched sample of 150 students who had attended one of two other junior high schools in the same geographic quadrant of the city whose school size, ethnic, and socio-economic make-up were comparable. Unfortunately, their study did not disaggregate course grades by subject area, so it is unclear whether the overall higher grade point averages of students who had attended the block scheduled junior high school applied to mathematics grades in particular.

Mathematics test scores: Canadian studies. With the possible exception of the study by Cobb, Abate, and Baker (1999), mathematics test scores are probably more valid than course grades for measuring the impact of semestered block scheduling on achievement. Because semestered block schedules became popular in Canada before they became popular in the United States, most early research on the schedule's achievement effects was conducted there.

Two relatively small scale investigations conducted in London, Ontario found no significant difference between the achievement of students studying under a semestered block schedule and that of students studying under a traditional schedule. As noted above, an early longitudinal study found that a semestered block schedule had no impact on students' mathematics achievement at the end of tenth grade (Smythe, Stennett, & Rachar, 1974; Stennett & Rachar, 1973). A second study compared test scores of 350 students in London, Ontario studying Grade 9 general level mathematics during the second semester under a block schedule to those of 309 students studying the same course under an all-year schedule. Because the test was administered approximately one month before the end of the school year, semestered classrooms had completed an average of only 74.4% of the curriculum, whereas all-year classrooms had completed an average of 82.9% of the curriculum. Nonetheless, scores between the two groups were nearly identical (Stennett, 1985).

Raphael, et al. (1985) performed a much larger scale study of the province of Ontario. They investigated performance on the Second International Mathematics Study

(SIMS) of Ontario students who were in Grade 12 or who were mathematics specialists in Ontario's (college preparatory) Grade 13. "Math specialists" were defined as students taking more than one mathematics course in Grade 13. The SIMS data were collected using a stratified random sample designed to be representative of the Ontario province as a whole. In all, 5280 students from 250 classrooms, 84 of which used a semestered block schedule, participated in the study. Student socioeconomic status was estimated based on student responses to SIMS questionnaire items asking each student's mother's and father's profession. According to this measure, students from classrooms using a semestered block schedule did not differ from students in all-year classrooms in socioeconomic status.

Achievement of students in Grade 12 classrooms using a semestered block schedule was compared to that of students in classrooms using a traditional year-long schedule on eight subscales: Number Systems, Algebra Computation, Algebra Other, Equations and Inequalities, Analytical Geometry, Trigonometry, Functions, and Probability and Statistics. Students in year-long classes outperformed students in semestered classes on all eight subscales, with differences on Number Systems, Algebra Other, Equations and Inequalities, and Analytical Geometry significant at the .05 level. Also, responses of Grade 12 students to an attitude questionnaire found that those in semestered classes generally had less positive attitudes towards mathematics than did their peers in year-long classes (Raphael, et al., 1985).

Achievement of Grade 13 math specialists in classrooms using a semestered block schedule was compared to that of Grade 13 math specialists in classrooms using a traditional year-long schedule on the eight subscales used for Grade 12 students, plus three additional subscales: Complex Numbers, Differentiation, and Integration. Students in year-long classes outperformed students in semestered classes on all eleven subscales. Differences were statistically significant at the .05 level for all subscales except Complex Numbers and Trigonometry.

Despite the careful sampling design, some Canadian educators contacted to provide context for understanding the Raphael, et al. (1985) study indicated that sampling students by mathematics class may have introduced unintentional bias in the analysis. They provided anecdotal reports that in semestered schools lower-ability students in Grades eleven and twelve were encouraged to "try" math courses in the fall, and if they didn't do well to retake the same course in the spring. Such students could have been sampled twice in the scheme used by SIMS. However, a decade later Marshall, Taylor, Bateson, and Brigden (1995) obtained similar results in a British Columbia study that avoided this particular pitfall by testing all tenth graders, regardless of the particular mathematics class in which they were enrolled.

Marshall, et al. (1995) reported results from British Columbia's 1995 Mathematics and Science Assessment. The assessment used a matrix sampling procedure, with each tenth-grade student taking one of four forms of the test. Each form contained 20 questions addressing knowledge covered in the British Columbia mathematics curriculum through tenth grade. The assessment tested 16,356 students who studied Mathematics 10 under a traditional all-year schedule, 6,461 students who studied Mathematics 10 under a semestered block schedule, and 1,703 students who studied Mathematics 10 under a "quarter plan" block schedule. Students who studied under the all-year schedule outscored those who studied under the semestered schedule, who in turn

outscored those who studied under the quarter plan. Results were statistically significant. The stability of the differences is perhaps best reflected in the fact that of the 80 items presented among the four forms of the test, Mathematics 10 students in all-year programs scored highest on 74 items, second on 5 items, and lowest on only 1 item. Semester students scored highest on 3 items, second on 68 items, and lowest on 9. Quarter students scored highest on 3 items, second on 7 items, and lowest on 70 items. Marshall, et al. reported similar though less extreme results on the same assessment for all-year, semestered, and quarter students enrolled in Mathematics 10A, a non-college-preparatory version of the Grade 10 British Columbia mathematics curriculum.

Wild (1998) provided further evidence to corroborate the findings reported by Marshall, et al. (1995). He reported participation rates and exam grades for all Grade 12 students enrolled in British Columbia schools in 1996-97. Mathematics 12 is the college preparatory mathematics elective for British Columbia Grade 12 students: nearly all students who wish to take advanced mathematics enroll in the course. Wild reported that during the 1996-97 school year, 51.6% of British Columbia students enrolled in all-year classes completed Mathematics 12 and took the provincial exam, compared to only 34.1% of students attending schools with a semestered block schedule and only 29.1% of students attending schools using a quarter plan. Further, the average marks on the provincial exam were higher for students in the all-year schools: 69.4% in all-year schools, versus 64.6% in schools using a semestered block schedule and 62.5% in schools using a quarter plan. The percent of students in each group who received an "A" on the provincial exam formed an even more striking pattern: 24.3% for all-year students, 14.1% for students using a semestered block schedule, and 10.7% for students using a quarter plan. The distinct pattern cannot be attributed to small sample sizes: in 1996-97 exams were completed by 8,407 students in all-year classes, by 8,936 students in semestered classes, and by 1,163 students in quarter-plan classes. Wild also reported results for 1995-96 that were similar to those for 1996-97.

Lastly, Wild (1998) summarized results of a nationwide study that had been conducted by the Council of Ministers of Education, Canada to investigate mathematics achievement of students at ages 13 and 16. Typically 500-1000 students were sampled in each Canadian province. A total of 12,881 13-year-olds participated in the study, 86% of whom were enrolled in all-year classes and 8% of whom were enrolled in semestered classes. Mathematics achievement was scored on a scale from lowest of 1 to highest of 5. Of 13-year-olds enrolled in all-year classes, 30% scored 3 or higher, versus 24% in semestered classes. A total of 11,079 16-year-olds participated in the study, 40% of whom were enrolled in all-year classes and 49% of whom were enrolled in semestered classes. Among 16-year-olds enrolled in all-year classes, 71% scored 3 or higher, versus 55% in semestered classes.

Taken as a whole, the Canadian studies provide a very clear picture of lower mathematics achievement in classrooms using a semestered block schedule. The one potential weakness shared by all the studies is the possibility of a school-level volunteer effect. Perhaps schools that adopted a semestered block schedule tended to have systematically weaker mathematics achievement before they adopted the schedule change. Raphael et al. (1985) did provide evidence that schools using semestered block schedules were similar in socio-economic to schools using traditional schedules. Further, the current author contacted both Wild and Bateson—one of the Marshall, et al. (1995)

authors—who provided anecdotal evidence that the schools adopting a semestered block schedule were not lower-achieving to begin with. Nonetheless, studies in the southern United States that have systematically investigated this issue have found that schools adopting a semestered block schedule did indeed have lower mathematics achievement before adopting the schedule than did schools which remained on a traditional schedule (North Carolina Department of Public Instruction, 1999; Texas Education Agency Research and Evaluation Division, 1999). The possibility that a similar pattern occurred in Canada cannot be entirely disregarded.

Mathematics test scores: United States studies. In recent years, several case studies have been published comparing mathematics achievement at particular sites in the United States under a semestered block schedule to achievement at the same site under a traditional schedule, but some were of very low quality and are not reviewed here.

Among the higher quality studies, one compared the achievement of students who completed algebra or geometry under a traditional schedule at two high schools in Dothan, Alabama in 1993-94 to that of students who took algebra or geometry from the same teachers under a semestered block schedule in the fall of 1994-95. All students were given a nationally normed standardized test for algebra at the end of the year in May, 1994, or at the end of the fall term in 1995. Although students using the traditional schedule scored higher on both subjects, the differences were not statistically significant (Lockwood, 1995).

Another study compared a sample of 355 students in Grades 8, 9, 10, and 11 who had attended a semestered junior high school to a matched sample of 355 students who attended one of two other junior high schools in the same geographic quadrant of the city whose school size, ethnic, and socio-economic make-up were comparable. Students were paired based on grade level, gender, ethnicity, and fifth grade scores on the Iowa Test of Basic Skills (ITBS). The “semestered block schedule group” of eighth and ninth graders in their study were still using the semestered schedule; the tenth and eleventh

graders had graduated from the semestered junior high school and were attending a traditionally scheduled high school. The study found that students who had attended the semestered junior high school scored lower on the mathematics portion of the ITBS than did the matched sample of students who had attended a regularly scheduled high school, but the differences were not significant at the .05 level (Cobb, Abate, & Baker, 1999).

Gruber and Onwuegbuzie (2001) conducted an investigation that closely parallels the current study in overall design. They compared achievement on the Georgia High School Graduation Test (GHS GT) of 146 high school students who graduated from a high school in the state of Georgia in the academic year 1996-97 to achievement on the GHS GT of 115 students who graduated from the same high school in the 1999-2000 academic year. Students who graduated from that particular school in 1997 attended high school on a traditional six-period schedule. In the academic year 1997-1998 the high school adopted a semestered block schedule. Therefore, the 1999-2000 graduating class received instruction via a semestered block schedule for three years. The students receiving three years of instruction under a semestered block schedule scored significantly lower on the mathematics portion of the GHS GT than did students who had attended high school using only a traditional schedule. The effect size of the difference was .52 standard deviations.

In the United States, three large-scale studies have compared mathematics test scores under a semestered block schedule to test scores under a traditional schedule. Their results are different from those of similarly large scale studies conducted in Canada. Two found little difference between schedule types, and one found a significant difference in Algebra test scores favoring the semestered block schedule.

The Texas Education Agency Research and Evaluation Division (1999) looked at the effects of schedule on student achievement at the 600 of Texas' 1070 high schools for whom complete demographic data was available. The demographic profile for schools on a semestered block schedule was different than that for schools with either alternating-day block schedules or traditional schedules. Semestered schools tended to be in larger districts, to have a larger percentage of ethnic minorities, and to be located in the least wealthy areas of the state. After controlling for these factors, there was no difference by schedule type in percentage of test-takers on campus who passed the Texas Academic Assessment System (TAAS) mathematics test in the spring of 1997, nor on any other academic measure. It should be noted that although the Texas study published both demographic data and conclusions, it did not contain a detailed description of the analysis methodology that produced those conclusions.

A second study analyzed results in Iowa and Illinois on the 1999 administration of the ACT Assessment, a test administered by ACT, Inc. and used by many colleges to assess high school students' general educational development and their ability to complete college-level work. The study reported data for 568 schools, including 351 using traditional all-year schedules, 161 schools using an alternating-day schedule, and 56 schools using a semester plan. After controlling for school demographic characteristics, there were no significant differences by schedule type in any area of achievement on the ACT, including mathematics (Pliska, Harmston, & Hackmann, 2001).

Zhang (2001) reported the latest in a series of studies that have been conducted by the North Carolina Department of Public Instruction to investigate the effects of semestered block scheduling in North Carolina high schools. In North Carolina, the percent of high schools using a block schedule grew from 1.6% in 1992-3, to 35% in 1994-5, to 64.8% in 1996-7, to 73.6% in 1997-8 (North Carolina Department of Public Instruction, 1999). Zhang compared student end-of-course test scores in two groups of schools. One group used a traditional schedule throughout the years 1993-2000, while the other group of schools adopted a semestered block schedule during the peak years of implementation, in 1995, 1996, or 1997. This yielded a sample of 214 Grade 9-12 high schools, consisting of 146 schools who had adopted a semestered block schedule and 68 schools who used a traditional schedule. The data set included scores from 640,000 end-of-course tests in Algebra 1, English 1, Biology, Economic Legal and Political Systems, and US History that had been completed at these schools between the years 1993 and 2000. For purposes of analysis, end-of-course scores were converted to "T-scores," with a mean of 50 and a standard deviation of 10. The reader should note that the term "T-score" as used by Zhang is not related to the *t*-test frequently reported in statistical literature.

Zhang reported the data in two ways, through an Analysis of Covariance (ANCOVA) and by graphically depicting change in student performance over time. The ANCOVA used the mean of school-level scores in 1997, 1998, 1999, and 2000 as a dependent variable and controlled for school-level variables reflecting percent low parent

education level, percent free and reduced lunch, percent non-white, and mean school-level subject area score in 1993 and 1994. Because block scheduling was not in widespread use in North Carolina in 1993 or 1994, Zhang used the mean of school-level 1993 and 1994 test scores as a covariate. Results of the ANCOVA found significant differences among schools only in Algebra 1 scores. The semestered block schedule schools outscored the traditionally scheduled schools with an adjusted mean of 48.2 points versus 47.2 points.

The effect size of the difference in Algebra 1 scores was small. Since T-scores have a standard deviation of 10, a 1-point difference translates into an effect size of 0.1. However, the ANCOVA obscures a more impressive trend visible in the descriptive data reported by Zhang. Since 1997, algebra scores at the schools using a semestered block schedule have been steadily rising, while those at schools using a traditional schedule have been falling. Although the semestered schools had generally lower Algebra 1 scores through 1997, by 2000 the semestered schools were outscoring the traditionally scheduled schools by approximately two Algebra 1 T-score points. It should also be noted that while semestered schools had a smaller percentage of minority students (in 2000, 34.4% versus 42.2% for traditional schools), they had lower scores on the two measures of socio-economic status. In 2000, the semestered schools had 30% of their students on free or reduced lunch, as opposed to 22% on free or reduced lunch for the traditionally scheduled schools. North Carolina's semestered schools had 49.2% of their students with a parent education level of high school or lower, as compared to 35.2% for traditionally scheduled schools.

Interaction of Semestered Block Schedule with Curriculum

As noted previously, researchers investigating block scheduling have often claimed that in order to be successful it is crucial to modify the mathematics curriculum. Generally, this claim has been based on surveys (King, et al., 1975; Sessoms, 1995) or on the testimony of teachers (Harter, 1994; Kramer, 1996).

The literature on achievement under a semestered block schedule generally supports this claim. Semestered block schedules have correlated with distinctly lower mathematics test scores across Ontario (Raphael, et al., 1985), across British Columbia (Marshall, et al., 1995; Wild, 1998) and in a case study reported in Georgia (Gruber and Onwuegbuzie, 2001). In the case of British Columbia, a provincially mandated curriculum and testing program prevented schools from changing the curriculum when they adopted a block schedule (Kramer, 1997). In the case of Ontario, the curriculum structure described by Raphael, et al. (1985) is quite similar to that in British Columbia, and there appear to be similar structural impediments to changing the curriculum. In the Georgia case study, the authors noted that there were “no notable changes in the curricula” when the school moved from a traditional schedule to a semestered block schedule. According to a personal communication from the authors (April, 2002), this specifically meant that although more courses were offered yearly under the block schedule, students in general did not take a larger number of mathematics courses than they had under the traditional schedule. Students took the same number of courses, and teachers compressed an unchanged curriculum into the shorter time available per mathematics course under the semestered block schedule.

Semestered block schedules correlated with higher mathematics test scores only in the case of Algebra 1 achievement in North Carolina. As reported by Zhang (2001) North Carolina differed from the other sites in that the algebra curriculum was modified under the semestered schedule. Specifically, under the semestered block schedule some students were able to split their algebra study over two courses, entitled Algebra 1a and Algebra 1b.

Interaction of Semestered Block Schedule with Teaching Methods

As noted previously, researchers reporting surveys and interviews of administrators, teachers and students using block schedules of any type have consistently emphasized the need to reduce lecture and increase the use of other means of teaching mathematics (Howard High School, 1994; King, et al., 1975; King, et al., 1978; Meadows, 1995; O'Neil, 1995; Reid, 1995a; Sturgis, 1995; Kramer, 1996).

At the sites where semestered block schedules correlated with lower mathematics achievement, there is evidence that such changes to teaching methods were limited. Raphael, et al. (1995) reported that teachers in Ontario's semester classes were more likely to use workbooks, individualized materials, and visual materials but the authors commented without further elucidation that "the differences, if real, (were) probably not very large (Raphael, et al., 1995, p. 43)." At the Georgia site described by Gruber and Onwuegbuzie (2001) teachers attended professional development in the summer before adopting the block schedule, but despite numerous staff changes there was no follow-up professional development at any time in the succeeding five years, and no systematic attempt to change teaching methods (personal communication, authors, April, 2002). In British Columbia, there was no allocation of planning time that would enable teachers to

change their teaching methods. In fact, British Columbia's schools kept planning time constant after switching to a block schedule by allocating to each teacher a full block-scheduled period for planning during some terms, and no in-school planning time during other terms (Bateson, personal communication, January, 1996; Reid, 1995a).

In North Carolina, where Algebra 1 scores increased under a semestered block schedule, planning time in block scheduled schools increased dramatically from one 50 to 55-minute period daily to one 90-minute period daily (Averett, 1994). Despite this, there may have been little change in teaching methods under the semestered schedule. Averett did report that in North Carolina, teachers used a wide variety of instructional practices in semestered classrooms, such as focusing on problem solving, conducting group discussions, and using performance assessments. However, more recent surveys contradict this impression. The North Carolina Department of Public Instruction (1997) reported instructional practices under the block schedule were mostly traditional, that is, lecture, students working at their desks, and small group work. Assessment practices consisted mainly of traditional paper and pencil tests. A survey of 2,167 North Carolina high school teachers, 1,036 of whom taught in traditional-scheduled programs and 1,131 of whom taught in block-scheduled programs, produced similar results (Jenkins, Queen, & Algozzine, 2001). There were few differences by schedule type in the use of most instructional strategies, with the exception that teachers in block scheduled classrooms used student coaching/peer tutoring slightly more often and used projects slightly less often. Both groups of teachers reported more extensive training in direct instruction/lecture than in any other teaching method.

Other authors have reported that under a semestered block schedule there was little change in methods of teaching mathematics (Shockey, 1997) or in teaching methods generally (Shortt & Thayer, 1997). This observation should not be surprising. Making appropriate adaptations to mathematics instruction under a block schedule will involve more than reducing the amount of teacher lecture and relying more on cooperative group work, individual projects, and peer tutoring. As Burrill (1997, p. 3) put it, “You can have students in cooperative groups working on trivial tasks. You can use manipulatives to do rote, meaningless procedures. A teacher can walk around encouraging students but never check their work or their thinking.”

Such implementations of “alternative methods” can have a negative effect on student learning, no matter what the schedule type. As part of the same testing program reported by Marshall, et al. (1995) students in British Columbia completed a survey asking how often various classroom activities occurred in math lessons. The British Columbia Ministry of Education (1995) reported that tenth graders with high scores on the test reported engaging in the following activities more often than did tenth graders with low scores on the test: “The teacher shows us how to do math problems,” “We copy notes from the board,” “We work from worksheets or textbooks on our own,” and “We discuss our completed homework.” In contrast, tenth graders with low scores on the test reported engaging in the following activities more often than did tenth graders with high scores on the test: “We work on math projects,” “We work together in small groups,” and “we check each other’s homework.”

In short, the forms of classroom activity that are most often used to replace lecture are not in and of themselves a panacea. Instead, the key to successfully limiting the

amount of lecture is probably to adopt the paradigm advocated by Hiebert, Carpenter, Fennema, Human, Murray, Olivier, and Wearne (1996) and make problem solving the basis for reform in curriculum and instruction. According to these authors, “analyzing the adequacy of methods and searching for better ones are the activities around which teachers (should) build the social and intellectual community of the classroom (Hiebert, et al., 1996, p. 16).” Centering classroom work around these kinds of activities will help to ensure that small group work and mathematics projects engage students in mathematical learning, not just trivial tasks.

At Suburban High School, the vehicle by which the mathematics faculty both redesigned the curriculum to match the timetable available under a semestered block schedule and redesigned instruction to center around mathematical problems was the IMP curriculum. The next section reviews the literature on IMP. It includes a description both of the purposes of the curriculum and of the research to date on achievement effects of IMP and similar curricula designed to implement the NCTM *Curriculum and Evaluation Standards* (NCTM, 1989).

The Interactive Mathematics Program (IMP)

IMP was written by a team of four authors, two of whom had earned doctoral degrees in mathematics and two of whom are classroom teachers of secondary mathematics (Fendel, Resek, Alper & Fraser, 1997). These authors began working in 1989 under a grant from the California Postsecondary Education Commission, and completed the curriculum with the assistance of funding from the National Science Foundation.

Although the authors began their work before the NCTM *Curriculum and Evaluation Standards* (NCTM, 1989) were published, early drafts of the *Standards* were available and from the beginning IMP was designed to comply with them. Thus, the language of the original California grant mirrored many of the key themes that were emphasized in the soon to be published *Standards*. Specifically, the IMP authors were engaged in developing a core high school curriculum that would replace the traditional Algebra I-Geometry-Algebra II/Trigonometry sequence and would set “problem-solving, reasoning, and communication as major goals; include such areas as statistics, probability, and discrete mathematics; and make important use of technology” (Key Curriculum Press, 1998, p. Section II, p. 9). Other goals adopted by the authors were to integrate algebra and geometry and to create a curriculum that would contain numerous small group and individual investigations. Further, in response to the *Standards*’ call to develop “mathematical power” in all students, the IMP authors designed a curriculum that could be used by students with a wide range of abilities in a heterogeneous classroom. In order to accomplish these goals, the authors explicitly acknowledged that they would follow NCTM’s (1989) advice to de-emphasize paper-and-pencil skills (Alper, Fendel, Fraser, & Resek, 1997).

The IMP curriculum consists of four textbooks, each containing five modular units. Each unit of the IMP curriculum generally begins with a central problem or theme. The problems are generally too complex for students to solve initially. Teachers guide students through a variety of smaller problems that develop the skills and concepts needed to solve the overarching unit problem. Numerous other long-term problems, called “Problems of the Week” are included in each unit (Key Curriculum Press, 1998).

Supplemental material is available in each unit to be used for additional practice or for extension, depending on student needs.

IMP Synergies With a Semestered Block Schedule

The characteristic of the IMP curriculum that first attracted the interest of the faculty at Suburban High School was its integration of algebra, geometry, trigonometry, probability, and statistics. Mrs. Sullivan, who chaired Suburban High School mathematics department at the time IMP was adopted, reported that she had long been concerned about the her students' lack of ability to use their algebra knowledge when they took her calculus classes. Generally, calculus students had taken Algebra 1 in eighth grade and Algebra 2 in tenth grade. She felt these students tended to "compartmentalize" their algebra knowledge into these two courses and were not fully prepared to apply it in calculus. A block schedule, with potentially longer gaps of time between classes, was likely to make this situation worse. In order to spread algebra study over more courses and better connect it to other areas of mathematics, the Suburban High School faculty decided to integrate the algebra, geometry, and trigonometry curricula when their school adopted a semestered block schedule. It was during discussions with nearby college faculty about the feasibility of writing their own integrated curriculum that Suburban High School teachers were first made aware of the possibility of utilizing the IMP curriculum.

A second aspect of the IMP curriculum that recommends it for use with a semestered block schedule is the active nature of IMP classrooms. According to the curriculum authors, maintaining an active classroom consistent with IMP's philosophy often means "replacing a teacher-led, whole-class discussion with a small-group activity that provides more immediate engagement for students (Alper, et al., 1998, p. 163)." To illustrate the way IMP accomplishes this, the authors presented the "Proof by Rugs" activity shown in Figure 1. After they work on the problem in small groups, some students present their ideas on why the diagrams constitute a proof. The IMP authors have found that when this occurs, all of the students are engaged in the problem and ready to listen to the reasoning. This is precisely the type of activity that the literature reviewed in the previous section indicates may need to be emphasized in order to improve mathematics achievement under a block schedule.

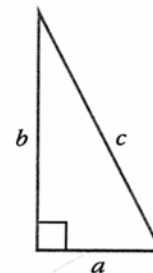
Figure 1. Sample IMP small group activity.

Classwork

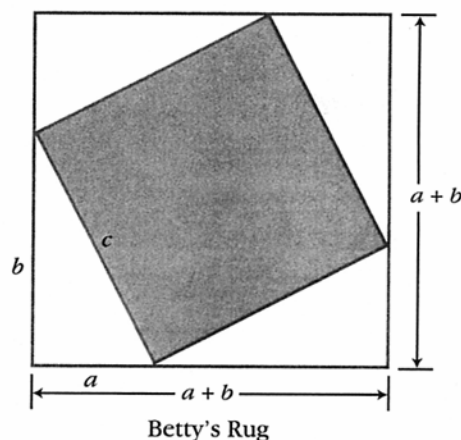
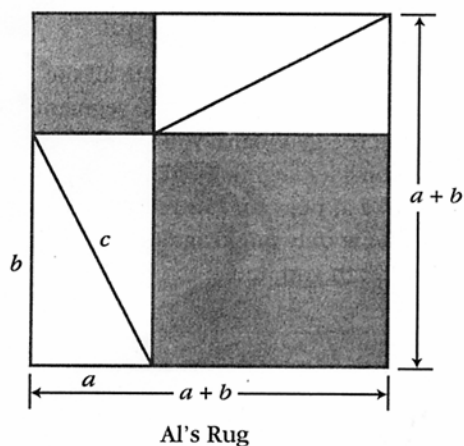
Do Bees Build It Best?

Proof by Rugs

Al and Betty have another game. They began with this right triangle, which has legs of lengths a and b and a hypotenuse of length c . Then they made the two square rugs shown below. Each rug has sides of lengths $a + b$, and the triangles within each square are the same as the single right triangle shown at the right.



When it's Al's turn, a dart drops on the square rug on the left. If it hits the shaded area, he wins a point. When it's Betty's turn, the dart falls on the square rug on the right. If it hits the shaded area, she wins a point. Assume that the darts always hit the rugs, but that they land randomly within the rug. In other words, all points on a rug have the same chance of being hit.



1. Is this a fair game? That is, is the chance of the dart landing on the shaded area the same for the two rugs? Explain your answer.
2. How do the two rugs demonstrate that the Pythagorean theorem holds true in general?

A related synergy between the IMP curriculum and a semestered block schedule involves planning and professional development. As stated previously, researchers have recommended that teachers be given both opportunities for professional development and additional planning time when adopting a block schedule (Brophy, 1978; King, et al., 1978; Watts & Castle, 1992; Averett, 1994; Meadows, 1995; Reid, 1995a; Canady & Rettig, 1995; Salvaterra & Adams, 1995; Kramer, 1997a). Precisely the same recommendation is made for teachers beginning to teach the IMP curriculum (Key Curriculum Press, 1978). Given the degree to which teaching the IMP curriculum naturally emphasizes the kinds of instruction that work well under a block schedule, learning to implement the IMP curriculum effectively can serve as the focus of professional development and planning time intended to improve mathematics instruction under a block schedule.

A final synergy between the IMP curriculum and the semestered block schedule may be the most important of all. Because the curriculum is composed of 20 distinct modules, it is relatively easy to match the course content to the time available per course under a block schedule, by distributing the modules appropriately among several consecutive mathematics courses. Thus, it is possible to take advantage of the schedule's structure by designing a curriculum in which content is covered over a larger number of courses, with less material contained in each individual course. As noted in a previous section, there are strong suggestions in the achievement literature that absent such an adjustment mathematics achievement is likely to decline under a semestered block schedule.

IMP Connections to Learning Theory

IMP's active and problem-centered approach to instruction is based on a constructivist theory of how students learn (Alper, et al. 1997). The IMP authors quote the following passage from the *Curriculum and Evaluation Standards* to provide a rationale for the active IMP classroom:

In many classrooms, learning is conceived of as a process in which students passively absorb information, storing it in easily retrievable fragments as a result of repeated practice and reinforcement. Research findings from psychology indicate that learning does not occur by passive absorption alone....Instead, in many situations individuals approach a new task with prior knowledge, assimilate new information and construct their own meanings....This constructive, active view of the learning process must be reflected in the way much of mathematics is taught....Our ideas about problem situations and learning are reflected in the verbs we use to describe student actions (e.g., to investigate, to formulate, to find, to verify) throughout the *Standards* (NCTM, 1989, p. 10).

A second theory that appears to have influenced the development of IMP is situated cognition (Brown, Collins, & Duguid, 1989; Goldman, Petrosino, & the Cognition and Technology Group at Vanderbilt, 1999). Situated cognition is based largely on two lines of research, namely investigations of transfer and investigations of expertise. Investigations of transfer have found that knowledge learned abstractly, in the absence of the context in which it is to be applied, is often "inert;" that is, when students need knowledge that they appear to have learned, they are unable to access it (Bransford,

Franks, Vye, & Sherwood, 1989). Investigations of expertise have found that experts tend to have rich, context-dependent schema for understanding and solving problems in their domain of expertise. Often, they develop this expertise by participating in a “community of practice” to obtain objectives of value. Expert tent makers often develop their knowledge as apprentices to more skilled tentmakers; expert mathematicians develop their knowledge working as graduate students with faculty mathematicians. Based on this perspective, advocates of situated cognition advocate having students work in groups on authentic problems that often take extended periods to solve. The influence of situated cognition on IMP can be seen in the curriculum’s use of extended problems and on the emphasis placed on cooperative groups. The authors intended for students to “explore open-ended situations actively, in a way that resembles the inquiry method used by mathematicians and scientists in their work” (Key Curriculum Press, 1998, section 2, p. 2).

Critiques of the IMP Curriculum

The IMP curriculum has generated considerable passion and controversy. In 1999, IMP was one of five middle and high school mathematics programs rated as “exemplary” by the U. S. Secretary of Education. Exemplary programs were selected by a three-tiered review, beginning with a review of each curriculum by four-person teams of teachers and others with expertise in education, followed by a similar review of data submitted to document each curriculum’s effectiveness, and culminating in a final selection of exemplary and promising programs by a panel of 14 educators, scientists, and policymakers. According to the panel’s report, “exemplary programs must be highly rated on quality, usefulness to others and educational significance and must provide *convincing* evidence of effectiveness in *multiple* sites with *multiple* populations (U. S. Department of Education, 1999, p. 1. Italics in the original).”

Soon after the report was released, a group of more than 200 mathematicians, educators, and scientists wrote an open letter to Secretary of Education Richard Riley, protesting the report (Klein, Askey, Milgram, Wu, Scharlemann, & Tsang, 1999). The letter, published as a full page advertisement in the *Washington Post*, criticized the panel for not including active research mathematicians. It further noted that one panel member had published an article claiming that teaching multidigit computational algorithms was a counterproductive and downright dangerous practice—in sharp contrast to a report published by the American Mathematical Society, which noted both the practical and theoretical importance of such algorithms. The letter noted that each of the letter’s authors had publicly criticized one or more of the exemplary programs. Finally, it requested that the Secretary withdraw the entire list of exemplary mathematics curricula for further consideration, and urged the secretary to include well-respected mathematicians in any future evaluation.

Wu’s detailed review and critique of the IMP curriculum. The debate about

the worth of curricula designed to implement the *Curriculum and Evaluation Standards* (NCTM, 1999) has continued, both in journals (Klein, 2000; Fey, 2000) and online. Among the numerous critical articles written, Wu’s review of the IMP curriculum (Wu, 2000), which was cited in the original open letter to Secretary Riley, stands out as providing the most detailed critique of the IMP program. In the context of a case study

designed to evaluate achievement effects of at a school using IMP, it is worthwhile to list Wu's concerns in detail.

Wu is a mathematics professor who teaches calculus at the University of California, Berkeley. He reviewed five IMP modules in 1992 and updated his review in March 2000, based on changes that had been made to the published form of the curriculum. Wu argued that the IMP curriculum needed to be judged on two criteria. These were its suitability for the approximately 15% of high school students who might consider pursuing a college degree in mathematics, science, or engineering, and its suitability for the remaining 85% who would pursue a different career or course of study.

For the 85% of students who either do not go to college or will not pursue scientific studies in college, Wu concluded that while IMP had serious flaws, he knew of no textbook series that was clearly superior and many series that were substantially worse. Thus, he recommended that for this group of students all teachers would do well to consult IMP often for supplementary materials to be used in the classroom.

Wu's criticisms of the IMP curriculum for the non-mathematical, non-science-oriented 85% of students were based on the perception that if some of these students should change their minds and subsequently pursue a mathematics or science program of study, they should have sufficient knowledge to do so. He listed five specific criticisms.

First, while he felt the curriculum promote understanding, he noted that technical fluency was also important, viewing it as key to mastering the language of mathematics. For this reason, he believed the IMP curriculum should spend more time on drills.

Second, he criticized IMP for failing to follow through on major mathematical ideas by presenting the summary formulas that can be derived from them. It should be noted, however, that the specific examples on which he based his critique were IMP's

treatment of derivatives, which was contained in a module that has been replaced in the current version, and IMP's failure to present the quadratic formula, which has been changed in the current version. The quadratic formula is discussed in the current version of IMP.

Third, Wu criticized IMP for abusing open-ended problems and de-emphasizing correct answers. He noted that many of the assignments were designed in such a way that almost any answer could be acceptable, leaving students no standard by which to judge good work from bad.

Fourth, Wu criticized the inclusion of too many mathematical puzzles that could reinforce the misconception that mathematics is nothing but a bag of cute tricks. He was particularly critical of the inclusion of such puzzles on tests, because these items were not likely to test whether a student had learned specific mathematics content well, but whether a student had happened to be inspired at the particular time encompassed by administration of the exam.

Finally, Wu criticized IMP for refusing to acknowledge that mathematics could be inspired by abstract considerations. He felt that there was an almost exclusive emphasis on real world problems. Historically, concepts like negative numbers and complex numbers have often been invented merely to satisfy the internal consistency of a mathematical system, and he felt that implying that all mathematics comes from real world applications would lead to a very biased view of the subject.

In contrast to his analysis for non-specialists in mathematics and science, Wu felt that IMP was far from adequate for the 15% of students considering further study in those fields. He had four primary criticisms.

First, he felt that IMP did not go far enough in abstracting key mathematical ideas. The presentation, he felt, stayed too close to the immediate problem situation and did not then extend ideas to their general applications and inter-connections across many situations.

Second, he felt that the mathematics and science specialists needed more technical drills, just as the non-specialists did. It is interesting to note that his concern for ensuring students develop operational fluency foreshadowed an objective that has been included in the NCTM's new *Principles and Standards for School Mathematics* (NCTM, 2000), but was missing from the *Curriculum and Evaluation Standards* (NCTM, 1989) which helped inspire IMP.

Third, Wu felt that the IMP curriculum lacked sufficient emphasis on precision. He felt the exposition was sometimes so chatty and informal as to lead to sloppiness. He felt that while there were some excellent discussions of proof, these were few and scattered. More importantly, students had insufficient opportunity to see models of rigorous proofs or to write rigorous proofs themselves. Also, some of the open-ended problems were designed in such a way that they could lead potential mathematics or science specialists to believe that an incomplete solution could be an acceptable solution to mathematical problems.

Finally, Wu felt that group activities were over-emphasized. He felt that too little attention was given to individual reflection on the mathematics.

Critiques of constructivism and situated cognition. Much of the criticism of IMP and similar curricula has been based on doubts about the learning theories underlying their design. Anderson, Reder, and Simon (1996) challenged some of the educational recommendations that have been made by advocates of constructivism and situated cognition. These authors did not attempt to debunk either constructivism or

situated cognition, noting that under some interpretations they themselves were “constructivists” and had been called so by mathematics educators. Further, one of the authors had previously written a review of situated cognition supporting its compatibility with modern information processing theory. It was the contention of the authors, however, that some of the more extreme proponents of both constructivism and of situated cognition had taken the theories too far, making claims that were contrary to evidence available from research in cognitive psychology. Among the issues raised by Anderson, et al. (1996), this review will address four that are particularly relevant to the IMP curriculum.

The first issue applies directly to the concern of Wu (2000) that there is too little drill in IMP and similar curricula. According to Anderson and his colleagues some constructivists and some advocates of situated cognition have claimed that cognitive tasks cannot and/or should not be decomposed into smaller subtasks. Anderson, et al. presented a large body of evidence demonstrating that cognitive tasks can indeed be broken down into subtasks, and that these subtasks can often be practiced independently of the larger task with fruitful results. They noted a related claim sometimes made by constructivists, that excessive practice or “drill and kill” could lead to routinization and of knowledge and drive out understanding.

This constellation of ideas may have influenced IMP, which according to Wu (2000) in its draft form contained almost no drill and in its published version contains less drill than most traditional texts. However, practice of cognitive subtasks decontextualized from their original context may be critical for making retrieval of those subtasks fluent or automatic. As described in Bransford, Brown, and Cocking (1999, p. 22), within the overall process of solving a problem there are a number of sub-processes that, for the expert, are fluent or automatic. Fluency is important because effortless processing places fewer demands on conscious attention. Since a person can attend to only a limited amount of information at one time, ease of processing some aspects of a task gives the person more capacity to attend to other aspects of the task.

The second issue raised by Anderson, et al. (1996) is the claim by some advocates of situated instruction that abstraction is of little use and that real learning occurs only in “authentic” situations. This idea is closely related to claims that learning seldom transfers between contexts. Because current performance will be facilitated to the degree that the context matches prior experience, the claim is made that the most effective training is apprenticeship to others in the performance situation; abstract instruction, in contrast, is viewed as relatively useless. Anderson and colleagues respond to this claim with evidence from a number of studies that have demonstrated the usefulness of abstraction.

It is clear that the authors of IMP had no intention of avoiding abstraction. As described by Alper, et al. (1997), they often used contexts primarily for motivational reasons, that is, to make the situation concrete enough for students to begin thinking about the problem. But the intent was to start with more concrete situations and build from there to the relevant abstractions. Once students get involved in IMP problems, the authors claim that it is the mathematics, not the context, that holds their attention.

Nonetheless, Wu (2000) claimed that IMP does not go far enough in its use of abstraction. It is possible that IMP, hailing in part from the tradition of situated cognition, was influenced by that philosophy to pursue abstractions less fully than would otherwise be the case. In the end, the question of whether or not the amount of abstraction in the IMP curriculum is sufficient is probably best addressed not by a review of curriculum content, but rather by studies like the current one, which investigate whether students who have utilized IMP are able to apply the knowledge they have learned so that they are successful in future courses.

A third issue raised by Anderson, et al. (1996) is the contention by some advocates of situated cognition that instruction needs to be done in a highly social environment. This is based on the ideas that (a) virtually all jobs are highly social in nature and (b) learning is closely associated with its context. Anderson and colleagues countered with evidence that research on cooperative learning has provided mixed results. They noted that, while useful, cooperative learning is not a panacea. Some learning, in particular drill to fluency in important subtasks, may be best accomplished in individualistic contexts. Both Wu (2000) and Alper, et al. (1997) have noted the extensive use of cooperative learning in IMP classrooms. Wu viewed this as a weakness; Alper and colleagues viewed it as a strength.

This author's view is that cooperative learning can be a powerful vehicle for improving understanding of mathematics, but only if it is introduced from the perspective of constructivism in addition to or instead of the perspective of situated instruction. The key to success is ensuring that mathematics is made "problematic" in the sense described by Hiebert, et al. (1996).

Allowing the subject to be problematic means allowing students to wonder why things are, to inquire, to search for solutions, and to resolve incongruities. It means that both curriculum and instruction should begin with problems, dilemmas, and questions for students. We do not use "problematic" to mean that students should become frustrated and find the subject overly difficult. Rather, we use "problematic" in the sense that students should be allowed to problematize what they study, to define problems that elicit their curiosities and sense-making skills (Hiebert, et al., 1996, p. 12).

Norms must be developed so that cooperative groups become the locus of debate about different approaches to and understandings of mathematics problems. Student groups must inquire, search for solutions, and resolve incongruities. Groups must analyze the adequacy of student methods and search for better ones. Further the class as

a whole must compare and contrast the adequacy of methods devised by different groups. Disagreement among students in the group, and contrasts between the approaches of different groups to the same problem, provide opportunities for the cognitive dissonance that Piagetian theory describes as being key to students making major advances (accommodations) in their understanding.

IMP provides a context in which cooperative groups can be used in this way, but it does not guarantee that they will be. Brombacher (1997) observed five teachers from three cities in the United States who were utilizing the IMP curriculum during its pilot phase. Teacher experience ranged from 4 to 14 years. One of the teachers was teaching his first IMP course; two teachers had taught both IMP 1 and IMP 2; one teacher had taught IMP 1, 2, and 3; and one of the observed teachers had taught all four IMP courses. The teachers Brombacher observed had volunteered to teach IMP and, in some cases, had requested permission to implement the program. In addition to describing many positive aspects of the curriculum, Brombacher listed a number of concerns, including the following:

In the classes that I watched, all the students sat in groups, but in only a few did I watch students working together on tasks. In the rest, the groups seemed to be forums for general discussion while the teacher was with the other groups. Although I certainly saw teachers trying to “restrain themselves” and in most cases with great success, I have to ask where the student debate was. I never saw any students engage in debate over mathematics. I did see some point out errors to others, but I had hoped to see students really wrestling with ideas and problems, reaching some common solution based on mathematical argument. Sadly I did not. (Brumbacher, 1997, pp. 103-104).

The fourth issue raised by Anderson, et al. (1996) is the claim by some constructivists that since all knowledge is constructed by the learner, direct instruction by the teacher is not a good way to assist student learning. Anderson and colleagues

claimed, in contrast, that while in some circumstances people are better at remembering information that they create for themselves, there is considerable research showing they can also remember what they are told. Anderson, et al. (1996) were concerned with the tendency of constructivism to devolve into pure “discovery learning” and noted that investigations of discovery learning have generally produced equivocal or negative findings. According to Anderson and his colleagues, when students cannot construct the knowledge for themselves, they need some instruction. In a similar vein, in their article that advocated enabling children to construct knowledge via problematizing the mathematics curriculum, Hiebert et al. (1996, p. 16) stated, “Our position is that the teacher is free, and obligated, to share relevant information with students as long as it does not prevent students from problematizing the subject.”

Two articles by researchers from Vanderbilt University provide a good perspective on the place of instruction in constructivist thinking today. The article titles evoke the main ideas: *New approaches to instruction: Because wisdom can't be told* (Bransford, et al., 1989) and *A time for telling* (Schwartz & Bransford, 1998). The first article, in claiming that “wisdom can't be told” noted that knowledge obtained by “being told” is frequently inert. People when prompted can frequently tell back what they learned, but they fail to use relevant information in unprompted problem-solving situations. The article reports a number of studies demonstrating that, in contrast, a problem-oriented approach to knowledge acquisition, like that of IMP, can lead to knowledge that is not inert. The second article reported a series of studies in which college undergraduates studied fundamental concepts in cognitive psychology. The studies found that there can be an appropriate “time for telling.” One group of students

analyzed raw data from psychological experiments that reported what information people remembered in various situations. Students were to look for fundamental patterns and principles that determined what would be remembered. Subsequently, they attended a lecture that organized the patterns they had found into a theoretical framework that uses schema theory and encoding theory to predict what people are likely or unlikely to remember. One week later, students were asked to predict outcomes for a hypothetical experiment that could be interpreted in light of the concepts they had studied. Students who had engaged in the problem solving and discovery task followed by lecture were much more likely to use the concepts spontaneously and successfully than were students who had read a summary of the relevant results of the experiments, followed by lecture. They were also much more likely to use the concepts spontaneously and successfully than were another group who spent extra time on discovery of the concepts, but did not receive the lecture.

In light of the view that often “wisdom can’t be told,” the problem-solving approach central to IMP can be viewed as a real strength that is likely to enhance students’ ability to utilize what they learn. However, there is also a danger that IMP teachers will fail to take advantage of appropriate “times for telling.” Wu’s (2000) criticisms that IMP fails to follow through on major mathematical ideas and fails to provide sufficient abstraction may reflect the results of missed opportunities for “telling”—opportunities that could potentially be addressed in student texts, in teacher manuals, or even in IMP teacher training, but perhaps have not been. The *Teaching Handbook for the Interactive Mathematics Program* (Greene, 2000) advises teachers that they will often need to “bite their tongues” to avoid robbing students of the “Ah-ha!”

experience by telling them the conventional super-formula that can answer a problem with which they are struggling. It provides advice on questioning strategies to move students towards appropriate discovery. But it does not address finding appropriate times to provide information or finding ways to ensure that students have sufficiently abstracted key concepts. Based on recent ideas expressed by Schwarz and Bransford (1999) and by Hiebert, et al. (1996) this may be a shortcoming of IMP as currently implemented.

Studies of Achievement under IMP and other Reform Mathematics Curricula

Because it is very difficult to conduct educational research that is both experimentally rigorous and externally valid, conclusions about the achievement effects of IMP and similar reform curricula will need to be based on the accumulated evidence provided by a large number of studies. The current study will be part of a growing body of research that is beginning to provide such evidence. According to Schoenfeld (2002), the data available so far seem to support the following findings:

1. On tests of basic skills, there are no significant performance differences between students who learn from traditional or reform curricula.
2. On tests of conceptual understanding and problem solving, students who learn from reform curricula consistently out-perform students who learn from traditional curricula by a wide margin.
3. There is some encouraging evidence that reform curricula can narrow the performance gap between whites and underrepresented minorities.

Schoenfeld's (2002) conclusions were tentative and were based largely on data from studies investigating elementary school curricula. This section of the literature review provides a more detailed account of studies in Schoenfeld's review that provided information about achievement in high school mathematics. In some cases, the current review references the original studies, whereas Schoenfeld referenced a forthcoming book containing the studies (Senk & Thompson, in press). The current review also discusses an important article by McCaffrey, Hamilton, Stecher, Klein, Bugliari, and

Robyn (2001) concerning achievement under IMP that was not included in Schoenfeld's review.

Studies of achievement under IMP's sister curricula. Curricula designed to implement the NCTM *Standards* (NCTM, 1989) were developed in the early- and mid-1990s. It takes three or four years after a multiyear high school curriculum has been implemented before students have completed enough of it for their achievement to be tested. Given this timeframe, the first detailed studies of mathematics achievement under the new high school curricula have only recently begun to be published. The current section reviews two such studies, one addressing the University of Chicago School Mathematics Project (UCSMP) *Advanced Algebra* (Senk, et al., 1993) text, and the other addressing the algebra content in the Core-Plus curriculum (Huntley, et al., 2000).

In 1993-94, Thompson and Senk (2001) evaluated the UCSMP *Advanced Algebra* (Senk, et al., 1993) text. Their study investigated only curriculum, not schedule, effects. Although the authors did not specify, the fact that schedule is not described makes it probable that all schools participating in their study used a traditional all-year schedule.

Advanced Algebra is the fourth in a six-book sequence designed in the late 1980s to improve the Grade 7-12 mathematics curriculum. The full series contains the books *Transition Mathematics; Algebra; Geometry; Advanced Algebra; Functions, Statistics, and Trigonometry with Computers; and Precalculus and Discrete Mathematics*. Like IMP, the UCSMP program was recognized by the Secretary of Education in 1999, but due to a lesser amount of available achievement data, the project was selected as "promising" rather than "exemplary."

Although originally developed before the *Curriculum and Evaluation Standards* (NCTM, 1989) were promulgated, *Advanced Algebra* was revised to be compatible with the *Standards*. The text is not as problem-centered as IMP, but *Advanced Algebra* spends much more time than do more traditional texts emphasizing applications and multiple representations of algebraic concepts. Before adopting IMP, teachers at Suburban High School had used *Advanced Algebra* to teach a course called Algebra 3/Trig. They viewed the text as an intermediate step between a traditional curriculum and a reform curriculum like IMP.

Thompson and Senk's (2001) experimental methodology was remarkably strong, given the real world constraints usually experienced by educational research. Although the research article was written by the authors of *Advanced Algebra*, to minimize

researcher bias the study data were collected and analyzed by an outside evaluator. Four schools from varying regions in the United States were recruited to participate in the study: one from a white middle-class suburb of Atlanta, one from a rural area in transition toward becoming a suburb of Chicago, one from a small semi-rural community in Mississippi, and one from an affluent suburb Philadelphia. At each of the four schools, two teachers each teaching two sections of second-year Algebra agreed to participate in the study.

Although students were not randomly assigned to classes, the study was a true experiment at the teacher level. Within each school, one participating teacher was randomly selected to teach UCSMP *Advanced Algebra*, while the other continued to teach the school's traditional Algebra 2 text. It turned out that there were three "traditional" texts used among the schools studied, representing the three most commonly used Algebra 2 texts at the time the study was conducted. Teachers selected to use the *Advanced Algebra* text received a minimal amount of professional development not received by teachers of the Traditional text, consisting of two one-day meetings in Chicago, one in the fall and one in the spring.

The experimental procedure resulted in eight classrooms using the *Advanced Algebra* curriculum, two in each school and eight classrooms using a traditional text, two in each school. All participating classes were either heterogeneous in schools with no tracking or designated "average" in schools with tracking. Altogether 150 students participated in UCSMP *Advanced Algebra* classrooms, and 156 students participated in control classrooms.

Data was analyzed comparing matched pairs of classrooms. At the beginning of the school year, each UCSMP *Advanced Algebra* classroom was matched to the control classroom in the same school it most closely resembled, based on pretest scores and demographic characteristics. There were no statistically significant differences between the two members of any matched pair on the pretest. Further, students in UCSMP classes resembled those in comparison classes in race and gender.

At the end of the school year, students were administered a UCSMP designed test to assess the core content of second-year algebra. The test contained a 36-item multiple-choice section addressing the topics of linear expressions, equations and inequalities; quadratic expressions, equations, and functions; higher degree polynomials and general properties of functions; powers, roots, exponents, and logarithms; variation; sequences and matrices; and trigonometry. In addition, the test contained an Advanced Algebra “Problem Solving and Understanding” subtest, consisting of six free-response items designed to measure students’ abilities to solve multistep problems. The six items were chosen because each was solvable using any of several strategies, including numeric, symbolic, and graphical methods, and each required students to explain their reasoning.

At the end of the year, each teacher was asked which of the questions on the multiple choice test was “fair” to his or her students, in the sense that the content it was testing had been covered in class. Then, each UCSMP classroom was compared to its matched-pair control classroom on a “fair test” comprised of only those items that both the UCSMP *Advanced Algebra* teacher and the matched control teacher reported they had covered during the year. This yielded a unique “fair test” for each experiment/control pair of teachers participating in the study. Thompson and Senk (2001) reported eight

comparisons of UCSMP *Advanced Algebra* classrooms to matched control classrooms on these “fair tests”. They analyzed this data by performing matched-pairs *t*-test with seven degrees of freedom using classroom as the unit of analysis. This yielded a significant difference in favor of UCSMP, $t=3.57, p=.009$.

In performing the matched-pairs *t*-test with seven degrees of freedom, Thompson and Senk (2001) can be criticized for using classroom rather than teacher as unit-of-analysis. Because the *Advanced Algebra* classes may by chance have had better teachers than the comparison classes, a more appropriate method would have been to use a hierarchical linear model with students nested within classrooms, nested within teachers, nested within schools. An alternate approach that while less able to detect differences between UCSMP and non-UCSMP classrooms would preserve the nominal significance level, would be to aggregate student scores to teacher-within-school, and then to perform a *t*-test on four matched pairs, each pair consisting of a UCSMP teacher and a non-UCSMP teacher within a school. Fortunately, data available within the article made it possible for the current author to perform this latter analysis. The result confirmed the statistically significant results reported by Thompson and Senk, yielding a *t*-statistic of 3.239 with three degrees of freedom, $p=.048$.

Looking in detail at differences between paired classrooms on the “fair test”, the *Advanced Algebra* class outscored its matched traditional class in seven of the eight comparisons. Differences between paired classrooms were significant for four of the eight pairs, in all cases favoring the UCSMP class.

Thompson and Senk (2001) also compared classrooms to their matched pair on a “Conservative” test, consisting of 15 out of the 36 original multiple-choice items that all

eight teachers said they had covered in class. While the original 36-item test had been designed to contain items testing a balance of skills, properties, uses, and representations, the majority of items on the Conservative test measured skills. On the Conservative test a matched-pairs t -test with seven degrees of freedom using classroom as the unit of analysis yielded a non-significant difference in favor of UCSMP, $t=1.843$, $p=.108$. As with the analysis of data on the “fair test,” the current author ran a second analysis on the “conservative test” reported by Thompson and Senk (2002) using a more conservative statistical procedure with teacher as unit-of-analysis and three degrees of freedom. The results confirmed Thompson and Senk’s conclusions, yielding a non-significant difference with a t -statistic of 1.516, $p = .227$.

However, looking in detail at differences between paired classrooms on the “conservative” test, the *Advanced Algebra* class outscored its matched traditional class in six of the eight comparisons. Differences between paired classrooms were significant for three of the eight pairs, in all cases favoring the UCSMP class.

Finally, Thompson and Senk (2001) compared classrooms to their matched pair on the 6-item Problem Solving and Understanding test. While all teachers in UCSMP *Advanced Algebra* classrooms reported that they had covered material on all six items, comparison teachers reported that they had covered between 50% and 83% of the items. Given the difference in opportunity to learn, it is not surprising that for the Problem Solving and Understanding test a matched-pairs t -test with seven degrees of freedom using classroom as the unit of analysis yielded a significant difference in favor of UCSMP, $t=4.97$, $p=.002$. As with the analyses of data on the “fair test” and on the “conservative test,” the current author ran a second analysis on problem-solving data

reported by Thompson and Senk (2002) using a more conservative statistical procedure with teacher as unit-of-analysis and three degrees of freedom. The results again confirmed Thompson and Senk's conclusions, yielding a t-statistic of 16.951, $p < .0005$.

Looking in detail at differences between paired classrooms, the *Advanced Algebra* class outscored its matched traditional class in seven of the eight comparisons. All seven differences favoring the UCSMP classes were statistically significant, while the one difference favoring a comparison class was not. As can be deduced from the number of significant differences, at least one *Advanced Algebra* classroom showed statistically significant higher performance than its matched pair in at least one class at each of the four schools participating in the study.

Huntley, et al. (2000) reported a study that was conducted in 1997 to evaluate algebra achievement under the Core-Plus Mathematics Project (CPMP). Core-Plus bears a closer resemblance to IMP than does the UCSMP. Its development was funded as part of the same NSF project that funded IMP, and, like IMP, Core-Plus was designed to implement the vision of the *Curriculum and Evaluation Standards* (NCTM, 1989). Like IMP, Core-Plus was among the five mathematics programs that in 1999 was recognized as "exemplary" by the Secretary of Education. The 3-year core curriculum is comprised of 21 connected units comprised of several multi-day units in which major ideas are developed through investigations and applied problems. Like IMP, Core-Plus integrates algebra, geometry, trigonometry, statistics and probability, and linear functions, and makes extensive use of graphing calculators. Probably the most significant difference between the two curricula is that each Core-Plus unit is designed around one overarching mathematical theme, whereas each IMP unit is designed around one central problem.

Topics in the Core-Plus curriculum are organized in a concept-then-skills-then-abstraction order.

Of the 21 units comprising the first three Core-Plus textbooks, seven deal primarily with algebra, while an additional three units apply and extend algebra concepts and skills in the context of studying other mathematical content areas. It was algebra achievement that Huntley and her colleagues set out to investigate. In the Spring of 1997, among 36 schools that were piloting the Core-Plus curriculum the researchers recruited six schools that would participate in the study, two from the Southeast, two in the Midwest, one in the South, and one in the Northwest. At each site, one or more classes completing the third Core-Plus textbook participated in the study, as did one or more comparison classes that were studying the third year of high school mathematics using a more traditional program. Each comparison class, selected from the same school or a nearby school, was to be comparable in ability to the Core-Plus class.

Unlike Thompson and Senk (2001), Huntley, et al. (2000) were unable to implement a random experimental design. Instead, they performed a quasi-experiment on already intact classrooms. At four of the six sites, they used eighth-grade test scores to ensure Core-Plus groups and comparison groups were comparable. At one of these four sites, students had comparable ability on entry into high school. At three others, the researchers used blocking techniques to match CPMP students with comparison-class students who had comparable mathematics achievement or aptitude scores in Grade 8. Of the two sites where Grade 8 test scores were unavailable, one had randomly assigned students to Core-Plus and control treatments on entry into Grade 9. At the remaining site, the researchers relied on repeated assurances from the school that the two groups were

indeed equivalent. At some sites both Core-Plus and comparison students were below average in prior achievement, at some sites both groups were above average in prior achievement, and at one site both groups were heterogeneous.

At each site, the researchers collected data over two days in April or May. At the time of testing, the Core-Plus group at five of the six sites had completed all algebra units in the first three textbooks before the time of testing; at the sixth site, where the class contained lower-ability students, they were just beginning the second of three algebra units contained in Course 3. Comparison classes used a wide variety of textbooks, including advanced algebra texts, a discrete mathematics with applications text, a text focusing on mathematics applications, and a text focusing on the use of mathematics in business settings. It should be noted that, since not all the comparison groups utilized algebra texts, some students in the comparison groups may have had limited opportunity to learn algebra content. This provides a possible alternate explanation of results obtained by Huntley, et al. (2000) and should be kept in mind when interpreting those results.

As noted previously, the Core-Plus researchers collected algebra achievement data using an assessment instrument that was also utilized for the current study. Their version used the form for Part 1 contained in Appendix A of this paper, plus three parallel forms. They used the form for Part 2 contained in Appendix A, plus one parallel form, and the form for Part 3 contained in Appendix A, plus two parallel forms. Forms of the assessment were randomly distributed among students at the time of testing.

For each part of the test, Huntley and her colleagues performed a simple comparison of the mean score across all forms of Core-Plus students to the mean score across all forms of Comparison students. On Part 1, Performance on Applied Algebra Problems with Use of Calculators, the Core-Plus students scored higher than the comparison students by about 0.46 standard deviations. Core-Plus students scored higher than comparison students at five of the six research sites. The authors reported a statistically significant difference ($t_{560} = 5.69; p < .001$). However, the reported t-test used student as unit-of-analysis, which can be criticized because student scores within a

curriculum program (Core-Plus or Comparison) at a given site might not be independent from one another. A more conservative procedure would be to aggregate data to the curriculum-within-site level, and perform a matched-pairs t -test with five degrees of freedom. The current author did so, and could not confirm statistical significance ($t=1.581, p=.175$.)

On Part 2, Performance on Algebraic Symbol Manipulation Without Use of Calculators, the Core-Plus students scored lower than the comparison students by about 0.54 standard deviations. On Part 2, Comparison students scored higher than Core-Plus students at all six research sites. The authors reported a statistically significant difference ($t_{575}=-6.50, p<.001$). As with Part 1, the current author re-analyzed the data by performing a more conservative matched-pairs t -test with five degrees of freedom, using program-within-site as unit of analysis. Again, the statistical significance could not be confirmed ($t=-2.455, p=.058$).

Part 3, Performance on Open-Ended Applied Algebra Problems with Use of Calculators, was completed by students working in pairs. The Core-Plus pairs scored higher than the Comparison pairs, by about 0.28 pair-level standard deviations. On Part 3, Core-Plus students scored higher than comparison at five of the six research sites. The authors reported a statistically significant difference ($t_{364}=2.77, p<.01$). As with Parts 1 and 2, the current author re-analyzed the data by performing a more conservative matched-pairs t -test with five degrees of freedom, using program-within-site as unit of analysis. Again, the statistical significance could not be confirmed ($t = .992, p = .367$).

Because the more conservative statistical analyses using curriculum-within-site as unit of analysis could not confirm statistical significance, it cannot be certain that the

differences reported by Huntley, et al. (2000) were non-chance. It is probably best to view the study by Huntley, et al. (2000) as a case study of an early implementation of the Core-Plus curriculum, which provided some indication that Core-Plus students were better than traditionally educated students at solving the types of algebra problems emphasized by the Core-Plus curriculum, but somewhat less good at solving the types of symbol manipulation problems emphasized by the traditional curricula.

In addition to the summary results for Parts 1, 2, and 3, Huntley, et al. (2000) reported details about how well students in the Core-Plus and Comparison classes performed on specific sub-skills within Part 1 and Part 2 of their test. Core-Plus students were much stronger than Comparison students at formulating algebraic models to describe a problem situation and at interpreting the meaning of an algebraic model presented in a problem situation. Core-Plus students were also better at what Huntley and colleagues called “representational fluency,” that is, translating among graphs, tables, and algebraic symbols to represent a function. Students in Comparison classrooms were much stronger than Core-Plus students at performing algebraic calculations without context or calculator access. However, when similar calculation problems were presented in context and calculators were available, the Core-Plus students were slightly stronger than those in Comparison classrooms.

Huntley, et al. (2000) made one last observation about how Core-Plus was implemented. At site 4, teachers supplemented the Core-Plus curriculum with materials that gave students more practice on traditional algebraic skills. Site 4 was the only site at which Core-Plus students matched the performance of control students on Part 2, Algebraic Symbol Manipulation Without Access to Calculators. This observation is

interesting in context of the current study, because although teachers at Suburban High School used a largely unmodified version of IMP for students who participated in the Algebra Achievement testing, in later years Suburban High School teachers supplemented IMP with materials similar in concept to those used at site 4 in the Core-Plus study.

Studies of achievement under IMP. Webb (in press) reported a series of studies designed to assess IMP's effects on students who utilized the program during its pilot years. His studies were conducted at nine schools between the years 1993 and 1997. Five of the schools studied were in California, two in the East, one in the Midwest, and one in a mountain state. In the series of studies, Webb addressed three questions. How did students who used IMP differ from comparable students enrolled in a traditional curriculum in the number of college-qualifying mathematics courses they studied during high school? How did students who used IMP differ from comparable students enrolled in a traditional curriculum in standardized test scores, as reflected on their high school transcripts? How did the achievement of students who used IMP differ from that of comparable students enrolled in a traditional curriculum in content areas where IMP has tried to increase emphasis, that is, probability and statistics and complex problem solving?

Webb addressed the question of mathematics course taking by examining the transcripts of 1,121 students who graduated from three high schools in California in 1993. At the time, they were the only schools at which students had had an opportunity to complete three years of IMP, generally in Grades 9 through 11. All three high schools served diverse student populations and offered a full range of mathematics courses, from

basic mathematics to Advanced Placement Calculus. All three schools offered students the option of enrolling in IMP or traditional mathematics, so at all three schools the IMP students were volunteers. Webb used prior test scores or course grades to ensure that students in the IMP group demonstrated prior mathematics ability that was basically comparable to that of students taking traditional mathematics courses.

Webb found that 64% of students who began their high school career in ninth grade by taking IMP Year 1 enrolled in four or more years of college-preparatory mathematics during high school, as compared to only 38% of students who began their high school career in Grade 9 by taking Algebra 1. A similar effect was observed in each of four ethnic groups that were represented across the three schools: Asian/Pacific Islander/Filipino, Black, Hispanic, and White. Webb reported that the result was statistically significant ($p < .01$) but did not account for possible within-school correlation when performing the statistical test.. Differences between the two groups in the proportion of students taking at least three years of college preparatory mathematics, or in the proportion of students taking pre-calculus or calculus, were negligible and not statistically significant.

Webb also looked at standardized test scores reported on student transcripts. At each school where data was available, Webb compared SAT mathematics test scores and/or Grade 11 CTBS mathematics achievement test scores of students who took IMP in ninth grade to those of students who took Algebra 1 in ninth grade, and of students who took IMP in tenth grade to those of students who took Algebra 1 in tenth grade. Results varied widely among the three schools, but only one contrast was statistically significant. At one of the schools, students who enrolled in IMP in ninth grade scored significantly

higher than did those who enrolled in Algebra 1 in ninth grade on SAT mathematics. This was true despite the fact that a larger percentage of students who enrolled in ninth grade IMP eventually took the SAT (34% for IMP students, versus 26% for Algebra 1 students).

Webb also addressed a question of great concern to parents in the school communities: Were students of particularly high ability likely to be harmed by IMP? To do so, at the school that had Grade 7 CTBS scores available, he created a matched sample of high-ability IMP students and high-ability Traditional students. For each group, Webb selected students who scored at the 76th percentile or higher on the Grade 7 CTBS, and were enrolled in their respective mathematics curriculum (IMP or Traditional) for at least 2.5 years. Fortuitously, this yielded 58 high-ability students in both the IMP and Traditional groups.

Regarding SAT scores, although differences between the two high-ability groups were not statistically significant, a larger percentage of IMP than of Traditional high-ability students took the SAT (83% versus 74%). Further, the IMP students had a higher mean SAT score, 544.8 versus 530.9. The high ability IMP students also had a higher grade point average than did the high ability Traditional students, both in mathematics and in all subjects excluding mathematics. Webb interpreted this to mean that involvement in IMP might be helping students in other courses as well. An equally likely possibility is that the difference in grade point averages indicates that, even though the two groups had similar prior test scores, there was some difference in attitude or ability that both made students more likely to volunteer for IMP and more likely to do better at school.

At the end of the 1995-96 school year, and again at the end of the 1996-97 school year, Webb tested student achievement on content emphasized by the IMP curriculum. He used three measures. In Grade 9, students completed modified versions of four statistics items that had been used by the Second International Mathematics Study (SIMS). In Grade 10, students completed two multi-step open-ended performance assessments prepared for the Wisconsin Student Assessment System. One of the items required some knowledge of probability and the other item required some knowledge of combinatorics. In Grade 11 students completed 10 multiple-choice items from a practice version of a quantitative reasoning test that was used by a prestigious university to screen its first-year students. The ten items focused mainly on data interpretation and sought evidence of how students used mathematics, probability, statistics, and computation to solve problems. Six high schools participated in this part of the study, although not every one of the six participated in every grade-level test. Webb completed all statistical analyses after controlling for eighth-grade test scores.

Grade 9 results appeared to demonstrate that opportunity to learn was the key to doing well on these assessments. At two of the three schools, the IMP students scored significantly higher on the-ninth grade probability and statistics test than did comparison students taking the traditional college-preparatory course (either Algebra 1 or Geometry, depending on the school). At the third school, the first year algebra course had been “enhanced” by the teachers to include a unit on probability and statistics. At that school the Algebra 1 students scored significantly higher than did the IMP students on the ninth grade probability and statistics test. The three effect sizes were +1.31, +.83, and -.76.

Tenth-grade results were similar, though a little more favorable to IMP. IMP students significantly outscored students in traditional classes on the open-ended problem at two of the three schools. Interestingly, one of the two schools where IMP students performed better utilized the UCSMP Algebra 1 and Geometry books for its “traditional” curriculum. At the third school, which had “enhanced” its geometry curriculum by adding a unit on problem solving and combinatorics, achievement of the IMP and Comparison groups was not significantly different, although the IMP students did slightly better. The three effect sizes were +1.04, +.74, and +.08.

The Grade 11 test was completed at two schools. In both cases, the IMP group scored better than did the Traditional group taking Algebra 2. The two effect sizes were +1.15 and +1.24.

Overall, Webb’s series of studies seemed to support the conclusion that IMP encouraged students to enroll in a larger number of college-preparatory mathematics courses, IMP did not harm and perhaps helped students on standardized tests, and IMP succeeded at teaching the content that was intentionally emphasized by the curriculum. The studies also demonstrated that other approaches could be equally successful at teaching probability, statistics, and other content emphasized by IMP, so long as teachers chose to give their students opportunities to learn such material.

McCaffrey, et al. (2001) provided an important additional perspective on how IMP can impact mathematics achievement. They looked at the relationship between instructional practices and achievement on the Stanford 9 test in 226 tenth-grade mathematics classrooms located in a large urban district that was pursuing reform under an Urban Systemic Initiative. One hundred eighty seven of the classrooms taught traditional algebra or geometry, while the remaining 39 used either IMP or a similar curriculum called College Preparatory Mathematics (CPM). The authors found that increased use of reform teaching practices consistent with inquiry-based instruction, as measured by teacher responses to 17 survey items that asked how frequently they engaged in various activities consistent with inquiry-based instruction, predicted better

achievement in classes using IMP or CPM, but was unrelated to achievement in classes using more traditional curricula.

The study by McCaffrey, et al. (2000) makes an interesting contrast to data provided by the British Columbia Ministry of Education (1995). As discussed earlier, that study found that tenth graders with low scores on a provincially administered mathematics assessment were more likely than students with high scores to report that “We work on math projects” and “We work together in small groups.” In contrast, the study conducted by McCaffrey, et al. found seemingly contradictory results. Their measure of “reform teaching practices,” was partly made up of questions that measured how frequently teachers made use of the same activities that had correlated with lower achievement in British Columbia. In contrast to the British Columbia results, McCaffrey and her colleagues found that “reform teaching practices” correlated positively with achievement in IMP classes.

It is possible that either the report by the British Columbia Ministry of Education (1995) or the report by McCaffrey, et al. (2000), or both, contained invalid or misleading results. The British Columbia study reported which classroom activities occurred “more frequently” in the classes of lower-achieving students, but provided no data on how “lower achieving” and “higher achieving” students were defined, on precisely how much “more frequently” the activities occurred, or on statistical significance of the findings. Further, the British Columbia report did not control for prior student ability in determining its results. The study by McCaffrey and her colleagues provided considerable detail on how their measure was defined and the magnitude of results, as well as a careful statistical analysis. The authors did control for prior ability. However, their measure of “reform teaching practices” may not have validly captured the construct they intended it to. For example, a teacher ostensibly using IMP who reports that his or her students “rarely” work on extended mathematics investigations or projects is probably not in fact utilizing the intended curriculum. In IMP classes, a low score on the “reform teaching practices” scale could in part be measuring lack of compliance with the course syllabus, rather than low implementation of inquiry based instruction.

Despite these concerns, it seems likely that both studies do in fact capture an aspect of reality. Probably, teachers in British Columbia engaging in small group work and extended projects were less successful than others in preparing their tenth-grade students for the provincially administered mathematics assessment. Probably, among teachers using IMP in the urban district studied by McCaffrey and her colleagues, teachers using these same instructional practices were more successful than others in preparing their tenth-grade students for the mathematics portion of the Stanford 9 test. As noted previously, the key difference may be the degree to which the “reform based” activities were used as a means to ensure students problematized the mathematics they studied (Hiebert, et al., 1996). The extended problems in IMP are designed to engage students in deep thinking about mathematics. Further, IMP’s problem-centered structure is likely to make it easier for teachers to use group work as a means to engage students in deep thought about mathematics, rather than as an end in itself.

As noted previously, adopting IMP makes it easier to adapt mathematics instruction in two ways that may be key to success under a block schedule: modifying the curriculum so that the amount of content covered in each course fits the schedule appropriately, and adopting a more reform oriented, inquiry based approach to teaching.

The study by McCaffrey, et al. (2000) provides evidence that, in addition, reform practices adopted to accommodate the block schedule are more likely to have a positive impact on student achievement in classrooms utilizing the IMP curriculum than in other classrooms.

In short, there is reason to believe that the IMP curriculum not only makes it easier to adapt mathematics instruction in ways that fit well with a semestered block schedule, but further the IMP curriculum makes it more likely these adaptations, once they have been made, will be successful in improving student achievement. In theory, then, one might expect to see a particularly positive impact on student mathematics learning at a school that adopts a semestered block schedule and at the same time adopts the IMP curriculum. The current study was designed to test this theory.

Chapter 3: Method

This chapter describes the research methodology used to investigate the joint effects of a semestered block schedule and the IMP curriculum on student mathematics learning at Suburban High School. First, the chapter provides a general description of the school community, including the students and teachers at the school. Then, the chapter describes the traditional curriculum and schedule that had previously been used at Suburban High School, and the IMP curriculum/semestered block schedule that were adopted. Then, the chapter describes data collection and statistical methodology for each of three approaches used to compare mathematics learning of students using the traditional schedule and curriculum to that of students using the block schedule and IMP curriculum:

1. analysis of an Algebra Achievement Test completed by students at the end of Grade 11,
2. analysis of student enrollment in mathematics courses as reported by student transcripts, and
3. analysis of student participation in and exam grades for Advanced Placement (AP) mathematics classes as reported on yearly School Profiles published by Suburban High School.

The chapter ends with a short description of qualitative data collected to help describe unique aspects of how the IMP curriculum and semestered block schedule were implemented at Suburban High School.

Participating Students

Participants in this study's analysis of algebra achievement include students at Suburban High School who were enrolled in the eleventh grade during the spring of 1997 (Traditional cohort), the spring of 2000 (First Reform cohort), or the spring of 2001 (Second Reform cohort). Participants in this study's transcript analysis include those who participated in the analysis of algebra achievement, plus students who used a traditional schedule and curriculum throughout high school, and graduated in the spring of 1996 or the spring of 1997. The analysis of Advanced Placement (AP) participation used information from additional years, analyzing transcripts of students who participated in other parts of the study, plus those of all other students who graduated between the spring of 1995 and the spring of 2001.

Suburban High School is the only high school in a small, relatively affluent school district in the Eastern United States. A number of colleges are located nearby, and the local population tends to be well educated. Traditionally, more than 75% of students at Suburban High School attend 4-year colleges after graduating from high school; counting both 2- and 4-year colleges, about 90% of Suburban High School students go on to college immediately after high school graduation.

The school generally enrolls between 900 and 1000 students each year, distributed as 200+ per grade, in Grades 9 through 12. Most students at Suburban High School have graduated from the district's only public middle school, which shares a campus with the high school.

In 1997, the Department of Education in the state where Suburban High School is located rated the school as being in the top quintile (20%) statewide in socio-economic status. In 2000 and 2001, the school was rated in the second-from-top quintile.

Since the 1991-92 school year, when students in the Traditional cohort were in sixth grade, students at the feeder middle school have used the *Visual Mathematics* curriculum (Foreman & Bennett, 1991). *Visual Mathematics* is an innovative curriculum designed to reflect the reforms characterized in the NCTM *Standards*. Published by a small non-profit company, *Visual Mathematics* is currently used only in a handful of districts across the United States. Because of their unusual middle school experience, students in both the Traditional cohort and students in the First Reform and Second Reform cohorts entered high school with a unique prior mathematics experience that may influence their readiness for solving the kind of problems emphasized by *Standards*-based curricula.

Participating Teachers

Mobility among mathematics teachers at Suburban High School is limited, so teachers are relatively experienced. Also, Mrs. Sullivan, who served jointly from 1984 through the spring of 1998 as mathematics supervisor for the school district and as high school mathematics department chair, had advocated adoption of both *Visual Mathematics* (Foreman & Bennett, 1991) at the middle school and the *Integrated Mathematics Program* (Fendel, Resek, Alper, & Fraser, 1997) at the high school. During that time, Mrs. Sullivan recruited teachers whose philosophy was broadly compatible with the teaching methods used in these two curricula.

Professional Development

Even before their school adopted the IMP, teachers at Suburban High School had received extensive professional development in problem-centered instruction. Each summer from 1993 through 1995, nearly all mathematics teachers from both the feeder middle school and the high school participated in 30 hours of professional development centering around the *Visual Mathematics* middle school curriculum. Once they had adopted the IMP curriculum, teachers at the high school received additional professional development to that curriculum. In general, teachers attended 30 hours of professional development on utilizing the IMP curriculum each summer for four consecutive

summers, beginning with a course in 1996 focusing on the first IMP textbook, and ending with a course in 1999 focusing on the fourth IMP textbook. Some teachers attended additional short workshops during the school year. Teachers who joined the mathematics faculty subsequent to 1996 also have enrolled in 30-hour courses focusing on each of the IMP textbooks, beginning with Book 1.

Planning Time

Under the traditional schedule, teachers at Suburban High School had available one 43-minute planning period per day. Under the semestered block schedule this was increased to a planning period consisting of one 80-minute block per day.

Traditional Schedule

Students in the Traditional cohort were enrolled in a 7-period day using a traditional mathematics curriculum throughout Grades 9-11. Each year students were enrolled in seven courses at a time, with each course meeting 43 minutes per day for the entire 180-day school year.

During the 1997-98 school year, when students in the Traditional cohort were in twelfth grade, Suburban High School adopted a block schedule school-wide. This study refers to the schedule adopted in 1997-98 as the “Pilot” block schedule, to distinguish it from a slightly different form of block schedule that was used in subsequent years. In 1997-98 students took 4 courses at a time, each course meeting 80 minutes per day over an 80-day semester. A 20-day spring session was reserved for special in-depth projects. This schedule change did not affect students in the Traditional cohort prior to the end of Grade 11, when they completed the Algebra Achievement test. However, the new schedule did affect their course-taking in twelfth grade. For this reason, when comparing

student course-taking under a traditional schedule and curriculum to course-taking under a block schedule and the IMP curriculum, this study used transcript data from the previous two cohorts of Suburban High School students: those who graduated in the spring of 1997, and those who graduated in the spring of 1996.

Traditional Curriculum

Suburban High School offered students in the Traditional cohort three “levels” of courses: Honors, College Preparatory (CP), and Academic Assisted (AA). For students taking Honors or College Preparatory courses, the core mathematics sequence was Algebra 1, Geometry, and Algebra 2. For Algebra I, they used *Algebra 1* (Larson, Kanold, & Stiff, 1995), published by D.C. Heath. For Geometry, they used *Geometry for Enjoyment and Challenge* (Rhoad, Malauskas, & Whipple, 1991), published by McDougal/Littel. For Algebra II, they used *Algebra 2* (Larson, Kanold, & Stiff, 1993), published by D.C. Heath. After completing the three core courses, students could enroll in Contemporary Mathematics, Algebra 3/Trigonometry, Functional Analysis, Discrete Analysis, Statistics, Calculus A/B and Calculus B/C. There were two distinctions between Honors and College Preparatory courses: 1) Honors students generally were given more challenging problem sets within any given unit; 2) Honors courses tended to be faster paced, covering a few extra concepts. For example, the Honors Algebra 2 class covered basic trigonometric functions, as well as arithmetic and geometric sequences, whereas the College Preparatory version of the same course did not.

For students in Academic Assisted courses, the first two core mathematics courses used texts entitled *Math Matters: An Integrated Approach* (Lynch & Olmstead, 1993a, 1993b) published by Southwestern. For the third course in the core sequence,

students could either enroll in a course using the third book of *Math Matters: An Integrated Approach* (Ebos & Zolis, 1987) or else enroll in a low level Algebra 2 course called “Algebra 2 Career/College Prep”, using the texts *Algebra* (McConnell, 1993) and *Geometry* (Coxford, 1991), published by Scott Foresman. Academic Assisted students who continued their math studies in twelfth grade could enroll in Contemporary Mathematics or Algebra 3/Trigonometry.

Reform Cohorts

Students in the Reform cohorts used a semestered block schedule and the IMP curriculum. The First and Second Reform cohorts were, respectively, the second and third group of students at Suburban High School to use the new schedule and curriculum.

The Semestered Block Schedule

The school piloted a semestered block schedule with ninth graders during the 1996-97 school year—the year before the students in the First Reform cohort entered high school. As noted above, during the 1997-98 school year when students in the First Reform cohort were in ninth grade, Suburban High School implemented a block scheduling school-wide. That year, students took 4 courses at a time, each course meeting 80 minutes per day for one 80-day semester. A 20-day spring session was reserved for special in-depth projects.

During the 1998-99 school year, when students in the First Reform cohort were in tenth grade and students in the Second Reform cohort were in ninth grade, the schedule was modified to consist of two 90-day semesters, as the 20-day spring session was eliminated. The schedule has remained unchanged since that time.

The IMP Curriculum

At the same time Suburban High School began piloting a semestered block schedule, the school also began piloting the IMP curriculum. Each of IMP’s four year-

long textbooks contains five modules. Individual modules are usually designed around a single over-arching problem whose solution requires a number of key mathematics concepts. Students spend several weeks working on sub-problems and related problems, developing the mathematics skills and knowledge needed to solve the module's central problem.

The four IMP textbooks cover most of the material contained in the traditional 3-year sequence of algebra 1, geometry, and algebra 2, plus some additional material generally contained in a trigonometry/pre-calculus course. In addition, units dealing with matrix algebra and/or units dealing with probability and statistics are included in each of the four textbooks.

All regular education students in the Reform cohorts used a sequence of four IMP courses to replace the traditional three-course core sequence. The IMP courses replaced Algebra 1, Geometry, and Algebra 2, at either the Honors, College Preparatory, or Academic Assisted level.

Suburban High School continues to distinguish among "levels" of courses. Different levels used the same IMP curriculum, but moved through it at differing speeds. The Honors level courses complete an entire IMP textbook (five modules) in each course; the College-Preparatory level courses finish four IMP modules per course, and the Academic-Assisted level courses finish three or four IMP modules per course. Table 1 displays the sequence each ability level followed as it completed the IMP modules as of 2000-2001. As Table 1 shows, Suburban High School adapted the IMP curriculum to the fewer hours available for instruction per course under a block schedule by completing less than one year's worth of material per course for all except Honors-level students.

Appendix E supplements Table 1 by providing a detailed description of the key concepts and skills within each IMP module as described on the Suburban High School 2000-2001 course syllabi.

It should be noted that, while the information in Table 1 is representative of the coursework students completed as they went through the Integrated Math sequence, the syllabi did change somewhat from year to year. In particular, in the earlier years of implementation teachers were less familiar with the IMP content, and course syllabi contained completed fewer modules per course than is reflected in the 2000-2001 data displayed in Table 1.

Also, teachers at Suburban High School have not always been able to complete the entire course syllabus. Partly because their students have the opportunity to study statistics either in an Advanced Placement or standard format after completing the IMP sequence, when pressed for time teachers have usually dropped one or more of the probability and statistics modules from their syllabus. According to teachers at the school, *The Game of Pig* and *The Pit and the Pendulum* have nearly always been taught as described in course syllabi, but *Is There Really a Difference?* and *Pennant Fever* have only sometimes been taught. *The Pollster's Dilemma* has generally not been taught, but it is the intent of the teachers that it will be in future years.

Table 1. Sequence of IMP modules completed by each ability group at Suburban High School

	Course in Which Module Was Completed		
	Honors	College Prep	Academic Assisted
IMP Textbook Year 1			
Patterns	Integrated Math 1	Integrated Math 1	Integrated Math 1

The Game of Pig	Integrated Math 1	Integrated Math 1	Integrated Math 1
The Overland Trail	Integrated Math 1	Integrated Math 1	Integrated Math 1
The Pit and the Pendulum Shadows	Integrated Math 1 Integrated Math 1	Integrated Math 2 Integrated Math 1	Integrated Math 2 Integrated Math 2
IMP Textbook Year 2			
Solve It!	Integrated Math 2	Integrated Math 2	Integrated Math 2 and 3
Is There Really a Difference?	Integrated Math 2	Integrated Math 3	-
Do Bees Build it Best?	Integrated Math 2	Integrated Math 2	Integrated Math 3
Cookies	Integrated Math 2	Integrated Math 3	Integrated Math 3
All About Alice	Integrated Math 2	Integrated Math 2	Integrated Math 4
IMP Textbook Year 3			
Fireworks	Integrated Math 3	Integrated Math 3	Integrated Math 4
Orchard Hideout	Integrated Math 3	Integrated Math 3	Integrated Math 4
Meadows or Malls?	Integrated Math 3	-	Integrated Math 4
Small World, Isn't It?	Integrated Math 3	Integrated Math 4	-
Pennant Fever	Integrated Math 3	-	-
IMP Textbook Year 4			
High Dive	Integrated Math 4	Integrated Math 4	-
As the Cube Turns	Integrated Math 4	-	-
Know How	Integrated Math 4	Integrated Math 4	-
The World of Functions	Integrated Math 4	Integrated Math 4	-
The Pollster's Dilemma	Integrated Math 4	-	-

Because students in different ability groups completed differing numbers of modules per course, there were some IMP modules that students in lower ability groups do not cover until twelfth grade, and others they did not cover at all. In general, by the end of Integrated Math 3, College Preparatory students completed most of the material usually contained in Algebra 1, Geometry, and Algebra 2. By the end of Integrated Math 4, College Preparatory students completed much of the material generally contained in a Trigonometry/Pre-Calculus course as well. The four modules deleted from the College Preparatory curriculum dealt primarily with probability, statistics, and matrix algebra.

Two of the modules that weren't addressed until Integrated Math 4 in Academic Assisted classes deal with concepts contained in the algebra achievement test utilized by this study. *All About Alice* deals extensively with exponential functions, and *Fireworks* presents extensive opportunities to work with quadratic equations. A third module, *Orchard Hideout*, covers key geometry concepts, and the fourth, *Meadows or Malls?* covers matrix algebra concepts that were left out of the College Preparatory classes. The modules that Academic Assisted classes never cover include three of the four dealing with probability and statistics, one dealing with matrix algebra, and nearly all of the Trigonometry/Pre-Calculus content. Some Academic Assisted students who wished to study the Trigonometry/Pre-Calculus content did so by enrolling in Integrated Math 4 College Preparatory after completing Integrated Math 4 Academic Assisted.

It should be noted while the Algebra topics tested on the Algebra Achievement test used by this study are addressed by IMP before the end of the Year 3 textbook, there is some review and extension of Algebra concepts in IMP Year 4. This is particularly

true of quadratic equations, which are among the topics addressed in the modules *Know How* and *High Dive*. Honors students completed both modules as part of Integrated Math 4, while College Preparatory students completed *Know How* in Integrated Math 4, but did not complete *High Dive*. For this reason, students in the Reform cohorts who took Integrated Math 4 in their senior year had not completed all of their Algebra study at the time the Algebra Achievement test was administered.

After completing the four Integrated Math courses, students could take Contemporary Mathematics, Functional Analysis, Discrete Analysis, Statistics, Calculus A/B and Calculus B/C. The Algebra 3/Trigonometry course, which had contained a mixture of Algebra review and more advanced topics that were now studied in Integrated Mathematics 4 College Preparatory, was discontinued.

Schedule and Curriculum: Summary

Table 2 provides the timetable followed by Suburban High School for implementation of the IMP curriculum and the semestered block schedule.

Table 2. Implementation of Semestered Block Schedule and IMP Curriculum at Suburban High School

Student Group	<u>1995-1996 and earlier</u>		<u>1996-1997</u>		<u>1997-1998</u>	
	<u>Curriculum</u>	<u>Schedule</u>	<u>Curriculum</u>	<u>Schedule</u>	<u>Curriculum</u>	<u>Schedule</u>
Grade 9 CP/AA	Traditional	Traditional	IMP	Block ^a	IMP	Block ^a
Grade 9 Honors	Traditional	Traditional	Traditional	Block ^a	IMP	Block ^a
Grade 10 CP/AA	Traditional	Traditional	Traditional	Traditional	IMP	Block ^a
Grade 10 Honors	Traditional	Traditional	Traditional	Traditional	Traditional	Block ^a
Grade 11 CP/AA	Traditional	Traditional	Traditional	Traditional	Traditional	Block ^a
Grade 11 Honors	Traditional	Traditional	Traditional	Traditional	Traditional	Block ^a
Grade 12 CP/AA	Traditional	Traditional	Traditional	Traditional	Traditional	Block ^a
Grade 12 Honors	Traditional	Traditional	Traditional	Traditional	Traditional	Block ^a
Student Group	<u>1998-1999</u>		<u>1999-2000</u>		<u>2000-2001 and later</u>	
	<u>Curriculum</u>	<u>Schedule</u>	<u>Curriculum</u>	<u>Schedule</u>	<u>Curriculum</u>	<u>Schedule</u>
Grade 9 CP/AA	IMP	Block	IMP	Block	IMP	Block
Grade 9 honors	IMP	Block	IMP	Block	IMP	Block
Grade 10 CP/AA	IMP	Block	IMP	Block	IMP	Block
Grade 10 honors	IMP	Block	IMP	Block	IMP	Block
Grade 11 CP/AA	IMP	Block	IMP	Block	IMP	Block
Grade 11 honors	Traditional	Block	IMP	Block	IMP	Block
Grade 12 CP/AA	Traditional	Block	IMP	Block	IMP	Block
Grade 12 honors	Traditional	Block	Traditional	Block	IMP	Block

^a In 1996-7 and 1997-8, mathematics instruction under the block schedule was conducted during two 80-day semesters.

A 20-day spring semester was reserved for special-interest courses. In 1998-9 and thereafter, mathematics instruction under the block schedule was conducted during two 90-day semesters. The 20-day spring semester was discontinued.

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.

Data Analysis: Eleventh Grade Algebra Tests

The Algebra Achievement test was intended to measure whether eleventh graders in the Reform cohorts differed from eleventh graders in the Traditional cohort in their understanding of algebra. To address this issue, the primary independent variable analyzed was TREATMENT. It could take on two possible values: 0 for the Traditional (1997) cohort, and 1 for the Reform (2000 or 2001) cohorts.

Covariate: Sixth Grade Test Scores

Beginning in the mid-1980s, students in the school district in which Suburban High School is located began taking a norm-referenced test called the “Comprehensive Testing Program (CTP)” published by the Educational Records Bureau. Until 1993 students completed basically similar tests in the spring of each year, from Grades three through ten. Students received scores in subjects across the curriculum, including mathematics computation, mathematics concepts, and general quantitative ability. In the spring of 1993, when students in the Traditional cohort were in seventh grade and students in the First and Second Reform cohorts were in fourth and third grade respectively, the Educational Records Bureau replaced the CTP II with the CTP III. The “quantitative ability” subtest was retained, but “mathematics concepts” and “mathematics computation” were combined into a single “mathematics” subtest. The Educational Review Board did not create an equated scale that could be used to translate the CTP III scale scores into CTP II scale scores.

While no scale-score equating was done, the Educational Records Bureau did rank student scale scores based on a national norm, with the norm recomputed yearly. As shown in Table 3, 1993 test scores in Suburban High School's district dropped in almost all grades, with a precipitous drop in some grades. The school district's *Testing Report* for 1993 notes that the drop may have been caused by the change from a test that had

Table 3. Yearly Median National Percentile Rank on Educational Records Bureau CTP Quantitative Ability Test

Grade	Year of Testing					
	1991	1992	1993	1994	1995	1996
3	82 nd	88 th	71 st	67 th	*	*
4	82 nd	86 th	82 nd	86 th	81 st	*
5	82 nd	88 th	88 th	91 st	91 st	88 th
6	76 th	87 th	75 th	87 th	86 th	90 th
7	83 rd	84 th	82 nd	85 th	84 th	88 th
8	*	87 th	54 th	81 st	88 th	85 th

* Data not available for this study

been used each year for the past nine years, to a new test adopted that year. For this reason, it is unlikely that the 1993 scores can be used as a valid control for analyses in this paper.

CTP III scores from the eighth grade, either the spring of 1994 for students in the Traditional cohort or the spring of 1997 and the spring of 1998 for the Reform cohorts, would be a good candidate to use as a control variable, even though there is some indication from median percentile ranks reported in Table 3 that scores in 1994 might still have been lower than in other years when students had more experience with the

particular type of test being used. Unfortunately, a number of student tests taken in 1998 were destroyed by a burst pipe before they could be graded. The destroyed tests included the eighth-grade tests taken by students in the Second Reform cohort.

Since seventh-grade scale scores from the Traditional cohort were probably invalid, and eighth-grade scale scores from one of the Reform cohorts were unavailable, this study used sixth grade-scores from the CTP II (level 4) as completed by students in the Traditional cohort and from CTP III (level E) as completed by students in the Reform cohorts. Four sixth-grade scores were found to be significant predictors of individual students' scores on the Test 1 and Test 2, the parts of the Algebra Achievement test completed by individual students. These four measures were Quantitative Ability, Reading Comprehension, Writing Mechanics, and Verbal Ability.

Test 3 was the portion of the Algebra Achievement test completed by students working in pairs. For each sixth grade measure available as a covariate, three alternate methods of describing the pair score were considered: the mean score for the pair of students completing the test, the maximum score of the pair taking the test (that is, the score of the more able student), and two scores consisting of the maximum and minimum score of the pair of students taking the test. Of the scores available, the best predictor was selected on the basis of the covariate or combination of covariates with the highest adjusted *R-square*. On this basis, the mean quantitative ability score for the two students taking the test was selected. After controlling for mean pair score on Quantitative Ability, none of the other covariates available were statistically significant, so they were not used in the final model.

Although not reported in Chapter 4, the analyses of Test 3 were run using the

alternate choice for covariate of the maximum quantitative ability score from the pair of students completing the test. The results of the alternate analysis were nearly identical to those reported in Chapter 4.

Scale scores on the CTP II taken in sixth-grade by students in the Traditional cohort have not been equated to scale scores on the CTP III taken in sixth grade by students in the two Reform cohorts. Nonetheless, both sets of scale scores are referenced to a “national percentile rank.” To control for prior ability, this study has matched sixth-grade scores based on national percentile rank. The legitimacy of this procedure depends on the assumption that a percentile rank in 1992, when students in the Traditional cohort took the CTP II, is comparable to the same percentile rank in 1995 or 1996, when students in the Reform cohorts completed the CTP III. That is, the assumption is that nationwide there was no large change in sixth-grade mathematics competency between 1992 and 1996. This assumption may be questioned: On average, national scale scores in mathematics on the National Assessment of Educational Progress increased by four points between 1992 and 1996 in both Grade 4 and Grade 8 (National Center for Education Statistics, 1997). For this reason, the Analysis section below supplements discussion of results when using sixth-grade test scores as a control by also reporting results without using sixth-grade test scores as a control.

Although sixth-grade scores were equated for this study by using national percentile rank, an unconverted percentile rank is not the best variable to use. In particular, analysis of Part 3, which was taken by students working in pairs, required computing the mean ability of a pair of students. Percentile rank is not an interval scale: for example, it takes a greater increase in ability to move from the 90th to the 95th

percentile than it does to move from the 50th to the 55th percentile. For this reason, sixth-grade scores were converted to z-scores before being used as a covariate. A z-score is the number of standard deviations a particular score is above or below the mean score; assuming ability is normally distributed, every percentile rank can be translated to a particular z-score. The z-scores were created in two steps: first, each national percentile rank was converted to a “national z-score”, defined as the z-score that would achieve that rank, assuming normal data. Then, in order to center the mean at zero for the Suburban High School data set, the “national z-scores” were reconverted to “Suburban High School z-scores” by subtracting the mean “national z-score” for all students used in this analysis, and dividing by the standard deviation.

Dependent Variables

The designers of the Algebra Achievement test used it to analyze results for three subscales (Huntley, et al., 2000):

1. “Applied Algebra Problems With Use of Calculators” consisting of all items on each of four forms they designed for Part 1;
2. “Algebra Symbol Manipulation Without Use of Calculators”, consisting of all items on each of two forms they designed for Part 2; and
3. “Open-Ended Algebra Problems With Use of Calculators”, consisting of three forms they designed for Part 3, each of which contained a single extended problem.

Subscales used for this study are necessarily somewhat different from those used by the Core-Plus authors, because the Core-Plus subscales used items from several forms for each part of the test, whereas this study utilized only one form for each of the three

parts of the test. Also, the pilot study indicated that one particular item on Part 1, Form C fit better on the subscale composed of Part 2 items than it did on the subscale composed of other Part 1 items, and one item on Part 2, Form A fit better on the subscale composed of Part 1 items than it did on the subscale composed of other Part 2 items.

Problem 1.2 on Part 1, Form C, asked students to write an equation for a line, given a graph of that line. The skill required was nearly identical to that tested by problem 14 on Part 2, Form A. In the pilot study for this proposed research, student scores on problem 1.2 correlated more highly with scores on Part 2 than with scores on other items in Part 1, and more highly correlated with problem 14 than with scores on any other item.

Problem 5 on Part 2, Form A asked students to identify an equation describing the relationship between the length and width of a rectangle, given that the length was four meters greater than the width. Except for the multiple-choice format of the question, the skill was similar to that required by items on Part 1 of the assessment that required students to formulate equations to describe algebraic situations. In the pilot study, student scores on problem 5 correlated more highly with scores on Part 1 than with scores on other items in Part 2.

Thus, the present study performed statistical analyses of the following three dependent variables:

Variable 1: Achievement on applied algebra problems in context, as measured by all items Part 1, form C except item 1.2, plus problem 5 form Part 2, Form A (hereafter referred to as Test 1). ;

Variable 2: Achievement on algebra symbol manipulation without context, as

measured by all items on Part 2, Form A except item 5, plus item 1.2 from Part 1, Form C (hereafter referred to as Test 2). ; and

Variable 3: Cooperative solution to an extended open-ended algebra problem, as measured by scores on the single extensive item in Part 3, Form A (hereafter referred to as Test 3).

Test 1, Test 2, and Test 3 were used to address, respectively, the first, second, and third research question posed in Chapter 1 of this study, namely:

- i. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to solve algebraic symbol manipulation problems? Do the results of this comparison differ depending on students' prior ability?
- ii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to interpret and solve challenging algebra problems presented in context? Do the results of this comparison differ depending on students' prior ability?
- iii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to collaboratively solve and communicate their solution to a complex open-ended algebra problem? Do the results of this comparison depending on students' prior ability?

Test reliability. For students participating in this study, Test 1 had a reliability (Cronbach's alpha) of .90, while Test 2 had a reliability of .89. Reliability for Test 3 was not computed, since Test 3 consisted of student scores on a single item.

Statistical Methodology: Rules for Establishing Confidence Intervals

Since this study examined three dependent variables, maintaining an experiment-wise error rate of 5 % requires a Bonferroni adjustment, assigning a Type I error rate to each of the three dependent variables of $.05/3=.0167$.

This study deals with the issue of statistical significance as follows. First, each variable is tested for a significant interaction with prior ability. For Test 1 and Test 2, prior ability is defined as the first principal component of the four grade-six ability scores, because this principal component correlated more highly with both Test 1 and Test 2 than does the quantitative ability score by itself. For Test 3, prior ability is defined as the mean quantitative ability of the two students who took the test.

If a Treatment-by-Ability interaction is not deemed significant, then Treatment alone is tested against each dependent variable, after controlling for prior ability. A 98.33% confidence interval is constructed for each the three effects (i.e., a 98.33% confidence interval around how much the Reform cohorts differed from the Traditional cohort on Test 1, on Test 2 and on Test 3). Thus, there is only a 1.67% chance that the true effect is outside the confidence interval, and a 95% probability that all three effects are actually within the reported confidence interval.

A 95% confidence interval is also reported for each of the three effects. For each test, there is only a 5% chance that the true effect is outside this confidence interval; overall, there could be as little as an 85% probability that all three effects are actually

within the reported confidence interval.

Statistical Models Used

The first two dependent variables (Algebra Problems in Context and Symbol Manipulation) consist of a student's average score on a large number of items and can be assumed to be on approximately a ratio scale. The pilot study found that residual scores on these variables, after entering controls, were approximately normally distributed. For ratio-scale data with normally distributed residuals, statistical methods based on the General Linear Model have optimum power and appropriate error rates. For this reason, when controlling for covariates the analysis of the first two dependent variables was performed using Ordinary Least Squares (OLS) Linear Regression, which is based on the General Linear Model. This methodology yields results identical to what would be reported by an Analysis of Covariance, or ANCOVA, but has the advantage of yielding effect sizes that can be interpreted. When not controlling for covariates, the analysis of the first two dependent variables was performed using an independent-samples t-test, which is mathematically equivalent to OLS regression.

The third dependent variable (Cooperative Solution to an Extended Open-Ended Algebra Problem) uses an ordinal scale of student scores, taking on possible values of 0 to 4. Analysis for this variable was identical to that utilized for the first two, except that Ordinal Regression was used instead of Ordinary Least Squares Regression. There are two varieties of Ordinal Regression that are commonly used with such data, Probit Regression and Logistic Regression. Unless the student ability distribution is very unusual, both types of regression will provide nearly identical results in terms of p-values, but the interpretation is slightly different. Each highlights a different and

important aspect of the data. Probit analysis provides an effect size that can readily be compared to output from the analyses of Test 1 and Test 2. Logistic regression analysis provides an odds ratio that is more easily related to student responses that were actually observed. In the interest of clearly explaining the observed results, both types of analysis are reported below.

Supplemental Analyses of Specific Skills

In addition to the omnibus statistical tests for differences in student achievement on Test 1, Test 2, and Test 3 a number of supplemental analyses are reported in this study. These analyses provided a finer-grained picture of how the Reform cohorts differed from the Traditional cohort on specific algebra skills contained within Test 1 and Test 2.

To facilitate the finer-grained analysis, two sub-scales of items were formed to examine the following specific skills:

1. Skill 1: Formulating Mathematical Models (Part 1, problems 1.1, 1.3, 1.5a, 1.5b, and Part 2, problem 5). Reliability as measured by Cronbach's alpha: .70
2. Skill 2: Interpreting Algebraic Models (Part 1, problems 4.1,4.2,4.3,4.4, and 4.5). Reliability as measured by Cronbach's alpha: .89

Differences between the Reform cohorts and the Traditional cohort on these two subscales were investigated using an independent-samples t -test. Items from Test 1 and Test 2 that were not on the specific subscales were examined individually. For dichotomous individual items, that is, items that were scored as right/wrong, a Pearson

Chi-square statistic was computed from a cross-tabulation table. On items for which students could receive partial credit, a Wald Chi-square statistic was computed from an ordinal Logistic Regression Analysis. The Logistic Regression Analysis is a generalization of the cross-tabulation method used to examine dichotomous items. Altogether, the supplemental analysis of specific skills compared the Reform cohorts to the Traditional cohort on 25 measures. To guard against over-interpreting results that occurred by chance, the Reform cohorts were deemed to be different from the Traditional cohort on one of the twenty-four measures if the statistical significance level for that measure reached a Bonferroni-adjusted $.05/25 = .002$ level.

Data Analysis: Student Transcripts

Because of the change to a semestered block schedule, no student at Suburban High School can complete mathematics courses entailing precisely the same number of hours as offered under the prior scheduling system. The semestered block schedule makes more courses available to each student, with each course lasting fewer hours. Students who allocate the same number of courses to mathematics as they would have done under the prior schedule will actually complete fewer hours of mathematics instruction. Students who enroll in more mathematics courses through semestered block scheduling than they would have under the prior schedule actually do receive more hours of mathematics instruction.

The semestered schedule affords students the opportunity to study more mathematics, if they are motivated to do so. In an extreme case, a student could hypothetically complete eight sequential mathematics courses over her/his high school career. Mathematics teachers believed that having students begin their study of high

school mathematics with the IMP curriculum would provide the motivation to study more mathematics. This belief is consistent with research to date. Webb (in press) investigated three schools in California that were the first to offer at least three years of IMP. He found that 64% of students at these schools who started IMP in Grade 9 actually completed at least four years of high school mathematics. In comparison, 38% of students at these schools who started a traditional sequence with Algebra I in Grade 9 actually completed at least four years of high school mathematics. This difference in course taking was significant at the .01 level.

The question remains: Did students really register in additional or more advanced mathematics under the IMP curriculum and semestered block schedule at Suburban High School than under the prior curriculum and schedule? The Secondary Research Questions listed previously break this larger question into the following more detailed questions:

- iv. How did students enrolled in a reform-based curriculum and a semestered block schedule differ from students enrolled in a traditional curriculum and traditional schedule in the number of registered mathematics class hours by the end of Grade 12?
- v. How did students enrolled in a reform-based curriculum and a semestered block schedule differ from students enrolled in a traditional curriculum and traditional schedule in participation in advanced courses, as measured by the number of registered hours in advanced mathematics classes by the end of Grade 12, by the number of students enrolling in Advanced Placement courses, and by scores on Advanced Placement tests?

This aspect of the study is addressed using two sources of data: student transcripts, and annual School Profiles published by Suburban High School.

The transcripts were used to determine course enrollment. For each student each year, the hours registered in mathematics class were computed by multiplying the number of mathematics courses in which the student was enrolled times the number of days per course times the number of hours per course per day

The number of hours registered in advanced mathematics classes was computed the same way, with an advanced class being defined as a class taken after the core requirements were completed. For the traditional curriculum, advanced classes were defined as those courses taken after successfully completing Algebra 2. For the reform curriculum, advanced classes were those courses taken after successfully completing Integrated Math 4.

Student transcripts were also used to determine the number of students enrolled in Advanced Placement courses each year, by grade level. Students completing Advanced Placement courses had the opportunity to earn college credit in mathematics by taking an Advanced Placement exam administered yearly by the College Board, and many students did so. This study obtained student grades on Advanced Placement exams from annual School Profiles published by Suburban High School.

Groups to be Compared

Students in the Traditional cohort who were tested in the spring of 1997 had used a traditional curriculum and schedule from ninth grade through the time of testing, when they were in eleventh grade. However, when those same students were in twelfth grade during the 1997-98 school year, Suburban High School adopted a pilot block schedule at

all grade levels. Since they were *not* enrolled in a traditional schedule in twelfth grade, it would not be appropriate to use these students' transcripts to address research questions iv. and v. Instead, this study used transcript data from the two preceding cohorts at Suburban High School: those students who graduated in the spring of 1996, and those students who graduated in the spring of 1997. Both of these cohorts of students were enrolled in a traditional schedule and curriculum throughout their high school tenure.

Complete transcript data was also available for the First Reform cohort, who graduated in the spring of 2001. As of this writing, students in the Second Reform cohort have not yet completed high school. Therefore, the transcript analysis in this report compared the registered hours of mathematics studied for students in the First Reform cohort, to the registered hours of mathematics for students in the Traditional cohorts who graduated in the spring of 1996 or the spring of 1997.

As noted previously, student transcripts were also used to identify the number of students enrolled in Advanced Placement mathematics courses each year by grade level and School Profiles were used to obtain student grades on Advanced Placement exams. Data on Advanced Placement enrollment was available each year from 1990-91 through 2000-01 and is reported. Data on Advanced Placement exam scores was available each year from 1994-95 through 2000-01 and is reported.

Statistical Methodology

Formal hypothesis testing could obscure important information about differences in the amount of mathematics that the Reform and Traditional students studied. This is because significance tests ask the question: "Is there a non-chance difference between the groups on some particular population parameter?" The parameter might for example

be an adjusted or unadjusted mean value (as reflected in t-tests or linear regression), or a median value (as in “nonparametric” tests like the Mann-Whitney U), or the odds of attaining a certain goal (as in Logistic Regression Analysis). Such population parameters report overall summaries for a set of data. They are likely to obscure possibly interesting patterns of differences between cohorts in detailed course-taking patterns. For example, reporting the mean number of hours enrolled in advanced mathematics courses might miss a pattern in which the bulk of students in a particular cohort enrolled in fewer hours, while a few students in that same cohort enrolled in substantially more hours. Other summary statistics have similar limitations.

Because of the limitations of traditional hypothesis testing, the analysis of the course registration data will consist of inspection of graphs depicting the registered mathematics hours for each group. In some cases, apparent differences between groups are tested for “statistical significance” to determine if they are likely to have occurred by chance. These post-hoc significance tests, which were conducted after identification of patterns from the graphs, should be viewed as exploratory analyses, not confirmatory hypothesis tests.

After examining differences between the AP exam scores of students who had used the semestered block schedule and the IMP curriculum and the AP exam scores of students in earlier years, a post-hoc Logistic Regression Analysis was done to see if the apparent differences between groups in the Advanced Placement Calculus BC exam could easily be dismissed as being chance variation. Because the statistical analysis of Advanced Placement Calculus BC exam grades was conducted after noticing a pattern in the data, it too must be viewed as an exploratory analysis.

Data Analysis: Documents and Key Informants

Research question vi states:

- vi. According to administrators and faculty who assumed critical responsibilities for implementing the shift to a semestered block schedule and reform-based mathematics curriculum, what school administrative policies and what unique aspects of the school and community affected mathematics course enrollment and mathematics achievement under the new schedule and curriculum?

Data collected from policy documents and key informants was sought in order to describe the key administrative and curricular decisions, as well as other variables that may have contributed to differences between the Traditional and Experimental cohorts in mathematics course-taking or in algebra achievement. Information gathered about these topics was used to inform the discussion in Chapter 5 of this study.

In addition to providing information to address the sixth research question, key informants also provided details about specific course content. This information was used to analyze student opportunity to learn specific topics of interest identified in the fine-grained analysis of specific skills within the Algebra Achievement test.

Summary of Method

This study compared the algebra achievement and mathematics course-taking patterns of two groups of students at a suburban high school in the eastern United States. The Traditional cohort used a traditional curriculum and schedule throughout high school. The Reform cohorts attended the same high school with primarily the same

teachers as the Traditional cohort, but they used a semestered block schedule and the IMP (Fendel, et al., 1997) curriculum.

The analysis of algebra achievement compared students on three subscales: “Algebra in Context,” “Symbol Manipulation,” and “In-depth Algebra Problem Solving.” These subscales were derived from an algebra assessment designed by the Core-Plus Mathematics Project (Huntley, et. al, 2000), and were administered to students in each cohort at the end of Grade 11. Items on the assessment were scored by a panel of experts using a scoring rubric, anchor items, and practice papers. Grade 6 test scores were used as covariates to control for differences between the two cohorts in initial ability. The analysis also investigated whether there is any interaction between treatment and prior ability as measured by sixth grade test scores.

In order to analyze mathematics course-taking student transcripts were reviewed to compare the number of hours that students from each cohort spent registered in mathematics classes in Grades 9-12. Hours registered in all mathematics classes, and hours registered in advanced mathematics classes were compared separately. Instead of formal hypothesis testing, a graphical display was used to detect important differences between the two groups in mathematics course-taking.

Yearly student enrollment in Advanced Placement courses was reported, using data obtained from student transcripts. Student grades on Advanced Placement exams administered by the College Board were obtained from annual School Profiles published by Suburban High School.

Chapter 4: Results

This chapter describes results from three quantitative analyses: the analysis of student scores on a three-part Algebra Achievement test taken at the end of eleventh grade, the analysis of mathematics course enrollment based on student transcripts, and the analysis of student performance on Advanced Placement mathematics exams, as reported in yearly School Profiles provided by the high school. Each analysis compared a Traditional group of high school students who used a traditional mathematics curriculum and year-long schedule, and a Reform group who used the IMP curriculum (Fendel, et al., 1997) within a semestered block schedule.

Algebra Achievement at the End of Eleventh Grade

In the spring of 1997, eleventh graders at Suburban High School completed a 3-part Algebra Achievement test that had been designed by the Core-Plus Mathematics project. Test 1 measured students' ability to understand and solve algebra problems presented in context. Test 2 was a traditional-style test of algebraic symbol manipulation. Test 3 measured students' ability to work in pairs to solve an extended applied algebra problem. Students tested in 1997 had used a traditional curriculum and schedule throughout Grades 9, 10, and 11. They form the Traditional cohort, to be compared to students in two Reform cohorts who attended Suburban high school in later years.

Two consecutive classes of eleventh graders at Suburban High School completed the same 3-part Algebra Achievement test in the spring of 2000 and the spring of 2001. Eleventh graders tested in 2000 and 2001 had completed courses that replaced the traditional algebra 1, geometry, algebra 2 sequence with IMP, and had used a semestered block schedule throughout high school. Students from these two cohorts comprise the Reform group, whose achievement was contrasted with that of the Traditional group.

For all students, testing was conducted over a two-day period. In the spring of 1997, all eleventh graders at Suburban High School were enrolled in a mathematics class, and testing was conducted in class. Once the block schedule was adopted, it proved to be

impossible to replicate this situation precisely, because there were a large number of juniors who were not enrolled in mathematics during any given semester. Therefore, for the one hour each day that was required to complete the testing conducted in 2000 and 2001, eleventh graders moved to a mathematics classroom or other classroom proctored by a mathematics teacher or a science teacher.

Results from Test 1: Understanding and Solving Algebra Problems presented in Context

Test 1 was constructed from 18 of 19 open-ended problems administered on the first day of testing, plus a single multiple-choice problem administered on the second day of testing. There were 196 students from the Traditional group who attended the first day of testing in the spring of 1997. There were 457 students from the two Reform cohorts who attended the first day of testing: 212 students tested in the spring of 2000, and 245 tested in the spring of 2001. Of the 663 students who attended the first day of testing, 43 did not attend the second day, and thus were missing a score on the one item from Test 1 that was administered on the second day of testing. For these 43 students, the score on that one item was estimated by using its predicted value, based on OLS regression of the 18 items from the test that the student had completed on the first day of testing.

Most items on Test 1 were scored using a rubric, with possible scores of 0, .5, 1, 1.5, or 2. The multiple-choice item was assigned a score of either 0 or 2. Thus, to obtain 100%, a student would need to get 38 out of 38 possible points.

Previous research indicated that Test 1 was likely to be very difficult for most eleventh graders. Except for one item, the test was identical to Part 1, Form C from the Core-Plus Algebra Test used by Huntley, et al. (2000). Early piloting of the four forms for Part 1 had shown Form C to be slightly more difficult than the others (personal

communication, R. S. Villarubi). Huntley, et al. reported testing results for two groups of students at each of six sites across the U.S. One group from each site had used Core-Plus Course Three, and the other group at each site had studied algebra 2 or some other traditional eleventh grade mathematics course. Student groups at the six sites varied in prior ability. Among all the groups tested, the group mean percent correct across the Part 1 forms ranged from a low of 25.2% to a high of 49.4%.

Test 1 results without controlling for prior ability. Table 4 reports differences on the percent correct between the Traditional and Reform groups at Suburban High School. As the table indicates, students in the Traditional group averaged 52.0% correct on Test 1. This result compares favorably to the scores of other groups who had taken similar tests. Students in the Reform group improved on this performance, as they averaged 58.2% correct. This difference is significant at the .002 level. Ninety-five percent confidence limits indicate that advantage in mean score due to being in the Reform group was probably between 2.3% and 10.2%.

Because this study looked at three related dependent variables, it used a Bonferroni adjustment, reporting a 98.33% confidence interval for each test. That way, one has a 95% confidence that actual differences between groups simultaneously lie within reported confidence intervals for all three dependent variables. As indicated by Table 4 there is a 98.33% probability that the advantage of the Reform group on Test 1 lies within the confidence interval (+1.5% , +11.1%).

The standard deviation for Test 1 scores in both the Traditional and Reform groups was 23%. Reform group students scored 6.2% better than did Traditional group students. This is an effect size of $6.2/23$, or approximately .27 standard deviations. Assuming that student abilities are normally distributed, then if a student would ordinarily score at the 50th percentile, improving his or her score by .27 standard deviations would move the student up to the 61st percentile.

Table 4. Difference Between Groups on Understanding and Solving Algebra Problems in Context

<u>Treatment</u>		<u><i>n</i></u>	<u>Percent Correct</u>		<u>Standard Deviation</u>		
Reform		457	58.2%		23.5%		
Traditional		196	52.0%		23.3%		
<u>Difference</u>	<u><i>t</i></u>	<u><i>df</i></u>	<u><i>p</i>-value</u>	<u>95% Conf. Interval</u>		<u>98.3% Conf. Interval</u>	
				<u>Lower Bound</u>	<u>Upper Bound</u>	<u>Lower Bound</u>	<u>Upper Bound</u>
6.2%	3.13	651	.002	2.3%	10.2%	1.5%	11.1%

As noted previously, results reported in Table 4 were computed after imputing scores on Part 2, item 5 for the 43 students who were administered all of Test 1 except that one question. Two alternative methods of handling the missing data yielded results very similar to those reported in Table 4. When the 43 students were simply dropped from the analysis, the mean score for the Traditional group was 52.7%, and the mean score for the Reform group was 58.4%. The 5.7% difference between groups was significant at the .005 level. Alternatively, a 1-parameter Item Response Theory model, also known as a Rasch model (Wu, Adams, & Wilson, 1998) was used to impute scores for the 43 students missing Part 2, Question 5 as well as for an additional 32 students who had scores on Part 2, Question 5 but not on the rest of Test 1. The Rasch modeling also incorporated multiple imputation (Rubin, 1987) to account for possible systematic differences in ability between students with missing data and students with complete tests. The Rasch model yielded an effect size of .25 standard deviations in favor of the Reform cohort, significant at the .008 level. However, because the Rasch model does not provide percent correct values for comparison to prior studies (Huntley, et al., 2000), this alternative approach was not deemed optimal. Details of the Rasch modeling are provided in Appendix F.

Test 1 results after controlling for prior ability. Grade 6 ability scores were available for 477 of the 663 students with scores on Test 1. This included 136 students from the Traditional cohort, 163 students from the First Reform cohort, and 178 students from the Second Reform cohort. There was no evidence of an interaction between prior ability as measured by the first principal component of the four sixth-grade scores, and the effects of Reform versus Traditional schedule and curriculum. After controlling for Treatment and for each of the four individual ability scores, the interaction had a t -value of .528 with 470 degrees of freedom, $p > .5$.

Table 5 reports differences between the Traditional and Reform cohorts on Part 1 scores, after controlling for sixth-grade ability. In computing the percent correct reported

in Table 5, all ability scores were centered at the mean score of all Suburban High School students in the Traditional or Reform cohorts who completed the Algebra Achievement test and for whom Grade 6 test scores were available. Thus, the percent correct for each group reported in Table 5 is the predicted score for a student whose ability in sixth grade was the average for this group of Suburban High School students.

Table 5 shows that controlling for Grade 6 ability makes little difference in results of the analysis for Test 1. The advantage of the being in the Reform cohort is 5.3%, or $5.3/23$, or approximately $.23$ standard deviations. Assuming that student abilities are normally distributed, then if a student would ordinarily score at the 50th percentile, improving her or his score by $.23$ standard deviations would move the student up to the 59th percentile. As before, alternate ways of handling missing data made little difference in the results. Of the 43 students for whom the response to Question 5 on the second day

Table 5. Difference Between Groups on Understanding and Solving Algebra Problems in Context After Controlling for Grade 6 Test Scores

<u>Treatment</u>		<u>n</u>		<u>Percent Correct for Average Ability Student</u>			
Reform	341			61.0%			
Traditional	136			55.7%			
Difference	t	df	p-value	<u>95% Conf. Interval</u>		<u>98.3% Conf. Interval</u>	
				Lower Bound	Upper Bound	Lower Bound	Upper Bound
5.3%	3.15	471	.002	2.0%	8.6%	1.3%	9.3%

of the testing was imputed, 34 had Grade 6 scores available. Dropping those 34 students changed the estimated effect of being in the Reform group from 5.3% to 5.0%. Using a Rasch analysis to impute data for those 34 students, plus the additional scores for an

additional 14 students who attended the second day of testing but not the first, yielded an effect size of .25 standard deviations, significant at the .004 level.

Results from Test 2: Algebraic Symbol Manipulation

Test 2 was constructed from 13 of 14 problems administered on the second day of testing, plus one problem administered on the first day of testing. There were 199 students from the Traditional group who attended the second day of testing in the spring of 1997. There were 443 students from the two Reform cohorts who attended the second day of testing: 207 students tested in the spring of 2000, and 236 students tested in the spring of 2001. Of the 642 students who attended the second day of testing, 32 had not attended the first day, and thus were missing a score on the one item from Test 2 that was administered on the first day of testing. For these 32 students, the score on that single item was estimated by using its predicted value, based on OLS regression of the thirteen items from the test that the student completed on the second day of testing.

Five items on Test 2 were multiple choice and were assigned a score of 0 or 2. The constructed-response item administered on the first day of testing was scored using a rubric with possible scores of 0, .5, 1, 1.5, or 2. The remaining eight items were scored using a rubric with possible scores of 0, 1, 2, or 3. Thus, to get 100%, a student would need to get 36 out of 36 possible points.

Previous research indicated that, like Test 1, Test 2 was likely to be difficult for most eleventh graders. Except for one item, the test was identical to Part 2, Form A from the Core-Plus Algebra Test used by Huntley, et al. (2000). Early piloting of the two forms for Part 2 had shown Forms A and B to have about the same level of difficulty (personal communication, R. S. Villarubi). Huntley, et al. reported that among the 12

groups of students who completed their test, the group mean percent correct for Part 2 ranged from a low of 15.3% to a high of 49.8%.

Test 2 results without controlling for prior ability. Table 6 reports differences on percent correct between the Traditional and Reform groups at Suburban High School. As the table indicates, students in the Traditional group averaged 58.2% correct on Test 2. This result compares favorably to the scores of other groups who had taken similar tests. Students in the Reform group averaged 53.8% correct. The 905% confidence limits for the difference between the Reform group and the Traditional group were between -8.9% and $+0.1\%$.

This study uses a Bonferroni adjustment, reporting a 98.33% confidence interval for each test to ensure that confidence intervals reported for all three dependent variables are simultaneously correct with 95% confidence. Table 6 shows that a 98.33% confidence interval for the difference between the Reform group and the Traditional group is $(-9.9\%, +1.1\%)$.

The standard deviation for Test 2 scores in both the Traditional groups was 26%. Reform group students scored 4.4% lower than did Traditional group students. This is an effect size of $-4.4/26$, or approximately -0.17 standard deviations. Assuming that student abilities are normally distributed, if a student would ordinarily score at the 50th percentile, decreasing his or her score by .17 standard deviations would move the student down to the 43rd percentile.

Table 6. Difference Between Groups on Algebraic Symbol Manipulation

Treatment		n		Percent Correct		Standard Deviation	
Reform	443	53.8%	25.6%	Traditional	199	58.2%	27.4%
Difference	t	df	p-value	95% Conf. Interval		98.3% Conf. Interval	
				Lower Bound	Upper Bound	Lower Bound	Upper Bound
4.4%	1.93	640	.073	-8.9%	+0.1%	-9.9%	+1.1%

As noted previously, results reported in Table 6 were computed after imputing scores on Part 1, item 2 for the 32 students who were administered all of Test 2 except that one question. Two alternative ways of handling the missing data obtained results that were similar to those reported in Table 6.

When the 32 students were simply dropped from the analysis, the mean score for the Traditional group was 58.2%, and the mean score for the Reform group was 54.3%. The 3.9% difference between groups was not significant ($p=.096$). Alternatively, a Rasch model (Wu, Adams, & Wilson, 1998) was used to impute scores for the 32 students missing item 1.2 of Part 1 as well as for an additional 43 students who had scores on item 1.2 of Part 1 but not on the rest of Test 2. The Rasch model yielded an effect size of -0.16 ; that is, the Reform cohort scored lower than the Traditional cohort by .16 standard deviations. The results did not meet this study's standards for statistical significance ($p=.066$). Details of the Rasch modeling are provided in Appendix F.

Test 2 results after controlling for prior ability. Grade 6 ability scores were available for 459 of the 642 students with scores on Test 2. This included 135 students from the Traditional cohort, 153 students from the First Reform cohort, and 171 students from the Second Reform cohort.

There was no evidence of an interaction between prior ability as measured by the first principal component of the five sixth-grade scores and the effects of Reform versus Traditional schedule and curriculum. After controlling for Treatment and for each of the four individual ability scores, the interaction had a t -value of .101 with 452 degrees of freedom, $p>.5$.

Table 7 reports differences between the Traditional and Reform cohorts on Test 2 scores, after controlling for sixth-grade ability. In computing the percent correct reported

in Table 7, all ability scores were centered at the mean score of all Suburban High School students in the Traditional or Reform cohorts who completed the Algebra Achievement test and for whom Grade 6 test scores were available. Thus, the percent correct for each group reported in Table 7 is the predicted score for a student whose ability in sixth grade was the average for this group of Suburban High School students.

Table 7 shows that controlling for Grade 6 ability provided a nearly identical point estimate for the difference between groups, with somewhat narrower confidence intervals. After controlling for Grade 6 ability, the disadvantage of being in the Reform cohort is 4.3%, or $-4.3/26$, or approximately -0.17 standard deviations. Controlling for

Table 7. Difference Between Groups on Algebraic Symbol Manipulation After Controlling for Grade 6 Test Scores

Treatment	<i>n</i>	Percent Correct for		95% Conf. Interval		98.3% Conf. Interval	
		Average Ability Student		Lower Bound	Upper Bound	Lower Bound	Upper Bound
Reform	324	55.9%					
Traditional	135	60.3%					
Difference	<i>t</i>	<i>df</i>	<i>p</i> -value	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-4.3%	- 2.09	453	.036	-8.5%%	-0.3%	-9.4%	+.6%

Grade 6 ability reduced the *p*-value of the difference between groups to .036. This value meets a .05 confidence level for hypothesis testing, but not the .017 confidence level required by the Bonferroni adjustment used in this study.

As before, alternate ways of handling missing data made little difference in the results. Of the 32 students for whom the response to Question 5 on the second day of the testing was imputed, 16 had Grade 6 scores available. Dropping those 16 students changed the estimated effect of being in the Reform group to 4.0%. with a *p*-value of

.063. Using a Rasch analysis to impute data for those 16 students, in addition to the scores for the 34 students who attended the first day of testing but not the second day, yielded an effect size of -0.16 standard deviations, $p = 0.038$.

Results from Test 3: Pairs of Students Solving an Extended Applied Algebra Problem

Test 3 was completed by students working in pairs on the second day of testing. Each pair was presented the following problem:

Many companies sell their products through long-distance telephone calls to customers. For example, the *CD Club* sells music compact discs all across the country from its headquarters in New York. Sales calls made by the *CD Club* last an average of about 4 minutes apiece. The club has bids from three possible providers of their long-distance telephone service:

- 1) *Apple Communications* will charge \$0.35 for placing each call and then \$0.15 per minute of time used in the call.
- 2) *Bell Telephone* makes a fixed charge of \$260 per week for access to its long-distance lines, but charges only \$0.10 per minute of time used by the calls.
- 3) *Capital Long Distance Services* will make a fixed charge of \$600 per week for unlimited use of its long-distance lines.

Question: For the *CD Club* the problem is to choose the long-distance service that is least expensive for their business. What advice would you give about the phone company to choose and why?

To solve the problem correctly, pairs of students had to:

1. realize that the “best” choice of provider would depend on how many calls the business made per week;
2. determine a correct algebraic equation for the weekly cost of using each provider, as a function of number of calls per week;
3. correctly solve the equations;
4. provide a clear and complete explanation to the business about which company it should choose, and under what conditions.

Pairs of students whose answer indicated they had correctly completed all four of these subtasks were assigned a score of “4”; pairs who completed three subtasks correctly received a “3”, pairs who completed two subtasks correctly received a “2”, and pairs who completed at least one subtask correctly received a “1”. The fully elaborated scoring rubric for Test 3 is contained in Appendix B.

To ensure that all test papers were comparable, only those submissions completed by two students were used in this analysis. Thus, responses that were submitted by either a single student or by a group of three students were dropped from the analysis. A total of 25 test papers were dropped for this reason.

This study analyzed 290 Test 3 papers. In the spring of 1997, pairs of students in the Traditional cohort completed 95 Test 3 papers. Pairs of students in the Reform cohorts completed 195 test papers: 99 submissions in the spring of 2000, and 96 submissions in the spring of 2001.

The study reported by Huntley, et al. (2000) used three forms for Part 3 and adapted a holistic scoring rubric that could be used across the three forms of the test with adjustments specific to each form. In contrast, the current study used only Part 3, Form A. It was thought that a rubric keyed more precisely to specific required elements in student responses to Form A would enable scorers to be more reliable in the codes they assigned. For this reason, the author in collaboration with teachers from Suburban High School developed a new rubric for scoring Test 3. Under the new rubric, it was somewhat more difficult to obtain a high score than under the old rubric. Because the scoring rubrics were different, it is not possible to directly compare scores on Test 3 at Suburban High School to scores obtained by students in previous studies.

Test 3 results without controlling for prior ability. Table 8 reports differences between the Traditional and Reform groups at Suburban High School, as determined by a Probit analysis. The Probit analysis assumes that pairs have an underlying normally

distributed achievement score, with the mean achievement score for all 290 test papers set to 0 and a standard deviation of 1. The table indicates that with achievement lower than .95 standard deviations below the mean, a pair would be expected to receive a score of 0; with achievement between -0.95 and 0.09 a pair would be expected to receive a score of 1; with achievement between 0.09 and 0.56 standard deviations above the mean

Table 8. Difference Between Groups on Test 3, an Applied Algebra Task

	Estimate	Stand. Error	χ^2 (1 df)	Sign.	95% Conf. Interval		98.3% Conf. Interval	
					Lower Bound	Upper Bound	Lower Bound	Upper Bound
Threshold								
Minimum for "1"	-0.95 sd	.097				-1.14	-0.76	-1.18
								-0.72
Minimum for "2"	0.09 sd	.086			-0.07	+0.26	-0.11	+0.30
Minimum for "3"	0.56 sd	.090				+0.38	+0.74	+0.35
								+0.78
Minimum for "4"	0.89 sd	.097				+0.70	+1.08	+0.66
								+1.12
Effect of Reform Group	0.28 sd	.13	4.3	.038	+0.02	+0.54	-0.04	+0.60

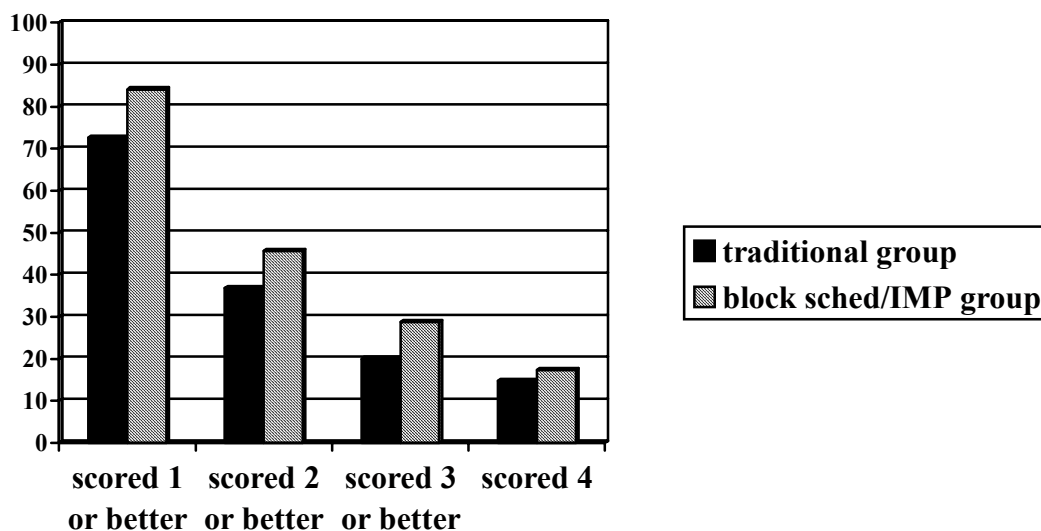
a pair would be expected to receive a score of 2; with achievement between 0.56 and 0.89 standard deviations above the mean a pair would be expected to receive a score of 3, and with achievement more than 0.89 standard deviations above the mean a pair would be expected to receive a 4.

On average, pairs in the Reform cohorts had an estimated achievement that was 0.28 standard deviations higher than that of students in the Traditional cohorts. Assuming that achievement is normally distributed, if a pair would ordinarily score at the 50th percentile, improving their score by 0.28 standard deviations would move the pair up to the 61st percentile.

The 0.28 standard deviation difference does not reach this study's standard for statistical significance ($p=.038$), although a 95% confidence interval indicates that the advantage of the Reform group is between +0.02 and +0.54 standard deviations. A Bonferroni-adjusted 98.3% confidence interval indicates that the advantage of the Reform group is probably between -0.04 and +0.60 standard deviations.

An alternate method of analyzing ordinal data, Logistic Regression Analysis, translates results into "the odds of getting a high score." Unless the distribution is very heavy-tailed, a Logistic Regression Analysis and a Probit analysis will obtain nearly identical significance levels. In this case, the Logistic analysis determined that being in a Reform cohort increased the odds of getting a high score by a factor of 1.61, significant at the .036 level. If the probability of an event is in the middle range, increasing the odds of success by a factor of 1.61 increases the event probability by about 11%. Figure 2 shows that odds of getting a high score were increased in about this way.

Figure 2. Percent scoring at each level on Test 3, an applied algebra task.



Test 3 results after controlling for prior ability. A pair's Ability score was

defined as the mean of the two Grade 6 Quantitative Ability scores for the two students who completed Test 3. Unfortunately, this created a problem. Of the 290 pairs who completed Test 3, there were 25 pairs who had no available Grade 6 Quantitative Ability score, and an additional 108 pairs who had available Grade 6 Quantitative Ability for one member of the pair, but not for the other. In order to use as much of the data as possible, an alternate definition of Ability was used. If both scores were available, Ability was defined as the mean of the pair's Grade 6 Quantitative Ability scores; otherwise, Ability was defined as the single Quantitative Ability score that was available. To ease interpretation, a new variable *zAbility* was created, defined as the z-score of Ability for the entire set of 265 Ability scores. That is, the mean Ability for all pairs was subtracted from each individual pair's Ability, and the result was divided by the standard deviation of Ability.

This alternate definition seemed acceptable for two reasons. First, because pairs were assigned within classrooms, their quantitative ability tended to be correlated. For the 157 pairs with both scores, the correlation was .354, $p < .00001$. Second, the variable *zAbility* was almost as good a predictor of Test 3 score for the 108 pairs with only one score available as it was for the 157 pairs with two mean scores available. For *zAbility* computed from one score, a Probit Analysis indicated that a 1 unit change in *zAbility*

predicted a 0.61 standard deviations change in achievement on Test 3. For zAbility computed from the mean of two scores, a Probit Analysis indicated that a 1 unit change in zAbility predicted a 0.71 standard deviations change in achievement on Test 3.

Because the pair zAbility estimate was more accurate for the 157 pairs with a Grade 6 Quantitative Ability for both students than for the 108 pairs with a Grade 6 Quantitative Ability score for only one student, the analysis of Test 3 while controlling for prior ability was performed twice: once with all 265 pairs with a zAbility estimate, and once for only the 157 pairs with the more accurate zAbility estimate. As will be seen below, results of the two analyses were surprisingly different. The difference may have important implications.

As noted above, zAbility could be estimated from at least one student score for 265 of the 290 pairs who completed Test 3. These included 84 pairs from the Traditional cohort, 92 pairs from the First Reform cohort, and 89 pairs from the Second Reform cohort.

For the set of 265 tests, there was no evidence of an interaction between prior zAbility and the effects of Reform versus Traditional schedule and curriculum. After controlling for Treatment and for zAbility, the interaction had a Wald Chi-square of 1.07 with 1 degree of freedom, $p=.301$. Nonetheless, Table 9 reports the full results of the Probit Analysis, including an interaction term. Table 9 is included for comparison to Table 11, below, which reports a significant Treatment-by-Ability interaction for the 157 pairs who had available a Grade 6 Quantitative Ability for both students.

Table 9. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability and Interaction (265 pairs for whom some ability estimate is available)

	Estimate	Stand. Error	χ^2 (1 df)	Sign.	95% Conf. Interval		98.3% Conf. Interval	
					Lower Bound	Upper Bound	Lower Bound	Upper Bound
Threshold								
Minimum for "1"	-1.28 sd	.116			-1.51	-1.05	-1.56	-1.00
Minimum for "2"	0.02 sd	.094			-0.17	+0.20	-0.21	+0.24
Minimum for "3"	0.60 sd	.099			+0.41	+0.80	+0.37	+0.84
Minimum for "4"	1.02 sd	.109			+0.80	+1.23	+0.76	+1.28
zAbility	0.79 sd	.133	35.47	<.0005	+0.53	+1.05	+0.47	+1.11
Effect of Reform Group	0.41 sd	.146	7.753	.005	+0.12	+0.69	+0.06	+0.76
Interaction	-0.16 sd	.154	1.07	.301	-0.46	+0.14	-0.53	+0.21

Because the interaction is not significant, Table 10 reports differences between the Traditional and Reform cohorts on Test 2 scores, after controlling for sixth-grade ability and with no interaction term. This analysis indicated that, after controlling for prior ability, pairs in the Reform cohorts on average scored 0.40 standard deviations higher than pairs in the Traditional cohort. Assuming the achievement of pairs is normally distributed, then if a pair of students would ordinarily score at the 50th percentile, improving their score by .40 standard deviations would move the pair up to the 66th percentile.

This 0.40 standard deviation difference was significant at the .006 level. A 95% confidence interval for the difference is (+0.11 sd, +0.68 sd.). A Bonferroni-adjusted 98.3% confidence interval for the difference is (+0.05 sd., +0.74 sd.).

Table 10. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability (265 pairs for whom some ability estimate is available)

	Estimate	Stand. Error	χ^2 (1 df)	Sign.	95% Conf. Interval		98.3% Conf. Interval	
					Lower Bound	Upper Bound	Lower Bound	Upper Bound
Threshold								
Minimum for "1"	-1.26 sd	.115			-1.48	-1.03	-1.56	-1.00
Minimum for "2"	0.04 sd	.094			-0.15	+0.22	-0.21	+0.24
Minimum for "3"	0.62 sd	.099			+0.42	+0.82	+0.37	+0.84
Minimum for "4"	1.04 sd	.109			+0.82	+1.25	+0.76	+1.28
zAbility	0.68 sd	.074	83.74	<.0005	+0.54	+0.83	+0.47	+1.11
Effect of Reform Group	0.40 sd	.145	7.53	.006	+0.11	+0.68	+0.05	+0.74

Eliminating two alternate explanations for Test 3 results. Two issues could raise concerns that Test 3 might not provide a fair comparison between students who had utilized IMP and a semestered block schedule and students who had utilized a traditional curriculum and schedule. First, it was possible that students who utilized the IMP curriculum had happened to practice a larger number of problems similar to the telephone company problem than had students who utilized the traditional curriculum. Since Test 3 contained only that one problem, this possibility if true would provide an alternate explanation of the results. Second, as noted in Chapter 3, the method of assigning students to pairs in the spring of 2000 differed from what had been the case in the spring of 1997. The way pairs were assigned could provide a second alternate explanation of the results. However, further analysis showed that both of these two alternate explanations were unlikely to be correct.

Test 3 required students to compare three phone companies, each differing in the fixed rate charged and the rate per minute, and decide which would be the best deal for a business. Problems of this type have been described as “typical” of IMP (Boaler, et al., 2002), raising the concern that perhaps students in the Reform cohorts had practiced a number of problems nearly identical to Test 3, whereas students in the Traditional cohort may not have had a similar opportunity. Such a situation might greatly reduce the validity of the comparison. However, a review of the IMP texts and of the traditional texts used at Suburban High School indicated that this was not the case. The current author examined the sections in the traditional algebra 1 text (Larson, et al., 1995) and algebra 2 text (Larson, et al., 2002) that addressed systems of linear equations, as well as the IMP modules addressing the same topic: *The Overland Trail*, *Solve It*, *Cookies*, and

Meadows or Malls. Although the IMP modules in general placed greater emphasis on word problems than did the traditional texts, problems similar to the “comparing costs” format of Test 3 did not occur more often in IMP. Neither the traditional algebra texts nor IMP modules reviewed contained any “comparing costs” problems as complex as the problem in Test 3. However, the traditional algebra 1 text did contain two worked examples and two student problems of the “comparing costs” type, using only two vendors. The IMP modules contained one problem of this two-vendor type, offered as an optional supplemental exercise to the unit *Meadows or Malls*, as well as one homework problem comparing the earnings of two students who got differing “starting bonuses” and differing earnings per hour, contained in the unit *Overland Trail*.

There was also some concern that the statistically significant difference on Test 3 between the Traditional cohort and Reform cohorts might be due to the fact that both students in a pair were always in the same mathematics class in 1997, but in 2000 pairs often consisted of students from different mathematics classes. However, as noted in Chapter 3 of this document the problem was largely corrected in 2001. For this reason, scores on Test 3 were reanalyzed, to see if the statistical significance was due to a particularly large effect in 2000.

The opposite turned out to be the case. That is, the statistical significance appeared to be largely due to a strong performance of pairs who completed the test in 2001, when students were requested to choose a partner who was enrolled in the same mathematics course or had most recently completed the same mathematics course as themselves. After controlling for prior ability, a probit analysis found that 2001 pairs outscored 1997 pairs by .44 standard deviations ($p=.008$). Meanwhile, after controlling

for prior ability, 2000 pairs outscored 1997 pairs by .37 standard deviations ($p=.027$).

Without controlling for prior ability, the results were as follows. Pairs completing Test 3 in 2001 outscored pairs completing the test in 1997 by .37 standard deviations ($p=.018$).

Pairs completing Test 3 in 2000 outscored pairs completing the test in 1997 by .19 standard deviations ($p=.224$). The strong showing by 2001 pairs makes it unlikely that the statistically significant advantage of the Reform cohorts on Test 3 was merely an artifact of the way students were assigned to pairs.

Test 3: Analysis of interaction for pairs with two Grade 6 ability estimates.

Table 11 reports results of a Probit Analysis for the 157 pairs of students who had Grade 6 Quantitative Ability scores available for both members of the pair. These 157 pairs included 44 pairs from the Traditional cohort, 54 pairs from the First Reform cohort, and 59 pairs from the Second Reform cohort.

Unlike Table 9, Table 11 shows a significant Treatment-by-Ability interaction. The next few paragraphs consider how being in the Reform cohorts affected student

Table 11. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability and Interaction (157 pairs for whom most accurate ability estimate is available)

	Estimate	Stand. Error	χ^2 (1 df)	Sign.	95% Conf. Interval		98.3% Conf. Interval	
					Lower Bound	Upper Bound	Lower Bound	Upper Bound
Threshold								
Minimum for "1"	-1.35 sd	.155			-1.65	-1.04	-1.72	-0.98
Minimum for "2"	0.00 sd	.118			-0.23	+0.23	-0.28	+0.29
Minimum for "3"	0.49 sd	.123			+0.25	+0.73	+0.19	+0.78
Minimum for "4"	0.93 sd	.135			+0.67	+1.20	+0.61	+1.26

zAbility	1.26 sd	.222	31.99	<.0005	+0.82	+1.70	+0.73	+1.79
Effect of Reform Group	0.62 sd	.209	8.84	.003	+0.21	+1.17	+0.12	+1.12
Interaction	-0.69 sd	.246	7.78	.005	-1.17	-0.20	-1.28	-0.10

