

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 1 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 1 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 1. Difference Between Groups on Understanding and Solving Algebra Problems in Context

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

As noted previously, results reported in Table 1 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 1. 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 2 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 2, 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 2 is the predicted score for a student whose ability in sixth grade was the average for this group of Suburban High School students.

Table 2 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 2. Difference Between Groups on Understanding and Solving Algebra Problems in Context After Controlling for Grade 6 Test Scores

				Percent Correct for Average Ability Student			
Treatment		<i>n</i>					
Reform		341		61.0%			
Traditional		136		55.7%			

Difference	<i>t</i>	<i>df</i>	<i>p</i> -value	95% Conf. Interval		98.3% Conf. Interval	
				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 3 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 3 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 3. Difference Between Groups on Algebraic Symbol Manipulation

Treatment <i>n</i>		Percent Correct		Standard Deviation	
Reform	443	53.8%		25.6%	
Traditional	199	58.2%		27.4%	

Difference	<i>t</i>	<i>df</i>	<i>p</i> -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 3 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 3.

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 4 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 4, 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 4 is the predicted score for a student whose ability in sixth grade was the average for this group of Suburban High School students.

Table 4 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 4. Difference Between Groups on Algebraic Symbol Manipulation After Controlling for Grade 6 Test Scores

Treatment	<i>n</i>	Percent Correct for					
		Average Ability Student					
Reform	324	55.9%					
Traditional	135	60.3%					
Difference	<i>t</i>	<i>df</i>	<i>p</i> -value	95% Conf. Interval		98.3% Conf. Interval	
				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 5 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 5. 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	Upper	Lower	Upper
					Bound	Bound	Bound	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

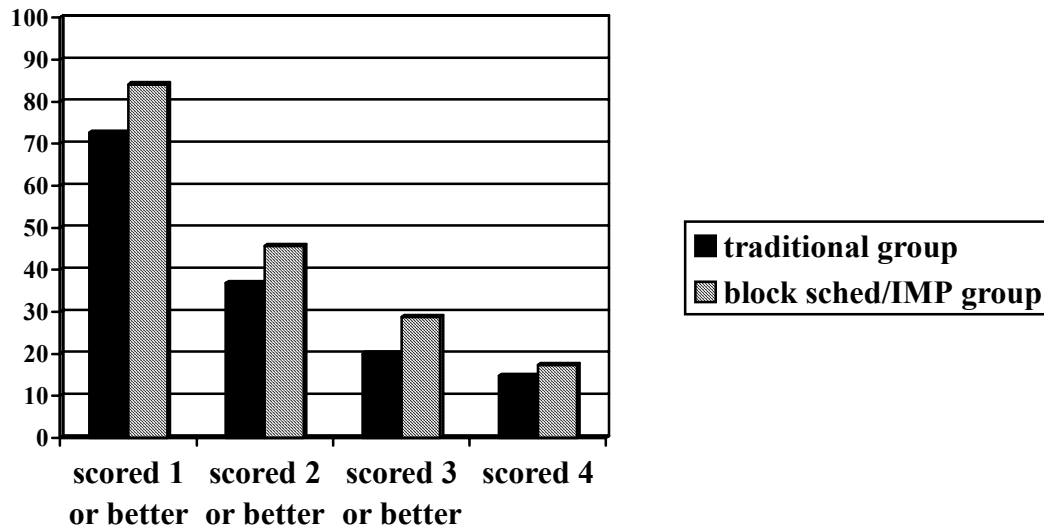
a pair would be expected to receive a score of 2; with achievement between 0.56 and 0.89 standard deviations above the mean a pair would be expected to receive a score of 3, and with achievement more than 0.89 standard deviations above the mean a pair would be expected to receive a 4.

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 1 shows that odds of getting a high score were increased in about this way.

Figure 1. 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 6 reports the full results of the Probit Analysis, including an interaction term. Table 6 is included for comparison to Table 8, 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 6. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability and Interaction (265 pairs for whom some ability estimate is available)

		Stand.	χ^2		95% Conf. Interval		98.3% Conf. Interval	
		Error	(1 df)	Sign.	Lower	Upper	Lower	Upper
	Estimate				Bound	Bound	Bound	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 7 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 7. Difference Between Groups on Applied Algebra Task, Accounting for Pair Ability (265 pairs for whom some ability estimate is available)

					<u>95% Conf. Interval</u>		<u>98.3% Conf. Interval</u>	
		Stand.	χ^2		Lower	Upper	Lower	Upper
	<u>Estimate</u>	<u>Error</u>	<u>(1 df)</u>	<u>Sign.</u>	<u>Bound</u>	<u>Bound</u>	<u>Bound</u>	<u>Bound</u>
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 6, Table 8 shows a significant Treatment-by-Ability interaction. The next few paragraphs consider how being in the Reform cohorts affected student

Table 8. 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 <i>df</i>)	Sign.	95% Conf. Interval		98.3% Conf. Interval	
					Lower	Upper	Lower	Upper
					Bound	Bound	Bound	Bound
					Bound	Bound	Bound	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

scores on Test 3 for different ability groups among the 157 pairs used to generate Table 11. However, this discussion should be treated with caution. The fact that the Treatment-by-Ability interaction detected in the analysis of these 157 tests was not confirmed in the analysis of all 265 pairs of students for whom some ability score was available raises the possibility that the interaction could be spurious.

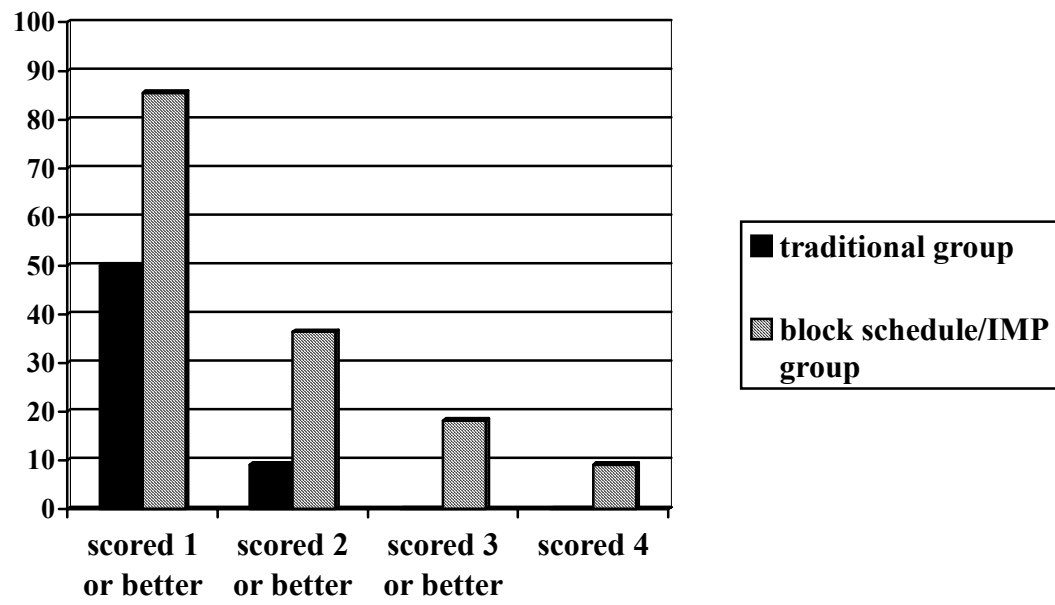
To analyze possibly differential effects of Treatment based on pair Ability, the 157 pairs with two Grade 6 Quantitative Ability scores were split into two groups. As noted above, these 157 pairs included 44 pairs in the Traditional group. Of the 44 Traditional group pairs, 22 had scores on zAbility below 0.13 and 22 had scores on zAbility above 0.13. Because there were a smaller number of students in the Traditional group than in the Reform, 0.13 was chosen as a cut-point to split the 157 tests into a “Lower Ability” group and a “Higher Ability” group.

The Lower Ability group contained 77 pairs. These included 22 pairs from the Traditional cohort, 26 pairs from the First Reform cohort, and 29 pairs from the Second Reform cohort. For these pairs, a Probit Analysis was performed using Test 3 score as a dependent variable and an indicator of Reform group vs. Traditional group as an independent variable. On average, Lower Ability pairs in the Reform cohorts had an estimated achievement that was 1.07 standard deviations higher than that of Lower Ability pairs in the Traditional cohort. This difference was significant (Wald Chi-square with 1 degree of freedom =13.01, $p=.0003$). Assuming that achievement is normally

distributed, then if a pair would ordinarily score at the 50th percentile, improving their score by 1.07 standard deviations would move the pair up to the 86th percentile. A 95% confidence interval indicates that for these low-ability pairs the advantage of the Reform group was between +0.49 and +1.65 standard deviations. A Bonferroni-adjusted 98.3% confidence interval indicates that the advantage of the Reform group was probably between +0.36 and +1.78 standard deviations.

Alternatively, a Logistic Regression Analysis determined that among these Lower Ability pairs being in a Reform cohort increased the odds of getting a high score by a factor of 5.99, significant at the .001 level. If the probability of an event is in the middle range, increasing the odds of success by a factor of 5.99 increases the event probability by about 36%. Figure 2 shows that for low-ability students, being in a Reform cohort increased the odds of getting a high score in about this way.

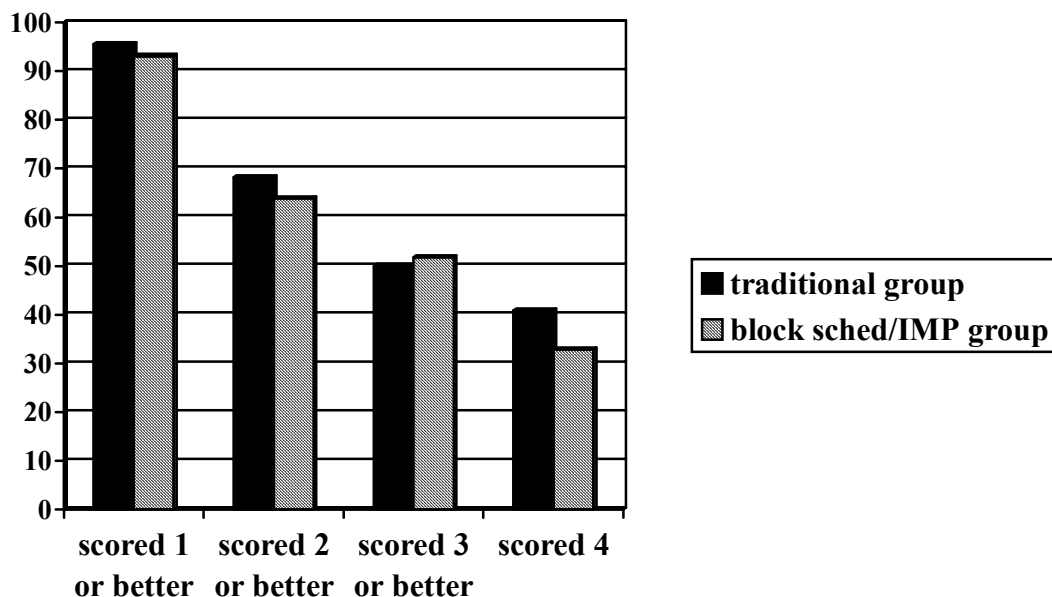
Figure 2. Low ability pairs: Percent scoring at each level on Test 3, applied algebra task.



The Higher Ability group contained 80 pairs. These included 22 pairs from the Traditional cohort, 28 pairs from the First Reform cohort, and 30 pairs from the Second Reform cohort. For these pairs, a Probit Analysis was performed using Test 3 score as a dependent variable and an indicator of Reform group vs. Traditional group as an independent variable. On average, Higher Ability pairs in the Reform cohorts had an estimated achievement that was 0.14 standard deviations lower than that of Higher Ability pairs in the Traditional cohort. This difference was not significant (Wald Chi-square with 1 degree of freedom = 0.273, $p > .5$). Alternatively, a Logistic Regression Analysis indicated that for these Higher Ability pairs membership in a Reform cohort decreased the odds of receiving a high score by a non-significant factor of 1.23 ($p > .5$). In the middle ranges, this would decrease the probability of getting a high score by about 5%. Figure 3 displays achievement on Test 3 for Higher Ability pairs in the Traditional cohort and Higher Ability pairs in the Reform cohorts.

Neither the statistical analysis nor Figure 3 showed any important difference between the Traditional and Reform cohorts in the achievement of pairs in the top 50% of prior ability on an applied algebra task. But could the interaction between Ability and Treatment indicate that in the Reform cohorts the achievement of pairs having the very highest ability suffered? Apparently not. A Probit analysis looked at the achievement of pairs in the top 25%: those with $z\text{Ability}$ greater than 0.80. This “Very High Ability” group contained 34 pairs. These included 11 pairs from the Traditional cohort, 9 pairs from the First Reform cohort, and 14 pairs from the Second Reform cohort. Results indicated that differences were non-significant ($p > .50$), but favored the Reform cohorts, indicating that on average Very High Ability pairs in the Reform cohorts had an

Figure 3. High ability pairs: Percent scoring at each level on Test 3, applied algebra task.



estimated achievement that was 0.16 standard deviations higher than that of Very High Ability pairs in the Traditional cohort.

Test 3: Opposite interaction pattern for pairs with one Grade 6 ability estimate.

As noted previously, no significant Ability-by-Treatment interaction was detected when analyzing the set of 265 tests with at least one Grade 6 Quantitative Ability score available. Given the strength of the interaction pattern for the 157 pairs for whom both students had Grade 6 Quantitative Ability scores, something very different appears to have been going on for the remaining 108 tests.

In fact, for these tests the pattern of interaction is non-significant, but appears to go in the opposite direction. Of the 108 pairs of students with only one Grade 6 Quantitative Ability score, 64 were Lower Ability, with a z_{Ability} less than 0.13. These included 23 pairs from the Traditional cohort, 25 pairs from the First Reform cohort, and 16 pairs from the Second Reform cohort. Among these students, a Probit Analysis indicated almost no difference between pairs from the Traditional cohort and pairs from the Reform cohorts, with a non-significant effect size of 0.01 standard deviations in favor of pairs in the Reform cohorts.

Of the 108 pairs of students with only one Grade 6 Quantitative Ability score, 44 had Higher Ability, with a z_{Ability} greater than 0.13. These included 17 pairs from the Traditional cohort, 13 pairs from the First Reform cohort, and 14 pairs from the Second Reform cohort. Among these students, a Probit Analysis indicated that pairs from the Reform cohorts outscored pairs from the Traditional cohort by 0.31 standard deviations. This difference was not statistically significant (Wald Chi-square=.878 with 1 degree of

freedom, $p=.349$), but the magnitude of the effect was greater than the 0.28 standard deviations reported for the full sample of 290 tests.

Test 3 interactions: A possible interpretation. Given the contrasting results on tests for a treatment-by-ability interaction for Test 3, depending on how “ability” was measured, it is quite possible that the interaction detected was spurious. The lack of any indication that there might be a treatment-by-ability interaction for Test 1 or Test 2 lends weight to the theory that the possible interaction for Test 3 was spurious.

If, however, the Test 3 interaction reflected a real pattern in student algebra learning, one possible explanation presents itself. Most often, students with Grade 6 Quantitative Ability scores were those who attended Suburban High School’s feeder middle school and used the *Visual Mathematics* curriculum throughout Grades 6, 7, and 8. Students lacking Grade 6 Quantitative Ability scores were those who had not done so. The very different interaction patterns for pairs with two Grade 6 Quantitative Ability scores, vs. pairs with only one Grade 6 Quantitative Ability score, raises the possibility of a synergistic effect between the *Visual Mathematics* curriculum and the IMP/block schedule treatment.

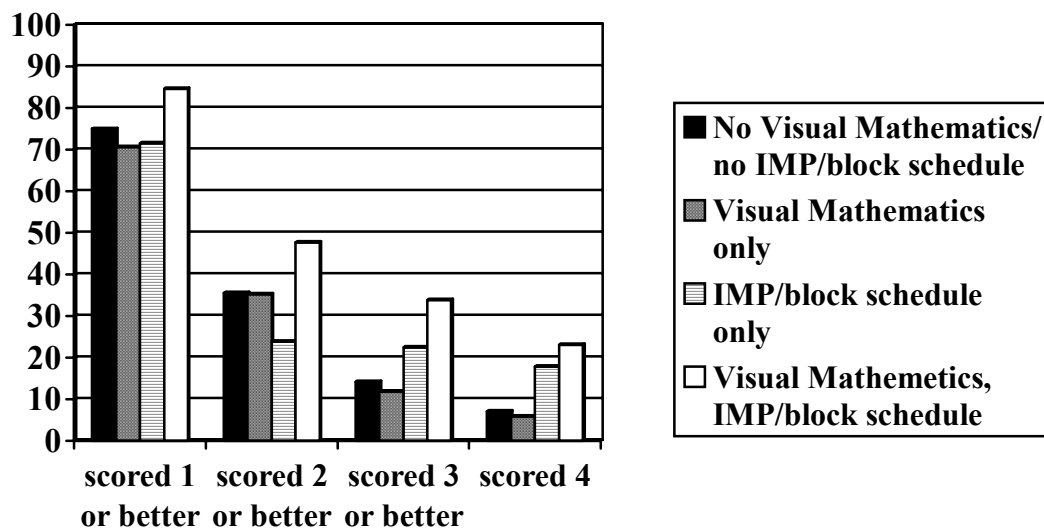
The data are consistent with the following explanation: Perhaps low-ability students were helped to solve problems comparable to the item in Test 3 if they used both the *Visual Mathematics* curriculum and the block schedule/IMP curriculum, but anything less was not sufficient. In contrast, perhaps high-ability students were helped to solve problems such as that presented in Test 3 if they used either the *Visual Mathematics* curriculum or the block schedule/IMP curriculum. Both experiences did not further enhance their achievement beyond that attributed to either curriculum.

The Suburban High School transcript files had Grade 6 transcripts available for students in the Traditional and First Reform cohorts, but not for students in the Second Reform cohort. For students with Grade 6 transcripts available, it was possible to obtain confirmatory information about whether they were enrolled in the middle school program that used *Visual Mathematics*. Too few pairs with sixth-grade ability estimates were available for this study to investigate how *Visual Mathematics* and the IMP/block schedule treatment interacted for high versus low ability groups, but it was possible to perform a tentative investigation of whether there was an interaction overall, ignoring student ability.

In Figure 4 pairs were coded as “No Visual Mathematics, No IMP/block schedule” if they were in the Traditional cohort and at least one member of the pair was lacking both a Grade 6 transcript and Grade 6 ability scores (28 pairs). Pairs were coded as “Visual Mathematics only” if they were in the Traditional cohort and both members of the pair had either a Grade 6 transcript or Grade 6 ability scores (67 pairs). Pairs were coded as “IMP/block schedule only” if they were in the First Reform cohort and at least one member was lacking both a Grade 6 transcript and Grade 6 ability scores (34 pairs). Pairs were coded as “Visual Mathematics, IMP/block schedule” if they were in the First Reform cohort and both members of the pair had either a Grade 6 transcript or Grade 6 ability scores (65 pairs). Figure 4 shows that pairs in which both members had used the *Visual Mathematics* curriculum, who furthermore were in the cohort using a block schedule and IMP, tended to outscore all other groups.

A Logistic Regression Analysis was run to see if this apparent difference could be dismissed as purely chance. Pairs who were in the First Reform *and* for whom both members had used the *Visual Mathematics* curriculum were compared to all other pairs in the Traditional or First Reform cohorts. Being in the “*Visual Math* plus IMP/block” group multiplied the odds of getting a high score by a factor of 1.91, significant at the .019 level. If the probability of an event is in the middle range, increasing the odds of success by a factor of 1.91 increases the event probability by about 16%.

Figure 4. Four groups: Percent scoring at each level on Test 3, applied algebra task.



The statistical significance level of 0.019 should be viewed as an indication that the pattern evident in Figure 4 cannot be dismissed as purely random. Nonetheless the reader should be cautious about giving the significance level too much weight. A 0.019 significance level does not meet the Bonferroni-adjusted 0.017 level used by this study. More importantly, the comparison of “*Visual Math* plus IMP/block schedule” to other groups was not among the pre-planned significance tests for this study and was made after looking at the data in order to explain an unusual interaction pattern. Any data set will have some patterns that occur by chance, and looking for patterns followed by statistical significance testing will often obtain spuriously “significant” results. Thus, the possibility of a synergistic effect between the *Visual Mathematics* curriculum and the IMP/block schedule must be viewed as an interesting possibility raised by this study, not as a confirmed hypothesis.

Specific Effects of Traditional Curriculum/Schedule

and IMP Curriculum/Block Schedule

This section provides details about the performance of the Traditional cohort and the Reform cohorts on specific questions or subsets of questions. The details illuminate the degree to which students in each group appear to have mastered specific skills underlying their scores on Test 1 and Test 2.

The discussion below generally reports “average score” on individual items. Because partial credit was given if, for example, a student made a minor arithmetic error but otherwise showed correct work, a student’s score was computed as that student’s score on the problem(s), divided by maximum possible score. See the scoring rubrics in Appendix B for details on how credit was assigned for each problem.

On individual 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 was computed from an ordinal Logistic Regression Analysis. This section also uses a *t*-statistic to compare groups on two subscales composed of multiple items. Altogether, 23 items and 2 subscales were investigated.

The results reported in this section are supplemental to the comparisons of overall scores on Test 1, Test 2, and Test 3 reported earlier in this chapter. A Bonferroni adjustment was used to keep an overall within-section alpha of .05, as insurance against putting too much weight on chance results. Since this section made 25 comparisons between the Traditional cohort and, a Bonferroni adjustment requires a significance level greater than .05/25, or approximately .002 to be relatively confident a difference was non-chance.

Evaluating expressions. Questions 3.1 and 3.2 on Test 1 required students to evaluate expressions presented in context, with access to graphing calculators. Question 7 on Test 2 required students to evaluate an expression presented without context or access to a graphing calculator. Figure 5 displays the three items.

Figure 5. Evaluating expressions.

The builders of a new American Eagle motorcycle plan to set a fixed price for the cycle—no rebates or bargaining allowed. The problem is finding the right price to charge. They estimate that their operating <i>costs</i> c and <i>sales income</i> I will be related to the number n of cycles that are made and sold: Operating costs: $c = 800n + 10,032,000$ Income from Sales: $I = 4500n - 0.2n^2$ Test 1, Question 3.1: What will the operating cost be if 5000 cycles are made and sold? Test 1, Question 3.2: What will the income from sales be if 2000 cycles are sold? Test 2, Question 7: If $s = 2$ and $t = -3$, then what is $2st + (s^2t)$?			
--	--	--	--

Table 12 shows that students in both the Traditional and the Reform cohorts did relatively well on all three problems, with little difference between groups.

Table 9. Scores on Evaluating Expressions

	Question 3.1	Question 3.2	Question 7
Traditional cohort	82.9%	80.1%	79.7%
	($n=196$)	($n=196$)	($n=199$)
Reform cohorts	85.4%	79.8%	81.9%
	($n=457$)	($n=457$)	($n=443$)
χ^2 Significance Level	.320	.806	.394

Simplifying expressions. Questions 1 and 6 on Test 2 required students to simplify expressions with integer coefficients presented without context or access to a graphing calculator. Figure 6 displays the question stems for these two multiple-choice items.

Figure 6. Simplifying expressions with integer coefficients.

Test 2, Question 1: Which of the following expressions is equivalent to $(4x^3)(5y)$? Test 2, Question 6: Which of the following expressions is equivalent to $-3(5x - 6)$?

Table 10 shows that on these problems students in the Traditional cohort tended to outperform students in the Reform cohorts. The advantage of Traditional cohort students on Question 1 was statistically significant.

Table 10. Scores on Simplifying Expressions with Integer Coefficients

	<i>n</i>	Question 1	Question 6
Traditional cohort	199	59.0%	61.3%
Reform cohorts	443	50.7%	57.6%
χ^2 Significance Level		.0004	.046

Questions 2 and 4 on Test 2 required students to simplify expressions with involving ratios, presented without context or access to a graphing calculator. Figure 7 displays the question stems for these two multiple-choice items.

Figure 7. Simplifying expressions involving ratios.

Test 2, Question 2: Which of the following expressions is equivalent to $\frac{125 + x}{25}$? Test 2, Question 4: If $t \neq 0$, which of the following expressions is equivalent to $\frac{8 - 5}{t - 2t}$?

Table 11 shows that there was essentially no difference between Traditional and Reform cohorts in students' ability to simplify expressions involving ratios. This skill was noticeably more difficult for students in both Traditional and Reform cohorts than was simplifying expressions with integer coefficients.

Table 11. Scores on Simplifying Fractional Expressions

	<i>n</i>	Question 2	Question 4
Traditional cohort	199	33.2%	24.5%
Reform cohorts	443	32.1%	28.4%
Chi-sq Significance Level		.696	.154

Solving linear equations. Three items on Test 2 required students to solve linear equations or inequalities without context or access to graphing calculators. Questions 8 and 9 presented linear equalities in an open-ended format, and Question 3 presented linear inequality in multiple-choice format. Figure 8 displays Questions 8 and 9, plus the question stem for Question 3.

Figure 8. Linear equations or inequalities without context or access to graphing calculators.

Test 2, Question 8: Solve the equation $3x + 4 = 5x - 2$ for x .	
	$-2x + 3y = 8$
Test 2, Question 9: Solve the system of equations	$x - y = 2$

Test 2, Question 3: Which of the following inequalities is equivalent to $4x - 7 > 3$?

Table 12 shows that students in both the Traditional and the Reform cohorts did quite well on the relatively easy Question 8, with essentially no difference between them in scores. Traditional cohort students were somewhat stronger than Reform cohort students at solving the system of linear equations in Question 9, but the difference was not statistically significant. On Question 3, students in the Traditional cohort outperformed students in the Reform cohorts, and the results were statistically significant.

Table 12. Scores on Solving Linear Equations, without Context or Access to Calculators

	<u>n</u>	<u>Question 8</u>	<u>Question 9</u>	<u>Question 3</u>
Traditional cohort	199	92.1%	58.6%	47.2%
Reform cohorts	443	92.7%	52.8%	38.8%
χ^2 Significance Level		.904	.152	.002

Two items on Test 1, Questions 2.3 and 2.4, required students to translate and solve linear equations presented in context, with a graphing calculator available. Figure 9 displays the two items.

Figure 9. Linear equations in context.

--

The Turtle Mountain Spring Company made plans about growth in their own share of the water business. They predicted that annual income from sale of their bottled water B and filters F would change over time according to the following formulas. Time t is in years since 1990 and income is in millions of dollars per year.

$$\text{Bottled Water Income: } B = 20 + 5t$$

$$\text{Filtering Devices Income: } F = 28 + 3t$$

Test 1, Question 2.3: When is the income from bottled water projected to reach \$50 million?

Test 1, Question 2.4: When does the Turtle Mountain Springs Company expect the two water products to give the same annual income?

Table 13 shows that students in both the Traditional and Reform cohorts performed reasonably well on these two items. Traditional cohort students performed somewhat better than did students in the Reform cohorts on Question 2.3 but the difference was not statistically significant. In contrast, Reform cohort students outperformed Traditional cohort students on question 2.4, and the advantage approached statistical significance.

Table 13. Scores on Solving Linear Equations in Context

	<i>n</i>	Question 2.3	Question 2.4
Traditional cohort	196	71.9%	60.5%
Reform cohorts	457	66.7%	71.2%
χ^2 Significance Level		.091	.003

The possible advantage of the Reform cohorts on Question 2.4 is consistent with results reported in previous research. Huntley, et al. (2000) found that students using Core-Plus Mathematics greatly outperformed Control students on this particular question.

It is unclear why students in the Reform cohorts appear to have been more advantaged on Question 2.4 than on Question 2.3. One possible explanation is that Question 2.4 can be solved by finding the intersection of two straight lines, a problem type that may be particularly amenable to solution using the graphing calculator skills emphasized by IMP and similar curricula.

Solving quadratic equations. Three items on Test 2, Questions 10, 11, and 13, required students to solve equations presented in a standard format, without context or access to graphing calculators. Figure 10 displays the three items.

Figure 10. Quadratic equations without context or access to a graphing calculator.

Question 10: Solve the equation $x^2 + 3x - 4 = 0$ for x. Question 11: Solve $(2x - 6)(x + 4) = 0$ for x. Question 13: Solve $\sqrt{9 + x^2} = 5$ for x.
--

Table 14 shows that students both in the Traditional cohort and in the Reform cohorts found these quadratic equations to be much more challenging than the linear

Table 14. Scores on Solving Quadratic Equations Presented without Context or Use of Graphing Calculators

	<i>n</i>	Question 10	Question 11	Question 13
Traditional cohort	199	39.7%	36.0%	21.8%
Reform cohorts	443	32.5%	33.1%	22.7%

χ^2 Significance Level	.143	.726	.747
-----------------------------	------	------	------

equations described earlier. Students in the Traditional cohort outscored students in the Reform cohorts on the Questions 10 and 11, which were presented in standard format. Students in the Reform cohort outscored students in the Traditional cohort on Question 13, which was presented in a non-standard format. None of the differences were statistically significant.

Two items on Test 1, Questions 3.3 and 3.4, required students to solve quadratic equations in a realistic context, with a graphing calculator available. The items are displayed in Figure 11.

Figure 11. Quadratic equations in context.

The builders of a new American Eagle motorcycle plan to set a fixed price for the cycle—no rebates or bargaining allowed. The problem is finding the right price to charge. They estimate that their operating <i>costs</i> c and <i>sales income</i> I will be related to the number n of cycles that are made and sold: Operating costs: $c = 800n + 10,032,000$ Income from Sales: $I = 4500n - 0.2n^2$ Test 1, Question 3.3: For what number of cycle sales will income be at least \$9,112,500? Test 1, Question 3.4: For what number of cycle sales will income be as large as possible, and what will that income be?

To answer Question 3.3 correctly, students needed to realize that the answer would not be a single number, but rather a range between the two roots of a quadratic equation. Nonetheless, they could receive partial credit by correctly identifying at least one of the roots. Rather than reporting mean scores for Question 3.3, the current study investigated the more interesting question of whether students understood that a single number was an insufficient answer. Table 18 reports the percentage of students answered the question completely (scoring 2 out of 2 points) or at least identified both roots (scoring 1.5 out of 2 points).

Question 3.4 required students to find the maximum of a quadratic function, and report both the value of this maximum (I) and the value of n at which the maximum was obtained. Students received partial credit (scoring 1 out of 2 points) if they correctly reported at least one of these two values. Rather than reporting mean scores for Question 3.4, Table 15 reports whether students were able to find the maximum, i.e., whether they scored 1 point or better.

Table 15. Percent of Students Showing a Good Understanding of Problems Involving Quadratic Equations Presented in Context, with Access to Graphing Calculators

	<i>n</i>	Percent who Identified 2 roots in Question 3.3	Percent who Identified Maximum in Question 3.4
Traditional cohort	196	7.1%%	18.4%
Reform cohorts	457	5.7%	24.5%

χ^2 Significance Level

.480

.102

Table 15 shows that only a very small percentage of students in either the Traditional or the Reform cohorts identified both roots in order to answer Question 3.3. Although a larger percentage of students found the maximum for Question 3.4 than identified both roots for Question 3.3, the numbers were still relatively small. Differences between cohorts on these two problems were not statistically significant.

Overall, the data in Table 14 and Table 15 indicate that students in both Traditional and Reform cohorts had substantial room for improvement in solving quadratic equations. However, this might be misleading. Differences in achievement between students in the Traditional cohort and students in the Reform cohorts might be due to a different mix of courses under the Reform and the Traditional program, or to different amounts learned in analogous courses under the Reform and the Traditional program. In order to assess differences due to both these sources, the Algebra Achievement test was administered to all eleventh graders, regardless of what coursework they had completed. Because of this, some students in each cohort had not completed coursework dealing with quadratic equations at the time they took the Algebra Achievement test. How did opportunity to learn affect students' ability to solve these quadratic equations? Table 16 presents achievement results for each of the five questions requiring students to solve quadratic equations, broken down by mathematics program status and by highest level algebra course completed prior to testing.

The number of students reported in Table 16 differs slightly from that in Table 14 and Table 15, because a few students whose transcripts were unavailable due to the fact that they could not be matched to a student id number were dropped from the data set.

Like Table 14 and Table 15, Table 16 reveals little difference in ability to solve quadratic equations between students in the Traditional and Reform cohorts, even when data is broken down by opportunity to learn.

Table 16. Achievement in Solving Quadratic Equations, Reported by Highest Level Algebra Course Student Enrolled in Through 11th Grade

	Mean	Mean	Mean	Pct who could i.d. 2 roots	Pct who found max
	Score on Quest. 10	Score on Quest. 11	Score on Quest. 13	Quest. 3.3	Quest. 3.4
Students who had not completed core curriculum.					
Traditional cohort	14.7%	7.8%	7.0%	0%	6.4%
	(n=43)	(n=43)	(n=43)	(n=47)	(n=47)
Reform cohorts	7.8%	9.1%	6.1%	0%	5.1%
	(n=153)	(n=153)	(n=153)	(n=158)	(n=158)

Non-honors students who had completed core curriculum (Alg 2 or Integrated Math 4)

Traditional cohort	19.2%	17.5%	9.2%	1.3%	2.6%
--------------------	-------	-------	------	------	------

	(n=80)	(n=80)	(n=80)	(n=76)	(n=76)
Reform cohorts	12.9%	11.7%	12.4%	1.4%	16.1%
	(n=140)	(n=140)	(n=140)	(n=143)	(n=143)

Honors students who had completed core curriculum (Alg 2 H or Integrated Math 4 H)

Traditional cohort	76.2%	73.3%	46.2%	19.1%	44.1%
	(n=70)	(n=70)	(n=70)	(n=68)	(n=68)
Reform cohorts	76.9%	78.5%	49.9%	15.5%	52.3%
	(n=147)	(n=147)	(n=147)	(n=155)	(n=155)

Students enrolled in programs other than Honors comprised roughly two thirds each of the Traditional and Reform cohorts. Even among Traditional cohort students who had completed Algebra 2 or Reform cohort students who had completed Integrated Math 4, only a small percentage could find the roots of a quadratic equation in either standard or non-standard format. Further, less than 10% of the non-Honors students could find two quadratic roots to solve the contextualized Question 3.3. Among non-Honors students, the only group with more than 10% finding the maximum for Question 3.4 consisted of those Reform Cohort students who had enrolled in Integrated Math 4 by the end of eleventh grade. The relatively poor performance of non-Honors students from both Traditional and Reform cohorts in solving quadratic equations, especially if the same pattern is observed at other schools, may be an area of concern for educators.

About three quarters of Honors students who had completed Algebra 2 topics (whether the IMP or the Traditional version) could find the roots of a quadratic equation presented in standard format. About half of these top students could solve a similar equation presented in non-standard format. About half of the Honors students could find the maximum value of a quadratic equation as required by problem 3.4. Less than 20% could apply their knowledge of quadratic equations to find two roots in the realistic context presented in Question 3.3.

The performance on Question 3.3 may alert educators to an area of concern. According to teachers at Suburban High School, they had not directly covered problems similar to Question 3.3. Nonetheless, it might be hoped that IMP's focus on flexible understanding of functions, combined with student practice in using graphing calculators to solve problems, might have enabled at least the Honors students to understand this

mathematics and solve this problem. As Table 19 shows, only 15% of IMP Honors students could do so, even fewer than the 19% of Algebra 2 Honors students in the Traditional cohort who could solve the problem.

Mathematics teachers at Suburban High School were able to shed some light about the reasons students performed as they did on Question 3.3. A common procedure for students using graphing calculators to solve equations of all types is to set one column of a table equal to one side of the equation, a second column of the table equal to the other side of the equation, and then to identify the position where the two columns have equal values. This procedure would lead to identification of one correct root for Question 3.3, but not a range of answers between two roots. In fact, 45.3% of all students in the Reform cohorts and 21.1% of all students in the Traditional cohort correctly identified one root. The percent reaching this particular partially correct answer was also high among honors students. It was produced by 52.6% of students in the Reform cohorts who had completed Integrated Math 4 Honors and 32.3% of students in the Traditional cohort who had completed Algebra 2 Honors.

The frequency with which students produced a point-value answer for problem 3.3, rather than a range between two values, may suggest that many were relying on the graphing calculator without paying sufficient attention to the underlying algebra needed to model the equation. It should be noted, however, that two characteristics of this particular test question may have inadvertently encouraged students to make this particular error. The first characteristic was an unfortunate aspect of the way Question 3.3 was worded. As can be seen in Figure 11, the question asked students to find “what number of cycles” solved the problem. The wording could perhaps have misled them

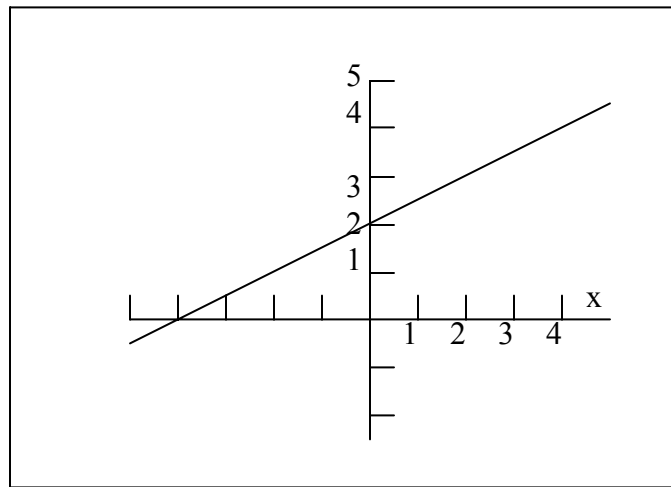
into looking for one specific number, rather than a range of values. Also, a close look at the numbers used in Question 3.3 indicates that solving the problem without use of a graphing calculator would require extensive arithmetic manipulations, accompanied by the possibility of making arithmetic errors. Thus, the arithmetic structure of the problem encouraged students to use the graphing calculator, and perhaps increased the frequency of responses based on an overly simplistic use of that particular tool.

Representational fluency: translating the graph of a line to an equation. Two of the problems on Test 2 required students to perform what traditionally is a well-practiced procedure in algebra 1: Given the graph of a line, write the equation for that line. Students were allowed to use a graphing calculator for one of these items, Question 1.2 from Part 1, Form C, but not for the other item, Question 14 from Part 2, Form A. The two items are displayed in Figure 12.

Figure 12. Translating the graph of a line into an equation.

Translating the Graph of a Line into an Equation

Question 1.2: Write an equation relating x and y that will produce this graph.



Question 14: What is an equation of the line shown in the graph below?

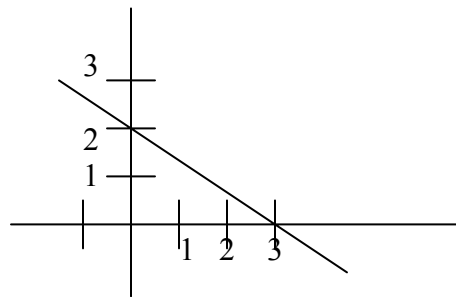


Table 17 shows that on these problems students in the Traditional cohort outperformed students in the Reform cohorts. The advantage of Traditional cohort students on Question 14 was statistically significant.

Table 17. Scores on Translating the Graph of a Line into an Equation

	Question 1.2	Question 14
Traditional cohort	63.0%	56.6%
	(n=196)	(n=199)
Reform cohorts	54.6%	39.5%
	(n=457)	(n=443)
χ^2 Significance Level	.025	<.000005

Discussions with mathematics teachers at Suburban High School about student performance on Question 1.2 and Question 14 led to a partial explanation of the Reform cohorts' weakness in this particular skill. According to the teachers, there is one module in each of the four IMP textbooks that extensively addresses the connection between linear data, linear equations, and the graph of a line. These modules are: *Overland Trail*, *Cookies*, *Small World, Isn't It?* and *The World of Functions*. However, in the *Overland Trail* and *Cookies* students most often translate from data or equations to the graph; procedures for translating from the graph of a line to its equation aren't emphasized until the module *Small World, Isn't It?* in the Year 3 textbook. *World of Functions* revisits this skill from another perspective. As reflected in Chapter 3, Table 2, during 2000-2001 Honors students did not complete *World of Functions* until they enrolled in Integrated Math 4. College Preparatory students did not complete either *Small World, Isn't It?* or

World of Functions until Integrated Math 4, and students taking Academic Assisted level courses did not complete these modules at any time.

This alone would limit opportunity to learn this concept for students in the Reform cohorts. But, in 1999-2000 the Integrated Math 4 syllabus was accommodating the fact that in previous years *Orchard Hideout* had not been completed in Integrated Math 3 College Preparatory. Therefore, in 1999-2000 *Orchard Hideout* was taught as part of Integrated Math 4 College Preparatory, and *World of Functions* was omitted.

Student achievement data lend some support to the theory that completing *Small World, Isn't It?* and/or *World of Functions* could have an effect on students' ability to translate the graph of a line to its equation. Table 21 reports the scores on Question 1.2 and Question 14 for students who at the time they were administered the Algebra Achievement test had completed neither *Small World, Isn't It?* nor *World of Functions*, for students who completed only *Small World, Isn't It?*, and students who completed both modules.

Table 21 provides some evidence that completing *Small World, Isn't It?* may have improved students' abilities to translate the graph of a line into its equation. It should be noted, however, that students who completed *Small World, Isn't It?* tended to be higher in mathematics interest and/or ability than those who did not.

Non-Honors students who completed only the module *Small World, Isn't It?* were similar in motivation and ability to non-Honors students who completed both *Small*

Table 18. Opportunity to Learn: Scores on Translating the Graph of a Line into an Equation for Reform Cohort Students who Completed Relevant Modules

	Question 1.2	Question 14
Completed neither	19.0% (n=146)	6.8% (n=143)
Completed <i>Small World, Isn't It?</i>	59.2% (n=65)	32.2% (n=61)
Completed both <i>Small World, Isn't It?</i> and <i>World of Functions</i>	61.1% (n=61)	45.0% (n=60)
Completed both <i>Small World, Isn't It?</i> and <i>World of Functions</i> (Honors Students)	86.0% (n=152)	74.1% (n=144)

World, Isn't It? and *World of Functions*. The two groups enrolled in the same courses,

but the syllabus differed, depending on when the student enrolled. Thus, Table 21 provides reasonably good evidence that completing *World of Functions* improved students' ability to solve Part 2, Question 14, which required students to translate the graph of a line into an equation without the help of a graphing calculator, but did not improve students' ability to solve Part 1, Question 1.2, which tested the same skill, but permitted use of a graphing calculator.

Nonetheless, even with full opportunity to learn, students in the Reform cohorts under-performed students in the Traditional cohort on this skill. Traditional cohort students who completed Algebra 2 Honors by the time of testing are an appropriate comparison group for Reform cohort students who completed the Honors version of both *Small World, Isn't It?* and *World of Functions*. Each comprises the top 35% of their respective cohort. On Question 1.2, Traditional Honors students scored approximately the same (86.9%) as Honors students in the Reform cohorts (86.0%), but on Question 14, Traditional Honors students scored better (89.6% vs. 74.1%).

Traditional cohort students who completed Algebra 2 by the time of testing are a fairly appropriate comparison group for students in the Reform cohorts who completed

either *Small World, Isn't It?* alone or both *Small World, Isn't It?* and *World of Functions*.

The Traditional students in this group comprise the middle 40% of the Traditional cohort, and the Reform students comprise the middle 29% of the Reform cohorts. On both questions requiring students to translate a linear graph into an equation, Traditional cohort Algebra 2 students outscored non-Honors students in the Reform cohorts, even when the Reform cohort students had completed both modules. On Question 1.2, Traditional cohort students who had completed algebra 2 scored 74.3%, versus 61.1 % for Reform cohort students who had completed both modules. On Question 14, Traditional cohort students who had completed algebra 2 scored 56.1%, versus 45.0% for Reform cohort students who had completed both modules.

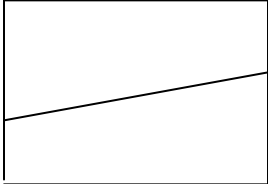
Representational fluency: Non-procedural tasks. Two additional items on the Algebra Assessment were designed to measure students' ability to translate information presented in a graphical format: Test 1, Questions 2.1 and 2.2. These items are displayed in Figure 13.

Figure 13. Translating graphs (non-procedural).

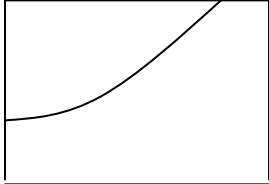
Recent reports about bacteria and parasites in city drinking water supplies have led to a booming business in sales of bottled water and water purifying devices. Water has become a big business.

The following graphs show three different predictions of how income from water sales might change over the next several years.

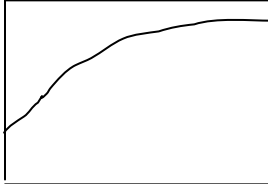
I



II



III



Test 1, Question 2.1: Which of the three graphs (I, II, or III) shows annual income rising:

at an increasing rate? _____

at a constant rate? _____

at a decreasing rate? _____

Test 1, Question 2.2: Income from sales of bottled water in 1990 was \$500 million. What is the best match between the three graphs and these equations relating sales income S in millions of dollars to time t in years since 1990?

$S = 500 + 60t$ _____

$S = -15t^2 + 290t + 500$ _____

$S = 500(1.2)^t$ _____

Table 22 shows that on these less procedural translation problems students in the Reform cohorts outperformed students in the Traditional cohorts. The advantage of Reform cohort students on Question 2.2 was statistically significant.

Table 19. Scores on Interpreting and Translating Line Graphs

	n	Question 2.1	Question 2.2
Traditional cohort	191	89.5%	59.7%
Reform cohorts	456	94.7%	73.8%
χ^2 Significance Level		.010	.00004

Question 1.4 on Test 1 required students to identify an exponential function presented in tabular form, and translate it into an equation. Figure 14 displays the question.

Figure 14. Exponential function in tabular form.

Test 1, Question 1.4: Write an equation relating x and y that will give the pairs of numbers in this table.					
x	-1	0	1	2	3
y	0.5	1	2	4	8

Table 20 shows that while this item was difficult for students in all cohorts, students in the Reform cohorts were about twice as likely to get the item correct as were students in the Traditional cohort. The advantage of Reform cohort students on Question 1.4 was statistically significant.

Table 20. Scores on Translating an Exponential Function

	<u><i>n</i></u>	<u>Question 1.4</u>
Traditional cohort	191	14.7%
Reform cohorts	456	28.3%
χ^2 Significance Level		.0001

Although students work with exponential functions in many IMP modules, the topic receives particular emphasis in three modules: *All About Alice*, *Small World, Isn't It*, *World of Functions*. Table 24 shows that students in the Reform cohorts who had completed a larger number of these modules were particularly advantaged in solving Question 1.4. The importance of opportunity to learn in determining student achievement is emphasized by the fact that students who completed only the modules *All About Alice* and *Small World Isn't It?* were those students enrolled in Integrated Math 4 College Preparatory before the 2000-01 school year, while those non-Honors students who completed the same two modules plus *World of Functions* were those students enrolled in the same Integrated Math 4 College Preparatory class during the 2000-01 school year, when *World of Functions* was added to the course syllabus. Although students in the two groups were of comparable ability and motivation, Table 24 shows that students who completed *World of Functions* scored much higher on Question 1.4 than did students who did not.

Table 21. Ability to Translate an Exponential Function to an Equation

	<u><i>n</i></u>	<u>Mean Score on Quest 1.4</u>
Traditional Cohort		
<u>Highest Enrolled Course</u>		
< Algebra 2	47	2.1%
Algebra 2	76	2.6%
Algebra 2 Honors	68	36.8%
Reform Cohorts		
<u>Relevant Modules Completed</u>		
none	57	3.5%
<i>All About Alice</i>	105	4.3%
<i>All About Alice</i> and <i>Small World, Isn't It?</i>	70	11.4%
<i>All About Alice</i> and <i>Small World, Isn't It?</i> and <i>World of Functions</i>	69	31.9%
<i>All About Alice</i> and <i>Small World, Isn't It?</i> and <i>World of Functions</i> (Honors Students)	155	59.7%

Formulation and interpretation of mathematical models. As described by Huntley, et al. (2000), many items on Test 1 asked students either to formulate an algebraic model or to interpret results from use of a given algebraic model. The present study used a total of 10 items that fit into this category.

Five of these items were combined to create a “Mathematical Formulation” subscale: Questions 1.1, 1.3, 1.5a, and 1.5b from Part 1, Form C of the Algebra Achievement test, and Question 5 from Part 2, Form A. The Mathematical Formulation subscale had a reliability (Cronbach’s alpha) of 70.4%.

Figure 15 displays Question 1.1. The remaining items are displayed in Appendix A.

Figure 15. Formulating an algebraic model.

Part 1, Question 1.1: The Watchdog Security Service provides home and business security systems with installation charge of \$150 and a charge of \$5 per week to monitor the system. Write a formula that gives the cost C of the service as a function of time t in weeks that the service is needed.

Table 22 reports differences between the Traditional and Reform groups on percent correct of the Mathematical Formulation subscale. As the table indicates, students in the Reform cohorts averaged 58.0% correct on the Mathematical Formulation subscale, significantly better than the 49.3% achieved by students in the Traditional cohort ($p<.0003$).

Table 22. *Difference Between Groups on Mathematical Formulation*

Treatment		<i>n</i>	Percent Correct		Standard Deviation	
Reform		419	58.0%		28.5%	
Traditional		191	49.3%		26.3%	

Difference of Scores				95% Conf. Interval		99.8% Conf. Interval	
<i>t</i>	<i>df</i>	<i>p</i> -value		Lower Bound	Upper Bound	Lower Bound	Upper Bound
8.8%	3.61	608	.0003	4.0%	13.5%	1.2%	16.3%

The standard deviation for Mathematical Formulation scores in the Traditional cohort was 26.3%. Reform group students scored 8.8% better than did Traditional group students. This is an effect size of 8.8/26.3, or approximately .33 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 .33 standard deviations would move the student up to the 63rd percentile.

Each of the five questions comprising Problem 4 on Test 1 called for students to interpret the meaning of algebraic calculations. For the present study, these five questions were combined into a single “Interpreting Algebraic Models” subscale. The Interpreting Algebraic Models subscale had an estimated reliability (Cronbach’s alpha) of 88.7%. Note, however, that because the five items making up the subscale are all part of the same super-problem, it is possible that this reliability estimate is somewhat inflated due to non-independence among student responses.

Figure 16 displays the first three questions in Problem 4. The remaining two questions are displayed in Appendix A.

Figure 16. Interpreting an algebraic model.

Several commercial airlines have non-stop flights from Los Angeles, California all the way to Sydney, Australia. It is a trip of 7500 miles and can take as long as 18 hours, most of the time out over the Pacific Ocean. Therefore, estimating flight time and fuel requirements is very important. One airline uses this formula to calculate flight time T in hours from expected wind speed W in miles per hour:

$$T = \frac{7500}{500 + W}$$

Test 1, Question 4.1: What will you learn from calculating T when $W=75$?

Test 1, Question 4.2: What will you learn from solving: $\frac{7500}{500 + W} = 14$?

Test 1, Question 4.3: What will you learn from solving: $\frac{7500}{500 + W} > 14$?

Table 23 reports differences between the Traditional and Reform groups on percent correct of the Interpreting Algebraic Models subscale. As the table indicates, students in the Reform cohorts averaged 49.9% correct, compared to 44.5% for students in the Traditional cohort. The advantage of Reform cohort students was not statistically significant ($p < .065$).

Table 23. Difference Between Groups on Interpreting Algebraic Models

Treatment		n	Percent Correct	Standard Deviation	
Reform		457	49.9%	35.0%	
Traditional		196	44.5%	32.7%	

Difference of Scores				95% Conf. Interval		99.8% Conf. Interval	
t	df	p -value		Lower Bound	Upper Bound	Lower Bound	Upper Bound
5.4%	1.85	651	.065	-2.9%	+11.2%	-3.7%	+14.3%

The standard deviation for Interpreting Algebraic Models scores in the Traditional cohort was 32.7%. Reform group students scored 5.4% better than did Traditional group students. This is an effect size of $5.4/32.7$, or approximately .17 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 .17 standard deviations would move the student up to the 57th percentile.

Summary of detailed differences between groups. Underlying the difference between the Reform cohorts and the Traditional cohort on Test 1 and Test 2 were apparent patterns of differences in specific sub-skills. Although the advantage of students in the Traditional cohort did not pass the Bonferroni-adjusted statistical significance test used in this study, they were observed to outscore Reform students on certain sub-skills within Test 2. When there was a difference between groups on questions involving specific, well-practiced procedures, students in the Traditional cohort tended to do better. Students in the Traditional cohort outperformed students in the Reform cohorts at the specific skill of writing an equation for a line, when presented a graph of that line—especially when this skill was performed without access to a graphing calculator. They also outscored the Reform cohorts on simplifying expressions, and on solving linear and quadratic equations when presented symbolically in standard form. However, they had no advantage in solving expressions involving ratios, or in solving quadratic equations when presented symbolically in a non-standard form, or in solving linear or quadratic equations when presented as part of a realistic problem context. Underlying their significantly higher achievement on Test 1, students in the Reform cohorts were noticeably better than students in the Traditional cohort at several particular sub-skills. They were consistently better at formulating a mathematical model to describe a problem situation. They also tended to be better at interpreting algebraic models, although the results were statistically significant. Finally, they were better at all three non-procedural items that measured their ability to translate among differing representations of functions. They were better at interpreting graphs on a Cartesian plane, and at estimating what type of equation might fit the graph. They were also better at recognizing an exponential function presented in tabular form and writing an equation for that function.

Effects of Measuring Achievement the end of Grade 11

Instead of at Other Possible Measurement Times

To put the results of the Algebra Achievement test in context, two additional analyses were undertaken investigating how the timing of the assessment might have affected those results. The first analysis compared scores on Test 1 and Test 2 of students in Reform cohorts who were enrolled in a mathematics course at the time of testing to scores on Test 1 and Test 2 of students in Reform cohorts who were not enrolled in any mathematics course during the semester in which they completed the Algebra Achievement test. The second analysis used student transcripts to see whether the Algebra Achievement test was likely to have captured most of the algebra learning students completed in high school, or whether in contrast students tended to study a significant amount of algebra in twelfth grade, after they had completed the test.

Possible underestimate of algebra achievement for those students not currently enrolled in mathematics. The current study measured student algebra achievement at the end of eleventh grade. The soonest students were likely to apply this knowledge was at the beginning of twelfth grade, at the end of a 3-month summer recess. For this reason, it is possible that a delayed test, administered at the very beginning of twelfth grade, would have been a better way to measure student algebra achievement. That way, all students would have experienced a summer recess during which they may have forgotten some of the mathematics they had learned. If much of what is forgotten is forgotten quickly, then those students in the Reform cohorts who were not enrolled in mathematics during the spring semester would likely already have forgotten much of what they were going to forget by the end of eleventh grade. Thus, students who had been enrolled in spring courses would have forgotten relatively more over the summer.

Data available from the current study provided one way to check whether the test was likely to have been biased against students not currently enrolled in mathematics. After controlling for prior ability, the achievement of Reform Cohort students enrolled in mathematics class at the time of testing was compared to that of Reform Cohort students who were not enrolled in mathematics at the time of testing on Test 1 (Understanding and Solving Algebra Problems presented in Context and Test 2 (Algebra Symbol Manipulation). The comparison indicated that on Test 2 there may have been such a bias.

On Test 1, Reform cohort students who were enrolled in mathematics class at the time of testing scored similarly to those not enrolled in mathematics class at the time of testing: 58.2% correct for those currently enrolled, compared to 58.3% for those not currently enrolled. On Test 2, Reform cohort students who were enrolled in mathematics class at the time of testing scored slightly higher than those not enrolled in mathematics class at the time of testing: 54.2% correct for those currently enrolled, compared to 52.0% for those not currently enrolled. However, for reasons not explained by data available in this study, the students enrolled in mathematics during the second semester of eleventh grade tended to have lower sixth-grade ability scores than students who were

not enrolled in mathematics during the second semester of eleventh grade. Therefore, after controlling for prior ability, Reform cohort students currently enrolled in mathematics class scored significantly higher on Test 2 than did Reform cohort students not currently enrolled in mathematics class ($p = .024$). On Test 1, after controlling for prior ability Reform students enrolled in mathematics class scored higher than did Reform students not currently enrolled in mathematics class, but the result was not statistically significant ($p = .335$).

A total of 28.5% of students in the Reform cohorts who completed the Algebra Achievement test were not enrolled in mathematics class at the time of testing. As administered, the test may have underestimated the achievement of these students, as compared to the achievement of other students completing the test.

Unmeasured effects of twelfth-grade coursework. Students at Suburban High School completed the Algebra Achievement test for this study at the end of Grade 11. Gathering student achievement data from the end of Grade 12 might have provided a better measure of the complete effects of the changes to the schedule and curriculum at Suburban High School. Unfortunately, it was not believed possible to obtain a valid measure of achievement at the end of Grade 12. The Algebra Achievement test was intentionally administered in such a way as to keep it a low stakes test in order to avoid creating any reason for faculty to contaminate the results by teaching to the test. Most student tests were not even scored nor results reported until several years after they were administered. However, while it seemed reasonable to expect valid results from a low-stakes test administered to students near the end of eleventh grade, it was deemed unlikely that participation and effort would be high quality if the test were administered near the end of twelfth grade, just before high school graduation. It was anticipated that this would be particularly problematic under the block schedule, since a large number of twelfth graders would not even be enrolled in mathematics class at the time the test was administered.

Student transcripts were analyzed to determine what percent of students were likely to have studied a significant amount of algebra in twelfth grade, subsequent to administration of the Algebra Achievement test. Because students in the Traditional cohort who were tested in the spring of 1997 did not utilize a traditional schedule in twelfth grade, their transcripts were not used in this analysis. Instead, transcripts from the two preceding cohorts, who graduated, respectively, in the spring of 1997 and the spring of 1998, were analyzed. Also, to analyze twelfth grade algebra course-taking for students using the semestered block schedule and IMP curriculum only transcripts from the First Reform cohort were utilized, because students in the Second Reform cohort had not completed twelfth grade at the time this investigation was conducted.

Under the traditional schedule and curriculum, some students studied Algebra 2 for the first time in twelfth grade, and others revisited the topic in a course called Algebra 3/Trigonometry that utilized the UCSMP text *Advanced Algebra* (Senk & Thompson, 1993). *Advanced Algebra* uses an alternative approach to traditional texts that past research has shown to be very successful (Thompson & Senk, 2001). During the 1995-96 school year, 36.4% of all twelfth graders studied algebra in one of the other of these two ways. During the 1996-97 school year, 34.3% of all twelfth graders studied algebra in one of the other of these two ways.

Of all students in the First Reform cohort who completed the Algebra Achievement test in eleventh grade, 32.5% subsequently enrolled in an Integrated Mathematics class during twelfth grade. Of the 27 students in the First Reform cohort who were administered the Algebra Achievement test and who eventually completed Integrated Math 4 Academic Assisted, 22 (81.5%) did so in twelfth grade. Of the 92 students in the First Reform cohort who were administered the Algebra Achievement test and who eventually completed Integrated Math 4 Academic Assisted, 41 (44.6%) did so in twelfth grade.

In summary, under both the Traditional and the Reform program, about one-third of all students at Suburban High School studied some form of algebra course during twelfth grade. The current study did not capture the effects of whatever algebra they learned at that time. It is particularly regrettable that most of the students in the Reform cohorts who were in Academic Assisted classes had not yet completed the core IMP sequence at the time of testing. Under the new program, Academic Assisted students were offered exposure to more sophisticated mathematics than had previously been the case, and it would have been desirable to measure the full effects of the program on this particular group.