*Data Analysis: Eleventh Grade Algebra Tests*

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

*Covariate: Sixth Grade Test Scores*

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

While no scale-score equating was done, the Educational Records Bureau did rank student scale scores based on a national norm, with the norm recomputed yearly. As shown in Table 1, 1993 test scores in Suburban High School's district dropped in almost

all grades, with a precipitous drop in some grades. The school district's *Testing Report* for 1993 notes that the drop may have been caused by the change from a test that had

*Table 1. Yearly Median National Percentile Rank on Educational Records Bureau CTP Quantitative Ability Test*

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

* Data not available for this study

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

CTP III scores from the eighth grade, either the spring of 1994 for students in the Traditional cohort or the spring of 1997 and the spring of 1998 for the Reform cohorts, would be a good candidate to use as a control variable, even though there is some indication from median percentile ranks reported in Table 1 that scores in 1994 might still have been lower than in other years when students had more experience with the particular type of test being used. Unfortunately, a number of student tests taken in 1998 were destroyed by a burst pipe before they could be graded. The destroyed tests included the eighth-grade tests taken by students in the Second Reform cohort.

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

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

Although not reported in Chapter 4, the analyses of Test 3 were run using the alternate choice for covariate of the maximum quantitative ability score from the pair of students completing the test. The results of the alternate analysis were nearly identical to those reported in Chapter 4.

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

Although sixth-grade scores were equated for this study by using national percentile rank, an unconverted percentile rank is not the best variable to use. In particular, analysis of Part 3, which was taken by students working in pairs, required computing the mean ability of a pair of students. Percentile rank is not an interval scale: for example, it takes a greater increase in ability to move from the $90^{th}$ to the $95^{th}$ percentile than it does to move from the $50^{th}$ to the $55^{th}$ percentile. For this reason, sixth-grade scores were converted to z-scores before being used as a covariate. A z-score is the number of standard deviations a particular score is above or below the mean score;

assuming ability is normally distributed, every percentile rank can be translated to a particular z-score. The z-scores were created in two steps: first, each national percentile rank was converted to a "national z-score", defined as the z-score that would achieve that rank, assuming normal data. Then, in order to center the mean at zero for the Suburban High School data set, the "national z-scores" were reconverted to "Suburban High School z-scores" by subtracting the mean "national z-score" for all students used in this analysis, and dividing by the standard deviation.

*Dependent Variables*

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

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

Subscales used for this study are necessarily somewhat different from those used by the Core-Plus authors, because the Core-Plus subscales used items from several forms for each part of the test, whereas this study utilized only one form for each of the three parts of the test. Also, the pilot study indicated that one particular item on Part 1, Form C fit better on the subscale composed of Part 2 items than it did on the subscale composed of other Part 1 items, and one item on Part 2, Form A fit better on the subscale composed

of Part 1 items than it did on the subscale composed of other Part 2 items.

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

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

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

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

Variable 2: Achievement on algebra symbol manipulation without context, as measured by all items on Part 2, Form A except item 5, plus item 1.2 from Part 1, Form C (hereafter referred to as Test 2). ; and

Variable 3: Cooperative solution to an extended open-ended algebra problem, as

measured by scores on the single extensive item in Part 3, Form A (hereafter referred to as Test 3).

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

i. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to solve algebraic symbol manipulation problems? Do the results of this comparison differ depending on students' prior ability?

ii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to interpret and solve challenging algebra problems presented in context? Do the results of this comparison differ depending on students' prior ability?

iii. How do students enrolled in a reform-based curriculum and a semestered block schedule compare to students enrolled in a traditional curriculum and traditional schedule in their ability to collaboratively solve and communicate their solution to a complex open-ended algebra problem? Do the results of this comparison depending on students' prior ability?

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

*Statistical Methodology: Rules for Establishing Confidence Intervals*

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

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

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

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

*Statistical Models Used*

The first two dependent variables (Algebra Problems in Context and Symbol

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

The third dependent variable (Cooperative Solution to an Extended Open-Ended Algebra Problem) uses an ordinal scale of student scores, taking on possible values of 0 to 4. Analysis for this variable was identical to that utilized for the first two, except that Ordinal Regression was used instead of Ordinary Least Squares Regression. There are two varieties of Ordinal Regression that are commonly used with such data, Probit Regression and Logistic Regression. Unless the student ability distribution is very unusual, both types of regression will provide nearly identical results in terms of p-values, but the interpretation is slightly different. Each highlights a different and important aspect of the data. Probit analysis provides an effect size that can readily be compared to output from the analyses of Test 1 and Test 2. Logistic regression analysis provides an odds ratio that is more easily related to student responses that were actually

observed. In the interest of clearly explaining the observed results, both types of analysis are reported below.

*Supplemental Analyses of Specific Skills*

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

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

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

Differences between the Reform cohorts and the Traditional cohort on these two subscales were investigated using an independent-samples *t*-test. Items from Test 1 and Test 2 that were not on the specific subscales were examined individually. For dichotomous individual items, that is, items that were scored as right/wrong, a Pearson Chi-square statistic was computed from a cross-tabulation table. On items for which students could receive partial credit, a Wald Chi-square statistic was computed from an ordinal Logistic Regression Analysis. The Logistic Regression Analysis is a generalization of the cross-tabulation method used to examine dichotomous items. Altogether, the supplemental analysis of specific skills compared the Reform cohorts to the Traditional cohort on 25 measures. To guard against over-interpreting results that occurred by chance, the Reform cohorts were deemed to be different from the Traditional cohort on one of the twenty-four measures if the statistical significance level for that measure reached a Bonferroni-adjusted .05/25 = .002 level.