

**Effectiveness of Stanford's EPGY Online Math K-5 Course
in Eight Title I Elementary Schools
in Three California School Districts, 2006-2007**

Patrick Suppes*, Paul W. Holland, Yuanan Hu*, and Minh-thien Vu***

15 July 2009

***Education Program for Gifted Youth, Stanford University**

****Paul Holland Consulting Corporation**

Abstract

Stanford University's Education Program for Gifted Youth (EPGY) conducted a randomized treatment experiment during the 2006-2007 school year to test the efficacy, for Title I students, of the EPGY Kindergarten through Grade 5 Mathematics Course Sequence (Math K-5). While the EPGY curriculum was originally developed for gifted students in any grade, the students in this study were not selected as gifted but were simply the students in the Title I schools. If we restrict attention to those EPGY students who were in the top half of the distribution of correct first attempts on the EPGY exercises (a measure of work and engagement in the EPGY curriculum) we see substantial and statistically significant improvements in the CST07 test scores over the scores of matched control students. The effects in second grade appear to be larger than those in grades 3 to 5.

Key words: Randomized treatment experiment, matching, EPGY, Correct first attempts, HLM, Proficiency Levels, CST, predicting future test results, Bayesian Classifiers, Mahalanobis distance, Learning curves.

Table of Contents

1. INTRODUCTION

Part I

2. METHODS

- 2.1 Research design
- 2.2 Data collection and the E/C pairings
 - Math CST scores
 - Course performance of the EPGY students

3. COMPARISON OF EPGY AND CONTROL STUDENTS ON THE 2007 MATHEMATICS CST

- 3.1 Paired t-test and related results for the Title I students in Grades 3 to 5
- 3.2 Weighting the correct first responses by the difficulty of the exercises
- 3.3 Results for second graders
- 3.4 District and school results
- 3.5 A Three-level, Hierarchical Linear Model (HLM) Analysis
 - Results for top half
 - Results for all 572 pairs
- 3.6 The effect of EPGY on changes in Proficiency Levels
- 3.7 The effect of EPGY on lower performing students
- 3.8 Summary of the effectiveness results
- 3.9 Justifying the use of CFA to stratify the students

Part II

4. REGRESSION MODELS ONLY FOR THE EPGY STUDENTS

- 4.1 CFA (WCFA) and CST06 as predictors of CST07
- 4.2 Results from a Title I school in district 2 that was outside the Effectiveness Study but which had EPGY students
- 4.3 Regressions adding the number of correct second-attempts, CSA, as a predictor of CST07
- 4.4 Correlation between CST06 and CFA
- 4.5 Correlation between *re-scaled* correct first-attempts and Math CST scores
- 4.6 Analysis of the 2006 to 2007 gain on Math CST by CST06 scores and CFA values

5. PREDICTING THE PROFICIENCY LEVELS OF THE EPGY STUDENTS USING MULTIVARIATE CLASSIFIERS

- 5.1 Minimum-distance classifier
- 5.2 Bayesian classifier
- 5.3 Learning curves

6. PREDICTION MODEL FOR EPGY STUDENTS ON THE MATH CST 2008

- OLS model
- HLM model
- 6.1 OLS and HLM prediction models using correct first-attempts
- Coefficients from California Standard Math Test 2006-2007
- Prediction result

6.2 OLS and HLM prediction models that include WCFA as a predictor

7. CONCLUSIONS

Acknowledgements

References

Appendix METHOD OF STATISTICAL ANALYSIS

A1 Effect sizes

Cohens' d

Our modification of d

A2 Three-level hierarchical linear model

A3 Binomial Analysis of Changes in proficiency level

1. INTRODUCTION

Stanford University's Education Program for Gifted Youth (EPGY) conducted a randomized treatment experiment (RTE) during the 2006-2007 school year to test the efficacy, for Title I students, of the EPGY Kindergarten through Grade 5 Mathematics Course Sequence (Math K-5). All eight participating schools had a full K-5 sequence of instruction. While the EPGY curriculum was originally developed for gifted students in any grade, the students in this study were not selected as gifted but were simply the students in the Title I schools.

This report is divided into two parts. Part I is concerned with measuring the effectiveness of the EPGY curriculum relative to the controls and consists of sections 2 and 3. In Section 2 we describe the research design and the data collection procedures for the RTE. Section 3 gives the main results of the RTE, focused on analyzing the Mathematics scores from the 2007 California Standards Test (Math CST07) for the EPGY (E) and control (C) students in several ways.

Part II reports several analyses we made that are not directly focused on measuring the effectiveness of the EPGY curriculum and consists of sections 4 to 6. Section 4 is devoted to developing linear regression models that relate the variables measured in the EPGY program to the Math CST07 scores for the E group. Section 5 presents the results of applying multivariate-normal covariate models to the E group's test results. In Section 6 we use several different regression models to predict the Math CST08 scores for EPGY using CST07 scores, CFA (WCFA) values and average latencies as predictors. The details of some of the statistical methods we used are briefly reviewed in the Appendix.

Part I

2. METHODS

2.1 Research design

The RTE was conducted with students in Grades 1 through 5 at eight Title I elementary schools located in three school districts within a 50-miles radius of Stanford University. Within each participating class, entering students were randomly assigned to one of two treatment groups, EPGY (E) or control (C). The random assignment process was done in the following way. Based on a prior measure of mathematical achievement, the students in a class were ranked from high to low. In each classroom, every two adjacent students in this ordering were considered a *pair*. In grades 3 to 5, the prior measure of mathematics achievement was the Mathematics score on the prior years (2006) California Standards Mathematics Test (CST06). Grade 2 students did not have a prior 2006 score because it was not administered in grade 1 in 2006, nor did Grade 1 students. Grade 1 and 2 students were administered the Stanford EPGY Mathematical Aptitude Test (SEMAT), discussed in Paek, Holland, & Suppes (1999) and this was used to form pairs ranked on prior mathematics achievement.

A computer algorithm then randomly assigned one member of each pair to the E condition, and the other member to the C condition. This random assignment was done 1000 times and, of these, the selected random assignment was the one that yielded the smallest sum of (a) the absolute difference in the *mean* prior test scores and, (b) the absolute difference in the variances, E versus C, across the pairs. This was done separately by grade. As the result of the combination of pairing and repeated random assignments, the mean and the variance of the prior test scores for the E and C groups were very close. Because of this, we were assured that at the start of the RTE the E and C groups were nearly evenly matched on prior mathematical achievement.

However, in addition to equalizing the prior mathematics achievement of the E and C groups, when attrition of various kinds occurred in the study, as it is certain to do in real schools, the pairings gave us a way to remove any bias it might introduce. If, for an example, an E-student did not have CST07 scores at the end of the study for some reason (didn't take the test, left the school, etc.) we simply delete his or her paired C student's CST07 data as well in order to maintain the close match we had initially on prior mathematics achievement. Similarly for C students with missing CST07 scores. The success of this matching-and-deleting approach to attrition is shown clearly in Figures 1 and 2 of section 2.2.

The logistics of the mathematics curriculum for the study were as follows. Students in the E group left their classrooms and went to the computer lab in their school where they worked for roughly 20 minutes a day, five days a week, under the supervision of a classroom teacher and an EPGY School Site Instructor. Students in the C group remained in the classroom during this time under the supervision of a classroom teacher and received an alternative treatment consisting of seatwork that was either worksheets from the adopted textbook or worksheets from the Renaissance Learning Accelerated Mathematics product, which was widely available in these districts. The control condition was the same in each of the schools. Additionally, both groups participated in the same basic mathematics instruction delivered by their classroom teachers during the school day. Scheduling and logistical details were determined on a school-by-school basis. Thus, the primary difference between the E and C students mathematics instruction was approximately 20 minutes a day of exposure to and work on the EPGY curriculum.

For the E students, their performance data and response latency on every exercise they attempted were logged into the EPGY Oracle database at Stanford. These data were used to compute various "EPGY"-variables, such as the *number of correct first attempts* (discussed later in this section) for every E student, but they are not available for the C students.

2.2 Data collection and the E/C pairings

After the initial pairing and random assignment, there were 1023 *pairs* in the study, with 27 in Grade 1, 194 in Grade 2 and 802 in Grades 3 to 5. Of the 1023 E students in these pairs, 919 completed *at least one EPGY exercise* as logged in the centralized EPGY Oracle database at Stanford. The remaining 104 E students (and their paired C students) were regarded as eliminated from the experimental study, though they may or may not have continued to participate in their class work. This left 919 pairs for which the E student did some work in the EPGY program; with 26 in Grade 1, 186 in Grade 2 and 707 in Grades 3 to 5. However, at the end of the school year, of these 919 pairs there were 742 pairs left *for which both students had scores on the Math CST07*, the outcome variable of the study; with 170 in Grade 2 and 572 in Grades 3 to 5. (The CST06 and CST07 were not administered to Grade 1 students so they could not be part of the effectiveness study, but their data are included in the second part of this report.)

Thus, starting with an initial group of 996 matched pairs of students in Grades 2 to 5, at the end of the study period there were 742 pairs left that had outcome measures for both members and for which the E student had completed at least one EPGY exercise in the EPGY curriculum. This represents attrition of about 25% of the Grade 2 to 5 students that were initially in the study. About 40% of this attrition is due to the E student not completing one of the EPGY exercises and 60% due to either the E or C student not having a CST07 score at the end of the study period.

Of the 742 pairs described above, the 572 in grades 3-5 are of special importance because they have CST06 scores, as well as CST07 scores, which can be used as a covariate in assessing the effect of EPGY relative to the control condition on the CST07 scores, as we discuss in detail in section 3.

CST Mathematics scores

At the end of the school year, students in Grades 2 – 5 took the 2007 Mathematics CST (CST07). For the pairs for which the E member completed at least one EPGY exercise the numbers of pairs with CST scores for 2006 and 2007, for each district and school, is shown in Table 1.

Table 1. Number of pairs with Math CST Scores for 2006 and 2007, by School and District.

Title I Schools		Math CST06		Math CST07		Math CST06 & Math CST07		
		Number of pairs*	Both members have scores**	Any member without score	Both members have scores***	Any member without score	Both members have both scores**	EPGY member has both scores**
All 8 Schools		919	632	287	800	119	572	619
District 1	All	526	353	173	486	40	333	348
	School A	144	87	57	136	8	82	84
	School B	102	77	25	94	8	72	78
	School C	145	103	42	140	5	99	102
District 2	All	174	108	66	128	46	92	104
	School E	97	58	39	60	37	50	56
	School F	77	50	27	68	9	42	48
District 3	All	219	171	48	186	33	147	167
	School G	113	104	9	93	20	87	103
	School H	106	67	39	93	13	60	64

*Total of all grades, 1 through 5.

**Only grades 3 through 5, Grade 1-2 students do not have CST06 scores.

***Includes Grade 2 students most of whom have CST07 scores, but no Grade 1 students.

The Math CST07 scores, and the related student proficiency levels, served as the outcome criteria for the effectiveness of EPGY compared to the control condition. Given that the Mathematics CST has been externally developed, validated, administered, and scored, there was no additional external evaluation instrument used in this RTE. In addition, when available, the Math CST06 scores served as a pretreatment covariate in some of our analyses. Figures 1 and 2

show the distributions (histograms) of the Math CST06 scores for all of the 572 pairs in Grades 3-5 that had both CST06 and CST07 scores for both members. In addition, the legends of these two figures give the median, mean, and standard deviation of the Math CST06 scores for these two groups. These graphs show the effectiveness of our method of removing the entire pair whenever one member of the pair was missing some data. Even after losing 25% of the pairs due to various causes, the means and standard deviations of the distributions of the CST06 scores for the remaining pairs are nearly identical, indicating that the two groups of 3-5 graders that we use for some of our main analyses of the effectiveness of EPGY are as similar on their prior mathematics performance as were all of the initial pairs were prior to the beginning of the experiment.

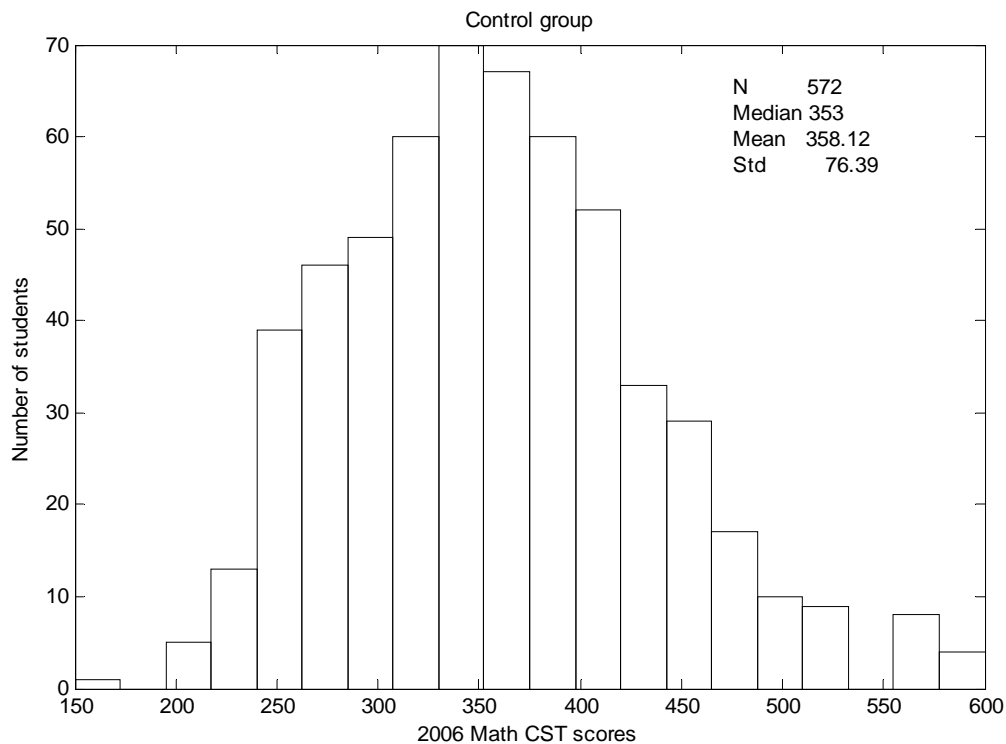


Figure 1: Histogram of the 2006 Math CST scores for the control group for the 572 matched pairs in which both members have both CST scores.

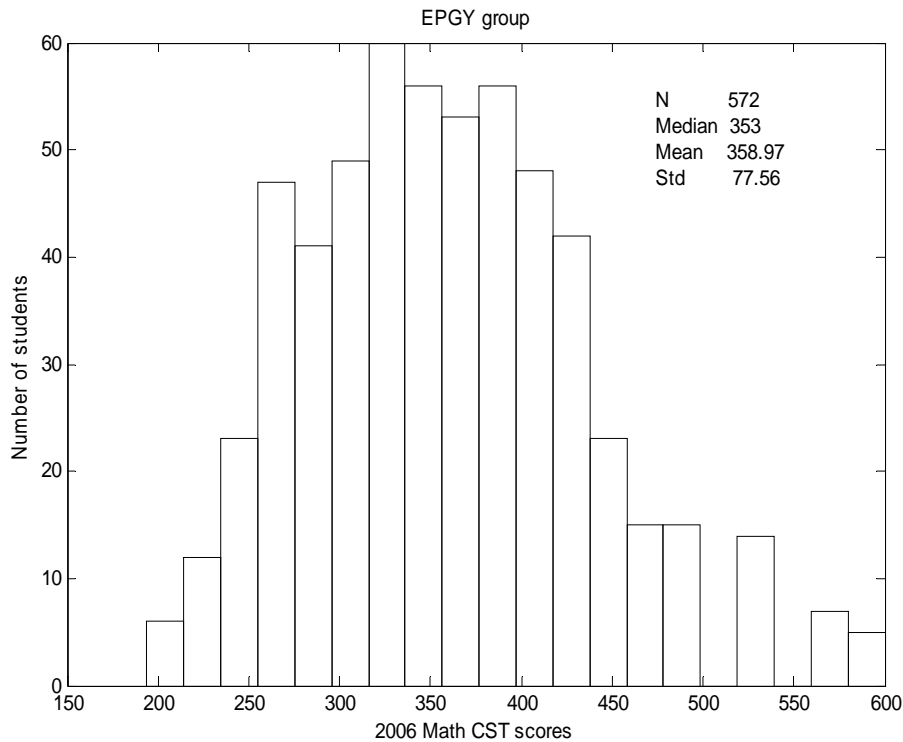


Figure 2: Histogram of the 2006 Math CST scores for the EPGY group for the 572 matched pairs in which both members have both CST scores.

Course performance of the EPGY students

The experimental treatment consisted of computer-presented mathematics exercises given to the students who then used the computer to respond to them. This results in a detailed record of student course performance in the EPGY curriculum that was available for analysis. Here we concentrate on a variable that plays an important role as a post-treatment covariate in many of our analyses. The *number of correct first attempts* (CFAs) is the total number of exercises for which the EPGY student got the correct answer on his or her first attempt at responding to it. Except for exercises with only two possible responses, students who made an error on their first attempt were immediately given a second opportunity to do the exercise. Thus, in addition to CFAs there is also the *number of correct second-attempts* (CSAs). If the first attempt was correct then there is no second attempt. For this reason, the interpretation of CSAs requires the value of

the CFAs as well. For example, a high value of CSA means that the student usually got the correct answer on the second try after missing it on the first try. A low value of CSA can mean that either the student tended to get the correct answer on the first try or that the student tended to get both the first and second attempt wrong. Our use of CSAs in this report is limited to a brief analysis reported in Section 4.3 in Part II.

Figure 3 shows the histogram of CFAs for the 919 EPGY students in Grades 1-5. The range of values in Figure 3 is from 2 to 4,917. All CFA values used in this report are cumulative over the period, September 15, 2006 to July 15, 2007.

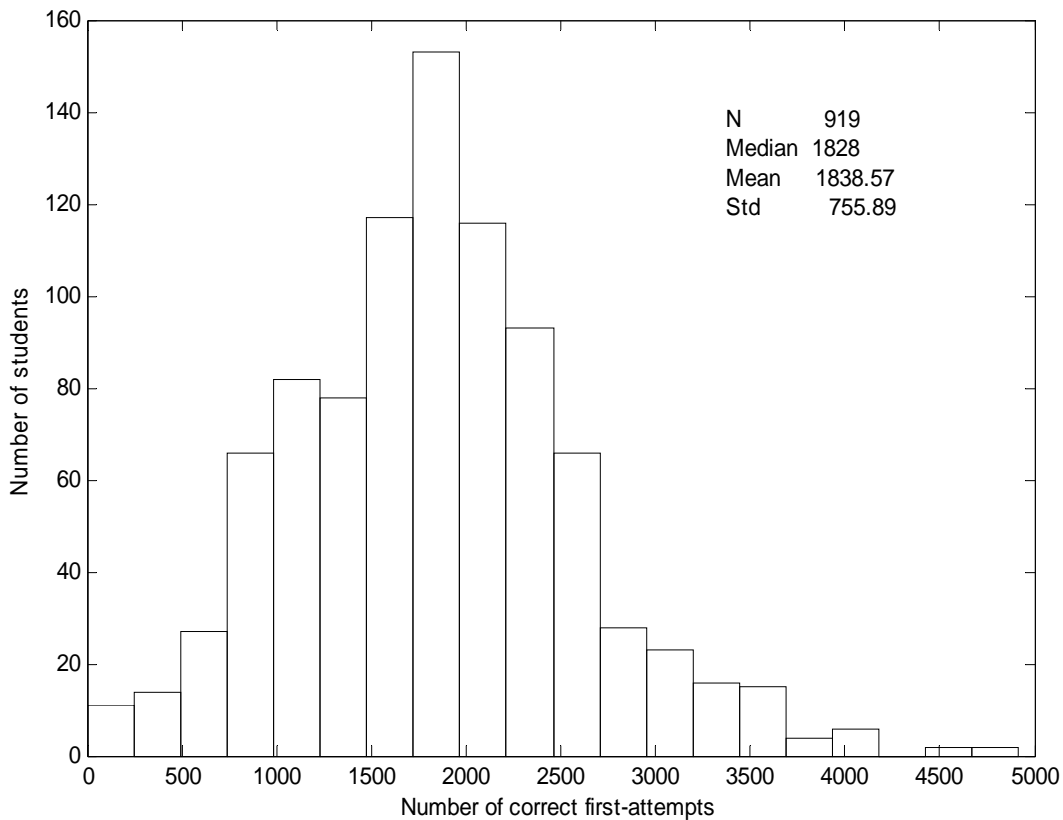


Figure 3: Histogram of CFA for all 919 EPGY students in Grades 1 to 5 at the 8 Title I schools who attempted at least one exercise.

Table 2 gives more detail about the distribution of CFA values in Figure 3. For each quartile of CFA values, sorted from lowest to highest, Table 2 gives the minimum, maximum, mean and standard deviation of the CFA values.

Table 2. Summary of the four quartiles of the distribution of CFA in Figure 3--Minimum, Maximum, Mean and Standard Deviation.

Quartile	Min	Max	Mean	SD
Q1 = Lowest Quartile N = 230	2	1325	903.2	310.6
Q2	1326	1825	1612.8	142.3
Q3	1828	2287	2033.3	128.5
Q4	2290	4917	2803.9	503.6

Figure 4 is a box and whiskers plot of the CFA values in Figure 3, separately by Grade.

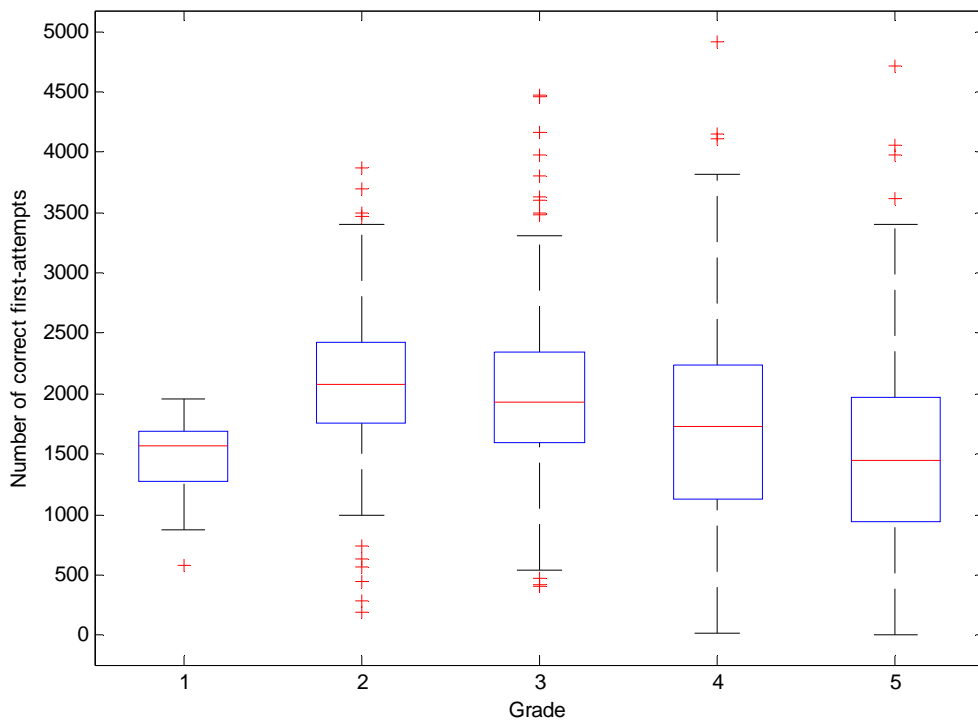


Figure 4. Box and whiskers plot of CFA by Grade for all 919 E students who attempted at least one EPGY exercise. The box contains the middle 50% of the distribution with the median marked in its middle. The whiskers indicate the extent of the rest of the data.

Figure 4 shows that there is a great deal of overlap between the five grades in the study with respect to the distributions of CFA. The spread tends to increase with grade, but the median values are all between CFA values of 1500 and 2200.

In subsequent analyses, we often use various subsets of the 919 EPGY students in Grades 1-5. In Table 3 we display the median, mean and standard deviation (SD) of the CFAs for some of these important subgroups.

Table 3. Descriptive statistics of CFAs for some subgroup of 919 EPGY students.

Subgroup of the 919 EPGY students who attempted at least one exercise.	Number of students	Median	Mean	SD
Both member of the pair in Grade 2 have a CST07 score.	170	2098	2117.9	554.9
Both members of the pair in Grades 3-5 have both a CST06 and CST07 score.	572	1808.5	1861.5	780.1
EPGY students in Grades 3-5 with CST06 and CST07 scores.	619	1797	1843.4	766.6

The CFA value for an E student is a measure of the amount of careful work done in the EPGY curriculum by that student. The higher the value of CFA the more exercises the student has attempted and the more of these that he or she got correct on the first try. Alternative measures, such as the sheer number of exercises attempted or the percent of correct first attempts out of all attempted exercises were not used because they do not capture the idea of the amount of *careful work* done in the curriculum as well as CFA does.

For this reason, we expected that students who had higher CFA values would benefit more from exposure to the EPGY curriculum than those with lower CFA values.

3. COMPARISON OF EPGY AND CONTROL STUDENTS ON THE 2007 MATHEMATICS CST

This section consists of seven subsections (3.1 to 3.7) that examine various comparisons of the E and C students on their CST07 test scores in Mathematics. Then, in subsection 3.8 we summarize our various finding on the effectiveness of the EPGY curriculum relative to the control. Finally, in subsection 3.9 we consider some issues that arise in the justification of stratifying the EPGY students on the value of their total correct first attempts during the 2006-07 study period.

3.1 Paired t-test and related results for the Title I students in Grades 3 to 5

We first restrict the analysis to the 572 pairs of Grade 3 to 5 students which had both CST06 and CST07 scores for both pair members and for which the E-member of the pair completed at least one EPGY exercise. The mean difference on the CST07 for all pairs is 0.05, with a standard deviation of the pair difference of 67.62. The paired t-test t-value for this difference is 0.02, with a 2-sided p-value of .98. Thus, over *all the pairs*, there is no significant difference between the performance of the E and C students. However, this analysis does not take into account the wide distribution of differences in the *amount of work on the EPGY curriculum by the E students* as measured by the CFA values illustrated in Figures 3 and 4. E students with few correct first attempts over the year are much less engaged in and working on the curriculum than are those with many correct first attempts. To illustrate the effect of CFA on the pair differences, we report most of our analyses by grouping the (E, C)-pairs by the CFA value for each E-member of the pair.

In Table 4 we give the results of paired t-tests for various subgroups of the pairs that are ordered by the E-member's CFA score. The group with the highest values of CFA is the top fourth, then the top third and then the top half. The mean differences for these successive groups are all positive and statistically significant beyond the 0.01 level and show a decreasing trend as more pairs are included in which the E-students CFA value are lower than those in the top fourth. The effect sizes corresponding to these t-tests range from 15 to 21 percent of the median standard deviation (across grades 2 to 5) for all California students. Next, starting in the fourth row of Table 4, we give results for the lower half of the distribution of the CFA values for the E-students. The groups are the top fourth of the bottom half, then the top third of the bottom half, and the top half of the bottom half. The mean differences for these three subgroups are all negative, but not statistically significant at the usual levels. These mean differences also show a decreasing trend as more low CFA scores are included in the comparison.

Table 4. Summary of the comparisons between E and C matched students in grades 3 to 5 on the paired t-test, grouped by the CFA score of the E student in each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Number of pairs	Mean of the CST07 difference for each pair	SD of the CST07 difference for each pair	Effect size relative to SD = 83.5*	2-sided Paired t test p-value
Top fourth ranked on the CFA of E	143	17.65	67.91	0.21	2.3×10^{-3}
Top third ranked on the CFA of E	191	16.90	64.18	0.20	4.0×10^{-4}
Top half ranked on the CFA of E	286	12.54	65.85	0.15	1.4×10^{-3}
Top fourth of the bottom half ranked on the CFA of E	71	-2.16	75.88	-0.03	0.81
Top third of the bottom half ranked on the CFA of E	95	-2.30	77.06	-0.03	0.77
Top half of the bottom half ranked on the CFA of E	143	-8.27	73.74	-0.10	0.18
All pairs	572	0.05	67.62	0.00	0.98

*The value of 83.5 is the median of the grade specific standard deviations (that range from 73 to 87) of the scaled scores for the Mathematics 2007 CST test for all of California for grades 2 to 5 and is used here as the denominator for the effect sizes.

To get an alternative view of the effect of stratifying over the entire range of CFA scores, Table 5 gives the corresponding results for the five quintiles (i.e., the fifths) of the distribution of the 572 pairs described in Table 4. In Table 5 we see that the mean difference between the pairs steadily increases as one goes up the quintiles starting with statistically significant negative differences in the first two (lower) quintiles, a non-significant negative difference in the third quintile and then increasingly positive differences that are statistically significant in the fourth and fifth (highest) quintiles.

Table 5: Summary of the comparisons between E and C matched students in grades 3 to 5 on the paired t-test, grouped by decreasing quintiles of the CFA score of the E student in each pair.

Quintile	Number of pairs	Mean of CST07 differences	SD of CST07 differences	Effect Size (relative to SD = 83.5*)	Two-sided paired t-test p-value
Highest CFA quintile	114	20.07	69.7	0.24	2.7×10^{-3}
4 th CFA quintile	114	14.85	61.1	0.18	1.1×10^{-2}
3 rd CFA quintile	115	-4.29	72.4	-0.05	0.53
2 nd CFA quintile	115	-13.88	66.4	-0.17	2.7×10^{-2}
Lowest CFA quintile	114	-16.36	60.5	-0.20	4.7×10^{-3}
All pairs	572	0.05	67.6	0.00	0.98

*The value of 83.5 is the median of the grade specific standard deviations (that range from 73 to 87) of the scaled scores for the Mathematics 2007 CST test for all of California for grades 2 to 5 and is used here as the denominator for the effect sizes.

Figure 5 shows the results of Table 5 in more detail. It displays the E minus C difference on the CST07 versus the CFA value for the E student for all 572 pairs. Also included is the regression line between the CST07 differences and CFA. This display shows that there is considerable spread in the CST07 differences but there is a slight, but statistically significant ($p < .0001$), upward trend with a slope of 0.0177 that we saw in Tables 5. Solving the regression equation, $CST07-dif = -32.984 + 0.0177CFA$, for the CFA value where the line is zero gives $CFA = 1863.5$. Comparing with Table 3 we see that this is about the mean of the CFA values for these E students and slightly larger than their median CFA value of 1808. Thus, when the CFA value exceeds 1863, the E students begin to have a slight edge on average when compared to their matched controls. As we see in Tables 4 and 5, this slight edge becomes a substantial average effect when all E-students whose CFA values are above the median are considered together.

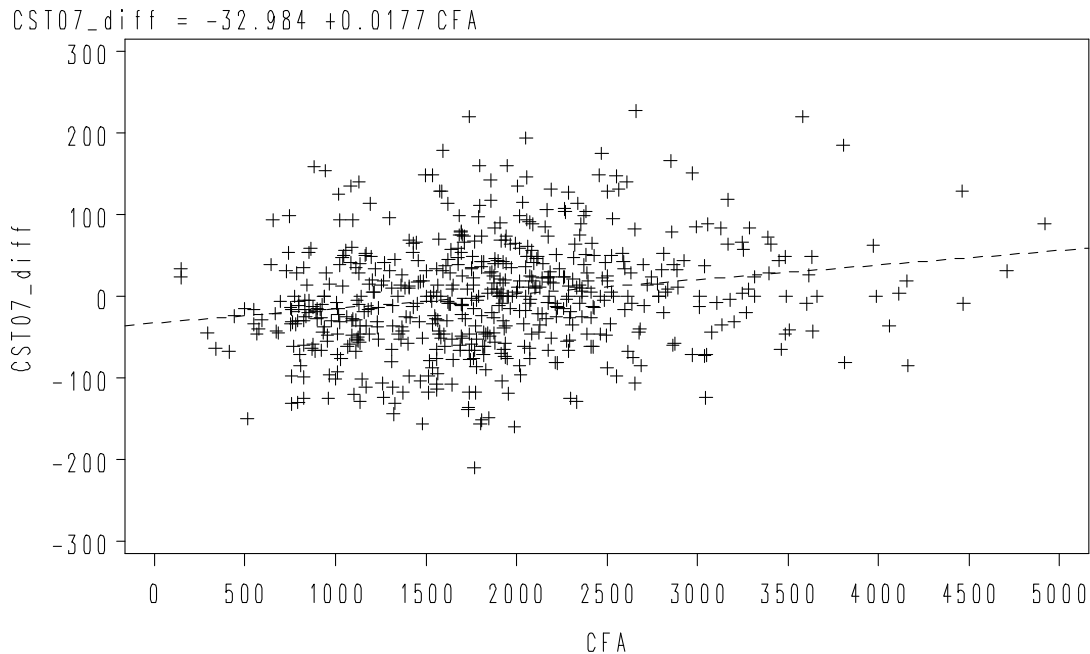


Figure 5. Scatter plot of E minus C difference on CST07 versus CFA of E student for 572 match pairs of students in Grades 3 to 5. R-square is 0.042, slope is 0.0177.

3.2 Weighting the correct first responses by the difficulty of the exercises

The *correct first attempts* used in the previous section do not reflect the difficulty of the attempted exercises. The actual difficulty of a presented exercise for a student depends on a number of factors—the prior performance of the student, the grade level of the exercise, etc. We considered the effect of a natural non-linear adjustment of the CFA score to be worth exploring to see if it changed the relationship between CFA and the CST07 differences. The new measure, the *weighted correct first attempts* (WCFA)

weights each exercise, j , that is worked correctly on the first attempt by student i , by the *adjusted grade-placement* (AGP_{ij}) of exercise j relative to student i . The AGP_{ij} is the difference between the grade-placement of exercise j (GP_j) and the school grade placement of student i . The AGP_{ij} can be negative when a student is placed in a higher grade than the level that he or she is working at in the EPGY curriculum. The value of $WCFA_i$ is

$$WCFA_i = \sum_j AGP_{ij} ,$$

where the summation is over all of the exercises that student i worked correctly on his or her first attempt.

We briefly consider a second scaling of CFAs by both the latencies on the correctly worked exercises and the AGP in Section 4.5 in Part II.

Table 6 is structured just like Table 4, but with the variable CFA replaced by $WCFA$. The results are very similar to those in Table 4, so we do not discuss them further.

Table 6. Summary of the comparisons between E and C matched students in grades 3 to 5 on the paired t-test, grouped by the WCFA score of the E student in each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Number of pairs	Mean of CST07 differences	SD of CST07 differences	Effect Size (relative to SD = 83.5*)	Two-sided Paired t-test p-value
Top fourth ranked on the WCFA of E	143	23.99	80.9	0.29	5.0×10^{-4}
Top third ranked on the WCFA of E	191	15.74	80.9	0.19	7.8×10^{-3}
Top half ranked on the WCFA of E	286	6.44	77.3	0.08	0.16
Top fourth of the bottom half ranked on the WCFA of E	71	-13.49	56.8	-0.16	0.05
Top third of the bottom half ranked on the WCFA of E	95	-10.71	61.0	-0.13	0.09
Top half of the bottom half ranked on the WCFA of E	143	-6.56	56.8	-0.08	0.17

*The value of 83.5 is the median of the grade specific standard deviations (that range from 73 to 87) of the scaled scores for the Mathematics 2007 CST test for all of California for grades 2 to 5 and is used here as the denominator for the effect sizes.

3.3 Results for second graders

The results for Grade 2, similar to those described above for Grades 3 to 5, are summarized in Table 7. Overall there is a slightly negative but statistically non-significant mean difference between the E and C groups. However, when we break out the data and stratify on the amount of work done by the E students as measured by their CFA values we find that the effect sizes in Grade 2 appear to be much larger than they were for Grades 3 to 5. As in the higher grades the mean differences are the most positive for the subgroup with the highest CFA values and they decrease as more students with lower values of CFA are included in the subgroup.

Unlike Tables 5 and 6, all of the mean differences in Table 7 are positive, although only the first three are statistically significant at the usual levels. Moreover, the effect sizes for the first three subgroups are two to three times larger than they were for those groups in the higher grades. We note from Figure 4 that the distribution of CFA values for Grade 2 are somewhat higher than for the higher grades, though there is considerable overlap.

Table 7. Summary of the comparisons between E and C matched students in grade 2 on the paired t-test, grouped by the CFA score of the E student in each pair.

Group of the 170 pairs of Grade 2 students in all 8 schools	Number of pairs	Mean of CST07 differences	SD of CST07 differences	Effect Size (relative to SD = 82.0*)	Two-sided Paired t-test p-value
Top fourth ranked on the CFA of E	42	53.33	76.3	0.65	5.0×10^{-5}
Top third ranked on the CFA of E	57	46.97	87.5	0.57	1.6×10^{-4}
Top half ranked on the CFA of E	85	28.25	95.4	0.34	7.7×10^{-3}
Top fourth of the bottom half ranked on the CFA of E	21	9.62	79.5	0.12	0.59
Top third of the bottom half ranked on the CFA of E	28	4.79	78.6	0.06	0.75
Top half of the bottom half ranked on the CFA of E	43	4.72	75.6	0.06	0.68
All pairs	170	-3.08	99.2	-0.04	0.69

*The value of 82 is the standard deviation of the scaled scores for the Mathematics 2007 CST test for all of California for grade 2 and is used here as the denominator for the effect sizes.

3.4 District and school results

Given the results of Sections 3.1 and 3.3, we restricted our detailed analysis of individual schools to pairs where the E student was in the top half of the CFA distribution. Because of the small number of students at some schools, we used all 800 pairs in which both members had CST07 scores (see first row of Table 1 for the students that are included in the 800). Of special note, the selection of the top halves of the CFA distributions was done separately for each school and each district. This accounts for the differences in the numbers of pairs in a district and in that district's schools in Table 8. Table 8 summarizes the comparisons between the EPGY and control students for each district and school.

Table 8. Summary of comparisons between matched E and C students for all districts and schools in the Effectiveness Study on the paired t-tests, for all pairs for which the CFA of the E student was in the top half of the distribution for that school or district.

District/School	Number of pairs	Mean of CST07 differences	SD of CST07 differences	Effect Size (relative to SD = 83.5*)	Two-sided Paired t-test p-value
District 1	243	18.74	79.3	0.22	3.00 x 10 ⁻⁴
School A	68	27.77	90.1	0.33	0.01
School B	47	25.43	79.3	0.30	0.03
School C	70	19.66	69.9	0.24	0.02
School D	58	6.72	81.9	0.08	0.53
District 2	64	8.20	83.9	0.10	0.44
School E	31	36.36	75.5	0.44	0.01
School F	34	13.36	79.2	0.16	0.33
District 3	93	14.48	72.5	0.17	0.05
School G	47	26.70	78.6	0.32	0.02
School H	47	5.26	58.0	0.06	0.54

*The value of 83.5 is the median of the grade specific standard deviations (that range from 73 to 87) of the scaled scores for the Mathematics 2007 CST test for all of California for grades 2 to 5 and is used here as the denominator for the effect sizes.

Table 8 shows that the E students performed consistently higher, on average, than the control students within each school and district when the pairs were restricted to those for which the E member was in the top half of students ranked by their CFA values. The effect sizes in Table 8 are all positive and most of the p-values for the t-test were smaller than the usual standard of .05.

3.5 A Three-level, Hierarchical Linear Model (HLM) Analysis

Results for the top half

There are 572 matched pairs in the study that had both CST06 and CST07 scores for both pair members and for which the E member had completed at least one EPGY exercise. The “top half” of these pairs refers to the 286 pairs for which the E member had a CFA value in the top half of the CFA distribution. In addition, these students stayed with the same teacher and school within the experiment. We first fit the HLM described in the appendix that has the form (see Appendix A2 for more details).

$$CST07_{ijk} = \gamma + \pi_1 CST06 + \pi_2 TX + u_k + v_{jk} + e_{ijk}$$

Where γ , π_1 and π_2 are fixed effects and u_k , v_{jk} and e_{ijk} are the random effects for schools, classes within schools and students within classes, respectively.

In these analyses, we treat the students as individual units and decouple the pairs. The distinction between E and C students is captured by the variable TX (1 for E, 0 for C).

The fixed effect estimates from HLM are given in Table 9.

Table 9. Fixed effects for the “top half” pairs of E and C students.

Parameter	Effect	Estimate	Std Err	DF	t-value	p-value
γ	Intercept	72.41	12.37	7	5.85	6.28×10^{-4}
π_1	CST06	0.80	0.03	561	28.30	$< 10^{-100}$
π_2	TX (Treatment)	10.41	3.89	561	2.68	7.61×10^{-3}

All of the fixed effects are statistically significant by the usual standards. On average, the effect of EPGY for the E students in the top half of the CFA distribution is 10.41 points on the CST07 scale. Compared to the median overall standard deviation of CST07 scores (83.5) used in the tables of section 3.1 this corresponds to an effect size of 12.5%. As we would normally expect, there is a strong correlation between the CST06 and 07 scores.

The estimated for CST06, 0.80, gives the statistical relationship between 2006 and 2007 math test scores. Students who differ by 1 point on CST06 differ by 0.80 on average on the CST07.

The random effects estimates from HLM are described in table 10.

Table 10. Random effects for the “top half” pairs of E and C students.

Notation	Level	Estimated variance	Std Err	z-value	p-value
u_k	School	160.19	193.93	0.83	0.20
v_{jk}	Classroom	923.95	267.61	3.45	3.00×10^{-4}
e_{ijk}	Student	2153.27	135.27	15.92	$< 10^{-50}$

The significant p-value for classrooms indicates the existence of significant variation among the classrooms in terms of CST07 scores. This is beyond what is expected due to variation in the prior years CST06 scores and is a possible indicator of teacher differences. The school differences are not significant.

Results for all 572 pairs

We did a similar HLM analysis for the data from all 572 matched pairs of E and C students who have both CST06 and 07 scores and the E student attempted at least one EPBY exercise. The results are summarized in Tables 11 and 12, below.

Table 11. Fixed effects for the 572 pairs.

Parameter	Effect	Estimate	Std Err	DF	t-value	p-value
γ	Intercept	77.10	9.30	7	8.30	7.21×10^{-5}
π_1	CST06	0.78	0.02	1131	38.30	$< 10^{-100}$
π_2	TX	-0.53	2.76	1131	-0.19	0.85

Again, we see that, overall, there is a negative but non-significant effect of the treatment when we do not restrict attention to those E students who worked carefully and steadily on the EPGY curriculum. This is also what we saw in the last line of Table 4, except that these results are adjusted for the small differences in the CST06 scores between the E and C students. The coefficient for CST06 is nearly the same as it was for the analysis of the “top half” pairs.

Table 12. Random effects for the 572 pairs.

Notation	Level	Estimated variance	StdErr	z-value	p-value
u_k	School	138.24	137.70	1.00	0.16
v_{jk}	Classroom	718.04	164.27	4.37	6.18×10^{-6}
e_{ijk}	Student	2170.08	93.45	23.22	$< 10^{-100}$

As we saw in Table 10 for the “top half” pairs, there are virtually no school effects but substantial classroom effects. The square root of the estimated classroom variance in Table 12 is 26.8 which means that net of the differences in their classroom CST06 scores, the classrooms had mean differences of plus or minus 27 or so points in the CST07 scores of the students.

As a check on the HLM results, we did an OLS regression with TX, CST06 and both school and classroom indicator variables as the predictor variables for all students and for the pairs where the E student was in the top half of the distribution of CFA. The estimates for the coefficients for TX and CST06 were nearly identical with those in Tables 9 and 11. In addition, the coefficients for the classroom indicator variables showed considerable variation in keeping with the significant classroom variation seen in the HLM results.

3.6 The effect of EPGY on changes in Proficiency Levels

We now turn to an analysis of changes in Math CST Proficiency Levels (PLs) between 2006 and 2007 for the 572 grade 3 to 5 matched pairs having both scores. The question addressed here is whether the positive changes in proficiency from 2006 to 2007 exceed the negative changes. For example, a student who moved from Proficient in 2006 to Basic in 2007 would represent a negative change. Table 13 gives the score boundaries for the PLs by grade for the 2007 Math CST. These score boundaries also apply to the CST06 Mathematics test.

Table 13. California classification of 2007 Math CST test scores by proficiency level.

Grade	Far below basic	Below basic	Basic	Proficient	Advanced
2	150–235	236–299	300–349	350–413	414–600
3	150–235	236–299	300–349	350–413	414–600
4	150–244	245–299	300–349	350–400	401–600
5	150–247	248–299	300–349	350–429	430–600
6	150–252	253–299	300–349	350–414	415–600
7	150–256	257–299	300–349	350–413	414–600

The main results are summarized in Table 14 and Table 15. These tables summarize the changes in both the PLs and in the corresponding test scores. We include the changes in test scores for comparison with the changes in PLs but do not discuss them because they merely echo the changes in the PLs.

Table 14. Result of binomial analysis of changes in proficiency level, grouped by the CFA score of the E student in each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Change in proficiency level		Change in test scores	
	p-value		p-value	
	E	C	E	C
Top fourth ranked on the CFA of E	+48 vs -15 $p=3.76 \times 10^{-5}$	+39 vs -28 $p=0.22$	+91 vs -52 $p=1.40 \times 10^{-3}$	+80 vs -62 $p=0.15$
Top third ranked on the CFA of E	+64 vs -24 $p=2.37 \times 10^{-5}$	+49 vs -36 $p=0.19$	+121 vs -70 $p=2.75 \times 10^{-4}$	+103 vs -87 $p=0.28$
Top half ranked on the CFA of E	+84 vs -50 $p=4.19 \times 10^{-3}$	+69 vs -62 $p=0.60$	+165 vs -120 $p=9.03 \times 10^{-3}$	+152 vs -133 $p=0.29$
Top fourth of the bottom half ranked on the CFA of E	+13 vs -22 $p=0.18$	+11 vs -26 $p=0.02$	+31 vs -40 $p=0.34$	+28 vs -42 $p=0.12$
Top third of the bottom half ranked on the CFA of E	+17 vs -32 $p=0.04$	+17 vs -34 $p=0.02$	+39 vs -56 $p=0.10$	+37 vs -57 $p=0.049$
Top half of the bottom half ranked on the CFA of E	+22 vs -53 $p=4.49 \times 10^{-4}$	+31 vs -48 $p=0.07$	+51 vs -92 $p=7.64 \times 10^{-4}$	+58 vs -83 $p=0.04$
All pairs	+120 vs -177 $p=1.12 \times 10^{-3}$	+130 vs -164 $p=0.05$	+258 vs -313 $p=0.02$	+263 vs -306 $p=0.08$

Table 14 summarizes the positive and negative changes in the PLs for the matched E and C students, with the pairs grouped by the CFA of each E member. For example, for the 143 pairs where the E member's CFA value was in the top fourth of the distribution there were 48 positive changes in the PLs between 2006 and 2007 and 15 negative changes for the E students. For the remaining 80 E students in this group, there was no change in the PLs. The 48/15 split was highly significant, with $p < 10^{-5}$. For the control students, the split was more even, 39 positive, 28 negative and 76 no change, and the 39/28 split has a p-value of only 0.22. Similar trends hold for the top third and top half of the CFA distribution, with the E students having significantly more positive than negative

changes in the PLs while the matched C students have more nearly even splits between positive and negative changes in PLs that are not statistically significant. The p-values for the results of the changes in PLs is discussed more extensively in the Appendix.

Table 15 is like Table 14, except the weighted variable, WCFA, replaces CFA in the analysis.

Table 15. Result of binomial analysis of changes in proficiency level, grouped by the WCFA score of the E student in each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Change in proficiency level		Change in test scores	
	p-value		p-value	
	E	C	E	C
Top fourth ranked on the WCFA of E	+21 vs -27 p=0.47	+19 vs -44 p=2.00×10 ⁻³	+63 vs -80 p=0.19	+51 vs -91 p=9.95×10 ⁻⁴
Top third ranked on the WCFA of E	+27 vs -51 p=9.00×10 ⁻³	+26 vs -64 p=7.66×10 ⁻⁵	+78 vs -113 p=0.01	+71 vs -119 p=6.12×10 ⁻⁴
Top half ranked on the WCFA of E	+46 vs -97 p=2.41×10 ⁻⁵	+47 vs -95 p=6.91×10 ⁻⁵	+110 vs -176 p=1.14×10 ⁻⁴	+109 vs -175 p=1.07×10 ⁻⁴
Top fourth of the bottom half ranked on the WCFA of E	+15 vs -30 p=0.04	+17 vs -21 p=0.63	+30 vs -41 p=0.24	+30 vs -41 p=0.24
Top third of the bottom half ranked on the WCFA of E	+22 vs -39 p=0.04	+25 vs -28 p=0.78	+39 vs -55 p=0.121	+43 vs -52 p=0.41
Top half of the bottom half ranked on the WCFA of E	+34 vs -50 p=0.05	+39 vs -41 p=0.91	+69 vs -73 p=0.80	+71 vs -72 p=1.00
All pairs	+120 vs -177 p=1.12×10 ⁻³	+130 vs -164 p=0.05	+258 vs -313 p=0.02	+263 vs -306 p=0.08

Table 15 show results that are similar to those in Table 14 except that now the E students have more even splits between positive and negative changes in PLs but the C students have *significantly more negative changes* than positive changes.

Table 16 summarizes the result of the binomial analysis of changes in the PLs for the pairs of students in the top fourth and top half of the CST06 score distribution.

Table 16. Result of binomial analysis of changes in proficiency level, grouped by CST06 scores.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Change in proficiency level	
	p-value	
	E	C
Top fourth ranked on 2006 Math CST	+6 vs -44 $p=3.21 \times 10^{-8}$	+10 vs -49 $p=2.71 \times 10^{-7}$
Top half ranked on 2006 Math CST	+34 vs -99 $p=9.91 \times 10^{-9}$	+42 vs -95 $p=6.92 \times 10^{-6}$

Table 16 illustrates the significant effect of *regression toward the mean* on the PLs. Because the CST06 and 07 scores are not perfectly correlated, by selecting the top students on their Math CST06, their CST07 scores tend to be lower on average in both the E and C groups. This effect explains much of what is seen in Table 16.

3.7 The effect of EPGY on lower performing students

The positive results we have seen so far are for E students who are in the top half of the distribution of CFA values among this sample of Title I students. Because there is a positive correlation between CST06 scores and CFA values, these students tend to overlap with the top half of students based on Math CST06 scores. The intersection is 150 of 286, which supports the view that these are among the best Title I students.

We now present results for the less able Title I students, as measured by their Math CST06 scores. We selected the students who constituted the bottom half of the Math CST06 score distribution. We then tested the effectiveness of EPGY with this group of students by considering the top fourth and the top half of such students as measured by their CFA values. The results are displayed in Table 17.

Table 17. Summary of the comparisons between E and C matched students in grades 3 to 5 on the paired t-test; the bottom half of the CST06 score distribution grouped by the CFA score of the E student in each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Number of pairs	Mean of the CST07 difference for each pair	SD of the CST07 difference for each pair	Effect size relative to SD = 83.5*	2-sided Paired t test p-value
Top fourth of CFA of bottom half of CST06	73	15.71	63.39	0.19	0.04
Top half of CFA of bottom half of CST06	146	8.49	59.85	0.10	0.09

Comparing Table 17 (the poor performing students on the CST06) to Table 4 (all of the students in the study) shows that the results for the lower performing students are similar to what we have seen earlier, but with smaller effect sizes and less statistical significance. The students in the top fourth show a larger EPGY effect than those in the top half of the CFA distribution, just as is seen in Table 4 comparing the top fourth to the top half of the CFA distribution.

Table 18 shows results that correspond to Table 9 for the binomial analysis of changes in the PLs for this lower performing group.

Table 18. Result of binomial analysis of changes in proficiency level, grouped by the CFA score of the E student in each pair, for the bottom half of the CST06 distribution.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Change in proficiency level	
	E	C
Top fourth of CFA of bottom half of CST06	+35 vs -6 p=4.87×10 ⁻⁶	+24 vs -13 p=0.1
Top half of CFA of bottom half of CST06	+62 vs -22 p=1.47×10 ⁻⁵	+46 vs -30 p=0.08

The results in Table 18 are very similar to those in Table 9 with the E students having significantly more positive than negative changes in PLs and the C students having more nearly even splits that do not reach high levels of statistical significance.

In summary, the results shown in Table 17 and Table 18 support the hypothesis that careful and diligent work by EPGY students, whose test scores have classified them as being in the bottom half of the Title I students in the experiment, can also show significant benefits, as measured by the Math CST07 results.

Just to look even more closely at the students with the lowest Math CST06 scores we examined the data of EPGY students in the top half on CFA of the bottom fourth on CST06. The number of students is now only 72. The difference between EPGY students and control students was not significant on the paired t-test. On the other hand, on the binomial analysis of change in proficiency levels, the results were significant, + 28 vs -10, with $p = 5.10 \times 10^{-3}$. The results for the control group were also positive but less significant, +26 vs -11 with $p = 0.03$.

3.8 Summary of the effectiveness results

Throughout Section 3, we have found that if we restrict attention to those EPGY students who were in the top half of the distribution of correct first attempts on the EPGY exercises we see substantial and statistically significant improvements in the CST07 test scores over the scores of matched control students. The effects in second grade appear to be larger than those in grades 3 to 5. Furthermore, positive effects also occur to a somewhat lesser extent for students who are less able mathematically, as measured by their 2006 CST mathematics scores, see subsection 3.7.

When we looked at changes in Performance Levels between 2006 and 2006 we saw that for the EPGY students with CFA values in the top half of the distribution there were significantly more positive changes than negative ones while for the matched control students the split was more even and not statistically significant.

In the context of general elementary-school mathematics student learning, this is not entirely unexpected. Students in these grades are presented with a math curriculum that is increasingly difficult and more complex. Whatever the particular math curriculum, for a student to do well he or she needs to work accurately and continually throughout the school year. The active engagement of doing hundreds of exercises, individually adapted to the level of each student, is probably the facilitating feature of the EPGY computer courses responsible for the positive results. In summary, what is important is to have a good measure of the work done in a curriculum, such as the number of correct first-attempts in the EPGY curriculum.

It is less clear what benefit, if any, there is for students in the EPGY program who do not work at it sufficiently as measured by their CFA values. Figure 5 shows that for low CFA values there are both positive and negative E minus C differences on CST07 scores, but Tables 4 and 5 both show that the average of these differences is *negative*, in favor of the C group when CFA is low. This indicates the importance of (a) identifying those students who are not engaged in and working on the EPGY curriculum and (b) attempting to focus their interest on it.

3.9 Justifying the use of CFA to stratify the students

From a theoretical perspective, it can be argued that it is inappropriate to stratify the (E, C)-pairs on the CFA-value of the E member because the value of CFA is student-determined and occurs *after* treatment assignment, that is, it is a *post-treatment covariate*, and only observed on the students in the E group. The primary concern is that by stratifying on the CFA value of E we are also potentially stratifying on those pairs where the E member *tends to be* a higher ability student in mathematics than the C member. In so far as the pairs are matched on their CST06 scores, this concern is muted because the CST06 and 07 tests are very similar, measure the same things and are well correlated year to year (for example, for the 572 pairs of Grade 3 to 5 students, the correlation between CST06 and CST07 is 0.73). However, it is still possible that there is a small selection effect of stratifying on the CFA score of the E member of a pair that could introduce some bias in favor of the E group.

To examine this concern more closely, we plotted the CST06 score differences for the pairs against the CFA values. This is displayed in Figure 6. There is a small positive slope of .0022 which is about an eighth of the size of the slope of the CST07 differences on CFA. However, neither the slope nor the whole estimated regression function is statistically significantly different from zero, $F_{(1,570)} = 2.61$, $p = 0.11$.

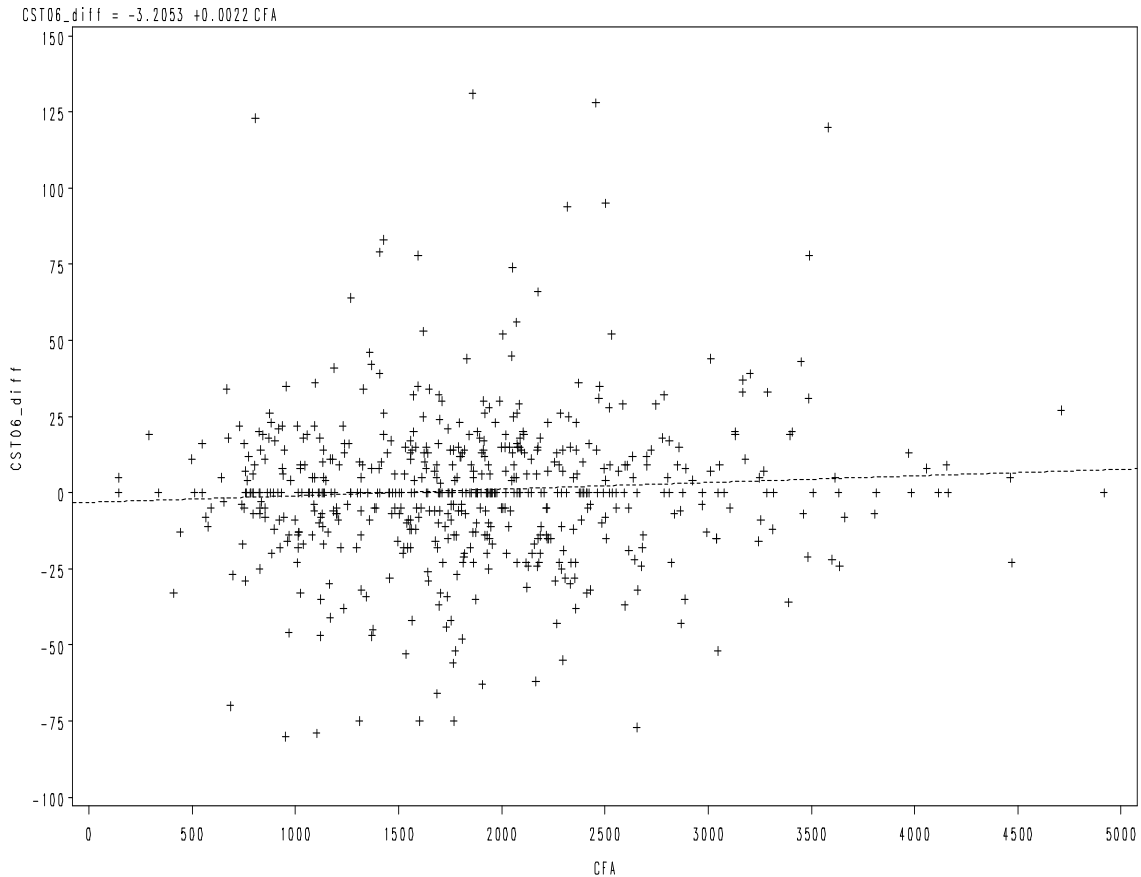


Figure 6. Scatter plot of E minus C difference on CST06 versus CFA of E student for 572 match pairs of students in Grades 3 to 5. R-square is 0.0046, slope is 0.0022.

Digging more deeply, in Table 19, we exhibit the means of the CST07 differences for the pairs in the top quartile and the top half of the CST06 scores values for the E member of each pair. While these differences are positive, they are not statistically significant at the usual levels. Thus, it is not simply differential mathematics ability that is responsible for the effects that we are seeing in the comparison between the E and C students.

For these reasons, we are not concerned that stratifying on the CFA values of the E members of a pair introduces an important bias in our analyses.

Table 19. . Summary of the comparisons between E and C matched students in grades 3 to 5 on the paired t-test, grouped by the CST06 score of the E member of each pair.

Group of the 572 pairs of Grade 3 to 5 students in all 8 schools	Number of pairs	Mean of the CST07 difference within each pair	SD of mean differences	Paired t-test p-value
Top quartile ranked on Math CST06	142	9.17	84.31	0.20
Top half ranked on Math CST06	281	3.41	75.50	0.45

Part II

4. REGRESSION MODELS ONLY FOR THE EPGY STUDENTS

In this section we examine several types of regression models for predicting CST07 scores from CST06 scores and various EPGY variables. These analyses necessarily restrict our focus on subsets of the 919 EPGY students in the study who attempted at least one EPGY exercise. Because we wanted both CST06 and CS07 scores, this restricted the sample further to the 619 E students who had both scores (see Table 1).

4.1 CFA (WCFA) and CST06 as predictors of CST07

This subsection reports several regression analyses that predict CST07 from CST06 and CFA or WCFA values. These analyses were done for all 619 students and separately for each school, as well.

The 2007 Math CST was modeled as a linear function of 2006 Math CST and the number of correct first-attempts using a regression model of the form

$$CST07_i = \beta_0 + \beta_1 CST06_i + \beta_2 CFA07_i + e_i$$

where $CST07_i$ is student i 's CST07 score, and $CFA07_i$ is the cumulative number of correct first-attempts of student i in 2006-2007.

The overall F-test for the model was used to determine if the regression was significantly different from a constant function. The t-test was used to examine the statistical significance of each covariate. The F-test results are shown under the "Model description" column in Table 20; and the t-test results are shown in the "Parameter Estimates" column.

As seen in Table 20, the regression models show a consistent positive relationship between CST07 scores and students' EPGY work for all schools, every district and every

school in the Effectiveness Study. The regression coefficients for CFA were positive and statistically highly significant with a very small p-value of less than 10^{-4} for all districts and all but two of the schools. These results show that, independent of where they started (i.e., their prior CST06 score), the more exercises a student did correctly on his or her first try, the higher CST07 scores he or she got.

Table 20. Regression model estimates of CST07 predicted by CFA and CST06.

School	Model description				Parameter estimates			
	N	F-value	p-value	R-square	CST06		CFA	
					Regression coefficient	p-value	Regression coefficient	p-value
All schools	619	441.7	$<10^{-100}$	0.59	0.74	$<10^{-100}$	0.03	$<10^{-10}$
District 1	348	261.3	8.18×10^{-70}	0.60	0.77	$<10^{-50}$	0.04	$<10^{-10}$
School A	84	55.1	$<10^{-10}$	0.58	1.10	$<10^{-10}$	0.01	0.38
School B	78	64.6	$<10^{-10}$	0.63	0.51	1.58×10^{-9}	0.05	1.23×10^{-8}
School C	102	87.7	$<10^{-20}$	0.64	0.65	$<10^{-10}$	0.06	$<10^{-10}$
School D	84	78.4	$<10^{-10}$	0.66	0.85	1.29×10^{-20}	0.01	0.14
District 2	104	85.0	$<10^{-20}$	0.63	0.64	$<10^{-20}$	0.02	3.00×10^{-3}
School E	56	65.4	$<10^{-10}$	0.71	0.64	$<10^{-10}$	0.05	4.97×10^{-6}
School F	48	43.8	$<10^{-10}$	0.66	0.60	$<10^{-10}$	0.01	0.15
District 3	167	164.1	7.18×10^{-40}	0.67	0.78	$<10^{-30}$	0.02	3.46×10^{-4}
School G	103	95.6	$<10^{-20}$	0.66	0.77	$<10^{-20}$	0.01	2.40×10^{-2}
School H	64	65.2	$<10^{-10}$	0.68	0.76	$<10^{-10}$	0.02	6.00×10^{-3}

One concern with the use of a variable like CFA is its skewed distribution, as seen in Figure 3. Moreover, its units are in a very different scale than the scale of the CST06 scores. To address both these issues we transformed both the CST06 score and the CFA values to Normal scores so that both variables mimicked the $N(0, 1)$ distribution. The results are given in Table 21.

Table 21. Regression model of CST07 predicted by N(0,1) values of CFA and CST06.

School	Model description				Parameter estimates			
					Normalized CST06		Normalized CFA	
	N	F-value	p-value	R-square	Regression coefficient	p-value	Regression coefficient	p-value
All Schools	619	441.71	$<10^{-100}$	0.59	58.0	$<10^{-100}$	20.30	$<10^{-10}$
District 1	348	261.3	8.18×10^{-70}	0.60	60.70	$<10^{-50}$	29.60	$<10^{-10}$
School A	84	55.11	$<10^{-10}$	0.58	86.70	$<10^{-10}$	8.20	0.38
School B	78	64.64	$<10^{-10}$	0.63	40.40	1.58×10^{-9}	34.90	1.23×10^{-8}
School C	102	87.68	$<10^{-20}$	0.64	51.00	$<10^{-10}$	47.80	$<10^{-10}$
School D	84	78.39	$<10^{-10}$	0.66	66.70	1.29×10^{-20}	9.60	0.14
District 2	104	85.02	$<10^{-20}$	0.63	50.40	$<10^{-20}$	16.50	3.00×10^{-3}
School E	56	65.41	$<10^{-10}$	0.71	50.30	$<10^{-10}$	41.50	4.97×10^{-6}
School F	48	43.83	$<10^{-10}$	0.66	47.50	$<10^{-10}$	10.40	0.15
District 3	167	164.14	7.18×10^{-40}	0.67	60.90	$<10^{-30}$	12.30	3.47×10^{-4}
School G	103	95.57	$<10^{-20}$	0.66	60.40	$<10^{-20}$	9.40	0.02
School H	64	65.17	$<10^{-10}$	0.68	59.70	$<10^{-10}$	17.10	6.00×10^{-3}

By and large the results shown in Table 21 are very similar to those in Table 20. By putting both CST06 and CFA onto similar scales we can compare the size of their coefficients more sensibly. Overall, the coefficient of CFA is about a third of the size of

the coefficient of CST06, but this varies from district to district. This means that the effect of a unit change in the value of CFA on the prediction of CST07 is about one third the size of a unit change in the value of CST06. Hence, while the CST06 is a better predictor of CST07, the effect of CFA is not negligible even when CST06 is included in the equation.

In Table 22 we show regression analyses like those of Table 20, but with CFA replaced by the difficulty-weighted WCFA, see section 3.2 for more on WCFA. In Table 23 we show the corresponding analyses when CST06 and WCFA have been normalized to mimic the $N(0, 1)$ distribution using Normal scores.

Table 22. Regression model of CST07 predicted by WCFA and CST06.

School	Model description				Parameter estimates			
					CST06		WCFA	
	N	F-value	p-value	R-square	Regression Coefficient	p-value	Regression Coefficient	p-value
All Schools	619	421.83	$<10^{-100}$	0.58	0.45	1.58×10^{-20}	0.03	$<10^{-10}$
District 1	348	217.84	$<10^{-60}$	0.56	0.46	3.02×10^{-9}	0.03	8.95×10^{-10}
School A	84	58.53	$<10^{-10}$	0.59	0.79	2.99×10^{-4}	0.03	0.06
School B	78	38.6	$<10^{-10}$	0.51	0.21	0.18	0.04	1.19×10^{-3}
School C	102	69.56	$<10^{-10}$	0.58	0.12	0.395	0.04	5.74×10^{-8}
School D	84	80.69	5.27×10^{-20}	0.67	0.65	3.66×10^{-7}	0.02	0.01
District 2	104	91.63	$<10^{-20}$	0.65	0.35	2.81×10^{-4}	0.03	2.21×10^{-4}
School E	56	50.18	$<10^{-10}$	0.65	0.30	0.029	0.04	7.60×10^{-4}
School F	48	45.14	$<10^{-10}$	0.67	0.41	3.41×10^{-3}	0.02	0.09
District 3	167	158.09	$<10^{-30}$	0.66	0.62	1.13×10^{-13}	0.01	3.13×10^{-3}
School G	103	95.97	$<10^{-20}$	0.66	0.62	3.17×10^{-9}	0.01	0.02
School H	64	56.52	10^{-14}	0.65	0.64	3.93×10^{-6}	9.00×10^{-3}	0.18

Table 23. Regression model of CST07 predicted by N(0,1) values of WCFA and CST06.

School name	Model description				Parameter estimates			
					Normalized CST06		Normalized WCFA	
	N	F-value	p-value	R-square	Regression Coefficient	p-value	Regression Coefficient	p-value
All Schools	619	421.83	$<10^{-100}$	0.58	35.70	1.58×10^{-20}	30.80	6.15×10^{-16}
District 1	348	217.84	$<10^{-6}$	0.56	36.10	3.02×10^{-9}	37.90	8.95×10^{-10}
School A	84	58.53	$<10^{-10}$	0.59	62.00	2.99×10^{-4}	32.30	0.06
School B	78	38.6	$<10^{-10}$	0.51	16.80	0.18	46.80	1.19×10^{-3}
School C	102	69.56	$<10^{-10}$	0.58	9.20	0.40	53.30	5.74×10^{-8}
School D	84	80.69	5.27×10^{-20}	0.67	51.20	3.66×10^{-7}	22.30	0.06
District 2	104	91.63	$<10^{-20}$	0.65	27.80	2.81×10^{-4}	35.20	2.21×10^{-4}
School E	56	50.18	$<10^{-10}$	0.65	23.90	0.03	52.70	7.6×10^{-4}
School F	48	45.14	$<10^{-10}$	0.67	32.20	3×10^{-3}	21.30	0.09
District 3	167	158.09	$<10^{-30}$	0.66	48.90	$<10^{-10}$	15.10	3.13×10^{-3}
School G	103	95.97	10^{-20}	0.66	48.80	3.17×10^{-9}	14.50	0.02
School H	64	56.52	$<10^{-10}$	0.65	50.00	3.93×10^{-6}	11.70	0.18

As in the previous comparison, the effects of normalizing made only small changes in the results, but the coefficients on the normalized values of WCFA are more comparable in size to those of the normalized values of CST06 than we saw in Table 21.

4.2 Results from a Title I school in District 2 that was outside the Effectiveness Study but which had EPGY students

There were 143 students with CST07 scores at this school. The average score was 342.10 points on the CST. The standard deviation was 73.79. The mean number of correct first-attempts responses was 1168.62 with a standard deviation of 517.63, and with a range from 129 to 2620 such responses.

A simple linear regression was used to examine the relationship between 2007 Math CST and the CFA values for these students. The result shows a strong positive relationship between these two variables. For every 100 correct first-attempts, students increased their 2007 Math CST by 4.92 points. This result is statistically significant: $p=2.39 \times 10^{-5}$.

4.3 Regressions adding the number of correct second-attempts, CSA, as a predictor of CST07

By including correct second-attempts as a predictor in the regression model described in Section 4.1, the enlarged model can perhaps explain more variability in the CST07 scores. Our model is thus

$$CST07_i = \beta_0 + \beta_1 CST06_i + \beta_2 CFA07_i + \beta_3 CSA07_i + e_i,$$

where $CST06_i$ is a student i 's CST Math score in 2006, $CST07_i$ is a student i 's CST Math score in 2007, $CFA07_i$ is the cumulative number of correct first-attempts of student i in 2006-2007, and $CSA07_i$ is the cumulative number of correct second-attempts of student i in 2006-2007.

This model, as in the case of that of Section 5.1, was applied to the data of 619 EPGY students. The R-square of the original model was 0.589. The R-square of this new

model was 0.600. Thus there was about a 1% increase in the fit. In term of the regression coefficients, correct first-attempts remained positively significant (0.04) with $p < 10^{-10}$. Correct second-attempt responses however was negatively significant (-0.15) at the p-value of 5.3×10^{-5} . The coefficient for CSAs is not surprising because students with a large value of CSA must have made an incorrect response on their first attempt. For this reason we would expect lower values of CSA to be associated with higher CST07 scores. As we mentioned earlier, the interpretation of CSA scores is complicated because of their connection to *incorrect* first attempts.

4.4 Correlation between CST06 and CFA

The data we consider here are given in Table 24.

Table 24. Data on CST06 and CFA.

Variable	N	Mean	Std	Minimum	Maximum
CFA	619	1843.44	766.61	145	4917
CST06	619	355.68	78.59	194	600

Because CST06 scores and CFA values in Section 4.1 are both positively related to CST07 scores, it is desirable to examine the correlation between these two variables for collinearity. We expect a weak positive relationship between the CST06 and CFA since EPGY students took the CST test in May 2006 before they began using EPGY program in November 2006. The Pearson correlation coefficient between these two variables confirms our hypothesis ($\rho=0.24$, $p<.0001$).

From the correlation and the two standard deviations, we may obtain the regression coefficient for predicting the CFA values from the students' CST06 scores. This is $(0.24)(776.61/78.58) = 2.37$.

However, this calculation does not take account of the fact that students were clustered within classrooms and schools. To examine the effect of this clustering of students we used a Hierarchical Linear Model to examine the relationship between CFA and CST06. In this model, CFA is the dependent variable and CST06 is the independent variable. This HLM model has three levels – school, classroom and students, and the intercept is modeled as random. The combined equation for this model is

$$CFA07_{ijk} = \delta + \delta_1 CST06_{ijk} + u_k + v_{jk} + e_{ijk},$$

where $CST06_{ijk}$ is 2006 Math CST score of student i in classroom j at school k , and $CFA07_{ijk}$ is correct first-attempts for student i in classroom j at school k

The results are presented in Table 25 and Table 26.

Table 25. Fixed effects of hierarchical model predicting CFA from CST06.

Parameter	Variable	Estimate	Std Err	DF	t-Value	p-value
δ	Intercept	1265.96	177.92	7	7.12	2.00×10^{-4}
δ_1	CST06	1.50	0.363	633	4.13	4.06×10^{-6}

The result of fixed effects shows a positive relationship between correct first-attempts and 2006 Math CST that is highly statistically significant. The estimate of 1.50 for δ_1 indicates that for every one point increased in the 2006 Math CST, there are 1.50 more exercises on average that students get correct on the first try.

Table 26. Random effects of hierarchical model predicting CFA from CST06.

Parameter	Level	Variance Estimate	Std Err	Z-value	p-value
u_k	School	97286	62721	1.55	0.06
v_{jk}	Classroom	112793	30325	3.72	9.98×10^{-5}
e_{ijk}	Student	421902	24753	17.04	$< 10^{-60}$

Significant p-values for classroom and student show there is substantial variance between classrooms and between students.

4.5 Correlation between *re-scaled* correct first-attempts and Math CST scores

The second re-scaling of the CFA we consider in this report is to divide $AGPA_{ij}$ by LAT_{ij} , the latency_{ij} of student i in correctly working exercise j . The intuition behind this is that faster work, ie., a shorter latency on correct first attempts, is a measure of greater mastery. We combine this with the reward for correct responses on exercises that are slotted for a higher grade-placement than the initial placement of the student that defines the WCFA value. We define:

$$\frac{WFCA}{LAT} = \sum_i \frac{AGP_{ij}}{LAT_{ij}}.$$

Table 27 gives the correlation of CFA, WCFA and $\frac{WFCA}{LAT}$ with CST07 for subsets of E

students. For CFA, the subset, A, is the top half of the E students based on their CFA values, for WCFA, it is the top half, B, of the E students based on their WCFA values,

and for $\frac{WFCA}{LAT}$ it is the top half, C, of the E students based on their $\frac{WFCA}{LAT}$ values.

Table 27. Pearson correlation coefficients for CFAs and two re-scaled CFAs for the top half of EPGY students as measured by these different re-scalings.

Group and selected variable	Pearson correlation coefficients between CST07 and selected variables	p-value
Group A:CFA	0.199	4.0×10^{-4}
Group B: CFA weighted by AGP	0.611	$< 10^{-30}$
Group C: CFA weighted by AGP and latency	0.543	3.57×10^{-20}

As can be seen in Table 27, the AGP weighting has a surprisingly large increase in correlation with Math CST07, from 0.20 to 0.61. Dividing by the latency *decreases* the correlation to 0.54, contrary to the intuition that faster means greater mastery.

That the AGP weighting of CFA turns out to be more highly correlated with the CST07 scores than the unweighted CFAs fits the intuition that giving greater weight to correctly worked exercises that are at a higher placement level in the curriculum produces a measure that is more closely related to mathematics achievement as measured by the CST07. The reader is reminded that EPGY’s K-5 Math course is highly individualized, so that the grade placement at which a student is working does not depend on the grade placement of other students in the class but does depend on his or her initial grade placement, the work he or she does in the curriculum and his or her mathematics achievement level.

The effects of the use of CFA or WCFA to select the top half of the E students can be seen in Figure 7, where CFA is the scale on the horizontal axis and WCFA the scale on the vertical axis. All EPGY students taking the Math CST07 Test in the study are used in the scatter plot.

Top Half:scatterplot

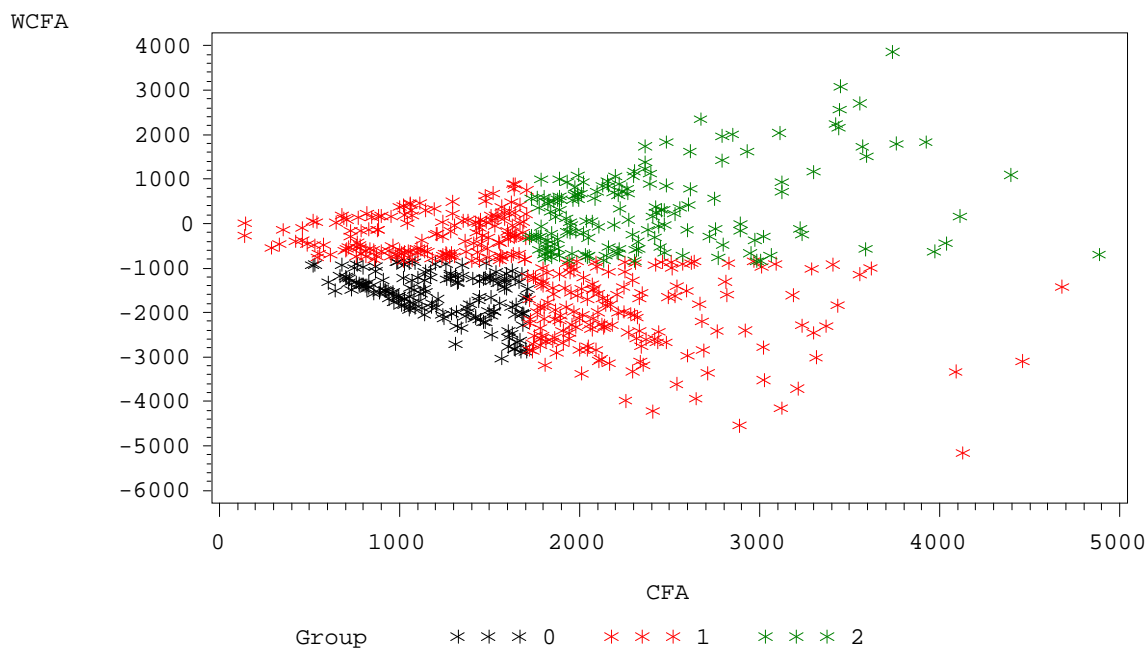


Figure 7. Scatter plots of the three groups of students: **green** for those who were in top half on both CFA and WCFA; **red** for those in top half of CFA or WCFA, but not both; and **black** for students in neither top half of CFA or WCFA.

Note that the re-scaling strongly affects the selection of the top half, as seen by the large number of red asterisks (*) in the scatter plot. The plot also shows that there is very little relationship between CFA and WCFA across the students, except for and strong increasing change of variance in WCFA as CFA increases.

4.6 Analysis of the 2006 to 2007 gain on Math CST by CST06 scores and CFA values

To examine the joint effect of CST06 scores and CFA values on the gains on the CST from 2006 to 2007, we grouped CST values into 50 point intervals and CFA values into 500 point intervals and formed the table of mean differences in the CST gains from 2006 to 2007. Cells with 5 students or less were eliminated (this removed 45 of the 619 E-students with both CST06 and 07 scores.) The results are displayed in Table 28.

Table 28 provides another view of the 2007 test results that support earlier evidence of the significant positive effect of students working diligently and carefully in the EPGY online course.

5. PREDICTING THE PROFICIENCY LEVELS OF THE EPGY STUDENTS USING MULTIVARIATE CLASSIFIERS

The data for this multivariate analysis consisted of information from the 619 EPGY students having both a CST06 and 07 score. Our goal was to predict the CST07 proficiency levels of the E students based on the following four covariates that are available for these students.

1. 2006 Math CST scores (x_1).
2. Number of correct first-attempts (x_2).
3. Adjusted final grade placement (x_3), i.e., the difference between a student's *final* grade placement in the EPGY mathematics curriculum and his or her actual classroom grade level.
4. Average latency on correct first-attempts (x_4), i.e. the student's average response time on all EPGY exercises answered correctly on the first attempt.

The simplest classifier is to use the 2006 proficiency-level classification as a prediction of the 2007 classification into proficiency levels. Table 29 gives the transition matrix of students' proficiency levels in 2006 and 2007. Each cell of the matrix contains two numbers, with the frequency at the top and the row percentage at the bottom. It shows that 54.1% of students were classified correctly in 2007 by their proficiency levels in 2006. Thus, a good classifier needs to do better than 54.1% correct classifications.

Table 29. Transition matrix for proficiency levels.

2006 level Frequency Percentage	2007 level				Total
	Advanced	Proficient	Basic	Below basic	
Advanced	92 65.25	36 24.83	15 10.00	2 1.09	145
Proficient	36 25.53	68 46.90	50 33.33	12 6.56	166
Basic	12 8.51	34 23.45	56 37.33	50 27.32	152
Below Basic	1 0.71	7 4.83	29 19.33	119 65.03	156
Total	141	145	150	183	619

5.1 Minimum-distance classifier

A minimum-distance classifier was the first classifier we used. First, we observed the outcomes and feature measurements in the training data set, and calculated the mean and covariance matrix for each class (proficiency level). Then, we computed the Mahalanobis distance from a unknown vector $\mathbf{x} = (x_1, x_2, \dots, x_4)^T$ to the k th class with mean $\boldsymbol{\mu}_k = (\mu_1, \mu_2, \dots, \mu_4)^T$ and covariance matrix P_k using the formula

$$d_M(x, \text{class } k) = \sqrt{(x - \boldsymbol{\mu}_k)^T P_k^{-1} (x - \boldsymbol{\mu}_k)}.$$

The vector of each student's properties as defined above was placed in the class whose multivariate distribution had the closest Mahalanobis distance to the vector.

We used all 619 students' information to construct the classifier.

Table 30 indicates that the model improves the classification significantly with 63.8% of students identified correctly, $p < 10^{-6}$. The comparison also shows that the classifier predicts 'Below Basic' and 'Advanced' much better than 'Basic' and

‘Proficient’. The worst case occurs at level ‘Basic’ with 47.33% of students classified correctly, which is still higher than the 37.33% in Table 29.

Table 30. Classification matrix of minimum distance classifier.

2007 Level Frequency Percentage	Predicted level				Total
	Advanced	Proficient	Basic	Below basic	
Advanced	112 79.43	24 17.02	5 3.55	0 0.00	141
Proficient	41 28.28	70 48.28	26 17.93	8 5.52	145
Basic	16 10.67	37 24.67	71 47.33	26 17.33	150
Below Basic	4 2.19	8 4.37	29 15.85	142 77.60	183
Total	173	139	131	176	619

5.2 Bayesian classifier

Instead of assigning a vector $\mathbf{x} = (x_1, x_2, \dots, x_4)^T$ to the class with minimum distance, a Bayesian classifier places it in the group with highest posterior probability. Bayes’ Theorem states that

$$P(H_j | E) = \frac{P(E | H_j)P(H_j)}{P(E)}$$

where H_j represents a one of several mutually exclusive hypothesis, $P(H_j)$ is the prior probability of H_j , $P(E|H_j)$ is the conditional probability of event E occurring given that H_j is true, this probability is called the likelihood of H on E , $P(E)$ is the marginal probability of event E , which can be calculated by $\sum_j P(E | H_j)P(H_j)$.

In our case, we wanted to predict to which classification a vector $\mathbf{x} = (x_1, x_2, \dots, x_4)^T$ belonged. We denote the posterior probability that x is assigned to class j as $P(\text{class } j | \mathbf{x})$, and assume a prior probability $P(\text{class } j)$ and a conditional distribution $f(x | \text{class } j)$ for class j . The posterior probability is computed in the following way:

$$P(\text{class } j | \mathbf{x}) = \frac{f(x | \text{class } j)P(\text{class } j)}{P(x)}.$$

Bayes' rule assigned the vector $x = (x_1, x_2, \dots, x_4)$ to the class j with maximum $P(x, \text{class } j)$.

Two main concerns here were to determine the prior and conditional distribution. The uniform prior was tried in our study. That is, we assumed a prior that the probability of classification had a uniform distribution $(1/4, 1/4, 1/4, 1/4)$. The conditional distribution $f(x | \text{class } j)$ was obtained from the estimated multivariate normal distribution.

$$f(x | \text{class } j) = (2\pi)^{-2} |\sum_{j,n}|^{-1/2} \exp\left\{-\frac{1}{2}(x - m_{j,n})' \sum_{j,n}^{-1} (x - m_{j,n})\right\}$$

where $m_{j,n}$ is the mean and $\sum_{j,n}$ is the covariance matrix for class j and n is the number of members in class j .

Table 31 provides the results obtained from the 619 students using the uniform prior. Compared to Table 29, the outcomes of the Bayesian classifier show significant improvement of classification with 66.6% of students classified correctly.

Table n Tables 31 and 30 show that the Bayesian model has much better performance for predicting the Basic level than the minimum distance classifier. There is no significant difference between the overall correct prediction rates of two classifiers.

Table 31. Classification matrix for Bayesian classifier using uniform prior.

2007 level Frequency Percentage	Predicted level				Total
	Advanced	Proficient	Basic	Below basic	
Advanced	104 73.76	29 20.57	8 5.67	0 0.00	141
Proficient	23 15.86	84 57.93	29 20.00	9 6.21	145
Basic	7 4.67	36 24.00	80 53.33	27 18.00	150
Below Basic	2 1.09	7 3.83	30 16.39	144 78.69	183
Total	136	156	147	180	619

5.3 Learning curves

The learning curve represents the relationship between the performance of the classifier and the size of training data. In our case, the correct classification rate of the test data was chosen to evaluate the classifier's performance. The mean learning curve was obtained by averaging over 1000 statistically independent runs. These methods are developed and described more extensively in Suppes and Liang (1998). The procedure for each run is as follows.

- (1) Form the test set by randomly selecting 10% of the elements from each class. The data remaining are available for training.
- (2) Form the original training set by randomly selecting samples with the same size for each class from the whole test population without replacement.
- (3) Construct the classifier on the current training set, obtain the predictions for the test data set, and compute the correct classification rate.
- (4) Generate new training sets by adding one trial of each class.

(5) Repeat step (3)&(4) till the training set size reaches the maximum.

In the above procedure, the classifier is updated recursively. With a vector $\mathbf{x} = (x_1, x_2, \dots, x_4)^T$ added to the old training data set containing n trials, the recursive learning rules for updating sample means and the sample covariance matrix for each class are the following:

$$m_{ij,n+1} = \frac{\delta_{j,n+1} x_{i,n+1} + \sum_{k=1}^n \delta_{j,k} m_{ij,n}}{\sum_{k=1}^{n+1} \delta_{j,k}}$$

$$S_{il,j,n+1} = \frac{\delta_{j,n+1} x_{i,n+1} x_{l,n+1} + (\sum_{k=1}^n \delta_{j,k} - 1) S_{il,j,n} + \sum_{k=1}^n \delta_{j,k} m_{ij,n} m_{lj,n} - \sum_{k=1}^{n+1} \delta_{j,k} m_{ij,n+1} m_{lj,n+1}}{\sum_{k=1}^{n+1} \delta_{j,k}}$$

where m_{ij} is the sample mean of feature x_i for class j , S_{il} is the covariance between two features x_i and x_l , n is the number of trials, and $\delta_{j,k}$ is an indicator function. If the k th trial falls in class j , it is 1. Otherwise, it is zero.

Figure 8 shows the mean learning curves for the two models we used for our study.

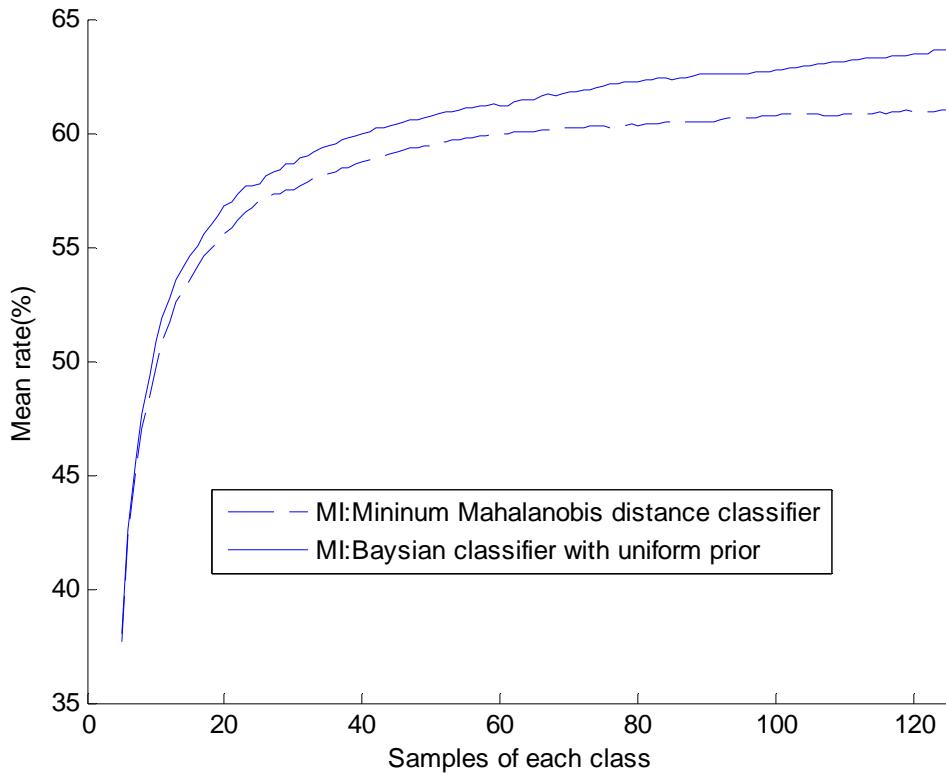


Figure 8. Learning curve for different models using test data set with fixed size (10% of each class).

The smallest training set in Figure 8 contains 5 samples for each class. The mean learning curves show that the performance of classifiers improves greatly when the training set size increases to 50 for each class. Both of them reach around 60% for the probability of a correct classification. Increasing the number further brings a small benefit. The mean rates are more than 61% for both of them when the number of samples reaches 100 for each category. The classifiers have similar learning ability in this case. The Bayesian classifier performs better than the Minimum Mahalanobis distance classifier.

Figure 9 represents learning curves obtained from test data with 219 trials. The procedure is the same as above except that the training data is formed by randomly

selecting 100 trials from each class in the step (1). In this case, the Bayesian classifier outperforms the minimum-distance model.

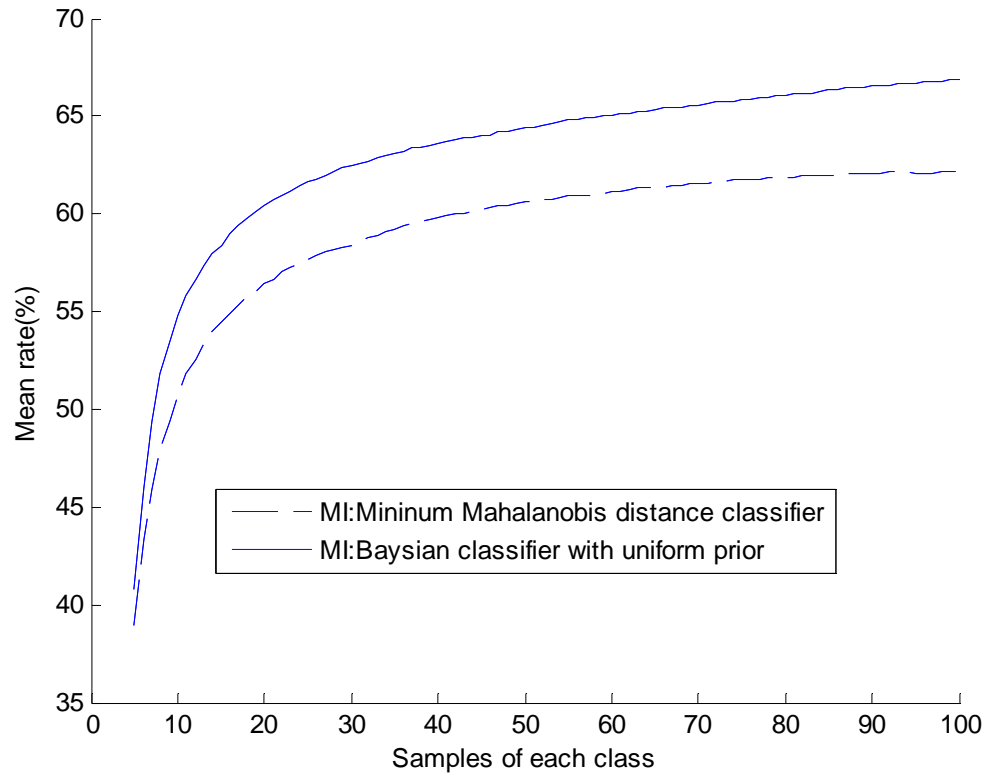


Figure 9. Learning curve for different models using test data set with fixed size (219 trials).

Compared to Figure 8, the classifiers get higher correct rates of predictions in Figure 9. The reason is that both classifiers have good performance in identifying “Advanced” students, who have higher test scores.

6. PREDICTION MODELS FOR EPGY STUDENTS ON THE MATH CST 2008

Here we predict the EPGY students' performance on the 2008 Math CST given their CST07 scores, their CFA values and their average latency per exercise from the beginning of the 2007-2008 school year. Two regression approaches were considered: Ordinary Least-square Regression (OLS) and the Hierarchical Linear Model regression (HLM).

OLS Model

For this prediction, we used the Ordinary Least-square Regression Model. Two steps were taken. The first step was to use the regression coefficients obtained in the 2006-2007 study.

$$CST07_i = \alpha + \beta_1 CST06_i + \beta_2 CFA07_i + \beta_3 LATENCY07_i + e_i$$

where $CST07_i$ is a student i 's CST Math score in 2007, $CFA07_i$ is the cumulative number of correct first-attempts of student i in 2006-2007, and $LATENCY07_i$ is the average latency in minutes of correct first-attempts by student i .

The second step was to predict 2008 Math CST scores by using the coefficient obtained above and substituting in the following equation.

$$CST08_i = \alpha + \beta_1 CST07_i + \beta_2 CFA08_i + \beta_3 LATENCY08_i + e_i$$

where $CST08_i$ is the *predicted* value of Math CST 2008 of student i , $CFA08_i$ is the cumulative number of correct first-attempts since July 15, 2007 until May 12, 2008 of student i , and $LATENCY08_i$ is the average latency in minutes of correct first-attempts by student i .

HLM Model

Similar to the steps above, we also used the Hierarchical Linear Regression Model of the 2006-2007 study, which took into account the variance between and within schools. The two-level model is presented as the following.

Level 1: Student level

$$CST07_{ij} = \pi_{0j} + \pi_1 CST06_{ij} + \pi_2 CFA07_{ij} + \pi_3 LATENCY07_{ij} + e_{ij}$$

where $CST07_{ij}$ is student i 's CST Math score in 2007 at school j , $CFA07_{ij}$ is the CFA value of student i at school j in 2006-2007, and $LATENCY07_{ij}$ is the average latency, in minutes, for correct first-attempts by student i at school j .

Level 2: School level. No predictor was available at this level and the coefficients are estimated directly in the 2007-2008 regression computation, allowed to be varying across schools.

$$\pi_{0j} = \beta_{00} + \nu_{0j}$$

Putting those two level models together, the HLM model in step 1 is

$$CST07_{ij} = \beta_{00} + \pi_1 CST06_{ij} + \pi_2 CFA07_{ij} + \pi_3 LATENCY07_{ij} + (\nu_{0j} + e_{ij})$$

The HLM model for prediction is

$$CST08_{ij} = \beta_{00} + \pi_1 CST07_{ij} + \pi_2 CFA08_{ij} + \pi_3 LATENCY08_{ij} + (\nu_{0j} + e_{ij})$$

6.1 OLS and HLM prediction models using correct first-attempts

Coefficients from California Standard Math Test 2006-2007

There were 619 students with CST06, CST07, CFA values, and latency values. The coefficients obtained by this regression model are presented in Table 32 and Table 33.

Table 32. Estimated coefficients of OLS model (2006-2007).

Parameter	Effect	Coefficient Estimate	Std Err	t value	p-value
α	Intercept	4.34	16.650	0.26	0.79
β_1	$CST06_i$	0.73	0.027	26.25	$< 10^{-100}$
β_2	$CFA07_i$	0.03	0.003	9.62	1.71×10^{-20}
β_3	$LATENCY07_i$	93.00	31.890	2.92	3.60×10^{-3}

Table 33. Parameter estimates of Hierarchical Linear Model (2006-2007).

Parameter	Fixed Effects	Estimate	StdErr	DF	t value	p-value
β_{00}	Intercept	-36.77	19.070	7	-1.93	0.10
π_1	$CST2006_{ij}$	0.71	0.027	608	26.22	$< 10^{-100}$
π_2	$CFA2007_{ij}$	0.04	0.004	608	10.87	$< 10^{-20}$
π_3	$LATENCY_{ij}^{07}$	175.20	32.990	608	5.31	1.53×10^{-7}
	Random Effects	Estimate	StdErr		z value	p-value
ν_{0j}	Between school variance	493.53	289.45		1.71	0.04
e_{ij}	Within school variance	2499.10	143.37		17.43	$< 10^{-60}$

Prediction result

The predicted 2008 test scores for 952 students with data in 2007 are shown graphically in Figure 10.

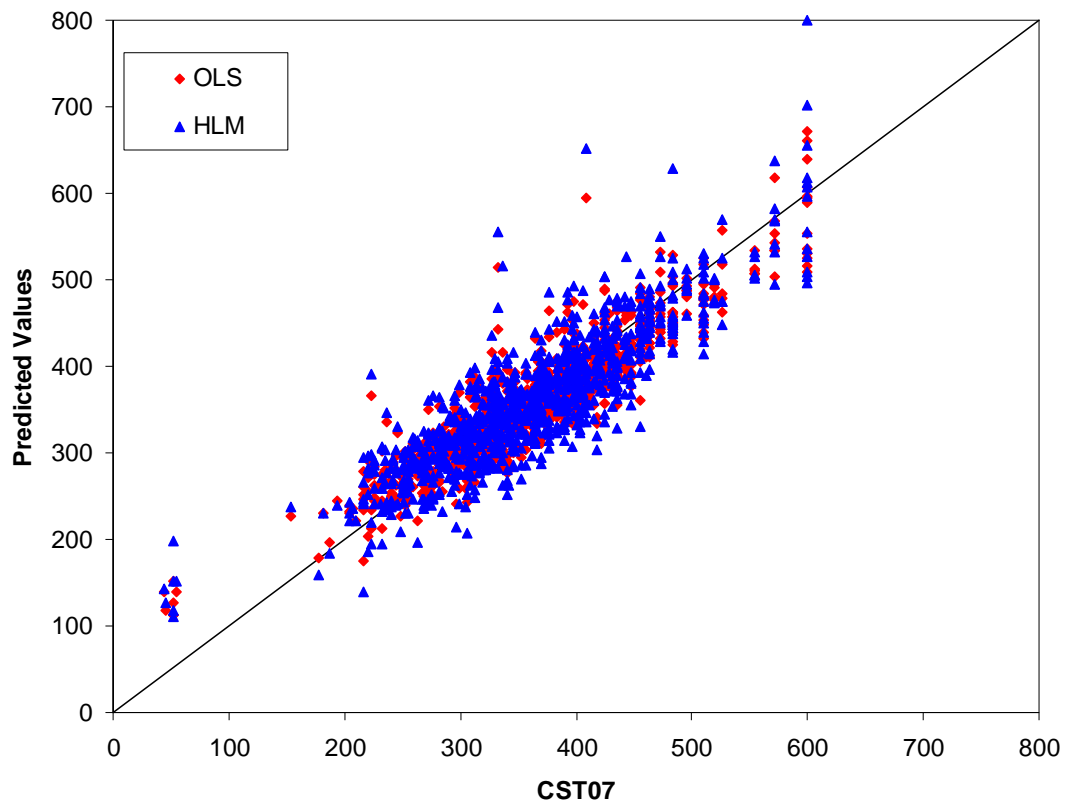
**Figure 10.** The comparison of predicted value 2008 CST for both the OLS and HLM regression models.

Figure 10 shows the predicted 2008 test scores from OLS and HLM models are closed to each other. Both groups clustered somewhat above the diagonal, as evidence of predicted improvement in scores. Table 34 summarizes the statistics of predictions. The t-test shows that the difference between predicted values obtained from OLS and HLM is not significant. Table 34 and Figure 11 show that the two normal distribution approximations of the prediction distributions are similar.

Table 34. Statistics of Effectiveness Study predictions obtained from the two different models.

Variable	N	Mean	Std	Minimum	Maximum
Prediction obtained from OLS	952	361.10	73.88	111.71	671.08
Prediction obtained from HLM	952	361.25	79.19	110.17	800.66

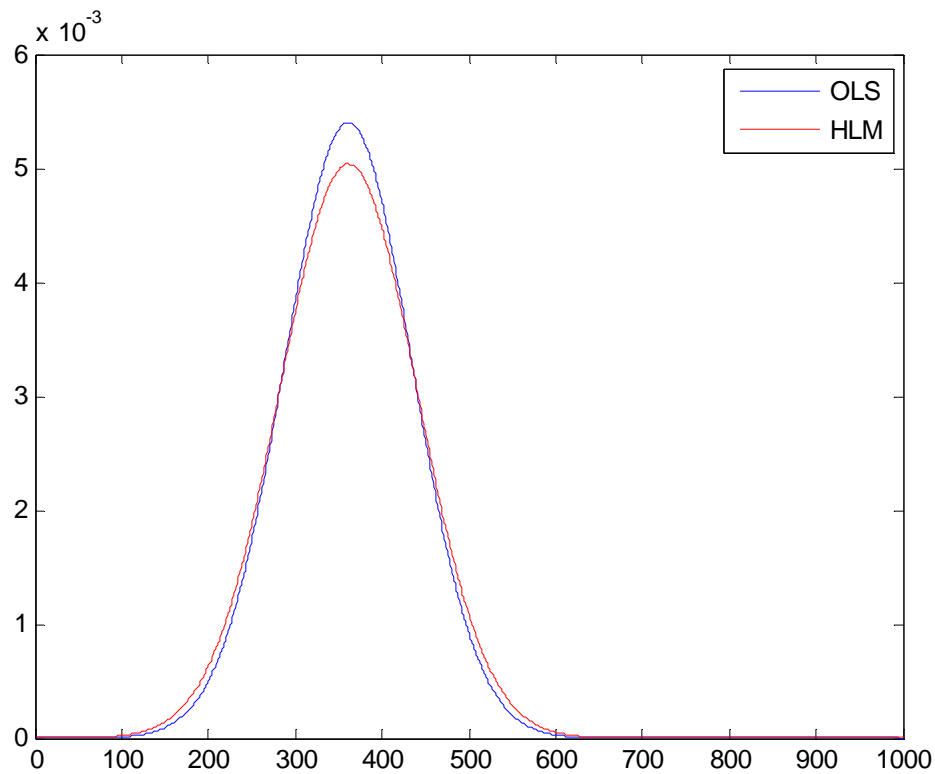


Figure 11. The normal-distribution approximations to the OLS and HLM predictions.

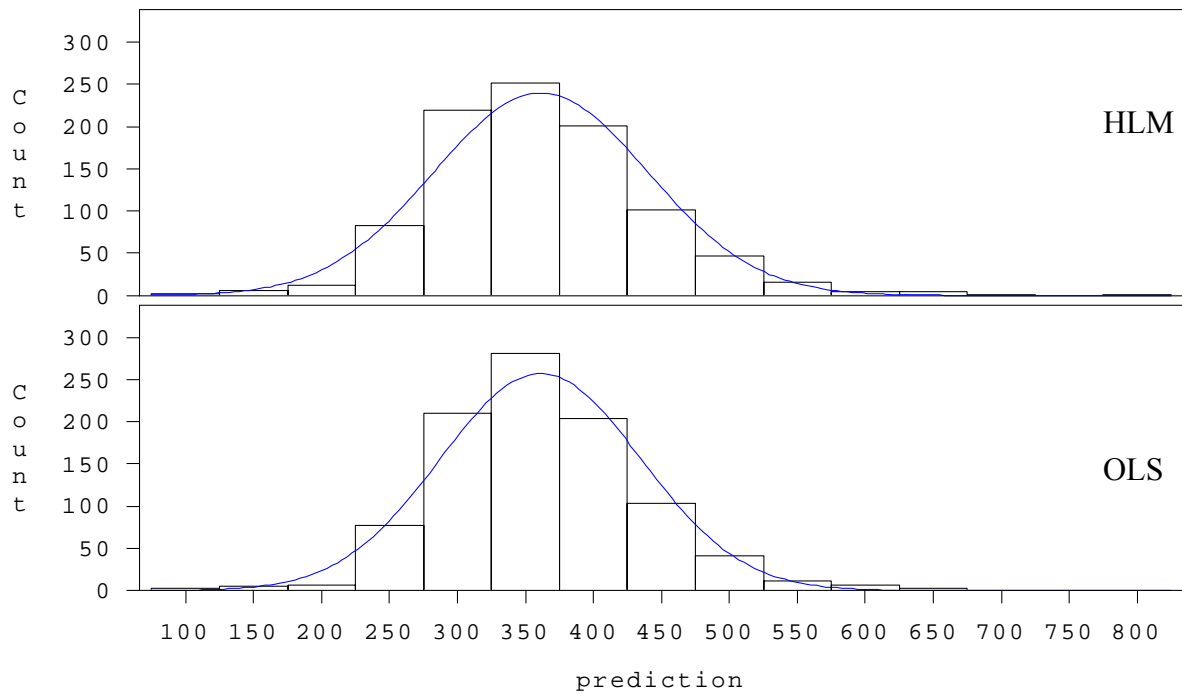


Figure 12. Histograms of the predictions from the OLS and HLM regression models, and their respective Normal approximations.

As can be seen in Figure 12, the normal-distribution approximations for the predictions do not reflect the slight skewness in the distributions of the predictions. This discrepancy is not directly relevant to the detailed predictions.

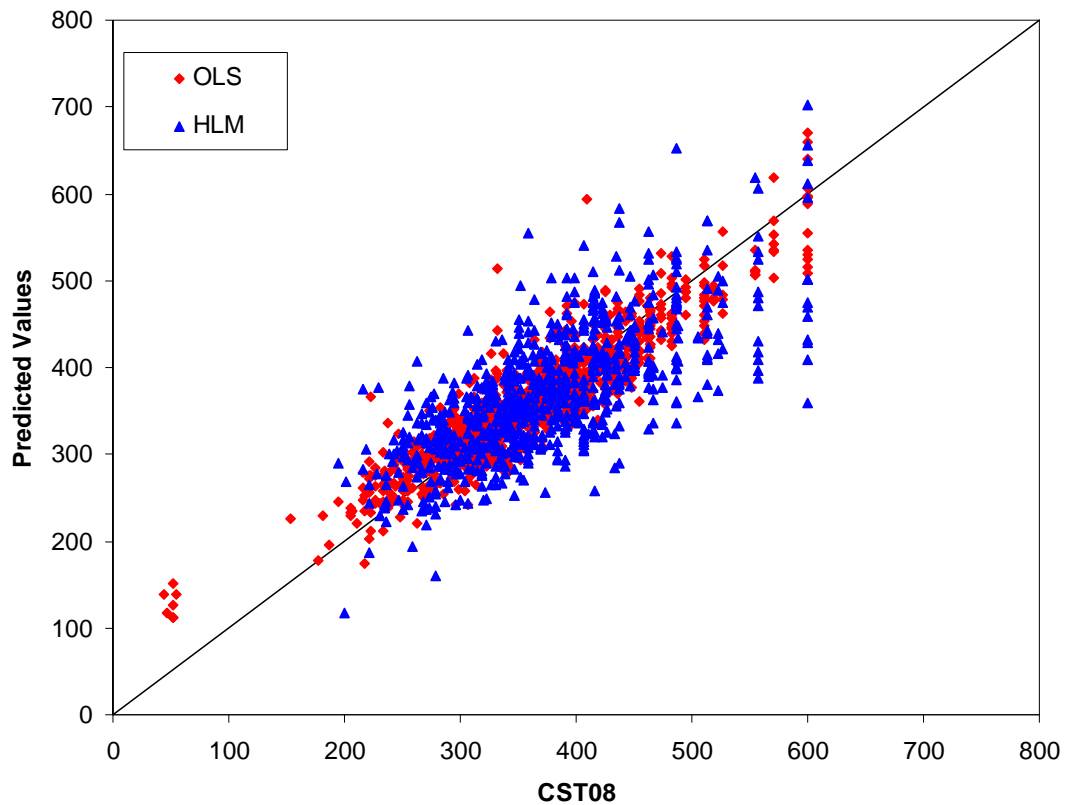


Figure 13. The comparison of predicted and actual 2008 CST.

Figure 13 shows the scatter plot of the predictions, with the actual math CST08 scores for 777 EPGY students shown on the horizontal axis and the predicted CST08 scores on the vertical axis. The figure suggests that the models under-predict, on average, actual scores. To examine this in more detail we formed the differences between the predictions and the actual values and then grouped their average values according to quartiles of the actual CST08 scores. This is displayed in Table 35.

Table 35. The difference between the actual CST08 scores and the predicted values when CFA is included as a predictor.

Quartile* of CST08	Number of students	OLS			HLM		
		Prediction – Actual CST08			Prediction – Actual CST08		
		Mean	Std	Std Err	Mean	Std	Std Err
Bottom quartile	194	25.58	40.46	2.90	25.64	44.33	3.18
3rd quartile	198	8.16	41.51	2.95	8.04	45.66	3.24
2nd quartile	184	-10.89	48.39	3.57	-10.67	52.46	3.87
Top quartile	201	-41.70	60.32	4.25	-38.22	64.22	4.53
All	777	-4.90	54.49	1.95	-3.96	57.45	2.06

*The numbers of students in each quartile are not equal because students with tied scores and at the border of two groups are assigned to the same group.

Both the OLS and the HLM predictions over-predict on average for the lower CST08 scores and under-predict on average for the higher CST08 scores. Some of this is simply due to regression to the mean because higher predictions are based, in part, on higher CST07 scores and lower predictions on lower CST07 scores. Overall, there is a slight under prediction for both methods.

Table 36. Number of students over-predicted or under-predicted for the predictions that include CFA as a predictor.

Quartile of CST08	Number of students	OLS		HLM	
		Under-predicted	Over-predicted	Under-predicted	Over-predicted
Bottom quartile	194	58	136	59	135
3rd quartile	198	86	112	84	114
2nd quartile	184	110	74	108	76
Top quartile	201	155	46	154	47
All	777	409	368	405	372

6.2 OLS and HLM prediction models that include WCFA as a predictor

We also considered OLS and HLM models with CST06, weighted correct first-attempts (WCFA, see section 3.2), and latencies as predictors. The coefficients obtained for these regression models are presented in Table 37 and Table 38.

Table 37. Estimated coefficients of OLS model (2006-2007) with WCFA.

Parameter	Effect	Coefficient Estimate	Std Err	t-value	p-value
α	Intercept	242.82	21.77	11.16	$< 10^{-20}$
β_1	$CST06_i$	0.45	0.047	9.61	1.81×10^{-20}
β_2	$WCFA07_i$	0.03	0.003	8.52	$< 10^{-10}$
β_3	$LATENCY07_i$	-76.72	28.07	-2.73	6.00×10^{-3}

Table 38. Parameter estimates of HLM model (2006-2007) with weighted correct first-attempts.

Parameter	Fixed Effects	Estimate	StdErr	DF	t-value	p-value
β_{00}	Intercept	231.94	22.66	7	10.23	1.84×10^{-5}
π_1	$CST2006_{ij}$	0.46	0.047	608	9.79	$< 10^{-20}$
π_2	$WCFA2007_{ij}$	0.025	0.003	608	8.30	$< 10^{-10}$
π_3	$LATENCY_{ij}^{07}$	-61.69	27.64	608	-2.23	0.03
	Random Effects	Estimate	Std Err		z value	p-value
ν_{0j}	Between school variance	224.28	137.95		1.63	0.05
e_{ij}	Within school variance	2697.58	154.69		17.44	$< 10^{-60}$

Figure 14 plots the predicted values against the actual values and suggests that these predictions substantially under predict the actual values. Comparing Table 39 with Table 34 shows that the predicted values from the new models are smaller than those from old models that used CFA rather than WCFA as a predictor. Detailed information is presented in Tables 40 and 41.

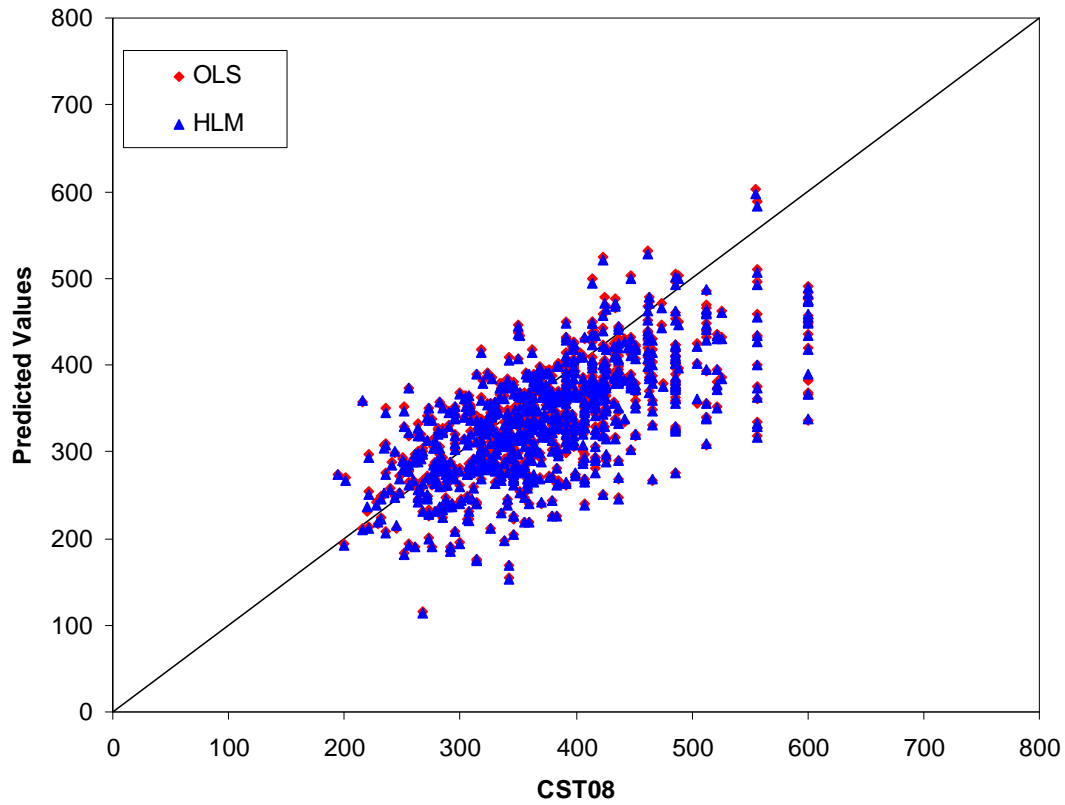


Figure 14. The comparison of predicted and actual 2008 CST when WCFA is used as one of the predictors.

Table 39. Statistics of Effectiveness Study predictions obtained from OLS and HLM models with weighted correct first-attempts.

Variable	N	Mean	Std	Minimum	Maximum
Prediction obtained from OLS	952	338.87	68.20	25.93	602.10
Prediction obtained from HLM	952	337.50	67.69	25.11	597.73

Table 40. The difference between the actual 2008 CST and the predicted values when WCFA is included as a predictor.

Quartile of CST08	Number of students	OLS			HLM		
		Prediction - Actual CST08			Prediction - Actual CST08		
		Mean	Std	Std Err	Mean	Std	Std Err
Bottom quartile	194	2.84	49.37	3.54	1.63	48.91	3.51
3rd quartile	198	-24.55	49.00	3.48	-25.89	48.60	3.45
2nd quartile	184	-43.21	46.32	3.41	-44.63	45.85	3.38
Top quartile	201	-77.33	65.35	4.61	-78.72	64.40	4.54
All	777	-35.78	60.72	2.18	-37.12	60.23	2.16

Table 41. Number of students over-predicted or under-predicted for the predictions that include WCFA as a predictor.

Quartile of CST08	Number of students	OLS		HLM	
		Under-predicted	Over-predicted	Under-predicted	Over-predicted
Bottom quartile	194	94	100	96	98
3rd quartile	198	130	68	135	63
2nd quartile	184	150	34	154	30
Top quartile	201	181	20	182	19
All	777	555	222	567	210

While both predictions tend to under predict the actual CST08 scores, those that included CFA values had 51% of the cases under predicted with an average under prediction of 4.9 CST points, while those that included WCFA had 71% of the cases under predicted with an average under prediction of 35.8 points.

7. CONCLUSIONS

A strong positive relationship between EPGY work and 2007 Math CST scores was found consistently for all Title I schools, each district, and each school. The more students worked carefully and in a sustained fashion, the higher the students scored on their 2007 Math CST. In particular, EPGY students in the top quartile or the top half, ranked by the number of correct first-attempts, performed significantly better than matched control students.

A clear graphic presentation representing these positive results for students is given in Table 28. All EPGY students whose number of correct first-attempts was greater than 2,000 (the mean number 1843.44) had higher test scores in 2007 than in 2006. Only 4 cells in the table below 2000 showed such an improvement, and these were all students with the lowest 2006 Math CST scores.

Acknowledgements.

For financial support to conduct this study, we are indebted to three corporations:

Tessera, Flextronics and *SanDisk*; as well as the following individuals: Bruce and Astrid McWilliams, Michael and Carole Marks, Tom and Johanna Baruch, and Tim Mott.

References

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) New York: Academic Press.
- Educational Testing Service (2008) California Standards Tests Technical Report, Spring 2007 Administration. Available from the CST website.
- West, B.T., Welch, K.B., and Galecki, A.T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. London, New York: Chapman & Hall/CRC.
- Paek, P., Holland, P., and Suppes, P. (1999). "Development and Analysis of a Mathematics Aptitude Test for Gifted Elementary School Students". *School Science and Mathematics*, 99: 228-247.
- Suppes, P. and Liang, L. (1998). Concept Learning Rates and Transfer Performance of Several Multivariate Neural Network Models. In *Recent Progress in Mathematical Psychology*. C.E. Dowling, F.S. Roberts, P. Theuns, eds. Mahway, NJ: Lawrence Erlbaum, pp. 227-252.

Appendix: METHODS OF STATISTICAL ANALYSIS

The basic question for the statistical analyses of Part I was how the students in the EPGY group performed relative to the students in the control group. Results for all schools together, individual districts and schools are given.

Three statistical approaches were used for comparison of experimental and control groups. First, paired sample t-tests were run to examine the difference between the EPGY and control groups at each level of aggregation. Effect sizes were also calculated using a modification of Cohen's *d* statistics (Cohen 1988). Second, a three-level hierarchical linear model (HLM) of student, classroom and school was used (West, Welch, and Galecki 2007). Third, student's changes in proficiency level were analyzed for statistical significance. The five proficiency levels used were those defined for the CST tests: Far Below Basic, Below Basic, Basic, Proficient, and Advanced.

A1 Effect sizes

Cohen's *d*

Cohen's *d* is an appropriate effect-size measure to use in the context of a t-test on means. "*d*" is defined as the difference between two means divided by the average standard deviations for those means. Because of the pairing, our two samples had the same size and thereby

$$d = \frac{mean_1 - mean_2}{\sqrt{(SD_1^2 + SD_2^2)/2}}$$

where $mean_i$ and SD_i are the mean and standard deviation for group *i*, for $i = 1, 2$. The standard interpretation of the effect size is that 0.2 is indicative of a small effect, 0.5 a medium and 0.8 a large effect size (Cohen 1988). These criteria are specific to

educational research, other fields, such as physics or quality control often use standards that imply much higher values of d as indications of the importance of an effect size.

Example. The mean of EPGY students was 432.44 with standard deviation of 63.87. The mean of control students was 385.44 with standard deviation of 63.44. So

$$d = \frac{432.44 - 385.44}{\sqrt{(63.87^2 + 63.44^2) / 2}} = 0.78.$$

An effect size of 0.78 is considered large, based on the standards stated above.

Our modification of Cohen's d

The problem with a statistic like d is that the standard deviations may vary from comparison to comparison depending on what the standard deviations are. We chose to use only two standard deviations as the denominators of d rather than letting them vary for each comparison. We used the standard deviation of the CST07 scores that were appropriate for the tests and grades we had in the study that were computed on all California students in that year. For second graders we used the standard deviation of all California second graders as obtained from the CST technical report ETS (2008). For the other comparisons that involved several grades pooled together, we used the median of the standard deviations of grades 2 to 5. Thus, all effect sizes in this report are mean differences divided by one these two standard deviations.

A2 Three-level hierarchical linear model

The basic question for this statistical analysis is whether students in the EPGY program had higher 2007 Math CST scores than those in the control group. With randomization at the student level, we expect 2006 Math CST scores to be equally distributed between treatment and control groups, but in any single randomization there

may be discrepancies between the distributions due to chance. In the model below, CST06 is included to increase the precision of the treatment effect. The dependencies among observations within classrooms and schools are also considered by modeling random effects (West, Welch and Galecki, 2007). This three-level hierarchical linear model of student, classroom and school is specified as follows.

Table A1. Three levels of student, classroom and school.

Level	Name	Factor and Covariate
I: $i = 1, 2, 3, 4, \dots$	Student within classroom	CST06 TX (1 for E and 0 for C)
II: $j = 1, 2, 3, 4, \dots$	Classroom within school	
III: $k = 1, 2, 3, 4, \dots$	School	

Level 1: Student level. Coefficients of CST06 and TX are fixed.

$$CST07_{ijk} = \pi_{jk} + \pi_1 TX + \pi_2 CST06 + e_{ijk}$$

Level 2: Classroom level. No predictor at this level and the coefficient of the intercept is allowed to vary within schools.

$$\pi_{jk} = \beta_k + v_{jk}$$

Level 3: School level. No predictor at this level.

$$\beta_k = \gamma + u_k$$

Putting these three models together, we have the following equation that relates students, classes, schools, covariates and the treatment condition (TX) to the outcome variable.

$$CST07_{ijk} = \gamma + \pi_1 CST06 + \pi_2 TX + u_k + v_{jk} + e_{ijk},$$

where γ, π_1, π_2 are the fixed effects and u_k, v_{jk}, e_{ijk} are the random effects for schools, classes within schools and students within classes, respectively..

A3 Binomial analysis of changes in proficiency level

The third type of statistical analysis is to compare the experimental and control groups on the *changes* in proficiency levels, as defined in the *No Child Left Behind* legislation and augmented for the CSTs in California. The null hypothesis in this case is that the numbers of positive changes (+1) in proficiency from the 2006 Math CST Test to the 2007 Test should be the same as the number of negative changes (-1). For example, a student who moved from Proficient in 2006 to Basic in 2007 counts as -1, and a student who moved from Proficient to Advanced counts as +1. The null hypothesis is that the chance of any student who changes his or her level moving up is the same as moving down, i.e., the probability of moving up or down is $\frac{1}{2}$, given a move was made. If N students in a unit changed, k of them moved up and $N-k$ moved down, then the probability of this occurring by chance (i.e., with $p = 0.5$) is:

$$P(x \geq k) = \sum_{j=k}^N P(x = j) = \sum_{j=k}^N \binom{N}{j} p^j (1-p)^{N-j} .$$