Educational Assessment, 18:3, 2013. pp. 162-181.

Effectiveness of an Individualized Computer-Driven Online Math K-5 Course in Eight California Title I Elementary Schools

Patrick Suppes
Stanford University

Paul W. Holland
Paul Holland Consulting Corporation

Yuanan Hu and Minh-thien Vu
Stanford University

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 technological and individualized EPGY Kindergarten through Grade 5 Mathematics Course Sequence, modified for the Title I schools. Restricting attention to students who were in the top half of the distribution of correct first-exercise attempts (a measure of work and engagement), we found substantial and statistically significant improvements in the 2007 California Standard Math Tests (CST07) scores compared to those of matched control students. Gains in second grade were larger than those in Grades 3 to 5. Less able students, as measured by their 2006 CST mathematics scores, also had, on average, larger gains.

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) in Suppes (1992, 1995a, 1995b). All eight participating schools had a full K-5 sequence of classroom instruction. Although the EPGY curriculum was originally developed for gifted students, its predecessor was developed specifically for Title I and other disadvantaged students, such as deaf students (Fletcher & Suppes, 1976; Jamison, Fletcher, Suppes, & Atkinson, 1976; Suppes, Fletcher, & Zanotti, 1976); American Indian students in Suppes, Fletcher, and Zanotti

Correspondence should be sent to Patrick Suppes, Center for the Study of Language and Information, Stanford University, 220 Panama Street, Stanford, CA 94305-4101. E-mail: psuppes@stanford.edu

(1975); and other types of compensatory education in Jamison, Suppes, and Butler (1970), as well as in Suppes and Morningstar (1970). Those in this study were not selected as gifted, and so received a revised curriculum, which omitted the more difficult optional parts, with, in addition, some adjustment in the learning parameters governing individual student motion in the course (Suppes & Zanotti, 1996).

We also stress that this present experiment was not a test of new curriculum content. The curriculum for Grades K–8 is not meant to be new in content. The course goals and standards of the California State Department of Education were closely followed. The innovation was to test the technological efficiency of computer-managed individualization of the curriculum to increase students' learning and general progress in the course.[1]

Beginning with the introduction, this article is divided into four sections. In the second section we describe the methods, including the research design and data collection procedures. The next section concerns the main results of the experiment. It is 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.

The final section contains summary conclusions of the experiment and then some policy implications of this study, which involves intensive use of educational technology, but with essential classroom management by teachers playing a critical role.

METHODS

Research Design of RTE

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-mile radius of Stanford University. Within each participating class, entering students were randomly assigned to one of two treatment groups, E or 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 year (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. So these students were administered the Stanford EPGY Mathematical Aptitude Test (SEMAT), discussed in Paek, Holland, and 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 1,000 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

[1]The first author's research in the application of computer technology to education began still earlier in Hawley and Suppes (1959, 1960a, 1960b); Suppes and McKnight (1961); Suppes and Hill (1962); Suppes (1962); Jamison, Suppes, and Wells (1974); and Suppes (1992).

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 most schools, the pairings gave us a way to remove any bias it might introduce. If, for instance, an E student did not have CST07 scores at the end of the study for some reason (did not take the test, left the school, etc.), we deleted 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, Holland (1988). 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.

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 min a day, 5 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. In addition, 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 min a day of exposure to and work on the EPGY technology-driven

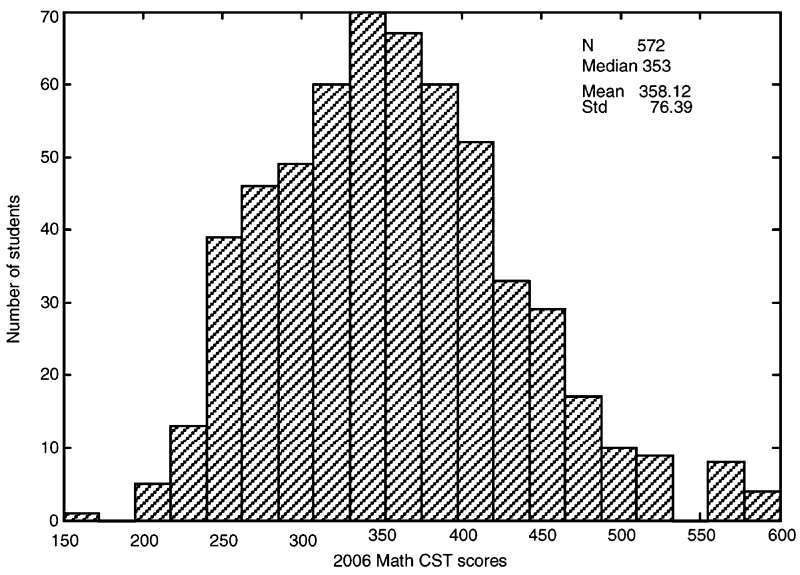

FIGURE 1 Histogram of the 2006 California Standard Math Test (CST) scores for the control group for the 572 matched pairs in which both members have both CST scores.

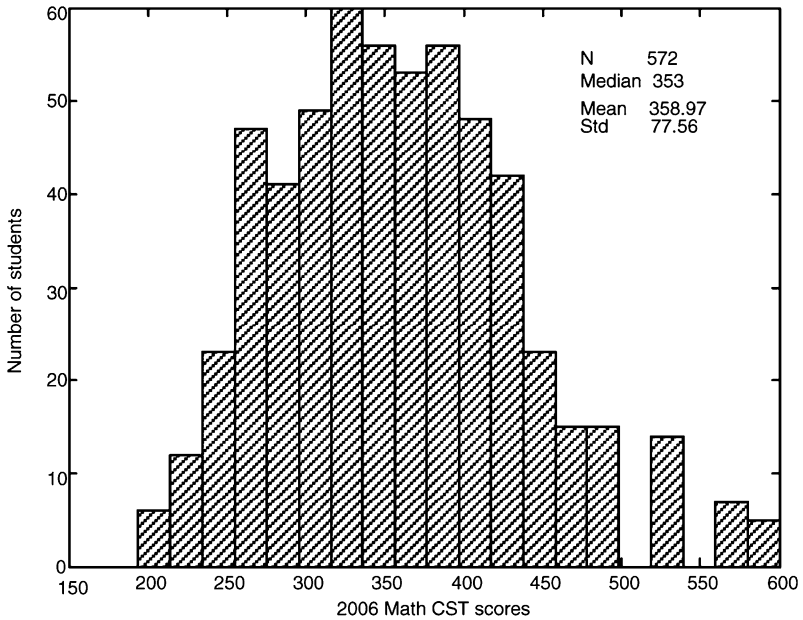

FIGURE 2 Histogram of the 2006 Math California Standard Math Test (CST) scores for the Education Program for Gifted Youth (EPGY) group for the 572 matched pairs in which both members have both CST scores.

version of the curriculum, whereas the C students continued for the same time studying the standard curriculum in the classroom. So the C and E students studied the standard mathematical curriculum approximately the same amount of time. Some initial Hawthorne effect may have benefited the E students, but the long length of the experiment being essentially the entire school yearly most likely minimized such possible effect, but we have no direct data in this study to support this view.

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 were not, of course, available for the C students.

Data Collection and the E/C Pairings

After the initial pairing and random assignment, there were 1,023 *pairs* in the study, with 27 in Grade 1, 194 in Grade 2 and 802 in Grades 3 to 5. Of the 1,023 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 classwork. 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, as they were not part of the effectiveness study.)

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 at least 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 just described, the 572 in Grades 3 to 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 later.

CST Mathematics Scores

At the end of the school year, students in Grades 2 to 5 took the 2007 Mathematics CST (CST07). The numerical data for each district and school are shown in Table 1.

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. Figure 1 (control) and Figure 2 (EPGY) show the distributions (histograms) of the Math CST06 scores for all of the 572 pairs in Grades 3 to 5 that had both CST06 and CST07 scores. In addition, the legend of each figure gives the median, mean, and standard deviation of the Math CST06 scores for that group. These graphs show the effectiveness of our method of removing the entire pair whenever one member of a 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 three to five grades that we used for some of our main analyses of the effectiveness of EPGY were as similar on their prior mathematics performance as were all of the initial pairs.

Course Performance of the EPGY Students

The experimental treatment consisted of computer-presented mathematics exercises given to the students, who then also 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 played an important role as a posttreatment covariate in many of our analyses. The *number of correct first attempts* (CFA) is the total number of exercises for which the EPGY student got the correct answer on his or her first attempt. 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.

TABLE 1
Number of Pairs With Math CST Scores for 2006 and 2007, by School and District

		Math CST06		*Math CST07*		*Math CST06 & Math CST07*	
Title I Schools	*No. of Pairs*[a]	*Both Members Have Scores*[b]	*Any Member Without Score*	*Both Members Have Scores*[c]	*Any Member Without Score*	*Both Members Have Both Scores*[b]	*EPGY Member Has Both Scores*[b]
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
School D	135	86	49	116	19	80	84
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

Note. CST06 = 2006 California Standards Mathematics Test; CST07 = 2007 California Standard Math Test; EPGY = Education Program for Gifted Youth.

[a]Total of all grades, 1 through 5. [b]Only Grades 3 to 5, Grades 1 to 2 students did not have CST06 scores. [c]Includes Grade 2 students most of whom had CST07 scores, but no Grade 1 students.

Table 2 gives details about the distribution of CFA values. 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.

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

TABLE 2
Summary of the Four Quartiles of the Distribution of CFA Values—Minimum, Maximum, Mean, and Standard Deviation

Quartile	*Min*	*Max*	*M*	*SD*
Q1[a]	2	1,325	903.2	310.6
Q2	1,326	1,825	1,612.8	142.3
Q3	1,828	2,287	2,033.3	128.5
Q4	2,290	4,917	2,803.9	503.6

Note. CFA = number of correct first attempts.

[a]Q1 = Lowest Quartile. $N = 230$.

TABLE 3
Descriptive Statistics of CFA for Some Subgroups of the 919 EPGY Students

Subgroup of the 919 EPGY Students Who Attempted at Least One Exercise	*No. of Students*	*Mdn*	*M*	*SD*
Both members of the pair in Grade 2 have a CST07 score.	170	2,098	2,117.9	554.9
Both members of the pair in Grades 3–5 have both a CST06 and CST07 score.	572	1,808.5	1,861.5	780.1
EPGY students in Grades 3–5 with CST06 and CST07 scores.	619	1,797	1,843.4	766.6

Note. CFA = number of correct first attempts; EPGY = Education Program for Gifted Youth; CST07 = 2007 California Standard Math Test; CST06 = 2006 California Standards Mathematics Test.

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 attempted and got correct on the first try. Alternative measures, such as the sheer number of exercises attempted, were not used because they do not capture the idea of the amount of *careful work* done. We expected that students who had higher CFA values would benefit more from exposure to the EPGY curriculum than those with lower values.

In using the CFA, we are not claiming it is necessarily the variable that is optimal for predicting the effect size of a given effort by a student to improve his knowledge and skills of elementary mathematics, but it is a natural measure of a sort widely used in classroom tests of mathematical competence, namely, the number of correct answers. In the present case, some correlated function might be a better predictor. For example, the learning rates in a sequence of learning models for responses of a student over several months might turn out to be better, or even a simple linear weighting of the later exercises might be so. We did not undertake the extensive statistical analysis required to study this problem with any thoroughness. But, we do believe our choice of CFA is reasonable, and very likely highly correlated with a variety of other choices of variables that make educational sense.

COMPARISON OF EPGY AND CONTROL STUDENTS ON THE 2007 MATHEMATICS CST

This section consists of seven subsections that examine various comparisons of the E and C students on their Math CST07 test scores. Then, in the next-to-last subsection we summarize our various findings on the relative effectiveness of the EPGY curriculum. Finally, in the last subsection we consider some issues that arise in the justification of stratifying the EPGY students on the value of their total correct first attempts.

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 who had both CST06 and CST07 scores for both pair members and for whom the E-member of the pair completed at least one EPGY exercise. The mean difference on the Math 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 two-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 shown in Table 2. 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 values are lower than those in the top fourth. The effect sizes corresponding to these t tests range from 15% to 21% of the median standard deviation (across Grades 2–5) for all California students.

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 nonsignificant negative difference in the third quintile and then increasingly positive differences that are statistically significant in the fourth and fifth (highest) quintiles.

There is considerable spread in the CST07 differences, but there is a slight, statistically significant ($p < .0001$), upward trend with a slope of 0.0177 that we saw in Table 5. Solving the regression equation, CST07-dif $= -32.984 + 0.0177$CFA, for the CFA value where the line is zero gives CFA $= 1863.5$. Looking back at 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 1,808. Thus, when the CFA value exceeds 1,863, the E students begin to have a slight edge

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	*No. of Pairs*	*M of the CST07 Difference for Each Pair*	*SD of the CST07 Difference for Each Pair*	*Effect Size Relative to SD = 83.5*[a]	*Two-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}
All pairs	572	0.05	67.62	0.00	.98

Note. E = Education Program for Gifted Youth; C = control; CFA = number of correct first attempts; CST07 = 2007 California Standard Math Test.

[a]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 Math CST07 test for all California students in Grades 2 to 5 and is used here as the denominator for the effect sizes.

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	*No. of Pairs*	*M of CST07 Differences*	*SD of CST07 Differences*	*ES (Relative to SD = 83.5[a])*	*Two-Sided Paired t Test p Value*
Highest CFA quintile	114	20.07	69.7	0.24	2.7×10^{-3}
4th CFA quintile	114	14.85	61.1	0.18	1.1×10^{-2}
3rd CFA quintile	115	−4.29	72.4	−0.05	.53
2nd 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	.98

Note. E = Education Program for Gifted Youth; C = control; CFA = number of correct first attempts; CST07 = 2007 California Standard Math Test.

[a]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 Math CST07 test for all California students in Grades 2 to 5 and is used here as the denominator for the effect sizes.

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.

Results for Second Graders

The results for Grade 2, similar to those described previously for Grades 3 to 5, are summarized in Table 6. Overall there is a slightly negative but statistically nonsignificant 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.

TABLE 6
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	*No. of Pairs*	*M of CST07 Differences*	*SD of CST07 Differences*	*ES (Relative to SD = 82.0[a])*	*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}
All pairs	170	−3.08	99.2	−0.04	0.69

Note. E = Education Program for Gifted Youth; C = control; CFA = number of correct first attempts; CST07 = 2007 California Standard Math Test.

[a]The value of 82.0 is the standard deviation of the scaled scores for the Math CST07 test for all California students in Grade 2 and is used here as the denominator for the effect sizes.

Unlike Table 5, all of the mean differences in Table 6 are positive. 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.

District and School Results

Given the earlier results, 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 7.

Table 7 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 7 are all positive and most of the p values for the t test were smaller than the usual standard of 0.05.

TABLE 7
Summary of Comparisons Between Matched E and C Students for Each District and School 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	*No. of Pairs*	*M of CST07 Differences*	*SD of CST07 Differences*	*ES (Relative to SD = 83.5[a])*	*Two-Sided Paired t Test p Value*
District 1	243	18.74	79.3	0.22	3.00×10^{-4}
School A	68	27.77	90.1	0.33	.01
School B	47	25.43	79.3	0.30	.03
School C	70	19.66	69.9	0.24	.02
School D	58	6.72	81.9	0.08	.53
District 2	64	8.20	83.9	0.10	.44
School E	31	36.36	75.5	0.44	.01
School F	34	13.36	79.2	0.16	.33
District 3	93	14.48	72.5	0.17	.05
School G	47	26.70	78.6	0.32	.02
School H	47	5.26	58.0	0.06	.54

Note. E = Education Program for Gifted Youth; C = control; CFA = number of correct first attempts; CST07 = 2007 California Standard Math Test.

[a]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 Math CST07 test for all California students in Grades 2 to 5 and is used here as the denominator for the effect sizes.

TABLE 8
Fixed Effects for the Top-Half Pairs of E and C Students

Parameter	*Effect*	*Estimate*	*SE*	*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}

Note. E = Education Program for Gifted Youth; C = control.

Three-Level, Hierarchical Linear Model (HLM) Analysis Results for the Top Half

The "top half" 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 that has the form.

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

where γ, π_1 and π_2 are fixed effects and u_k, υ_{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 8.

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 Tables 4 and 5, 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 estimate 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-effect estimates from HLM are described in Table 9.

The significant p value for classrooms indicates the existence of significant variation among the classrooms in terms of Math CST07 scores. This is beyond what is expected due to variation

TABLE 9
Random Effects for the Top-Half Pairs of E and C Students

Notation	*Level*	*Estimated Variance*	*SE*	*z Value*	*p Value*
u_k	School	160.19	193.93	0.83	.20
υ_{jk}	Classroom	923.95	267.61	3.45	3.00×10^{-4}
e_{ijk}	Student	2153.27	135.27	15.92	$<10^{-50}$

Note. E = Education Program for Gifted Youth; C = control.

TABLE 10
California Classification of Math CST07 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

Note. CST07 = 2007 California Standard Math Test.

in the prior CST06 scores and is a likely indicator of teacher differences. The school differences are not significant.

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 10 gives the score boundaries for the PLs by grade for the Math CST07. These score boundaries also apply to the Math CST06 test.

The main results are summarized in Tables 11 and 12. These tables show 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 11 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 the 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, so their changes are not statistically significant.

TABLE 11
Result of Binomial Analysis of Changes in Proficiency Level, for the Top Half Ranked by 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 p Value*		*Change in Test Scores p Value*	
	E	*C*	*E*	*C*
Top half ranked on the CFA of E	+84 vs. −50, $p = 4.19 \times 10^{-3}$	+69 vs. −62, $p = .60$	+165 vs. −120, $p = 9.03 \times 10^{-3}$	+152 vs. −133, $p = .29$

Note. CFA = number of correct first attempts; E = Education Program for Gifted Youth; C = control.

TABLE 12
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 Grades 3 to 5 Students in All 8 Schools	*No. of Pairs*	*M of CST07 Difference for Each Pair*	*SD of CST07 Difference for Each Pair*	*ES Relative to SD = 83.5*[a]	*Two-Sided Paired t Test p Value*
Top half of CFA of bottom half of CST06	146	8.49	59.85	0.10	.09

Note. E = Education Program for Gifted Youth; C = control; CST06 = 2006 California Standard Math Test; CFA = number of correct first attempts; CST07 = 2007 California Standard Math Test.

[a]The value of 83.5 is the standard deviation of the scaled scores for the Math CST07 test for all California students in Grades 3 to 5 and is used here as the denominator for the effect sizes.

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 12.

Comparing Table 12 (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 13 shows results for the binomial analysis of changes in the proficiency levels for this lower performing group, which correspond to Table 11 results for the top half.

TABLE 13
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 p Value*	
	E	*C*
Top half of CFA of bottom half of CST06	+62 vs. −22, $p = 1.47 \times 10^{-5}$	+46 vs. −30, $p = .08$

Note. CFA = number of correct first attempts; E = Education Program for Gifted Youth; CST06 = 2006 California Standard Math Test; C = control.

Summary of the Effectiveness Results

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 Math 2006 CST scores.

When we looked at changes in Performance Levels between 2006 and 2007, 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, whereas 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 most 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. Tables 4 and 5 both show that the average of the 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, a strategy that is familiar to many good teachers.

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 *posttreatment covariate* and observed only 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 (e.g., for the 572 pairs of Grades 3–5 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 3. 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, $F(1, 570) = 2.61$, $p = .11$.

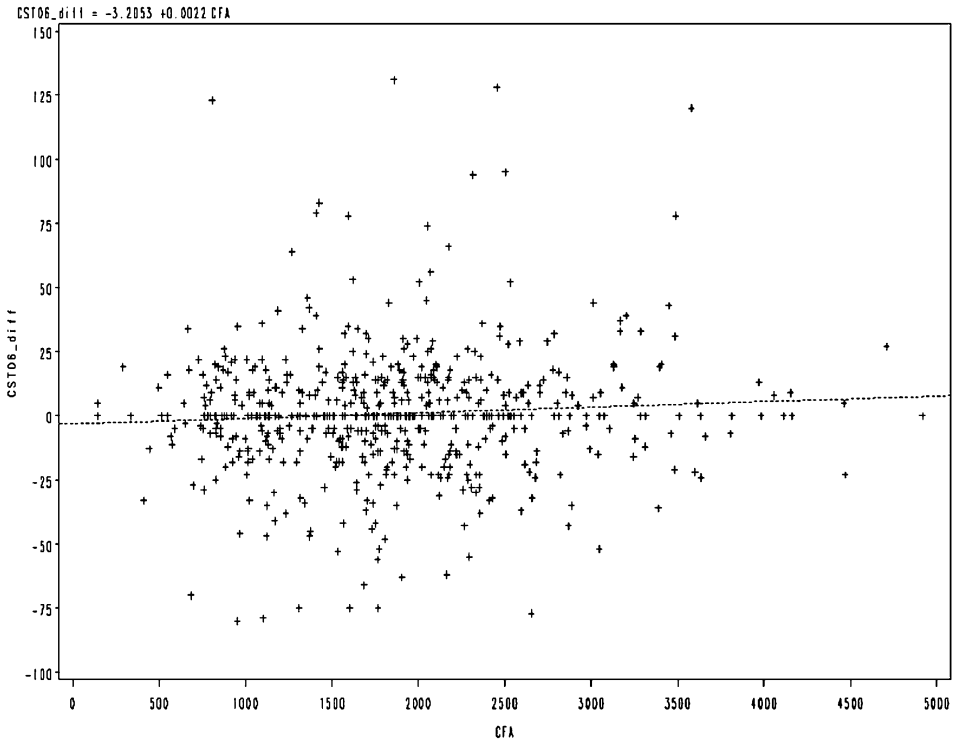

FIGURE 3 Scatter plot of Education Program for Gifted Youth (E) minus control difference on 2006 Math California Standard Math Test (CST06) versus number of correct first attempts (CFA) of E student for 572 match pairs of students in Grades 3 to 5. *Note.* $R^2 = .0046$, slope = 0.0022.

Digging more deeply, in Table 14 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. Although 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

TABLE 14
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	*No. of Pairs*	*M of the CST07 Difference Within Each Pair*	*SD of M Differences*	*Paired t Test p Value*
Top half ranked on Math CST06	281	3.41	75.50	.45

Note. E = Education Program for Gifted Youth; C = control; CST06 = 2006 California Standard Math Test; CST07 = 2007 California Standard Math Test.

not concerned that stratifying on the CFA values of the E members of a pair introduces an important bias in our analyses.

SUMMARY CONCLUSIONS AND POLICY IMPLICATIONS

A strong positive relationship between EPGY work and 2007 Math CST scores was found consistently for each Title I district and school. The more students worked carefully and in a sustained fashion (i.e., had more correct first attempts), the higher the students scored on their Math CST07. In particular, EPGY students in 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 Figure 4. All EPGY students whose number of correct first attempts was greater than 2,000 (the mean number being 1,843.44) had higher test scores in 2007 than in 2006. Only four cells in the table below 2000 showed up arrows for such an improvement, and these were all students with the lowest 2006 Math CST scores.

Policy Implications

The policy implications of these results are evident. Contrary to the overly pessimistic view of educational technology now held by many persons participating in or following the current national debate on the effectiveness of technology, statistically significant outcomes of technologically driven experiments can be obtained. However, positive results, such as those reported here, do not justify a general endorsement of the use of technology in school instruction. As in most such matters, the devil is in the details, as we have shown in the study reported here.

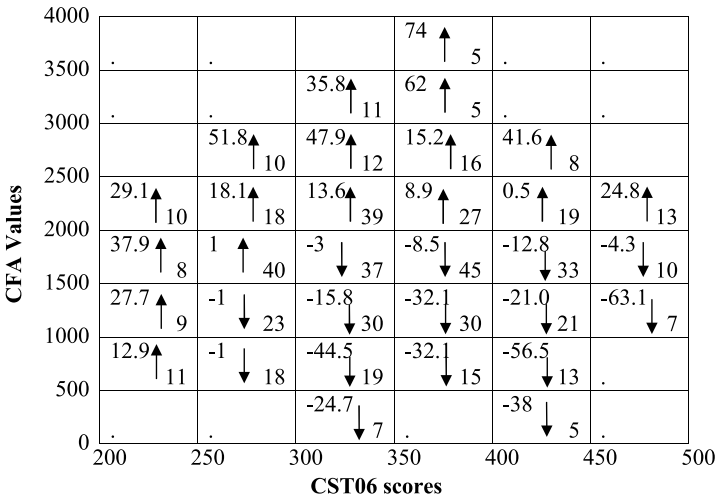

FIGURE 4 CST07 − CTS06 mean differences for combinations of 2006 Math California Standard Math Test (CST06) scores and number of correct first attempts (CFA) values.

Analysis of the work reported supports two policy conclusions about the use of educational technology in schools. The first concerns the technology itself, and the second its implication, especially in Title I schools populated with underachieving and low socioeconomic students.

The main positive features about the use of computers for direct delivery of all or part of a course is that three desirable features can enhance the learning of students at all levels. The first is immediate feedback and reinforcement to the student, as each individual exercise is worked. In teaching elementary mathematics, the focus of this study, many exercises have several steps. Feedback can be quickly given on the answer to each step by a computer program, something that is not possible for a teacher to do in a classroom of 20 or more students. Second, when an error is made, the computer program can provide a concrete helpful hint, without giving the correct answer. This is something a good teacher can also do well in working with an individual student, but not for a classroom of students. Third, with such technology, the progress of each student through the course can be highly individualized, with the selection of each exercise being made by a learning model whose parameters are separately and dynamically continually estimated for each student. (The detailed methods for making such estimates is too technical to describe here, but see Malone, Macken, & Suppes, 1979; Malone, Suppes, Macken, Zanotti, & Kanerva, 1979; Rosenthal & Suppes, 2002; Suppes, 1967; Suppes et al., 1976; Suppes, Hyman, & Jerman, 1967.) In one way or another, the present technological experiment, with careful attention by use of randomized control methods to support the validity of the results, is needed to demonstrate the potential effectiveness of technology in instruction.

This description of the positive features of computer-based instruction might seem to suggest that the role of teachers is drastically reduced, when such instruction is provided to students. But this is most certainly not the case. The role of the teacher is changed, but not reduced in importance. Even when most, if not all, of the direct instruction is given by a computer program, the teacher's role remains of critical importance.

First, the teacher retains the role of manager of the students and their activities. This has many dimensions, especially salient in many Title I classrooms. It often takes a surprisingly short time to judge the management skills of the teacher in such classrooms. More often than should be the case, near chaos reigns too much of the time. Second, the teacher must be a troubleshooter, doing everything from replacing broken earphones to answering a request for help from an individual student, who has not been listening to the audio instruction and feels lost. Doing such troubleshooting replaces giving routine mini-lectures to the class as a whole. Third, and in many ways, the most essential role of the teacher is as a motivator of the students to work regularly and carefully. Without such efforts by the students, little will be accomplished, no matter how well organized and well thought out the math course is, or any other of the same type. Of course, the computer-driven presentations of concepts and exercise should be attractive and motivate students. But in Title I classrooms, with many students living in chaotic home environments, a teacher who can personally motivate students and provide a safe environment for learning is a critical component. This claim is supported in Tables 8 and 9 by the highly significant p values of the variation in the classroom (or teacher) variable in the three-level hierarchical linear model. The quality of teachers matters as much as the quality of the computer-based curriculum.

The results of this randomized treatment experiment support the hypothesis that a computer-based math course, used primarily as a supplement to regular classroom instruction, can improve test scores of elementary-school Title I students. This improvement, not surprisingly, is greater

for students who worked harder. A policy recommendation that follows naturally is that it would be desirable to have a better understanding of the instrumental factors that produce such habits of hard work. Moreover, can these instrumental factors, such as the attractiveness of the computer-driven curriculum or the ability of the teacher to motivate students, be increased significantly by deliberate focused efforts to do so? The long and extensive experience in American schools, with focused teacher training in specific skills, supports positive expectations for such programs, as does the shorter history of educational technology since 1960. The continued increase in computational power and the development of software, ever better adapted to individual students, particularly support such expectations. But to this positive note must be added the realism reinforced directly by the present study. Adding technology will have little effect on learning if students, just as in the past, do not work carefully and regularly, especially Title I students who are so often lacking family support.

A Policy Conclusion

The results of the technological random-control experiment reported on here in detail is that use in schools of the kind of computer programs used in this experiment required two features of management to be of optimal use. First, a school policy framework for students and teachers to feel confident are safe in their environment, with willingness on all sides to define and encourage the independence of each student to feel safe and free to work independently while learning how to work productively. Second, the teachers and the school administration must themselves learn how to use the computer programs that optimize the opportunities for student learning. Creating such an environment should be a major policy goal for all concerned with the quality of schooling.

ACKNOWLEDGMENTS

This research was supported in part by the U.S. Department of Education's Institute of Educational Studies under Grant R305A080464 for much of the statistical analysis and writing of the article.

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

California Standards Tests (CSTs) Technical Report Spring 2007 Administration. (2008). Retrieved July 15, 2013 from http://www.cde.ca.gov/ta/tg/sr/documents/csttechrpt07.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Fletcher, J. D., & Suppes, P. (1976). The Stanford project on computer-assisted instruction for hearing-impaired students. *Journal of Computer-Based Instruction, 3*, 1–12.

Hawley, N. S., & Suppes, P. (1959). Geometry in the first grade. *American Mathematical Monthly, 66*, 505–506.

Hawley, N. S., & Suppes, P. (1960a). *Geometry for Primary Grades. Book 1*. San Francisco: Holden-Day.

Hawley, N. S., & Suppes, P. (1960b). *Geometry for Primary Grades. Book 2*. San Francisco: Holden-Day.

Holland, P. (1988). *Causal inference, path analysis and recursive structural equations models* (Program Statistics Research, Tech. Rep. No. 88-81).

Jamison, D., Fletcher, J. D., Suppes, P., & Atkinson, R. C. (1976). Cost and performance of computer-assisted instruction for education of disadvantaged children. In J. Froomkin, D. T. Jamison, & R. Radner (Eds.), *Education as an industry* (pp. 201–240). Cambridge, MA: NBER, Ballinger.

Jamison, D., Suppes, P., & Butler, C. (1970). Estimated costs of computer assisted instruction for compensatory education in urban areas. *Educational Technology, 10*, 49–57.

Jamison, D., Suppes, P., & Wells, S. (1974). The effectiveness of alternative instructional media: A survey. *Review of Educational Research, 44*, 1–67.

Malone, T. W., Macken, E., & Suppes, P. (1979). Toward optimal allocation of instructional resources: Dividing computer-assisted instruction time among students. *Instructional Science, 8*, 107–120.

Malone, T. W., Suppes, P., Macken, E., Zanotti, M., & Kanerva, L. (1979). Projecting student trajectories in a computer-assisted instruction curriculum. *Journal of Educational Psychology, 71*, 74–84.

Paek, P., Holland, P., & Suppes, P. (1999). Development and analysis of a mathematics aptitude test for gifted elementary school students. *School Science and Mathematics, 99*, 228–247.

Rosenthal, T., & Suppes, P. (2002). *Gifted students' individual differences in computer-based C, programming course.* Stanford, CA: Education Program for Gifted Youth, Stanford University.

Suppes, P. (1962). Mathematical logic for the schools. *The Arithmetic Teacher, 9*, 396–399.

Suppes, P. (1967). Some theoretical models for mathematics learning. *Journal of Research and Development in Education, 1*, 5–22.

Suppes, P. (1992). Instructional computers: Past, present, and future. *International Journal of Educational Research, 17*, 5–17.

Suppes, P. (1995a). Arithmetic and Geometry, Grades 1 [CD-ROM/Web-based]. Stanford, CA: Education Program for Gifted Youth, Stanford University.

Suppes, P. (1995b). Arithmetic and Geometry, Grades 3–6 [CD-ROM/Web-based]. Stanford, CA: Education Program for Gifted Youth, Stanford University.

Suppes, P., Fletcher, J. D., & Zanotti, M. (1975). Performance models of American Indian students on computer-assisted instruction in elementary mathematics. *Instructional Science, 4*, 303–313.

Suppes, P., Fletcher, J. D., & Zanotti, M. (1976). Models of individual trajectories in computer-assisted instruction for deaf students. *Journal of Educational Psychology, 68*, 117–127.

Suppes, P., & Hill, S. A. (1962). The concept of set. *Grade Teacher, 79*, 51, 86, 88, 90.

Suppes, P., Hyman, L., & Jerman, M. (1967). Linear structural models for response and latency performance in arithmetic on computer-controlled terminals. In J. P. Hill (Ed.), *Minnesota Symposia on Child Psychology* (pp. 160–200). Minneapolis: University of Minnesota Press.

Suppes, P. & McKnight, B. A. (1961). Sets and numbers in grade one, 1959–1960. *The Arithmetic Teacher, 8*, 287–290.

Suppes, P., & Morningstar, M. (1970). Technological innovations: Computer-assisted instruction and compensatory education. In F. Korten, S. Cook, & J. Lacey (Eds.), *Psychology and the problems of society* (pp. 221–236). Washington, DC: American Psychological Association.

Suppes, P., & Zanotti, M. (1996). Mastery learning of elementary mathematics: Theory and data. P. Suppes & M. Zanotti (Eds.), *Foundations of probability with applications* (pp. 149–188). New York, NY: Cambridge University Press.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: A practical guide using statistical software.* London, UK: Chapman & Hall/CRC.

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, as well as for 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).[2] Second, a three-level hierarchical linear model of student, classroom and school was used (West, Welch, & Galecki 2007). Third, students' 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.

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.

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 deviations of the Math CST07 scores of all California students in that year that were appropriate for the tests and grades we had in the study. For second graders we used the standard deviation of all California second graders as obtained from the CTSs Technical Report (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 of these two standard deviations.

[2]Cohen's *d* is an appropriate effect-size measure to use in the context of a *t*-test on means. The *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).