Computers & Education 71 (2014) 23–32

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Positive technological and negative pre-test-score effects in a four-year assessment of low socioeconomic status K-8 student learning in computer-based Math and Language Arts courses

Patrick Suppes*, Tie Liang, Elizabeth E. Macken, Daniel P. Flickinger

Education Program for Gifted Youth, Stanford University, United States

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form
18 September 2013
Accepted 19 September 2013

Keywords:
Elementary education
Evaluation methodologies
Intelligent tutoring systems
Interactive learning environments
Evaluation of CAL systems

ABSTRACT

Motivated by the Federal Title I program to improve the Math and Language Arts learning of underachieving students of low socioeconomic status, the Education Program for Gifted Youth (EPGY) at Stanford University has developed computer-based online Math and Language Arts courses for such students in elementary and middle schools. Using several large student samples, the four-year statistical assessment of state test performance is the focus of this report. The main statistical conclusion is that sustained and careful computer-based work, guided by motivated teachers, can be done by many, when taught on an individualized basis, at their current level of competence. The gains made by individual students are, to a large extent, monotonically increasing in their amount of net correct computer-based course work, and to an even larger extent monotonically decreasing as pre-test scores rise, a result that favors technological support of the more underachieving students.

1. Introduction

This introduction covers briefly the objectives and background of this study.

1.1. *Objectives of the study*

The first, and certainly the most important, objective of this study is to assess the effectiveness of a supplementary technologically driven computer-based instruction program in Math and Language Arts for underachieving low socioeconomic status K-8 students.

The second objective is to assess with a hierarchical linear model (HLM) the classroom and school effects on student learning.

The third objective is to report on the use of complex computer software to assess the grammatical correctness of students' written work in the Language Arts course. This is the most sophisticated software used in the study.

1.2. *Background of the study*

Title I federal funding to public schools, as part of the 1965 Elementary and Secondary Education Act (ESEA), is committed to narrowing the gap between underachieving students of low socioeconomic status (SES) and their middle-class peers. In school year 2006–2007, Title I served more than 17 million students; about 60% were in Grades K-5 and 21% in Grades 6–8. Beginning in 1992, the Education Program for Gifted Youth (EPGY) at Stanford University has been developing computer-based distance-learning courses in Mathematics and Language Arts for Title I students. These two computer-based courses are actually derived from computer-based course work that began, even before Title I, at Stanford in 1960 in the Institute for Mathematical Studies in the Social Sciences, which the first author directed from 1960 to 1992. An account of this early work can be found in (Suppes, Jerman, & Brian, 1968; Suppes & Morningstar, 1972; Suppes, 1978, 1981; 1989).

Given the well-recognized difficulty of developing instructional regimes that measurably improve the performance of Title I students, EPGY Math and Language Arts courses have tried to do this by using important technological features: (*i*) progress is individualized for each

* Corresponding author.
E-mail address: psuppes@stanford.edu (P. Suppes).

0360-1315/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compedu.2013.09.008

student according to a stochastic motion driven by a learning model embedded in a computer program; (*ii*) responses to exercises are continually made by students seeing supporting visual displays and hearing associated audio lectures, ordinarily no longer than 90 s; (*iii*) hints, tailored to a student's wrong answers, are given by a computer program; (*iv*) immediate reinforcement determined by computer evaluation of individual responses is given at the end of each exercise.

All computer-based courses offered by EPGY use computer programs as the primary instructional resource. The computer presents students with multimedia lessons that introduce and illustrate concepts. Students have considerable control over the presentation and may review as often as they wish. The brief lectures are kept intentionally short and focused, usually ranging from thirty to ninety seconds.

Multimedia presentations are followed by exercises, which range from questions with exact mathematical answers, to ones in which students compose sentences in Language Arts. Their written sentences are evaluated for the correctness of their grammar and meaning.

Student work is evaluated by the computer and students get immediate individualized feedback. Incorrect answers are typically followed by individualized tutorial hints, in which students are given further instruction and asked to try again. These tutorial hints try to mimic the behavior of an expert tutor guiding a student to an understanding of an exercise for which the student has given an incorrect answer. This often requires different hints for different incorrect answers to the same exercise.

As a student progresses through an EPGY course, the software, using the results of prior assessments, individualizes the student's sequence of exercises. As a result, students who readily master a concept move quickly on to a new one, while students who need more, receive additional instruction and practice. Moreover, material that a student has trouble mastering is reviewed with a higher degree of frequency than material that the student learns quickly.

The Math course content has been correlated with the standards for a number of states, and also with those of the National Council for the Teachers of Mathematics, and the Common Core State Standards for Mathematics. Similar correlations have been done for the Language Arts courses.

2. Materials and methods

Subsection 2.1 describes the large samples of students that participated in this study. Subsection 2.2 describes the computer-based Math and Language Arts courses used in this study. The courses themselves were not created as part of this study. Subsection 2.3 describes methods of analysis used to study the data of students in the two courses. Subsection 2.4 concerns methods of assessment of student performance.

2.1. Sample sizes of student study groups

The data analyses that follow are for three Title I sample groups which include twelve study groups shown by year or multiple years, for example, Memphis Math, 08–09. The sample size of the study groups is shown in Table 1. One Title I sample group is made up of Memphis math elementary- and middle-school students, located in the Memphis City School System in Tennessee in 2007–2011. Four single-year and two multi-year studies are included for this group of students.

The second Title I sample group is made up of Memphis Language Arts (LA) students located in Memphis in 2009–2011 (two single-year and one multi-year studies are included for this group of students). The third Title I sample group is made up of the math students from seven California middle school districts in 2008–2011.

Table 1 shows the sample size of each study group. As would be expected, data from the same students are used in more than one sample group, mainly when single-year sample groups are included in multi-year sample groups.

2.2. Math and Language Arts courses

The most important materials of this study are the versions of the EPGY Math and Language Arts courses used in this study. These are the versions used by many students other than those in the study during 2007–2011. We emphasize these courses were not created, changed or modified for the purposes of this study.

Table 1
Sample sizes for each study group.

Sample group I	Sample size
Memphis Math, 07–08	11,397
Memphis Math, 08–09	27,500
Memphis Math, 09–10	20,821
Memphis Math, 10–11	19,438
Memphis Math, 08–11	4570
Memphis Math, 07–11	1659
Sample group II	**Sample size**
Memphis LA, 09–10	11,872
Memphis LA, 10–11	14,024
Memphis LA, 09–11	5149
Sample group III	**Sample size**
IES California Math, 08–09	724
IES California Math, 09–10	1445
IES California Math, 10–11	881

2.2.1. EPGY Math course

The EPGY Math course used in this study consists of six strands: Integers, Fractions, Measurement, Geometry, Logic, and Statistics. Each of the strands distributes concepts across a series of lectures and classes of exercises. There are approximately 4000 classes and 26,000 exercises in all.

The Integer strand begins in Grade K with number concepts, counting, number relations, and addition, subtraction, and estimation concepts. As the grades progress, students practice computation with the 4 operations based on an understanding of place value and laws of arithmetic. Upon completion of Grade 8, students have practiced computations with positive and negative integers, operations with exponents, and operations with absolute value; they have studied factors and multiples, prime numbers and products of prime factors, GCF and LCM, expanded notation, and operations with scientific notation. They improve computation response times at the beginning of each session in the form of "Math Races".

The Fraction strand begins in Grade K with fraction concepts and continues in the early grades with improper fractions and mixed numbers, equivalent fractions and fraction inequalities. By the end of Grade 8, students have learned operations with positive and negative fractions, equivalent fractions, decimals, ratios, percents, squares and square roots of perfect squares, terminating and repeating decimals and laws of the real number system.

The Measurement strand begins in Grade K with comparisons of width, height, and other dimensions. Students learn to tell and measure time, measure lines and angles with simulated rulers and protractors, count money, estimate distance, study linear, liquid, and weight measures, and work with Celsius and Fahrenheit thermometers.

The Geometry strand begins in Grade K with exercises about identifying plane and solid figures and noting their properties. As students progress through the grades, they study geometric relations, learn the properties of simple 2- and 3-dimensional figures, and study hierarchies of properties of plane figures. They name, measure, and compare angles, work with intersections of lines and planes, study perpendicular and parallel lines, and compute perimeter and area of plane figures, including the circumference and area of a circle and sectors of a circle. They also study rotations, line, mirror, and rotational symmetry, and mirror reflections. They compute volume and surface area for simple regular figures, and make applications of the Pythagorean Theorem.

In the Logic strand, students start by studying simple patterns, comparing the size and shape of plane figures, choosing the figure that does not belong, performing simple sorting tasks, and answering questions that involve simple reasoning about everyday life using informal applications of logical rules. As they progress they reason about relations of numbers and sets of numbers, including inequalities, and study more formal aspects of logical reasoning including conjunctions, disjunctions, denials, and conditionals. They study rules of formal reasoning and practice drawing conclusions from sets of premises, many related to science experiments.

In the Statistics strand, students study both statistics and probability. In Grade K they make picture graphs. They go on to do simulated experiments, collect data and represent it in various kinds of displays, compute measures of central tendency and measures of spread, including variance, note effects of outliers, and draw conclusions. They consider the merits of various samples as being biased or not, judge interview questions, and compare various display options for their appropriateness. They also study concepts of probability, and compute probabilities for independent and dependent events.

2.2.2. EPGY Language Arts course

The EPGY Language Arts course spans Grades 2–6, consisting of some 700 lectures and 14,000 exercises, organized in four strands: Parts of Speech, Sentence Structure, Sentence Composition, and Paragraphs. Each of the strands distributes concepts across a series of lectures and sets of exercises. The Parts-of-Speech strand provides practice identifying and using the grammatical categories of English words in context. The Sentence-Structure strand offers practice analyzing and identifying properties of Standard English syntax. The guided sentence-writing practice of the Sentence-Composition strand reinforces concepts taught in the Parts-of-Speech and Sentence-Structure strands, as described in more detail below. The Paragraph strand teaches and gives practice in the writing of basic types of paragraphs: narrative, opinion, information, and persuasion.

For three of the four strands presented in this section, the course teaches each new concept in a short lecture, and then gives the student a series of exercises, usually multiple-choice, which let the student develop and demonstrate mastery of that concept. The distinctive Sentence Composition strand, which gives students immediate and detailed feedback on sentences they compose, uses a different exercise evaluation method, and is described in Subsection 2.4.3.

Concepts addressed in the Parts-of-Speech strand include the basic English word classes both for open vocabulary (nouns, verbs, adjectives, and adverbs) and for basic functional vocabulary (articles, helping verbs, prepositions, pronouns, demonstratives, conjunctions, etc.). Students get practice working with spelling, inflectional rules and their exceptions (irregular plurals for nouns, tensed forms for verbs, and comparative adjective forms), and learn to use productive word-formation rules such as adding the "-ly" suffix to form adverbs from adjectives. Lectures and exercises also teach the use of contractions and abbreviations, provide practice with often-confused word pairs such as "lie" and "lay", and present relations among words including synonymy and antonymy.

The Sentence-Structure strand addresses the syntactic rules governing the construction of well-formed phrases and sentences in Standard English, including types of sentences (statements, questions, and commands), phrase types and how they are combined to form larger phrases and clauses. Lectures and exercises also teach simple and complex predicates, modification both in nominal phrases and in verb phrases, the proper use of punctuation and capitalization, and the construction and use of conjoined phrases.

In the Paragraph strand, lectures and exercises address the basic types of paragraphs, their internal structure (lead sentence, supporting sentences, and concluding sentence), and concepts to guide the student both in understanding and in composing well-constructed paragraphs, including choice of a main idea, the use of detail, adoption of a point of view, and choice of style appropriate for the intended reader.

2.3. Methods of analysis used in the courses

The methods described in this subsection were used in constructing the courses over many years. They are not part of the assessment analysis as such.

2.3.1. *Theoretical underpinnings for optimizing individual learning*

When one reflects on the problem of organizing a curriculum for optimizing individual learning, it becomes clear that the real problem is to understand as thoroughly as possible the nature of student learning, and, second, to have detailed ideas about what is important for students to learn. As regards the former, a formal development of the theory behind EPGY's approach to the organization of its computer-based mathematics courses is given in a number of the first author's publications, most notably (Suppes, 1969; Suppes & Morningstar, 1972; Suppes & Zanotti, 1996). The implementation of the learning model is what determines for each student which area of the course to work on next, what constitutes mastery, when to give different types of instruction, and how frequently to review.

As regards the organization of the curriculum into concepts, it is important to recognize that not all concepts are equal in importance, so the expected time devoted to mastery should vary. Addition of positive integers, for example, is much less important in the fourth grade than addition of fractions. But how should the expected time be allocated and on what intellectual basis? There is, unfortunately, only a very limited literature on this important matter in the theory of curriculum. Given this situation, the most feasible approach, and the one which we had taken at first, was to use as initial data the curriculum guidelines set by various state and local school systems, and then to count the empirical frequency of exercises in various widely used textbooks. From this starting point, we have used the experience of the more than 60,000 students who have worked through the courses to identify areas of deficiency and to address them. This includes an average of over two million exercises worked per month during the past year by Title I students.

2.3.2. *The conceptualization of motion algorithms in EPGY courses*

The stochastic motion of each student in the K-8 Math and Language Arts courses has a quite detailed technical specification that we do not give here. But we can describe its main features. First, why is the motion stochastic rather than a simpler deterministic algorithm? The answer is that the stochastic or probabilistic aspect of the motion provides an appropriate method of smoothing, both of student progress in a given concept class, or in the selection of which strand is to follow the current one. Another aspect is that it keeps students from being inappropriately conditioned to one subject always following another. The real world does not have that restricted simplicity, nor should student learning.

Second, in such stochastic motion, a decision must be made after each exercise about what the student should do next. The structure of the current courses provides for a student to work a sequence of exercises in a given concept class before leaving that class for another. But, it is critical to recognize the striking individual differences in student learning. Consequently, there is a learning model attached to each concept class. A student who quickly masters a concept is required to do only a sequence of three or four exercises. If the mastery criterion is met, the student is moved up to the next concept. If he or she does not reach mastery level, the student is either set to repeat this concept class, or to move down to a lower class.

Third, the stochastic motion permits a natural adaptation of movement in the courses to individual learning differences. For example, one student may learn geometric ideas faster than arithmetic ones, another student the reverse, or one student learns parts of speech faster than composition, but another student the opposite. It is impossible to estimate in advance these subtle student differences, but by continually assigning different amounts of time to each strand of a course, depending upon the student's own work, accommodations of such differences can be made. This accommodation is made in the courses by a probabilistic algorithm that is computed anew after a student has finished each concept class. Note that this is a fine example of a delicate adjustment for each student that could not possibly be done on such an individual basis in the group setting of a standard classroom.

This discussion has been at a qualitative and rather general level. It can easily be expanded by the references already given to results on a massive amount of computations on student learning data.

2.3.3. *Theory of optimizing individual learning*

The EPGY program uses a theory focused on optimizing individual learning. The theory consists of two parts; the first one includes learning models of mastery, which capture the nature of student learning, and the second part is on student course trajectories, which give information about student temporal progress. These two components of the courses are described below.

2.3.3.1. Learning models of mastery. Learning models (Suppes, 1969; Suppes & Morningstar, 1972, Ch. 4; Suppes & Zanotti, 1996) have been applied to EPGY courses to decide if a student's performance on a given class of exercises satisfies some criterion. Mastery is defined recursively in terms of the student's probability of making an error on exercise $n+1$ given a correct or incorrect response on exercise n; mastery is achieved when the error rate for a given strand on exercise $n+1$ is less than a pre-determined value in the current system. Three models have been used. The equations and meaning of each model are presented using the notation.

$A_{0,n}$ = event of incorrect response on trial n,
$A_{1,n}$ = event of correct response on trial n,
x_n = sequence of correct and incorrect responses from trial 1 to n inclusive,
$P(A_{0,n})$ = marginal probability of an error on trial n.

Model I

$$(i)\ P\left(A_{0,n+1}\middle|A_{0,n}x_{n-1}\right) = (1-\omega)P\left(A_{0,n}\middle|x_{n-1}\right)+\omega,$$
$$(ii)\ P\left(A_{1,n+1}\middle|A_{1,n}x_{n-1}\right) = (1-\omega)P\left(A_{1,n}\middle|x_{n-1}\right)+\omega.$$

In Model I, we assume reinforcement of the actual response made, whether it is correct or incorrect. **Model II** differs only by having two learning parameters ω_0 for an incorrect response and ω_1 for a correct response. **Model III**, like Model I, has a learning parameter, ω, to reinforce whichever response was made, but it also has parameter $\alpha,\ 0<\alpha<1$, that operates on each trial, independent of the response made, to reinforce the correct response by multiplying the probability of the incorrect response by α, which is less than 1.

As an example, the mean learning curve derived from Model III is $P(A_{0,n+1}) = \alpha^n q$, where q is the probability of an error on trial 1, and α is the learning parameter just discussed (Suppes & Zanotti, 1996). Fig. 1 shows the estimated mean learning curve based on data from 1283 students working in the fraction strand. It shows how the probability of incorrect responses reduces as students finish more exercises. As can be seen, the fit of the theoretical learning curve to the mean data is quite good.

2.3.3.2. Learning trajectories of individual students. To develop a theory of prediction for individual students' progress through the curriculum, the use of learning trajectories in computer-based courses was proposed in the mid-1970s. The first article in the series was by Suppes, Fletcher, and Zanotti (1976). Continued discussions and applications can be found in (Larsen, Markosian, & Suppes, 1978; Malone, Suppes, Macken, Zanotti, & Kanerva, 1979; Suppes, Macken, & Zanotti, 1978; Suppes & Zanotti, 1996). The research has shown that a student's trajectory through a computer-based course can be described well by a power function $y(t) = bt^k + \alpha$, where t is computer time, $y(t)$ is grade-placement as a function of time, the parameter α is the grade-placement position at which the student started the course, and b and k are both measures of the student's pace through the curriculum. They determine the shape of a student trajectory. The parameter k measures how curved a trajectory is. When k is close to 1, b is the approximate fraction of the course completed per computer time. Fig. 2 shows three 1992–1993 examples of computer-time trajectories. The x-axis shows the computer time the students spent on the course and y the grade-placement gains of the students. Fig. 2 illustrates well the diversity of student trajectories. The fits to the theoretical power function are pretty good, but unlike a mean curve averaged over the data from many students, far from perfect.

2.4. Methods and measures of assessment

2.4.1. Effect size of gains or losses

2.4.1.1. Effect size as a function of post-test –pre-test. It is intended that EPGY curriculum is developed to enhance student competence, as measured by state test scores. To address this point, Title I students' performance in end-of-year state test results has been monitored for four years 2007–2011. The major EPGY student-performance variable used for assessment purposes is *diff*, the difference between the number of correct first-attempts and the number of incorrect first-attempts at answering exercises; *diff* is a net measure of a student's quantitative amount of correct work. The importance of subtracting the amount of incorrect work in the definition of *diff* is that it properly evaluates students who are mainly just quickly guessing and therefore get a lot of correct answers, but even more incorrect ones. Another factor of this approach, in the present setting of testing a large number of students, is that even a small gain can be statistically highly significant. The use of *diff* came from a detailed study of some underachieving students, as measured by effect size, but with a large number of correct responses, as just described. For details of using just the number of correct responses as the measure of work, see (Suppes, Holland, Hu, & Vu, 2013).

The methodology used here was to compute effect size as a function of the present and past state tests. So, after a year or more of technological intervention, in the form of computer-based instruction, as measured by annual state test, not designed by this project, the student's gain or loss was measured for its effect size. Details are given later, and also amplified in the Supplementary materials.

To measure the quantity of group differences in average gains or losses, each study was divided into four groups, ordered from 1 to 4 by the mean amount of net correct work *diff*. So for example, every student in Group 1 had a smaller *diff* than any student in Group 2. The 1-step average effect size is an average of three 1-step effect sizes for Group 1 vs Group 2, Group 2 vs Group 3 and Group 3 vs Group 4. The average 2-step effect size is an average of two 2-step effect sizes for Group 1 vs Group 3 and Group 2 vs Group 4. Finally, the 3-step effect size measures the difference of test-score gains or losses between Group 1 and Group 4. More generally, for a given sample group, every student in Group i has a smaller *diff* than any student in Group$j$, if $i < j$, $1 \le i < j \le 4$. The effect size was calculated using the formulas below, (Cohen, 1992),

$$d = \frac{\overline{x}_j - \overline{x}_i}{s_p},\ \textit{where}$$

$$s_p = \sqrt{\frac{(n_i-1)s_i^2 + (n_j-1)s_j^2}{n_i+n_j-2}}.$$

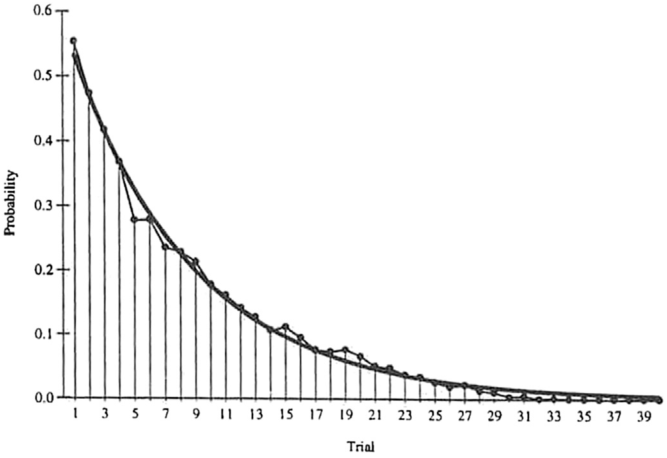

Fig. 1. Mean learning curve of Model III for the fraction-strand of students at grade level 3.90 (sample size = 1,283, $q = 0.533$, $\alpha = 0.885$), (Suppes & Zanotti, 1996, p. 166).

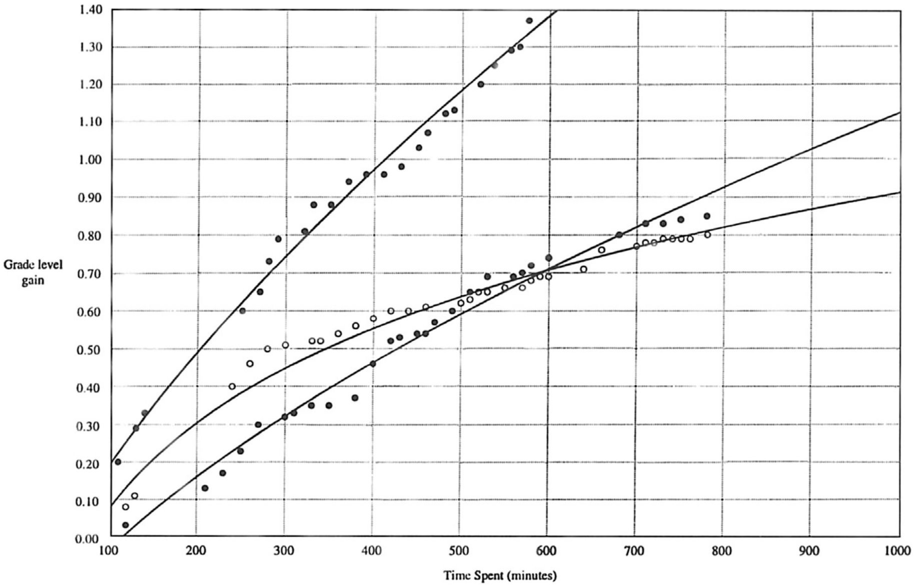

Fig. 2. Three examples of individual student trajectories, (Suppes & Zanotti, 1996, p. 173).

Here $\overline{x}_i$ and $\overline{x}_j$ are the means of gains and losses in each group, d is the effect size, s_i and s_j are the corresponding standard deviations, n_i and n_j are sample sizes for the two groups i and j of interest, and s_p is the pooled standard deviation, as defined by the equation (Hedges, 1981). Note that the unit of effect size is one standard deviation, or the weighted average of two.

2.4.1.2. Gains and losses as a function of two variables. In assessing the gains or losses in achievement from the technological intervention of computer-based instruction, it is necessary to recognize from the beginning of any analytic thinking about this assessment that there are many other causes affecting the performance of students, many of which have been studied extensively.

In the assessment considered in this subsection we restrict ourselves to two causal variables. The first is the relevant prior test score, which is intended to measure the achievement of a student at the time the test is taken. This measure itself is meant to be a measure of the student's effective learning of Mathematics and Language Arts to date, and is therefore a measure of the effect of many relevant causes on each student, as of the date of this pre-test. The second causal variable is the measure *diff* of the net correct work done in interacting with an instructional computer program.

When we focus on these two causal variables, we plot relevant past test scores on the x-axis of a graph and relevant *diff* measures of work done in a course on the y-axis. For students, with coordinates (x_s, y_s), we plot on a two-dimensional contour map, the value of the function $f(x_s, y_s)$ measuring this gain or loss on the post-test. We use the points (x_s, y_s) on the gain graph to draw contours on the graph such that each contour curve is meant to represent the curve on which gains or losses of a given value, such as 15, are represented. Students whose gains are zero do not change their relative positions in going from the pre- to the post-test. Average shows a loss each year. The objective of most interventions is to make these losses gains, or, at least, to reduce the losses.

To draw the contours of the map of this gain function, to get relatively smooth understandable results, we average over neighboring points to get an expected gain from a given set of students. In practice, we divide the contour surface into cells each containing gain data, on at least 50 students in the case of Memphis, and at least 10 students for the smaller samples of IES California Math. Hereafter, we often refer only to gains, but these should be interpreted algebraically, as being either positive or negative.

Two extreme results could be obtained: all contour curves are in fact straight horizontal lines, which would show that only *diff* had a causal effect. Contrariwise, if all contour lines were vertical then the contour lines connecting equal gains would show that only the level of past achievement, as reflected in the pre-test score had a causal effect on the post-test score, of each student. Of course, such extremes are not found in the empirical data.

2.4.2. Hierarchical Linear Model for investigating school and class effects

The focus of our study of the second objective of this article was to apply a three-level Hierarchical Linear Model (HLM) (West, Welch, & Galecki, 2007) that was fit to various data sets for the purpose of investigating school and class effect sizes on students' performance.

The model is specified as follows:

Level 1: Student level;

$$Score2_{ijk} = \pi_{jk} + \pi_1 Score1 + e_{ijk}$$

Level 2: Classroom level;

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

Level 3: School level.

$$\beta_k = \gamma + u_k$$

Putting these three models together, we have the following equation that relates students, classes, and schools to the outcome variable.

$$Score2_{ijk} = \gamma + \pi_1 Score1 + u_k + \upsilon_{jk} + e_{ijk},$$

where Score 1 is year-1 test score, Score 2 is year-2 test score, γ, π_1 are the fixed effects (effects that do not vary across schools, classes and students) and $u_k, \upsilon_{jk}, e_{ijk}$ are the random effects for schools, classes within schools and students within classes, respectively.

2.4.3. *Computer analysis of sentence compositions in Language Arts*

The EPGY Language Arts curriculum includes a strand for sentence composition in each of Grades 2–6. This strand gives students exercises and immediate feedback when writing complete sentences. In each of the roughly 750 individual exercises distributed throughout the five grades, students are presented with a statement and then a question which they answer by composing a sentence using a word list specific to each exercise. Each answer is evaluated for correctness, both in terms of grammaticality and content, with grammatical errors identified by a combination of fixed sentence lists and grammar-checking software that parses answers by using the broad-coverage English Resource Grammar (Flickinger, 2000; Flickinger, 2011) developed at Stanford over the past two decades as part of the international DELPH-IN consortium (www.delph-in.net), and extended by software specifically developed for this course. The parser used with the grammar in this evaluation step is the PET engine (Callmeier, 2000). For each exercise, when the first attempt is incorrect, the student is asked to make a second attempt, guided by the error analysis presented.

The LA curriculum was in regular use within the Memphis elementary public school system for the two school years 2009–2011, with weekly usage ranging between 5000 and 20,000 students.

3. Results and discussion

3.1. *Results on gains and losses*

We begin with results on effect sizes for both courses. We then turn to gains or losses as a function of two causal variables, one positive and one negative.

3.1.1. *Results on effect size as a function of post-test–pre-test*

The results in Table 2 show that as the number of steps measuring *diff* increased, effect sizes increased strictly monotonically except in two studies in IES California Math. Detailed computations for each step in each dataset are given in the Supplementary materials.

3.1.2. *Results for gains and losses as a function of* diff *and pre-test score*

In order to show visually the relationship between *diff* and the state-test score gains or losses, four plots representing Memphis Math (single year), Memphis Math (multi-year), IES Math (single year) and Memphis LA (multi-year) were selected to display state-test score gains or losses as a function of the starting-year test score and the course-performance variable *diff* measuring net correct work. Figures for each dataset and the computations on average test score gains or losses can be found in the Supplementary materials. Some of the important results are shown in Fig. 3. The test-score gains, shown in white, and losses, in light grey, are plotted as a function of *diff* on the *y*-axis, and students' starting year test scores on the *x*-axis. To guarantee robust statistics for each cell, for the Memphis data those cells with the number of students less than 50 were eliminated; for the more restricted IES data, cells with less than 10 were eliminated. The average gain or loss in test scores from one year to another is shown in the middle of each cell of a graph.

Fig. 3(A) presents a one-year analysis of Memphis Math students in 2009–2010, which is the best single year average effect-size. Fig. 3(B) shows the three-year Memphis analysis for 2008–2011. Fig. 3(C) presents data for a much smaller California population of middle-school students. The overall weaker results probably reflect the well-known greater difficulty of middle-school math for Title I students. Fig. 3(D) displays a two-year (2009–2011) study of Memphis LA students. In this study, the greatest improvement generally in state test scores was for the students with lower past scores, in some ways a desirable outcome.

Table 2
Effect size by number of steps of *diff* levels for both Math and Language Arts courses.

Study group	Eff size (1 step)	Eff size (2 steps)	Eff size (3 steps)	Sample size
Memphis Math, 07–08	0.140	0.253	0.407	11,397
Memphis Math, 08–09	0.208	0.398	0.566	27,500
Memphis Math, 09–10	0.292	0.690	0.737	20,821
Memphis Math, 10–11	0.075	0.151	0.210	19,438
Memphis Math, 08–11	0.235	0.485	0.645	4570
Memphis Math, 07–11	0.111	0.135	0.309	1659
Memphis LA, 09–10	0.182	0.408	0.492	11,872
Memphis LA, 10–11	0.044	0.070	0.124	14,024
Memphis LA, 09–11	0.198	0.409	0.526	5149
IES California Math, 08–09	0.078	0.347	0.204	724
IES California Math, 09–10	0.069	0.014	0.220	1445
IES California Math, 10–11	0.058	0.123	0.187	881

Note: LA = Language Arts, Eff Size = Effect Size.

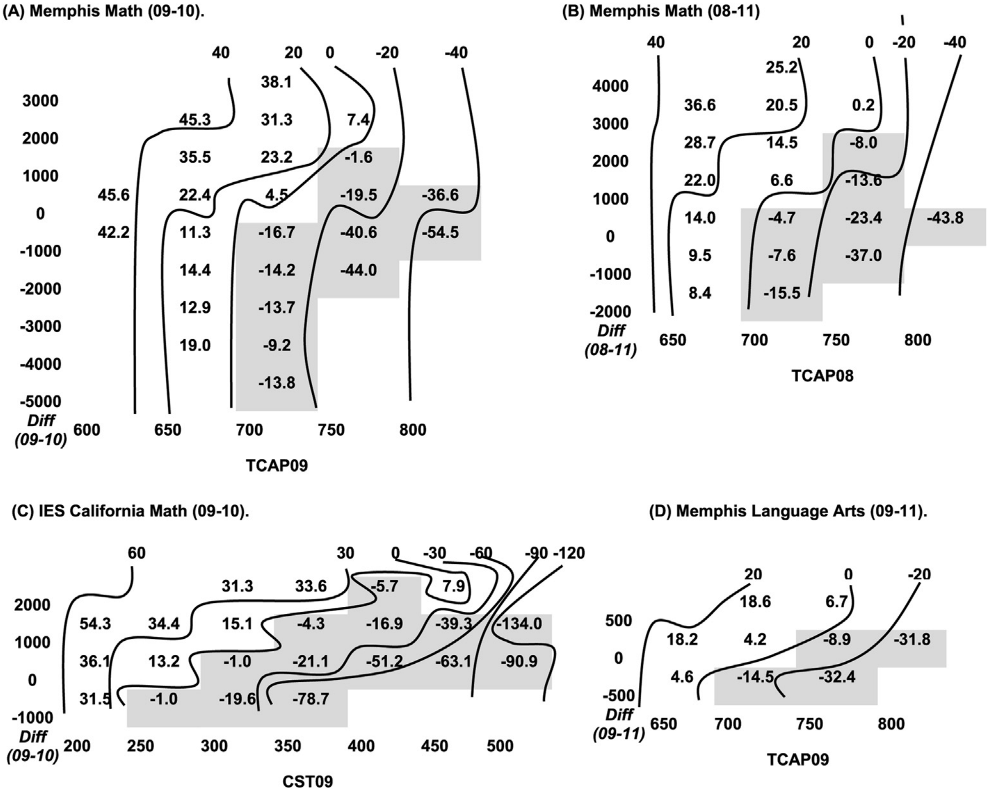

Fig. 3. Analysis of test-score gains or losses as a function of the starting-year test score and the course variable (*diff*) for four studies, three in Math and one in Language Arts.

These four plots of Fig. 3 show that the averages of students' score gains increased strictly monotonically as *diff* increased while holding the test score column constant, with five exceptions in Memphis Math, 09–10, and one exception in IES California Math (09–10). The surprise is that strict decreasing monotonicity holds true for pre-test scores as negative causal variables for all 12 sample groups. For the four contour maps of Fig. 3, in the case of decreasing row-monotonicity there are only two exceptions, which occur in Fig. 3(A) for Memphis Math, 09–10. The total data for all 12 sample groups are shown in the Supplementary materials. There are a total of 18 reversals of strict increasing monotonicity for the causal effect of *diff*, which is just 12.6% of the 142 vertical pairs with the same x-coordinate. In the case of the negative effect of pre-test scores, the results are even stronger: only 6 exceptions to strict decreasing monotonicity, which is 4.2% of the 139 horizontal pairs with the same y-coordinate. (For a systematic general discussion of negative causes, see Suppes (1970, pp. 43–47).) Detailed data for all 12 sample groups are given in the Supplementary materials. These negative results definitely suggest that if resources are limited, highly individual presentations using technological interventions, especially emphasizing the use of computer programs to individualize each student's trajectory in a course, will benefit the more underachieving Title I K-8 students the most.

Even though these prima facie negative effects of pre-test scores raise more questions than they answer, we are not able to pursue further in the framework of this article either the causes or further implications of these surprising negative findings.

The contour curves of equal expected gain are neither purely horizontal nor purely vertical. This fact shows that both causal variables are effective, but even for the mean results shown here the curves are complicated and too complex to satisfy any simple mathematical equations.

3.2. Results on HLM school, class and student effects

In Table 3 the 95% confidence intervals of the estimates are shown in parentheses; all the results are significant. Of course, the most precise, i.e., smallest, confidence intervals are for the students rather than classes or schools, because of the large sample sizes. Concerning our second objective, Table 3 suggests that the role of teachers as class managers and motivators is critical, especially in the case of Title I students in maintaining order, creating a disciplined classroom environment and, when necessary, providing cognitive and emotional help to troubled students, who are having special difficulties in learning how to work effectively. More detailed study of these teacher-effects would be desirable. An excellent meta-analysis of 66 past studies of this topic has been published by Roorda, Koomen, Split, and Oort (2011).

3.3. Computer grammatical analysis of student sentences

The table below summarizes the number of sentences attempted by the students, with most of the usage occurring from January 2010 through May 2011, and with the answers divided into those evaluated as correct or incorrect, and further separated according to answers given on the first attempt and on the second attempt, when required.

Table 4 shows that students produced a correct answer on their first attempt in one of every two exercises (1,446,253, or 50.6% of the total), and improved to 57.3% correct after the second attempt. Interestingly, the 1.6 million correct answers consist of only 12,137 distinct

Table 3
School, class and student effects from HLM analyses.

Study group	School		Class		Student	
	SS	Parameter	SS	Parameter	SS	Parameter
Memphis Math, 07–08	127	33 (24, 50)	1124	63 (55, 74)	11,397	342 (333, 352)
Memphis Math, 08–09	150	32 (24, 44)	2838	69 (61, 78)	27,500	578 (567, 588)
Memphis Math, 09–10	148	328 (261, 425)	3389	136 (124, 151)	20,821	659 (645, 673)
Memphis Math, 10–11	156	92 (71, 123)	3623	82 (72, 94)	19,438	633 (619, 647)
Memphis LA, 09–10	130	209 (163, 277)	2500	59 (49, 72)	11,872	586 (570, 603)
Memphis LA, 10–11	153	34 (26, 48)	3328	24 (18, 35)	14,024	521 (508, 535)
IES California Math, 08–09	3	207 (47, 416)	30	360 (210, 754)	724	1266 (1143, 1411)
IES California Math, 09–10	7	96 (34, 775)	103	154 (87, 350)	1445	2397 (2226, 2589)
IES California Math,10–11	8	409 (165, 2184)	37	134 (71, 344)	881	1366 (1244, 1506)

Note: SS represents sample size and numbers in the parentheses under parameter estimates are their 95% confidence intervals.

answers (where duplicate answers are merged), while the 1.2 million incorrect answers consist of more than 180,000 distinct answers, showing much more variation than correct answers do. Similarly, while there were only 2323 correct answers which appeared just once each in all of the corpus, students produced a much larger proportion of unique incorrect answers, more than 40% of the total number of distinct incorrect answers.

Of the nearly three million individual answers, many are repeated with high frequency, but even with the strong exercise-specific limits on words to choose from, many of the answers are unique or rarely repeated. The extent of variation is markedly different in correct and incorrect answers, with 6588 distinct correct answers (roughly half) appearing fewer than 10 times, contrasted with 166,505 distinct incorrect answers with this low frequency (over 90%). This great variation in the range of distinct incorrect answers underscores the importance of using a linguistically sophisticated parser which automatically evaluates and diagnoses grammatical errors in these incorrect answers.

The length of the students' answers is normally under 10 words, with an overall average length of 5.45 words per distinct correct answer, and 5.70 words per incorrect answer. The exercises are designed to need no more than 20 words per answer, and normally no more than 10 words, with an average length of 6.77 words for the expected correct answers.

As students progress through the five grades, the exercise-specific vocabulary lists grow in complexity, as do the sentences that the students are asked to compose. By Grade 6, these sentences include subordinate clauses, relative clauses, and conjoined noun phrases, verb phrases, and clauses. These recursive syntactic devices quickly lead to the wide variation summarized above for the particular sentences that students compose, and accurate analysis of errors in these sentences depends on the linguistically sophisticated grammar implementation used for this course.

4. Conclusion

The statistical analyses reported support the conclusion that EPGY online computer-based courses can provide an effective way to improve the Math and Language Arts learning of underachieving students of low socioeconomic status in elementary and middle schools. The test-score results, as a function of *diff*, indicate that the difference of score gains between more and less industrious students is highly significant. The variation in classroom effects supports this hypothesis.

Finally, concerning our third objective of reporting on our intention to construct computer programs to evaluate the grammatical correctness of the written work of individual students, our initial effort of handling almost 3 million sentences strongly suggests this is a hopeful direction for solving an important problem in the teaching of Language Arts. It is simply not practical for an elementary-school teacher with about 30 students in the classroom to evaluate written sentences of each student on a daily basis. Sophisticated computer software offers a realistic method for solving this important problem.

We believe that our positive results, obtained by extensive use of educational technology and supported by detailed statistical assessment of their significance, can contribute to the broader discussion of how to improve elementary-school and middle-school education, especially for underachieving students of low socioeconomic status.

This article is technical and detailed. It seems necessary to convince those responsible for making and implementing instructional decisions and policies that a serious case can be made for the extensive use of information technology in public-school instruction. Assistance from teachers in managing the classroom and motivating students is also needed. The emphasis here has been on Title I students who, perhaps more than any other large group, need methods of instruction that can help them not fall as far behind in school, as has been the case in much of the past and still too often today. Technology alone will not solve this problem, but we believe we have shown that, even

Table 4
Language Arts sentence composition exercises for Grades 2–6 in 2010–2011.

Total	Correct			Incorrect			All
	1st attempt 1,446,253	+	2nd attempt 190,729	1st attempt 709,489	+	2nd attempt 509,995	
	1,636,982			1,219,484			2,856,466
	57.3%			42.7%			
Distinct	12,137			180,795			192,932
Unique	2323			79,580			81,903
	19.1%			44.0%			

Note: The "Distinct" row gives the total number of different (non-duplicate) answers presented by students, while "Unique" gives the number of answers that only appear exactly once in the whole set of student answers.

with large numbers of students, effective use of computers can provide positive results of improved learning under rigorous conditions of assessment.

Acknowledgments

We would like to thank Yuanan Hu and Minh-thien Vu for their work on students' performance assessment in EPGY. The research study at Stanford involving the six California school districts was funded by the Institute of Education Sciences (IES), U.S. Department of Education under Grant R305A080464.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2013.09.008.

References

Callmeier, U. (2000). A platform for experimentation with efficient HPSG processing techniques. *Natural Language Engineering, 6*(1), 99–108.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering, 6*(1), 15–28.

Flickinger, D. (2011). Accuracy vs. robustness in grammar engineering. In E. M. Bender, & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing* (pp. 31–50). Stanford, CA: CSLI Publications.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.

Larsen, I., Markosian, L. Z., & Suppes, P. (1978). Performance models of undergraduate students on computer-assisted instruction in elementary logic. *Instructional Science, 7*, 15–35.

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.

Roorda, D. L., Koomen, H. M. Y., Split, J. L., & Oort, F. J. (2011). The influence of affective teacher-student relationships on students' school engagement and achievement: a meta-analytic approach. *Review of Educational Research, 81*, 493–529.

Suppes, P. (1969). Stimulus-response theory of finite automata. *Journal of Mathematical Psychology, 6*, 327–355. German translation: In M. Balzer & W. Heidelberger, (Eds.), Zur Logik empirischer Theorien. (pp. 245–280). Berlin, New York: Walter de Gruyter, 1983.

Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North Holland Pub. Co.

Suppes, P. (1978). *The future of computers in education*. In *Computers and the Learning Society: Hearings before the Subcommittee on Domestic and International Scientific Planning, Analysis and Cooperation, of the Committee on Science and Technology, U.S. House of Representatives, Ninety-Fifth Congress, First Session, October 4, 6, 12, 13, 18 and 27, 1977* (Vol. 47). Washington: U.S. Government Printing office. pp.548–569.

Suppes, P. (1981). Future educational uses of interactive theorem proving. In P. Suppes (Ed.), *University level computer-assisted instruction at Stanford: 1968–1980* (pp. 165–182). Stanford, CA: Stanford University, Institute for Mathematical Studies in the Social Sciences.

Suppes, P. (1989). Review of the O.D. Duncan, notes on a social measurement: historical and critical. *Journal of Official Statistics, 5*(3), 299–300.

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., Holland, P. W., Hu, Y., & Vu, M.-T. (2013). Effectiveness of an individualized computer-driven online math K-5 course in eight California title I elementary schools. *Educational Assessment, 18*(3), 162–181.

Suppes, P., Jerman, M., & Brian, D. (1968). *Computer-assisted instruction: Stanford's 1965–66 arithmetic program*. New York: Academic Press.

Suppes, P., Macken, E., & Zanotti, M. (1978). The role of global psychological models in instructional technology. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 229–259). Hillsdale, NJ: Erlbaum.

Suppes, P., & Morningstar, M. (1972). *Computer-assisted instruction at Stanford, 1966–68; data, models, and evaluation of the arithmetic programs*. New York: Academic Press.

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

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