Instructional Science 30: 79–110, 2002.

Gifted students' individual differences in distance-learning computer-based calculus and linear algebra

ERIC W. COPE[1] & PATRICK SUPPES[2]
Education Program for Gifted Youth, Stanford University,
[1] *e-mail: cope@epgy.stanford.edu*
[2] *e-mail: suppes@epgy.stanford.edu*

Received: 3 March 2000; in final form: 11 April 2001; accepted: 8 May 2001

Abstract. We examine student performance in computer-based calculus and linear algebra courses offered by Stanford University to pre-college students of high mathematical ability. Our analysis puts special emphasis on modeling student performance over time and on capturing long-term trend effects. The sequential nature of students' responses to course exercises is characterized through the use of stochastic and nonlinear models. We find that student performance varies widely within this group for a variety of different measures, including error rates, times to completion, progress rates, and latency of response. In addition, we measure the informational efficiency of the courses through a Markov order analysis of student response sequences. For the performance measures studied, the 75th and the 25th percentile of sampled values differ by a factor of approximately two. We also find that there is little correlation among the performance measures, which suggests that student performance in this ability range may not be well characterized by any single performance measure or ability parameter.

Keywords: calculus, computer-assisted instruction, distance learning, gifted education, linear algebra, nonlinear analysis, stochastic process models

Introduction

Over the past decade, there has been a phenomenal rise in the number of computer-based mathematics courses available, and this increase has made possible the collection of large amounts of data on student performance. Standard educational assessment techniques can now be greatly augmented by computers' ability to record exact response times to exercises as well as *sequences* of responses and response times. Emphasis is placed on the latter because with modern data analysis techniques, we can begin to explore temporal dependencies in student performance over an entire course or beyond. Thus a more global picture of the overall performance of a student, as well as of the pedagogical effectiveness of the courses themselves, can emerge. In addition, sequential data can also shed new light on the cognitive processes underlying the learning of mathematics.

With this in mind, we studied individual student differences in performance in a group of computer-based advanced-placement calculus and university-level linear algebra courses, which were offered by the Education Program for Gifted Youth (EPGY) at Stanford University. As the name of the program suggests, these students had all demonstrated a high degree of mathematical ability before beginning the courses. Their level of ability was evidenced in particular by their age at the time they began the courses – the average age of students beginning the first calculus course was 14.9 years old, well before the age most students begin an advanced-placement calculus course. These courses were run by the students on their home computers, and the software was equipped to record the details of their actions while running the course and then transfer the data to Stanford.

This study examines these data and compares performance characteristics across the student population in each of these courses. We find that student learning rates and patterns vary dramatically, and often range over an order of magnitude. These observations are all the more striking considering that students are selected for this program if they score within a narrow band of standardized test results: most students in this study broadly scored within the top three percent nationally in mathematical ability. These differences suggest that in the high-ability range, learning patterns and rates are highly diverse, and therefore student needs vary widely.

In addition, we present a number of methods of data analysis, which, though long in use and familiar in the statistical literature, have not been widely applied in educational measurement. Central to our analysis is the characterization of overall student performance according to their progress "trajectory", and its underlying stochastic process. Nonlinear methods of time series analysis are used to study its dynamics. We also examine the short-range dependencies in the answer sequences in the course, and test the stationary character of these dependencies. In addition, we also give the more usual characterizations of student performance – exercise and test scores – as well as the distribution of response times, and compare all of these measurements with each other. We found that when these measures are taken together, they provide a high-dimensional characterization of student ability which cannot be reduced to a single performance measure or ability parameter.

There are other important individual differences in gifted students that have been reported in the literature, but were not accessible in our computer-based, but nonexperimental environment. A primary example is achievement motivation, recently reviewed in detail by Dai, Moon & Feldhusen (1998). We also have not studied the role of individual differences in pairings of students solving together complex mathematical tasks, as reported in Fuchs, Fuchs, Hamlett & Karns (1998). A component of general intelligence that was studied throughout the twentieth century is gifted spatial ability, as separate from general mathematical giftedness. Some striking individual differences have been found, as reported, for example, in Humphreys & Lubinski (1996) and Gohm, Humphreys & Yao (1998). Many, but certainly not all, mathematicians feel that by the time able students reach the calculus and linear algebra, easy spatial visualization of concepts and problems is very beneficial. In the research studies known to us, none have reported on spatial abilities related to student performance at this level. Our current data do not permit us to make such a report, but we consider the subject an important one for the future.

The EPGY calculus and linear algebra curriculum

EPGY has offered computer-based multimedia courses in mathematics since 1992. The project's primary focus is to make advanced mathematics instruction available to very bright students who ordinarily might not have access to such curricula. The three-course sequence in Advanced Placement (AP) calculus was the first set of courses produced by EPGY, and the early success of students completing these courses on the Advanced Placement examinations encouraged the subsequent production of several university-level courses that assume the calculus; Linear Algebra was the first of these to be produced. The performance of students on an early version of the calculus series, which at the time was called Dfx, was studied in an earlier paper (Suppes & Ager, 1995).

The instructional model is quite similar in each of the four courses studied here (Calculus A, B, C, and Linear Algebra). Each course is primarily conducted on personal computers at students' homes or schools, using multimedia CD-ROM-based software. Students are presented with a series of lessons, often corresponding to sections in a companion textbook, which are generally composed of an audio lecture accompanied by a synchronized video presentation. Lectures are usually followed by a series of exercises and quizzes based on the contents of the lesson. In some cases, concepts are presented to students through a step-by-step series of questions that reveal important identities and relationships. In addition, Calculus A and B students also complete exercises on the derivation system, a symbolic computation environment which allows students to make use of specified rules and operations to complete problems in the manner of a proof. Screen shots of sample exercises are displayed in Figure 1.

Periodically, students take online and written exams, which are graded by EPGY instructors, and all students are required to take a written final examination, which is submitted to EPGY via mail. Students are given three days to

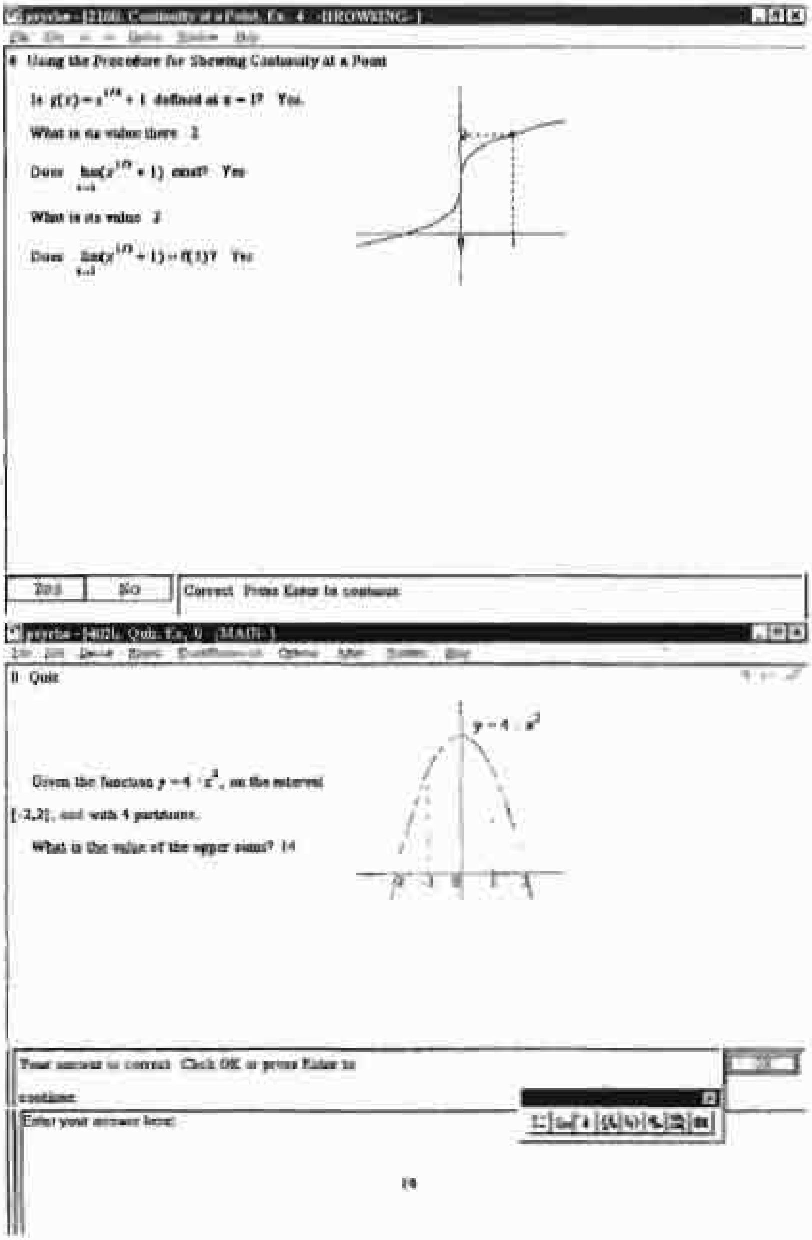


Figure 1. Sample exercise screens from the Calculus A and B courses.

complete the final exams as they are of substantial length and difficulty. Each course in the calculus sequence was designed to be run within a 3–4 month time period so that it would be feasible for a student to complete enough of the sequence in an academic year before the Advanced Placement exams in mid-May. The Linear Algebra course was designed to take 4–5 months on average for students to complete.

In addition to the software, students maintain regular contact with their assigned EPGY instructors. At a minimum, students must submit a data file to their instructors every week via email. This data file allows instructors to assess an individual student's strengths and weaknesses in detail and provide help when necessary. It is these data which provide the basis for most of the results of this study. If students have questions or trouble with the course material, they may contact their intructors by phone or attend optional help sessions through the "virtual classroom", an online conferencing environment which allows simultaneous discussion among instructors and students using a shared whiteboard. These sessions are used extensively as the AP examinations approach to give students any additional preparation that may be needed.

While the presentation of EPGY's courses is designed to closely replicate a classroom setting, the software allows students to gain a much greater degree of freedom and control over the information than is usually possible. Students may repeat lectures and lessons as many times as they wish, and may run the software at any time of day or day of the week. Students may start their courses on the first of any month and are often given leave from the courses for vacations, illness, or other mitigating circumstances. Students have up to a year of active participation to complete their courses successfully. We found that this level of freedom helps produce the wide differences in student performance that are evident from the data we report.

Student selection

The students in this study were selected from a pool of students who successfully completed courses begun on or after January 1, 1997, through February, 1999. The earlier date was selected because few important modifications or additions to the courseware have been made after that time. Also, improvements in EPGY's data-keeping system have allowed us to keep detailed records in an easily accessible format since then. We selected for particular study those student records which were relatively complete, i.e., missing less than five percent of response records. Unfortunately, this cut the number of students under consideration by a large percentage, since students often

Table 1. Breakdown of total student numbers in each course by age and sex. Ages are calculated at the time students began the courses

Student Demographics

	Calculus			Linear
	A	B	C	Algebra
Total	54	24	23	23
Male	42	20	19	21
Female	12	4	4	2
Average Age	14.9	14.9	16.3	16 3
Standard Deviation	1.7	1.7	1.4	1.2
Minimum Age	10.4	11.7	12.0	12.2
Maximum Age	18.1	17 2	18.4	18.0

do not complete their courses or fail to send in data regularly (A single missed report might disqualify a student from consideration). Also, in many cases, student data were corrupted, either because the clocks in the students' computers were incorrectly set, or email reports were not processed successfully. Once these selections were made, the student sample included a total of 54 students in Calculus A, 24 in Calculus B, and 23 in both Calculus C and Linear Algebra, for a total of 103 different students.[1] This limited the number of complete student performance records to a greater extent than we had hoped, although we were still able to observe a high degree of variation in student performance among the remaining population.

Students applying to the EPGY program must present evidence that they score within the top fifteen percent of mathematical ability, based on standardized test scores, including the SAT, PSAT, ACT, and AP exams, or other commonly recognized achievement and aptitude tests. In our sample, it was the case that most (over 90%) of the students submitted test scores that placed them in the upper 3% of mathematical ability. However, this characterization of students' ability is incomplete for a variety of reasons. First, these tests vary widely in scope and intent, and are normed against different populations. In addition, these tests in many cases were not intended for students of a young age as many were in our sample (see Table 1); other tests of basic skills suffer from ceiling effects when applied to gifted populations. Because the students in our sample submitted a wide variety of different standardized tests, we did not consider a detailed study of performance and admissions test data to be appropriate.

Instead, we consider students' age at the time they started the courses as a more relevant characterization of their abilities. As a rule, EPGY allows any student who is ready to begin calculus by their junior year of high school, based on demonstrated ability, to begin the calculus sequence. Table 1 displays the age distribution of the students. The average age of students in all courses is no higher than 16.3, which normally corresponds to sophomore or junior standing in high school. It is very rare for students to begin a university-level linear algebra course at this age, which is not usually taken by students until their second or third year of college. Thus it is age, rather than entering test scores, that we use as a more reliable overall statistic in classifying students' giftedness. In our study, we often found (as will be shown later) that there is a *negative* correlation between age and various performance measures we study, which gives further credence to the appropriateness of this measure.

Main results of individual differences

Trajectories

The first measure of performance that we studied was students' rates of progress throughout the courses. This can best be seen by plotting a chart of the student's position in the course as a function of time, which we denote as a "trajectory". Progress was measured both in terms of hours of computer time spent answering the exercises of the course, as well as the number of days since the course began. The first type of trajectory, which we label a "computer-time" trajectory, gives an accurate assessment of how many hours were required to complete the course exercises, regardless of how many days it took since the student began the course. The second type, or "calendar-time" trajectory, measures how many days it took for the student to reach any point in the course, regardless of how much computer time was required. From a theoretical viewpoint, the first measure is of greater interest, not only because the reporting system allowed us to measure computer time with a higher degree of precision, but because this measurement gives us a much finer-grained picture of the rates of student progress at various points in the course. Calendar-time trajectories are of greater interest from an administrative standpoint, since differences in overall completion times affect instructional resource allocations.

Computer-time trajectories

In order to plot the computer-time trajectories, the individual exercises in each course were first ordered according to their position in the course, where

Table 2. Total number of unique questions per course, along with the average, minimum, and maximum number of questions answered by each student

Number of Course Exercises				
	Calculus			Linear Algebra
	A	B	C	
Per Course:	1380	1142	539	403
Per Student:				
Average	1498	1162	501	351
S.D.	271	82	26	31
Minimum	1187	1017	463	314
Maximum	2401	1322	550	415

this position was denoted as the percentage of the course exercises a student must have completed to reach a given point. All exercises were given equal weight in this percentage ordering, i.e., the individual exercises were evenly spaced in the [0, 100] percent interval of overall course completion. The measurement of these times also included any exercises a student may have repeated; at various points in the courses, the program sets students back to the beginning of a lesson to repeat the material if they did not correctly answer enough of the questions in that lesson. The total numbers of distinct questions per course are displayed in Table 2.

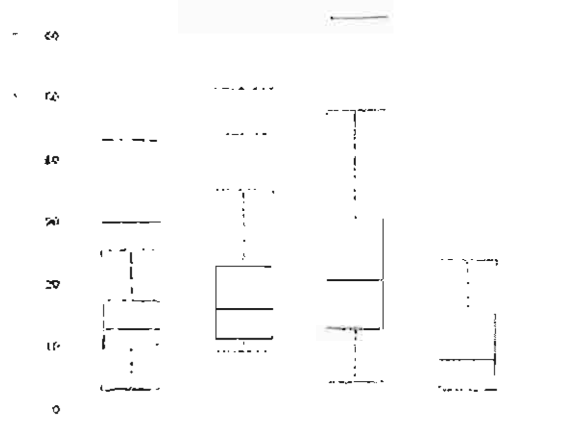
In this table, we see that the number of questions that each student actually answers is quite variable, especially in Calculus A and B. Not only are some students frequently required to repeat material, but some students are often allowed to skip exercises once they have demonstrated that they have mastered the relevant material. This often happens at the beginning of a course, where students who have taken EPGY courses in the past are able to skip introductory exercises relating to the software itself that they have seen before.

Exercise completion times were recorded to the accuracy of a second. Only the time between when students are first presented with questions to when they answer is recorded. It should be noted that students normally have two chances to answer questions correctly: if the first answer is wrong, information is displayed on the type of answer format expected by the computer, and students are then given the chance to re-enter their answers. Because it is also impossible to tell if a student is attending to the question during this entire time interval, the measurement includes any possible time

Table 3. Completion times for each course, in hours of computer time. Note the wide degree of variation in the distributions plotted in the boxplots. Minimum and maximum times differ considerably; the 75th and 25th percentiles generally differ by factors of 2 (For an explanation of the boxplot notation, see Appendix A)

Total Computer Time to Complete Courses (Hours)

	Calculus			Linear
	A	B	C	Algebra
Average Time	13.9	19.3	22.3	10 2
S.D.	6.7	11.2	13 9	6.7
Maximum	42.6	50.8	62.7	23.3
75%	16.7	22.1	28.3	14.5
Median	12.5	15 8	20.3	7.9
25%	9.3	10.9	13.1	5.2
Minimum	2.9	9.0	4.2	3.2



off-task, such as getting up from the computer to answer the telephone, take a break, etc. We rejected any data sets where it appeared that an inordinate amount of time was taken on any single question. Time spent watching lectures, doing homework, or using the derivation system is not included in these trajectories.

Table 3 shows the distribution, mean, and standard deviation of the completion times in hours for the students studied in this sample. The differ-

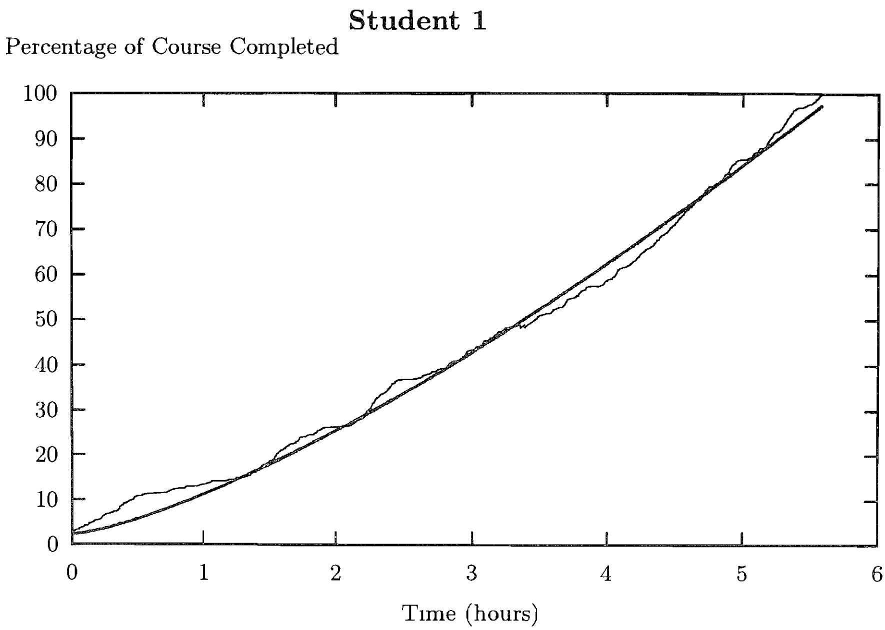
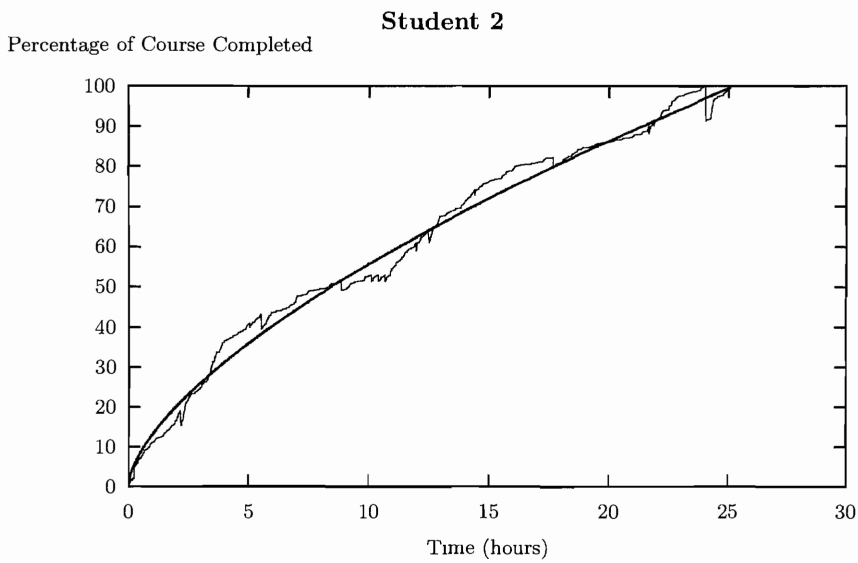


Figure 2. (continued on next page) Three sample student trajectories, with corresponding fitted power functions. Student 1 is a Calculus A student with a steadily increasing rate of progress. Student 2 is another Calculus A student with slower progress. Note that at several points in the course, Student 2 was required to review lesson material. Student 3 is a Calculus C student whose rate of progress is highly variable, and not as steadily increasing or decreasing as with Students 1 and 2. Also note that Student 3 began the course near the 10-percent mark, a common event in Calculus C, as students who have taken prior EPGY courses are permitted to skip the introductory exercises. We see that the fits are quite good in the cases of Students 1 and 2 In general, the fits for all students were no worse than that of Student 3.

ences between the minimum and maximum time for each course differs by at least a factor of 5, and at most, in the case of Linear Algebra, by a factor of 23. The 75th and the 25th percentiles differed by a factor of around two, a factor that also arose in a prior study of EPGY student performance (cf. Stillinger & Suppes 1994). Students were often able to answer questions extremely quickly, sometimes hundreds of exercises in a few hours of computer time.[2] As can be seen from the distributions, even discounting the most extreme scores, the range of student completion times were spread quite widely in each course. There was also no significant correlation between completion times and student age.

Three sample trajectories are plotted in Figure 2. In the first, we see a student who completed the course very quickly (all questions were completed in less than 6 hours). Also, the rate at which the student answered the



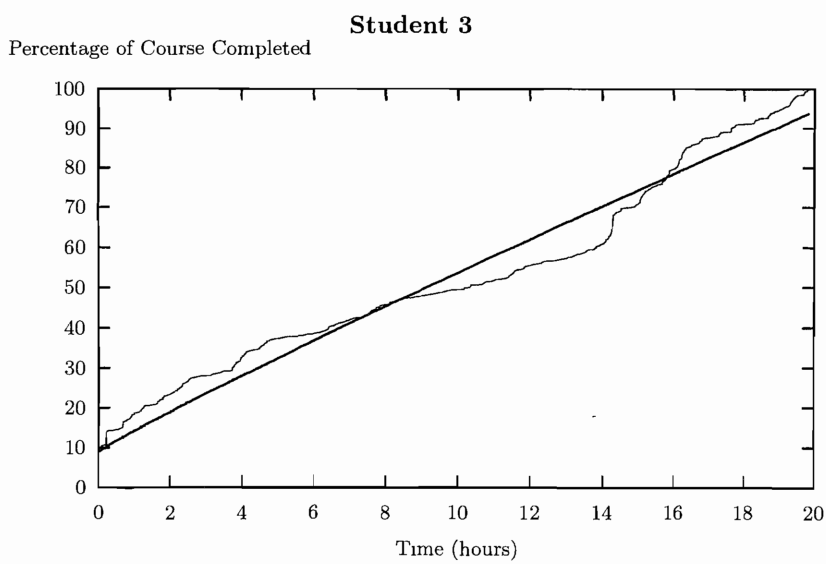


Figure 2. Continue.

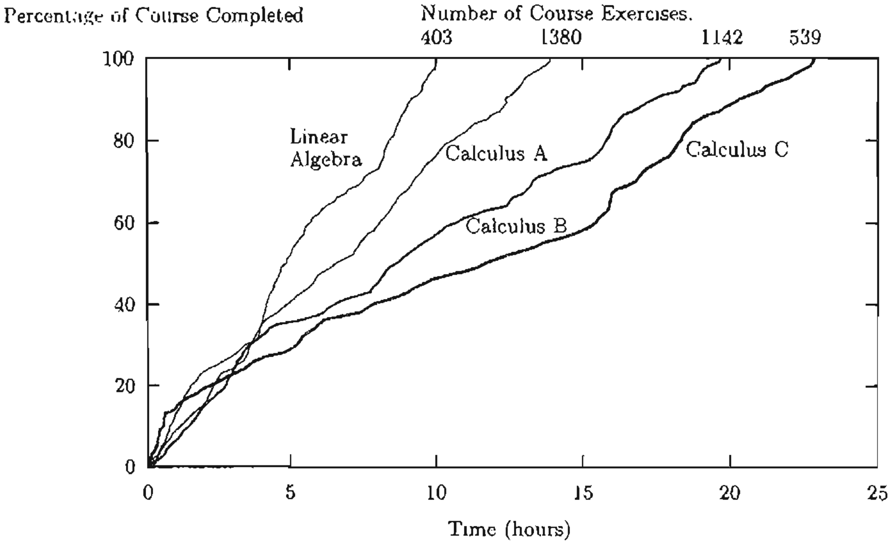


Figure 3. "Average" trajectories for each course. Curves were found by averaging total student latencies over each percentage block of exercises in the courses. The number of course exercises is listed above the trajectory for each course on top of the graph. Note the negative correlation between time and number of exercises for the calculus courses.

questions increased as he went through the course, which is, surprisingly, not unusual among the student performances analyzed here. The second trajectory shows the progress of a student who took much longer to complete the questions, and whose pace gradually slowed as the course went on. Such progress is more in line with our expectations for most students in the course. Also, the student often had to repeat lessons before the computer allowed him to progress. Finally, in the third trajectory, we see variable rates of progress for a student in the Calculus C course. This is not at all unusual for students in this course, as well as in Linear Algebra. Many students consistently found sections of the course to be of varying difficulty, which accounts for the variable changes in progress rate. This indicates that sections of the course may be too easy or too difficult for students given their overall progress in the course. Figure 3 gives a comparison of the "average" trajectories of students in each of the four courses studied here. We see that overall completion times vary greatly, although *not* in proportion to the number of course exercises.

The curves in Figure 3 were derived by averaging the amount of time students spent on each percentage subset of questions in the course, where these average times were then concatenated to give a composite trajectory of all students in the sample. In general, students take the longest amount of computer time to complete Calculus C, and the shortest amount for Linear

Algebra; interestingly, the ordering of course completion times does not correlate with the number of exercises in the courses. This runs counter to our experience that students actually take the longest amount of calendar time to complete Linear Algebra (see "Calendar-time trajectories" below). Also, the inflections in these curves show the relative levels of difficulty of each section in the course. We can see that although the curve for Calculus A (M040) is quite smooth, the curves for Calculus B, C, and Linear Algebra clearly show periods where student progress slows and accelerates. This indicates that the difficulty of the questions in these courses is unevenly distributed, and that perhaps certain sets of problems should be made easier or more difficult to match the average level of student capability at those points.

Prior research on student trajectories has shown that student progress through computer-based courses may be described by a power function of the form $y(t) = bt^k + c$, where t represents time, $y(t)$ represents course position as a function of time, and b, k, and c, are scale, shape, and position parameters, respectively (cf. Macken, van den Heuvel, Suppes, and Suppes (1976), Malone, Suppes, Macken, and Zanotti (1979), Suppes (1992), Suppes, Fletcher, and Zanotti (1975), Suppes, Fletcher, and Zanotti (1976), Suppes, Macken, and Zanotti (1978), and Suppes & Zanotti (1996)). A derivation of this model from basic, yet substantive, assumptions regarding the nature of information acquisition may be found in Suppes & Zanotti (1996).

The parameter k is important, since the shape of the power function is most sensitive to values of this parameter. Figure 4 shows power curves with various values of k to illustrate the degree to which changes in the value of this parameter can affect the shape of the curve. Note that each of these values of k was actually realized by some student in the sample (although values of b also varied greatly for these students). We should note that student review of material after all lessons were completed, as shown at the tail of Student 2's trajectory shown in Figure 2, was not counted as part of the trajectory in this fitting procedure; however, repetition of exercises in the middle of the course (as in the intermediate ranges of that same trajectory) were included. Tables 4 and 5 show the distributions of k and b for students in each course.

In fitting these curves, the value of c was set to the point in the course from which the student began, in case this was not the beginning. The remaining two parameters were fitted using a nonlinear least squares fitting procedure, by solving for the zeroes of the partial derivatives with respect to b and k the following sum for each student:

$$\sum_{\iota=1}^{n} \left(y(t_\iota) - bt_\iota^k - c\right)^2$$

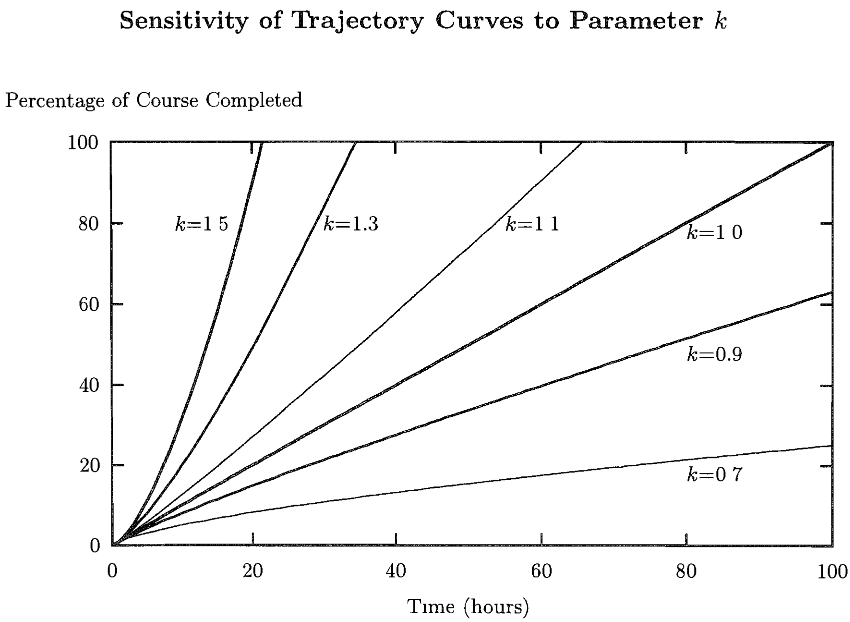


Figure 4. Shapes of various theoretical trajectory curves for different values of k. $b = 1$, $c = 0$ for all curves Curve shape is highly sensitive to changes in the value of k.

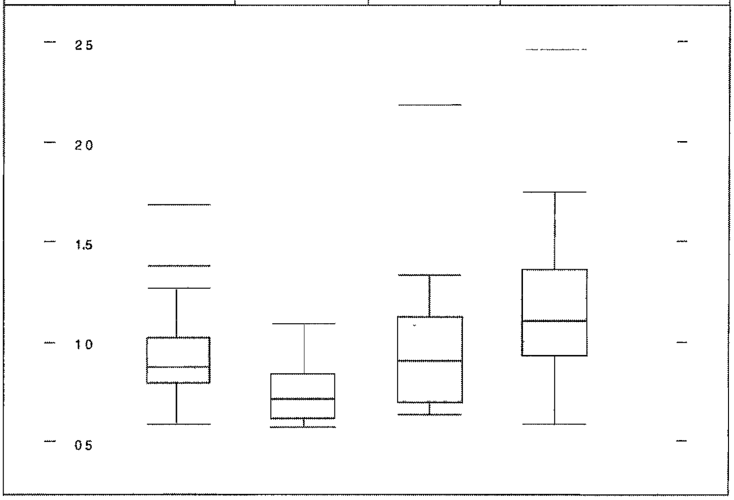
where n is the total number of exercises completed by the student, t_i is the cumulative amount of computer time elapsed on completing exercise i, $y(t_i)$ is the student's actual course position at time t_i, and c is given. Analytical solutions are available for the minimizing value of b for any given k; line search techniques were then used to find the minimizing value of k.

Correlating the parameter settings of the best-fit parameter curves with other data, we found no significant correlation with student age or average test or final exam score in all four courses. As we might expect, there was usually a negative correlation between values of b and k for each student and average student response latencies; surprisingly, however, in Calculus B, a *positive* correlation was found between k and the average latencies. In this case, a few data points where high k values were correlated with high average latencies had extremely low b values; indeed the b values had a strong negative correlation for this course in relation to the other courses (r = -0.74). This shows that, contrary to our expectation, the value of b for a parameter setting can sometimes be as important as k in determining the shape of trajectories.

Table 4. Estimated parameter values of *k* for each course. Note again the wide variation between minimum and maximum values. The 75th and 25th percentiles also differ by 0.2–0.4 in each course

Values of Parameter *k*

	Calculus A	Calculus B	Calculus C	Linear Algebra
Average	0.91	0.74	0 97	1.18
S.D.	0.19	0.14	0 34	0.40
Maximum	1.68	1.09	2.19	2.47
75%	1.02	0.83	1.11	1.34
Median	0.88	0.72	0.91	1.10
25%	0.80	0.62	0.70	0.96
Minimum	0.59	0.58	0.64	0.59



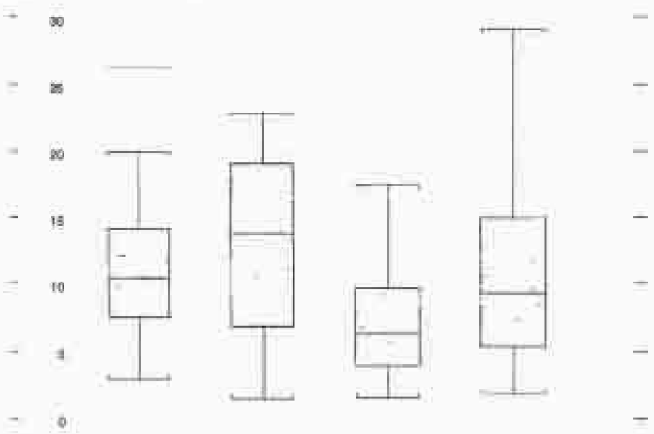
Calendar-time trajectories

The total length of time required for students to complete the courses, measured in days, also varied greatly. Figure 5 above shows a typical progress curve for a student who worked on the course consistently. Long flat areas show periods of days in which the student did little or no work. Table 6 shows the minimum, maximum, and average student completion times for each course. While the average completion times are very close to the suggested

Table 5. Estimated parameter values of b for each course. The large maximum value for Linear Algebra has been omitted from the boxplot. Again, the 75th percentile differs from the 25th percentile by about a factor of 2 for each course

Values of Parameter b

	Calculus A	Calculus B	Calculus C	Linear Algebra
Average	11.3	13.4	7.3	15.4
S.D.	4.7	6.3	4.4	21.0
Maximum	26.3	22.9	17.4	103.9
75%	14.2	19.0	9.6	16.7
Median	10.5	13.8	6.4	9.5
25%	7.7	7.0	4.1	5.4
Minimum	2.9	1.4	1.5	1.9



progress schedules given by EPGY, the minimum and maximum completion times again differ by about an order of magnitude. Because the program was entirely self-paced, students could complete the course in as little or as much time as could be allowed.[3] It should be noted, however, that students were allowed to take vacations from the course; the statistics we report here would include this time away. This accounts for some of the high maximum times,

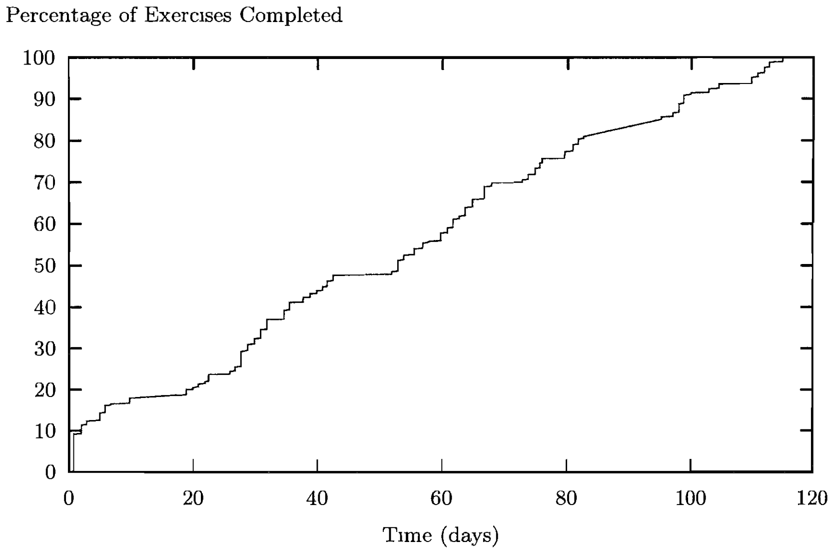


Figure 5. Typical student progress trajectory measured in calendar time for Calculus A.

and why students were allowed to continue past one year, which is normally the maximum time students are allotted to complete any EPGY course.

What is most interesting about the calendar times is the nearly complete lack of correlation with computer times, with the exception of the Linear Algebra course, where there is a *negative* correlation. This shows that students who take longer to answer questions are just as likely to complete the course in the same number of days as those who answer questions quickly. See Appendix B for correlations with other performance measures.

Error rates

Overall student error rates were also studied; in general, these did not vary greatly for the calculus courses, although there was greater variation in Linear Algebra. Table 7 shows the minimum, maximum, and average values of average error rates on computer-based exercises. Table 8 gives summary statistics for average test and final exam scores for each course. Calculus C has online tests, which were counted among the regular computer-based exercises; Linear Algebra has a midterm but no final exam. Average test scores for these two courses were not reported.

There was generally a correlation found between students' average exercise scores, test scores, and final exam scores. For Calculus A and B, the correlations were stronger between exercise and test scores (r = 0.42 and

Table 6. Completion times for each course, in days of calendar time. Again, minimum and maximum time differ by orders of magnitude. Average completion times are close to EPGY's suggested progress schedules

Course Completion Times (Days)

	Calculus			Linear
	A	B	C	Algebra
Average	110	100	103	155
S.D.	69	45	89	94
Maximum	466	188	395	373
75%	121	133	107	218
Median	92	90	85	117
25%	76	54	53	85
Minimum	30	30	30	33

0.32, respectively); for Calculus C and Linear Algebra, strong correlations were found between final exam scores and average exercise scores (r = 0.53 and 0.56, respectively). (Recall that no average test scores were available for Calculus C and Linear Algebra.) In addition, as was expected, a negative correlation was found between the number of exercises completed and average exercise score in Calculus A and B ($r = -0.74$ and –0.26, respectively). These are the courses where students are regularly set back in the courses to review lessons should their exercises scores be too low.

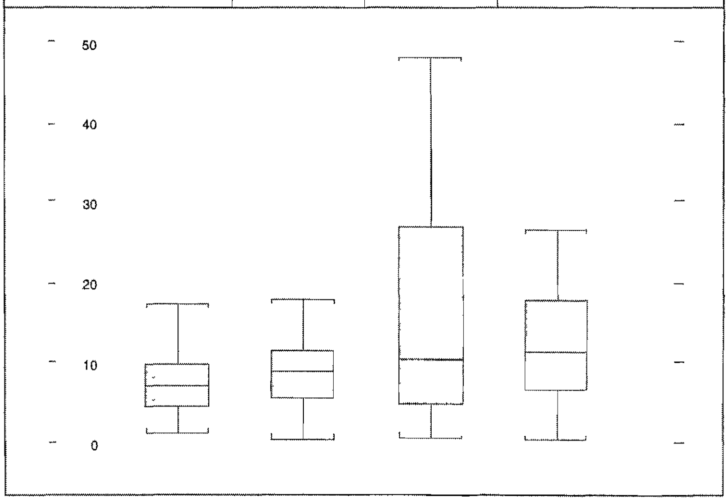
Also, there was in most cases a slight *negative* correlation between age and exercise and test scores. That is, younger students generally had better test scores than older students. In Calculus A, these correlations were not significant ($-0.06 < r < 0$); in Calculus B there was a negative correlation between age and test scores ($r < -0.31$ for final exams), but a slight positive correlation with exercise scores ($r = 0.11$). In Calculus C and Linear Algebra the correlations were generally insignificant, although there was a negative correlation between age and exercise scores in Linear Algebra ($r = -0.33$). Again, see Appendix B for a complete listing of correlations.

This last finding deserves comment, especially in light of our earlier remarks on using age as an indicator of giftedness. One would certainly expect positive correlations between various performance measures and students' age. The opposite result may be explained by some bias that may arise in the selection process. These students were not approached by EPGY

Table 7. Student error rates for computer-based exercises as a percentage. Again, the 75th percentile differs from the 25th percentile by about a factor of 2 for each course, except Calculus C, where the ratio between the percentiles is well over 4. The boxplots reflect the large degree of spread in the error rates for this course, which consistently appears to be the most difficult of all the four courses by the performance measures studied here

Error Rates, Computer-Based Exercises

	Calculus			Linear
	A	B	C	Algebra
Average	7.6	8.8	14.9	12 3
S.D.	3.8	4.3	12.6	7.4
Maximum	17.3	17.8	48.1	26.5
75%	9.9	11.2	24.0	17.5
Median	7.6	9.0	10.6	11 5
25%	4.7	5.8	5.3	6.8
Minimum	1.2	0.5	0 5	0.4



to participate in these courses; rather, they or their parents usually contacted EPGY because they felt that their educational needs were not adequately being met. It has been our experience that the greater these perceived needs are, the earlier will a substitute educational program be sought; thus the youngest students who come to EPGY to study mathematics well beyond

Table 8. Student error rates in online and offline tests, and "take-home" final exams, expressed as percentages

Student Error Rates, Tests and Final Exams					
		Calculus			Linear
Exercises		A	B	C	Algebra
Test Scores	Average Error Rate (%)	17.7	16.3	n/a	n/a
	Standard Deviation	7.7	7.3	n/a	n/a
	Minimum Error Rate	4.4	3.6	n/a	n/a
	Maximum Error Rate	34 7	37.2	n/a	n/a
Final Exams	Average Error Rate	13.9	8.9	18.0	4.4
	Standard Deviation	9.5	7.7	15.0	5.3
	Maximum Error Rate	42.0	33 0	60.0	22.0
	Minimum Error Rate	0.0	0.0	0.0	0.0

their current grade level are usually the very brightest. We plan to pursue follow-up studies to test this hypothesis.

Markov analysis of student error rates

Order analysis

A question which arises in examining the sets of student responses to questions is whether correctness of response is dependent on prior responses, i.e., on the recent history of correct responses. If such a dependence exists, then information from prior questions may influence student error rates on current questions. This may or may not be desirable, depending on the context in which the question is asked. In the context of assessment or evaluation of student knowledge within an accelerated curriculum, or immediate review of lecture material, the maximum efficiency of information transmission is optimal. That is, questions on an examination should cover the maximal amount of material in the fewest number of questions. In a computer-based curriculum, adaptive movement algorithms and problem-selecting techniques may be used to achieve a high degree of efficiency based on the current set of student responses. While such algorithms feature prominently in the design of EPGY's K-7 mathematics curriculum, they are not in any meaningful sense part of the calculus and Linear Algebra courses. The questions are fixed, although students may be forced to review sections where they got a low score. A study of this sort to find dependencies may indicate how well-tailored to average student learning rates these questions are.

On the other hand, in cases where questions are meant to elucidate a concept or theorem by forcing students to complete steps in an overall (guided) derivation, such a dependence is highly desirable. If the sequential presentation of the questions is critical to the student's ability to correctly answer all the questions, then information must be carried over from one question to another. Thus the conceptual dependence of questions on prior questions should be reflected in a dependence in student errors on responses to prior questions. While the dependence may be strongest on the question immediately prior to the current question, there may also be dependence to arbitrary numbers of questions in the past. The measure of the order of dependence can thus serve as a measure of the complexity or the conceptual interrelation of steps in a derivation.

Unfortunately, it was impractical to sort the questions by their pedagogical intent, as might be desired from the above comments. Rather, we examined the entire sequence of student responses for each student for dependencies, without regard for the types of questions (i.e., evaluative, explanatory) that were asked. For the test of first-order dependence, the null hypothesis is that the sequence of correct answers is a Bernoulli phenomenon, independent of prior responses. Predicting a correct response would then be like predicting the outcome of a coin toss. Against this we test the hypothesis that there is a dependence on the correctness of the prior response. Second-order dependence, where the response to the next question depends on the result of the prior *two* questions, may then be tested by comparing the result to the hypothesis of first-order dependence. Higher-order dependencies are treated in an analogous fashion, by comparing them to the hypothesis that the order is one less than the order under consideration.

To perform this analysis, we treat the series of student responses as a Markov chain. According to this model, the state space is defined in terms of the recent history of student responses. If the model has order $n \geq 1$, then the state space is the space defined by the Cartesian product $\{0, 1\}^n$, where 0 denotes an incorrect and 1 a correct response. Thus, in the case of a second-order Markov chain, we would use the notation '(1,0)' to denote that the prior two responses were correct and incorrect, respectively (where a 1 denotes a correct response and a 0 denotes an incorrect response). By looking at the entire sequence of responses in terms of their correctness for a given student, transition probabilities may be estimated for any order of interest according to the frequencies of such transitions in the student history. Based on this assignment of probabilities, statistical tests may be applied to find the highest order that may meaningfully be ascribed to the sequence of correct responses.

Statistical tests for determining the order of this Markov process are found in Suppes & Atkinson (1960). To summarize, a χ^2 test is used, where for comparing the Bernoulli hypothesis to the first-order hypothesis, χ^2 is calculated as

$$\chi^2 = \sum_{i,j} n_i \frac{\left(\frac{n_{ij}}{n_i} - \frac{n_j}{N}\right)^2}{n_j/N},$$

where $i, j \in \{0, 1\}$, N is the total number student responses, n_{ij} represents the number of state transitions from state i to getting response j; $n_i = \sum_j n_{ij}$; $n_j = \sum_i n_{ij}$. The degrees of freedom for the χ^2 test are equal to $(m-1)^2$, where m is the number of states. Since $m = 2$ for this case, there is one degree of freedom in this test.

For testing the second-order hypothesis against the first, χ^2 is calculated as:

$$\chi^2 = \sum_{i,j,k} n_{ij} \frac{\left(\frac{n_{ijk}}{n_{ij}} - \frac{n_{jk}}{n_j}\right)^2}{n_{jk}/n_j},$$

where $k \in \{0, 1\}$, n_{ijk} is the number of state transitions from state "ij" (recalling the state notation described earlier) to getting response k, $n_{ij} = \sum_k n_{ijk}$, $n_{jk} = \sum_i n_{ijk}$, and n_j is defined as before. For this test there are $m(m-1)^2 = 2$ degrees of freedom. Note that this latter test can easily be generalized to tests of any order.

In calculating the numbers of response transitions, all reported student responses were used, including student review. The transitions between groups of exercises were not used in the calculations, since these points formed a natural break in the series of exercises.

The sequence of responses for the students in this study showed that there is usually a Bernoulli or first-order dependence among responses. Using 95% certainty as our criterion for these tests of hypotheses, most students in Calculus A and B showed a second-order dependence or less, while most students in Calculus C and Linear Algebra showed a first-order dependence or less. Rarely did any student show more than a second-order dependence. Table 9 shows the percentages of students falling into each category.

From this, we see that for the overall set of exercises, the order dependency is low, often independent. Since the majority of questions in the courses are quizzes and review exercises, this is a welcome finding, since this indicates that the questions in a lesson are approximately independent of each other. This is not to say that questions in an individual lesson are not conceptually related; they most certainly are. However, the information content of sequential questions does not overlap. Thus even if a student is told if their answer is

Table 9. Percent of students whose response sequences displayed Bernoulli, first, and second or higher-order dependence on prior responses

Results of Order Analysis				
Course	Bernoulli	First-Order	Second-Order	Third-Order+
Calculus A	31%	35%	25%	9%
Calculus B	33%	45%	22%	0%
Calculus C	49%	42%	9%	0%
Linear Algebra	92%	8%	0%	0%

wrong (as always happens), or whether an explanation of the question occurs after an incorrect response (as generally does not happen with quiz and test questions), the questions cover a varied and wide enough range of information that this does not affect performance on later questions. Thus from an evaluative standpoint, the questions seem to cover an appropriate range of information.

Stationarity

The above results on order are predicated on the stationarity of the sets of state transitions over time. That is, the probability of responding correctly or incorrectly on a given question with a given recent response history should be independent of where in the course the student is. To test this hypothesis, a χ^2 test of homogeneity was used, modified from the one given in Suppes & Atkinson (1960). For each block of 100 consecutive responses, the transition frequency between states was compared to the overall frequencies for the entire set of responses. To determine if there are significant differences in the overall sets of responses, a χ^2 value for each row in the transition table for first-order states as

$$\chi_i^2 = \sum_{t,j} n_i(t) \frac{\left[\frac{n_{ij}(t)}{n_i(t)} - \frac{n_{ij}}{n_i}\right]^2}{n_{ij}/n_i},$$

where $i, j \in \{1, 2\}$, t ranges over the number of blocks of 100 consecutive responses, $n_{ij}(t)$ is the number of responses j while the student is in state i during block t, $n_i(t) = \sum_j n_{ij}(t)$, and n_{ij}, n_i are defined as before over the entire set of responses. The overall χ^2 value is the sum of χ_1^2 and χ_2^2. The number of degrees of freedom is then tallied as $m(m-1)(T-1)$, with $m = 2$ as before and T being the number of complete blocks of 100 exercises. Again,

Table 10. Numbers of students whose response sequences were stationary, based on a 95% confidence criterion on the χ^2 test of homogeneity

Results of Stationarity Analysis		
Course	Stationary	Non-stationary
Calculus A	31	23
Calculus B	14	10
Calculus C	15	8
Linear Algebra	21	2

in tabulating the numbers of state transitions, transitions between exercises spanning different lesson groups were not counted.

Using again a 95% confidence criterion in our χ^2 test, the student response sequences were stationary for the majority of students. Table 10 shows the numbers of students in each course whose response sequences were stationary. In general, most students had stationary response sequences, although in the calculus sequences a substantial number did not. This shows us that in these courses, the distribution of exercises where first-order Markov properties apply are not evenly distributed through the courses.

Response latencies

Another performance measure of interest was the distribution of the lengths of time students spent answering questions in the course. In the current study we have found that a lognormal distribution fits the data well. Before we look to the shapes of these distributions in detail, we first discuss the overall average latencies of students in each course.

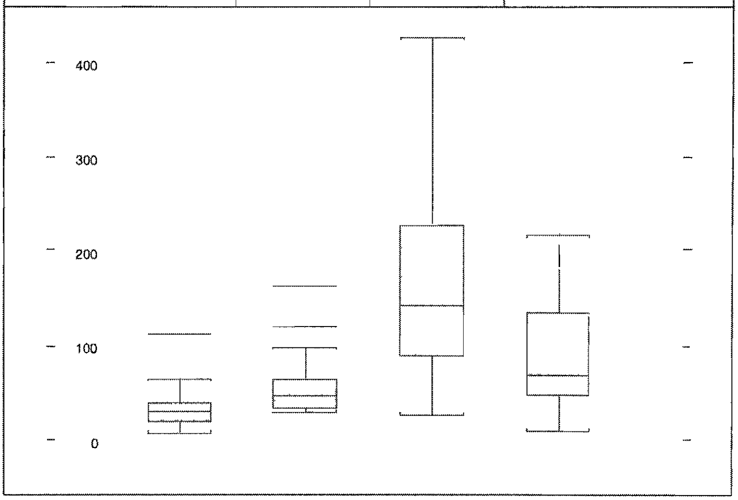
Average latencies per student

We have found that average latencies of students vary widely both within and across courses. These averages were obtained by observing many hundreds of response times for each student. In general, the time required for students to answer each question becomes longer as the courses are more advanced; although the average latencies for Linear Algebra are less than that for Calculus C. Table 11 shows the distribution, average, and standard deviation of the average latencies for students in each course. The maximum times are many times greater than the minimum times; there is also a high degree of variability in relation to the average across all courses.

Table 11. Distributions of latencies per student in each course, in seconds. Note again the factor of two separating the 25th from the 75th percentile in each distribution

Student Average Latencies

	Calculus			Linear
	A	B	C	Algebra
Average	33.5	57.8	158.1	94.0
S.D.	16.9	33.3	96.7	62.0
Maximum	112.2	163.5	427.5	217.7
75%	39.6	62.9	206.3	131 9
Median	30.7	57.8	142.8	70.1
25%	21.0	33.8	96.0	49.1
Minimum	7.7	30.2	27.8	9.1



In general, no strong correlations were found between average latencies and student age, average percentages of correct response on computer-based exercises, and overall student calendar time in the course. However, a negative correlation between age and average latencies was observed in Calculus B ($r = -0.31$), and in Linear Algebra, a positive correlation between overall calendar time and average student latency ($r = 0.48$).

Distributions of latencies per student

Prior research has indicated that a gamma distribution fits the distributions of average latencies well when individual exercises are taken as the unit of analysis (cf. Stillinger & Suppes (1994)). In this study, we find that this does not carry over to the distribution when the individual student is taken as the unit of analysis, as we have done here. The lognormal distribution fit the observed distribution of latencies for each student much better than did the gamma distribution, as was previously found. The density function of the lognormal is given by:

$$f(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\sigma^2}} \exp \frac{-(\ln x - \mu)^2}{2\sigma^2} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Thus, a random variable X has a lognormal distribution with parameters μ and σ if and only if $\ln X$ has a normal distribution with parameters μ and σ. It should be cautioned that μ and σ do not retain the same meanings for the lognormal distribution: μ is not the mean nor is σ the standard deviation. Rather, σ should simply be understood as a shape parameter and μ a scale parameter. Figure 6 shows two representative latency histograms with fitted lognormal curves for two students in Calculus B.

Table 12 shows the minimum, maximum, and average parameter settings for best-fit lognormal distribution parameters for students in each course. Note that small variations in parameter μ change the scale of the distribution significantly. As can be seen in the table, there is quite a large difference in minimum and maximum μ-values in each course, especially in Calculus C and Linear Algebra. This parallels the findings in the last section of the order-of-magnitude differences in the averages of student latencies.

A few remarks on how these distributions were fitted are in order. The parameters were first estimated using the maximum likelihood estimators for the lognormal:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln X_i}{n}, \quad \hat{\sigma} = \left[\frac{\sum_{i=1}^{n} (\ln X_i - \hat{\mu})^2}{n} \right]^{1/2}$$

where X_i represents the ith of n total observations of student latencies. From this point, an iterative binary line-search technique was employed to find any local points with lower χ^2 values; this process was halted when the estimates of the bounds of the best parameter settings were both within 0.01 of the previous estimates. For the most part, the maximum-likelihood estimates gave very good approximations to the final estimates; generally they were within 1% of the minimum found through the search technique. The latencies considered were those ranging between 1 and 300 seconds; any values

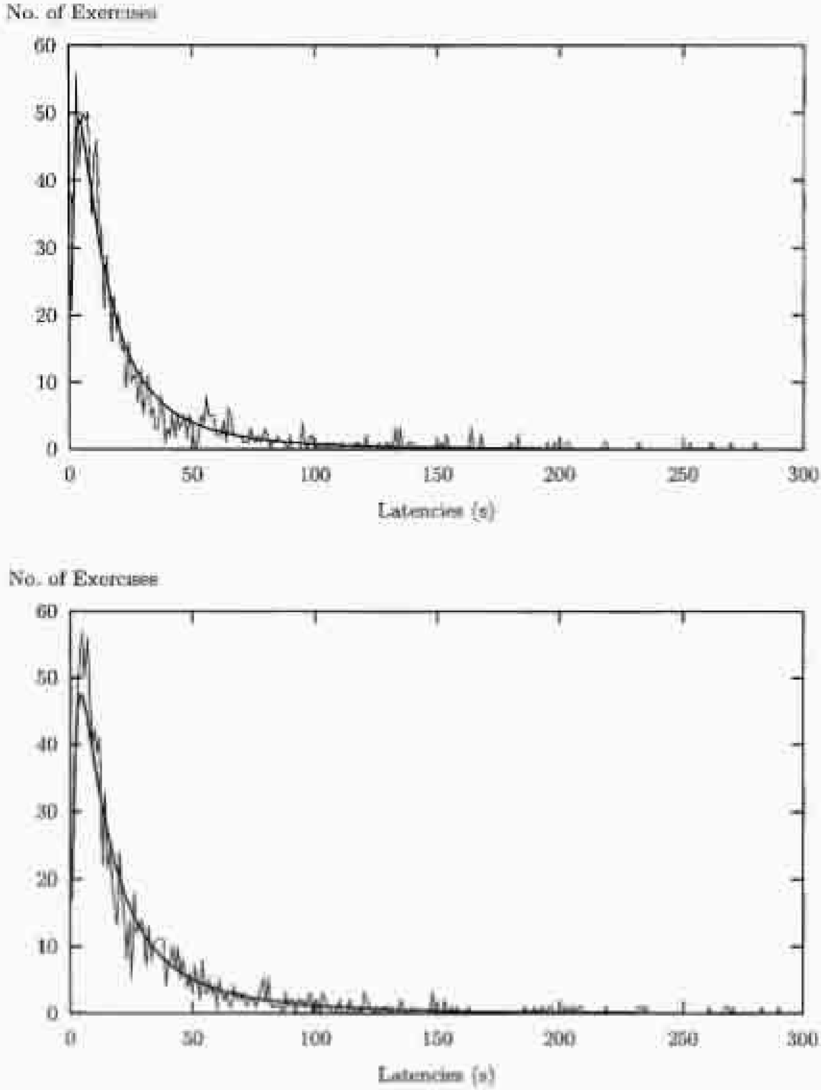


Figure 6. Two sample histograms of student latencies in the Calculus B course with fitted lognormal density curves. The distribution in the upper figure has parameters $(\mu, \sigma) = (2.66, 1.10)$; for the lower figure, $(\mu, \sigma) = (2.80, 1.15)$.

Table 12. Minimum, maximum, and average parameter settings for best-fit lognormal distribution curves to student latency histograms

Fitted Lognormal Parameter Values to Latency Distributions

Parameter		Calculus A	Calculus B	Calculus C	Linear Algebra
μ	Average	2.62	2.82	3.80	3.17
	S.D.	0.35	0.35	0.59	0.75
	Minimum	1.76	2.05	2.13	1.66
	Maximum	3.35	3.47	4.68	4.40
σ	Average	1.15	1.29	1.51	1.24
	S.D.	0.10	0.12	0.18	0.13
	Minimum	0.94	1.10	1.13	1.02
	Maximum	1.34	1.45	1.82	1.60
χ^2	Average	6.45	4.66	7.38	4.79
	S.D.	2.14	1.45	3.07	1.65
	Minimum	2.93	2.47	2.26	2.33
	Maximum	11.91	8.36	14.61	8.36

over 300 were discarded. Such values accounted for less than 1% of the total observed latencies; by discarding these, we also filtered out a large number of responses where students were interrupted for significant periods of time from solving the problem at hand. The χ^2 values were calculated by grouping the observations up to 100 seconds in groups of 5, and the observations above 100 seconds in groups of 25; this insured that enough (> 5) observations would be in each "bin" for the χ^2 statistic to be meaningful. Thus there were 28 "bins" altogether, giving us $28 - 2 = 26$ degrees of freedom for the χ^2 statistic since two parameters were estimated from the data. Therefore, even the worst fit among all the distribution fittings ($\chi^2 = 14.61$) was well within a 95% confidence region that the fitted densities matched the observed frequencies.

Conclusions

The one outstanding feature of our statistical analyses is the high degree of variation in student performance. While our data sample of students was small, the consistently high deviations in measured values leads us to believe that, at the high end of student achievement levels, many different learning

Table 13. Ratios of 75th to 25th percentile for many performance measures studied in this report

Ratios of Performance Measures
75th to 25th Percentile

	Calculus			Linear
Measure	A	B	C	Algebra
Computer Time	1.8	2.0	2.2	2.8
Parameter k	1.3	1.3	1.6	1.4
Parameter b	1.8	2.7	2.3	3.1
Calendar Time	1.6	2.5	2.0	2.6
Error Rate	2.1	1.9	4.5	2.6
Response Latency	1.9	1.9	2.1	2.7

rates and patterns are manifest. Generally, we have found that the upper quartile differs by a factor of approximately two from the lower quartile of performance on such measures as total completion time, trajectory parameter values, and calendar time to completion, and error rates. Table 13 shows the ratios between the 75th and 25th percentiles for a number of the measures studied in this report. This degree of variation suggests that an individualized curriculum may best meet the needs of students in this ability range.

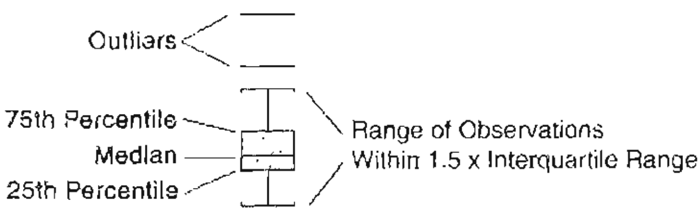
Also of note is the low degree of correlation between the performance measures studied in this report (see Appendix B for tables of these correlations in each course). Notably, performance measures do not generally correlate with student age, and where some degree of correlation is present, the correlation is usually negative with the level of merit of each performance measure. Thus, younger students often perform better than older students. In some cases, a degree of correlation is to be expected, e.g., the best trajectory parameter fits generally correlate with each other and with total computer time, and the number of exercises correlates with the average exercise scores (as students are often required to repeat exercises when their scores are poor.) Also, in most courses, a slight positive correlation was found between the average exercise scores and the final exam scores. Otherwise, the correlations found are consistent with the hypothesis that the performance measures collected here are approximately independent of each other, indicating that a full characterization of student performance cannot be reduced to one or two parameters.

As we have seen from the Markov order analysis of the student trajectory data, we can find quantitative measures of the efficiency of information presentation in the course exercises, and the stationarity analysis suggests methods of identifying sections of the courses where these order properties are suboptimal. Other findings in this report may be used to evaluate and improve upon the existing courses in a number of ways. First, the average trajectory measures in each course can help us to identify parts of the course that are relatively too easy or too hard given their context in the course. Second, the use of order analysis can help to ensure that questions are fulfilling their pedagogical intent, be they evaluative questions or questions designed to introduce a new concept. Third, error analysis allows us to identify challenging and trivial questions in the course, and ensures that a proper distribution of questions of varying levels of difficulty is present throughout the courses.

Notes

1. Some students took more than one of these courses.
2. It should be mentioned that students are allowed to "browse" ahead in the course to preview exercises before they take them officially; however, students are not allowed to browse ahead to preview quiz questions. In all, though, it could not be determined from the available data whether students had previewed the exercises or not. Also, all students with extremely low completion times had respectable error rates on the exercise questions.
3. Tuition was billed monthly for the calculus courses, however, regardless of how much material students covered. Thus, students (and their parents) had a strong incentive to finish the courses early.

Appendix A: Guide to boxplots



Appendix B: Correlation matrices

	Student Age	Avg Test Score	Final Exam	Calendar Time	No of Exercises	Avg Ex Score	Computer Time	Parameter *b*
Calculus A								
Average Test Score	−0.02							
Final Exam Score	−0.06	0.09						
Calendar Time	−0.07	−0.06	0.17					
Number of Exercises	0.00	−0.47	0.10	−0.03				
Average Exercise Score	−0.04	0.42	−0.11	−0.02	−0.74			
Computer Time	0.01	−0.09	0.14	−0.05	0.09	−0.14		
Trajectory Parameter *b*	−0.10	0.38	−0.03	0.11	−0.29	0.20	−0.52	
Trajectory Parameter *k*	0.07	−0.20	−0.11	0.09	0.09	0.15	−0.31	−0.40
Calculus B								
Average Test Score	−0.20							
Final Exam Score	−0.31	0.13						
Calendar Time	0.00	−0.15	−0.17					
Number of Exercises	−0.09	−0.06	−0.41	−0.10				
Average Exercise Score	0.11	0.32	0.19	−0.12	−0.27			
Computer Time	−0.27	−0.09	0.08	0.00	0.12	0.00		
Trajectory Parameter *b*	0.13	0.07	0.06	0.08	−0.13	0.04	−0.71	
Trajectory Parameter *k*	0.07	0.15	−0.03	−0.18	−0.15	0.06	0.29	−0.76
Calculus C								
Average Test Score	−0.56							
Final Exam Score	−0.03	0.25						
Calendar Time	0.12	−0.23	0.07					
Number of Exercises	−0.08	−0.14	0.20	−0.05				
Average Exercise Score	−0.12	0.23	0.53	0.04	0.12			
Computer Time	−0.01	−0.17	0.08	−0.03	0.12	−0.14		
Trajectory Parameter *b*	0.05	0.01	−0.07	0.21	−0.31	−0.01	−0.33	
Trajectory Parameter *k*	0.22	0.10	0.15	−0.06	0.30	0.40	−0.47	−0.45
Linear Algebra								
Final Exam Score	−0.17							
Calendar Time	0.20	−0.32						
Number of Exercises	−0.14	0.01	−0.46					
Average Exercise Score	−0.33	0.56	−0.32	0.21				
Computer Time	0.06	−0.13	0.49	−0.25	−0.27			
Trajectory Parameter *b*	−0.05	−0.50	0.34	0.34	0.11	−0.50		
Trajectory Parameter *k*	−0.06	−0.41	0.21	0.21	0.27	−0.41	0.12	

References

Dai, D.Y., Moon, S.M. & Feldhusen, J.F. (1998). Achievement motivation and gifted students· a social cognitive perspective. *Educational Psychologist* 33: 45–63.

Fuchs, L.S., Fuchs, D., Hamlett, C.L. & Karns, K. (1998). High-achieving students' interactions and performance on complex mathematical tasks as a function of homogeneous and heterogeneous pairings. *American Educational Research Journal* 35: 227–268.

Gohm, C.L., Humphreys, L.G. & Yao, G. (1998). Underachievement among spatially gifted students. *American Educational Research Journal* 35· 515–531.

Humphreys, L.G. & Lubinski, D. (1996). Assessing spatial visualization. In C.P. Benbow & D. Lubinski, eds, *Intellectual Talent*, pp. 116–140. Baltimore, UP: Johns Hopkins.

Macken, E., van den Heuvel, R., Suppes, P. & Suppes, T. (1976). *Home-Based Education: Needs and Technological Opportunities.* Washington, DC: National Institute of Education.

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.

Stillinger, C. & P. Suppes. (1994). Gifted students' individual differences in computer-based algebra and precalculus courses. Stanford, CA: EPGY Technical Report. Available at http://epgy stanford.edu/Other/index.shtml?research.html.

Suppes, P (1992). Estes' statistical learning theory: past, present and future. In A.F. Healy, S.M. Kosslyn & R M. Shiffrin, eds, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, pp. 1–20 Hillsdale, New Jersey: Lawrence Erlbaum.

Suppes, P. & Ager, T. (1995) Computer-based advanced placement calculus for gifted students. *Instructional Science* 22: 339–362

Suppes, P. & Atkinson, R C. (1960). *Markov Learning Models for Multiperson Interactions.* Stanford, CA· Stanford UP.

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* 2: 117–127.

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: Lawrence Erlbaum.

Suppes, P, & Zanotti, M. (1996). Mastery learning of elementary mathematics: theory and data. In P Suppes, ed , *Foundations of Probability with Applications*, pp. 149–188. Cambridge University Press.