Monitored Design of an Effective Learning Environment for Algebraic Problem Solving

Kenneth R. Koedinger

Erika L. F. Sueker

Human-Computer Interaction Institute

Carnegie Mellon University

Pittsburgh, PA 15213

Please send correspondence to Ken Koedinger at the address above. Phone: 412-268-7667.

Email: koedinger@cmu.edu.

Abstract

We describe a formative design experiment in which we adapt and evaluate the Practical Algebra Tutor, "PAT," for integrated use in developmental algebra classrooms at the college level. PAT is a software learning environment that presents students with real-world problem situations, modern mathematical representational tools to analyze these situations, and constant background support from a "cognitive tutor"-- an intelligent computer tutor based on the ACT theory of cognition (Anderson, 1993). In two colleges, students who used PAT solved a performance-based assessment better than those who did not. This assessment required the use of mathematical representations to analyze a real-world problem situation and captured the reform objectives of the PAT approach, which are consistent with new national standards for mathematics. Changes over three semesters that increasingly integrated the technology into the classroom curriculum doubled student performance; students in control classes at one university had average scores of 30% after the first semester while students in experimental classes were averaging 67% by the third semester. Use of PAT provided impetus for universities to reform their courses and these reforms appeared to contribute to the observed increases in student learning across semesters.

Monitored Design of an Effective Learning Environment for Algebraic Problem Solving

Formative design experiments are research efforts that have the goal of developing, testing and disseminating innovations, similar to the way engineers develop new products (Collins, 1992; Brown, 1992). Collins (1992) discussed the use of this design methodology for the development of effective educational innovations. He describes eight features necessary for an effective instructional design experiment, such as employing multiple assessments, including instructors as co-investigators, and using formative evaluation methods. Brown (1992) expanded this notion of a design experiment. She stressed that success in the learning sciences requires going beyond laboratory instructional experiments to face the complex issues of engineering instructional solutions in real learning environments. She emphasized the need to guide instructional design by theoretical principles and to evaluate the design in terms of the instructional objectives of multiple stakeholders. We present an elaboration of Collins' and Brown's design experiment approach which we call "monitored" design, to emphasize the importance of both (a) identifying problems and objectives and (b) continual measurement and monitoring of achievement of these objectives.

Table 1 presents a summary of our extensions to the design experiment methodology. These extensions are based on our experience in an ongoing design experiment in which our goal is to develop a new classroom environment for learning algebra at the college level. The design began with the addition of a technological innovation, a cognitive tutor (Anderson, Corbett, Koedinger, & Pelletier, 1995), to a traditional algebra classroom at two universities. Following that, the technological innovation was integrated more fully into the curriculum, although the traditional course format was retained. Finally the entire algebra course was redesigned with the cognitive tutor at the heart of the reform effort. We report on how these changes led to continued student performance increases across three semesters.

Our discussion of this project will follow the 5 major components of monitored design summarized in Table 1: (1) goals and problem identification, (2) theory, (3) design solution, (4) formative evaluation, and (5) summative evaluation. The remainder of this section will summarize

the design problem and goals, the underlying theory that guides design, and the proposed design solution. Following this introduction, we describe the formative evaluation process that has tested and refined this design solution in studies across three semesters. We will then return to draw lessons from this design experiment in terms of the monitored design components and features summarized in Table 1.

Insert Table 1 about here

The Design Problem: Challenges of Developmental Mathematics

A design experiment should begin with a careful assessment of the current learning environment, in particular, the existing needs and problems of the stakeholders-- those who will benefit from and be involved with the development of the design. In the case of designs for educational settings, the stakeholders include administrators and instructors as well as students. In this experiment we focused on the design problem of developmental mathematics courses.

Developmental mathematics courses, defined as college-level courses prior to calculus, have been the fastest-growing post-secondary mathematics courses in the nation. From 1965 to 1985, total enrollment in college mathematics courses increased 60%, while enrollment in college courses in high-school-level algebra and geometry increased by 250% (Madison and Hart, 1990). College algebra courses present unique challenges for instruction, but instructors currently have limited resources with which to meet these challenges.

Based on input from instructors, administrators, and publishers, we identified a profile of the college algebra student population that helped to focus our efforts at instructional innovation. Developmental mathematics students tend to have one or more of the following characteristics: (1) high variability in age and prior mathematical background; (2) a prior history of failure or frustration in mathematics; (3) a fragmented, proceduralized understanding of mathematics; (4) a lack of appreciation for the direct importance of mathematics to their lives; and/or (5) poor study

skills, particularly, little awareness of how crucial it is to spend time working through problems if one is to succeed in mathematics. Thus we need to design an instructional solution that can adapt to the individual background of students, that offers a success experience, that interrelates various mathematics skills during problem-solving, that makes mathematics meaningful, and that can compensate for poor study skills. Also, despite their poor mathematics skills, many students enrolled in developmental courses declare college majors that depend heavily on mathematics (e.g., economics, biology, physics). Thus we need an instructional solution that prepares students for higher-level mathematics courses and that focuses on the application of mathematics to a variety of fields. Our design solution aims at these objectives.

A design solution must also address the needs of university administrators as stakeholders in the problem of developmental algebra. Universities are reluctant to spend resources on teaching algebra or other remedial courses because passing algebra in high school is a requirement for entering college. However on the mathematics placement test at one of our university research sites, up to 60% of incoming freshman place into algebra. Then the remedial classes are often based on the prevailing belief that the students' high school teachers must have done a poor job, so the college courses re-explain and re-teach the same traditional algebra. Students are bored and disillusioned, thinking (correctly) "I've had this before," and emerge having only relearned the same partial knowledge with which they entered. The remedial classes are often worth fewer credits which further undermines motivation and lowers administrators' financial incentive to invest in such courses.

Another aspect of needs assessment for learning environments is identifying the cognitive processes with which students have the greatest difficulty. In a summer pilot study prior to Study 1, below, we discovered that students were performing much better on relatively obscure symbolic skills, than on applied processes required for both workplace competence and future academic success (e.g. creating and interpreting mathematical representations of situations to make appropriate recommendations). Such skills are important goal outcomes of our design solution.

Theory: ACT and the Inductive Nature of Mathematical Knowledge Acquisition

Cognitive tutors are a type of intelligent computer tutoring system well suited to the unique challenges of teaching developmental algebra (Anderson, Corbett, Koedinger and Pelletier, 1995). They have at their core a fine-grained computer model of the knowledge, skills, and strategies involved in competent student problem solving in the subject-matter domain. That is, within the tutor is a simulation, developed through psychological research, that can solve mathematics problems in the ways we expect students to solve the problems, as well as simulate the kinds of mistakes that students make.

A cognitive tutor for algebra called PAT (Practical Algebra Tutor) has been developed through a collaborative effort between cognitive psychologists, computer scientists, and educators at the Pittsburgh Advanced Cognitive Tutoring (PACT) center of the Human Computer Interaction Institute at Carnegie Mellon University. Cognitive tutoring technology is grounded in the ACT theory of cognition, perhaps the most comprehensive, empirically tested and influential of contemporary theories of human learning (Anderson, 1993). Three key premises of the theory are particularly relevant to developmental mathematics. (1) Effective learning occurs in the context of a problem-solving activity. Thus the tutor software is a content-rich tool that gives students the opportunity to learn algebraic skills and apply them to relevant problems at the same time. In traditional algebra courses, students learn algebraic skills in isolation and then, often in an extra step following the "coverage" of each topic, attempt to apply them to standardized word problems. This leads to poor retention, as evidenced by the large number of people who passed algebra in high school but still place into developmental algebra when they get to college. It also creates a sense of arbitrariness where students believe that algebra has no relevance to their lives. (2) People learn by doing, not by watching or listening. Thus the tutor technology allows each student to actively participate in the problem-solving process. The role of the instructor changes from one who imparts the correct set of facts and procedures to a facilitator who helps students construct their growing algebraic knowledge. (3) Learning events occur at a grain size well captured by

ACT's production rule notation. That is, the tutor responds to the student's actions at the level of individual production rules. Production rules include an "If" part that defines the conditions that will cause the production to activate, and a "Then" part that defines the action to be taken when the specified conditions are met. By writing production rules that match the steps students take while doing algebra, we try to characterize not only the kinds of mathematical concepts, skills and strategies students need to learn (this is the "Then" part of a production), but also the contexts in which such knowledge elements are effectively employed (the "If" part). Thus, production rule analysis seeks to specify instructional objectives not in terms of decontextualized topics, facts, or procedures as in traditional instructional design, but in terms of performance knowledge as it is situated in specific problem solving contexts.

The Design Solution

Classroom reform. The current emphasis in mathematics reform efforts, both at the high school and the post-secondary levels, is on the use of multiple representations and computational tools to solve real-world problems (NCTM, 1989; AMATYC, 1995). Recently the Mathematics Association of America recommended that quantitative reasoning is a necessary goal for college graduates. Their recommendations stated that rote, passive learning of mathematical facts and procedures is not enough to expect of educated adults. Rather, colleges and universities should expect every graduate to be able to use different mathematical methods to the solve real-world problems, to interpret and draw inferences from mathematical models such as formulas, tables and graphs, and to represent mathematical information in several ways (MAA, 1996). They also stated that mathematics should be taught in context and instructional materials should be practical and should demand active student involvement.

The PAT curriculum. The PAT curriculum is consistent with these reform efforts. The emphasis in PAT is less on symbol manipulation and more on symbolization as way to unleash the power of mathematics for problem solving. This is the reason that we chose to use PAT in our client-centered design solution. Figure 1 shows the PAT screen after completion of a problem.

While PAT covers most of the traditional content of beginning algebra (e.g., linear functions, systems of equations), these skills are introduced in the context of authentic, realistic problem-solving contexts. The problem in Figure 1 involves using systems of equations to compare the costs of two car rental companies, Avis vs. Hertz. One goal is to find the circumstances under which one company is more cost effective than the other (Answer: rent from Hertz when traveling less than 750 miles). This is a useful task in the real world and one to which we would like students of mathematics to apply their skills.

Insert Figure 1 about here

Customized features. Like much computer-based instruction, cognitive tutors allow for self-paced use. Unlike other approaches, the cognitive model within the tutor facilitates three additional forms of customized instruction: (1) Individualized feedback and advice can be supplied as a student solves a problem; through a technique called "model tracing" (see Anderson, Boyle, Corbett and Lewis, 1990), the tutor uses a student's behavior to identify the chosen solution path and prompts the student appropriately; (2) The number and types of problems are individually selected to maximize learning opportunities; using "knowledge tracing" (see Corbett and Anderson, 1992), the tutor follows the student's knowledge acquisition process, compares it to the skills embedded in the learning model, and selects appropriate problems based on a student's successes and failures in applying each individual skill and strategy; and (3) it is easy to customize the tutors for particular courses; this is important both because there is great variation between different basic mathematics courses and because college instructors want to put their own signature on the courses they teach.

Multiple representations. PAT provides computerized computation and visualization including a tabling tool, (see Fig. 1, upper-right) a graphing tool (lower-left) and an equation solving tool (lower-center). By allowing manipulation of algebraic functions through the use of multiple representations, the tutor promotes deeper understanding of the meaning of the symbolic

expressions (Kozma, 1993). One component of the college algebra curriculum that is critical to future mathematics success is skill in judging when two algebraic functions are equivalent. With the multiple representations that are in our current system a student could, for example, use a table to check when the two functions produce the same output for the same input, see where the graphs of the functions intersect, or use symbolic manipulation to solve the equation when the two functions are set equal.

Support for inductive problem-solving. Instruction in PAT promotes instance-based investigation, a flexible problem-solving strategy often used by competent problem solvers to support their use of abstract mathematical formalisms as they attempt to make sense of a difficult problem (Koedinger and Anderson, 1990; 1997). PAT promotes the use of this strategy in two ways. One is by asking students to first investigate concrete instances of unknown values, e.g., "How much would you owe each company if you drove 500 miles?" (see Fig. 1, row 2 of the Worksheet). After answering the questions, they attempt to formulate the abstract rule, (see Fig.1, "Formula" row of the Worksheet). Instance-based investigation can lead students to solutions, particularly in problems requiring a discrete decision (e.g., from which company does it cost less to rent a truck?) as do many real-world problems. The other type of instance-based investigation is based on Koedinger and Anderson's (1997) work, where they found that students' prior ability to solve arithmetic word problems and their general inductive capabilities can be used to support learning of algebraic symbolization as a generalization of arithmetic. This "inductive support" strategy is further supported in the optional Pattern Finder window of PAT, which is not shown in Figure 1. This tool asks students to think about how they would figure out the answer using small numbers, e.g. "How much would it cost to rent from Avis if you drove 2 miles? If you drove 3 miles? 4 miles?" and helps them see a pattern by substituting a variable for the small numbers.

Previous Evaluations and Overview of the Studies

Many laboratory studies have investigated the effects of cognitive tutors on student learning, and the success of these systems has also been demonstrated in real classrooms (see Anderson et al., 1995; Koedinger, Anderson, Hadley & Mark, in press). In a laboratory study with the LISP programming tutor, students who used the tutor performed three times faster than students who did not (Anderson et al., 1995). Two classroom studies with different geometry proof tutors showed that students tutor-supported classes scored one standard deviation higher than students in regular classes (Anderson, et al., 1995, Koedinger & Anderson, 1993b). In addition, significant gains in student attitudes and in positive instructor behaviors, including acting and being seen by students more as a facilitator than as an authority, have been reported in classrooms using a cognitive tutor (Schofield, Evans-Rhodes and Huber, 1990). In this paper we focus on a cognitive tutor called the Practical Algebra Tutor, or "PAT", and it's use at the college level. In a previous evaluation of PAT in high school algebra classes, students using the tutor with a reform curriculum performed 100% better on assessments of authentic problem solving and representation use, and 15% better on standardized tests than students who followed a traditional curriculum without the tutor (Koedinger, et al., in press).

In this experiment, we monitored outcomes across three semesters in classes using PAT and in comparable classes that did not use PAT. Results of each semester provided feedback on earlier designs and direction for design improvements. We present three studies performed in successive semesters: Fall 1995, Spring 1996, and Summer 1996.

Study 1: Fall 1995

After some pilot work in the Spring and Summer of 1995, this first study was performed in the Fall. PAT was implemented in a total of 11 classes at two universities. Within University P, instructor and curriculum were controlled for in that the same instructor taught the comparison class using the traditional curriculum and the experimental class with the traditional curriculum and

the cognitive tutor. Within University L different instructors taught the experimental and comparison classes, but the same traditional curriculum was used in both.

We were interested in two major outcomes of student learning: (1) algebraic manipulation skills, the goal of a traditional curriculum, and (2) use of these skills as well as alternative algebraic representations (e.g., tables and graphs) to analyze and solve real-world problem situations, the goal of the PAT curriculum. To assess these outcomes we used a performance-based assessment that targeted students' problem-solving skills, qualitative reasoning, and ability to communicate effectively about mathematics, as well as traditional algebra exams.

Study 1 differed from the previous evaluation of PAT at the high school level (Koedinger et al., 1997) in two ways: (1) in the current study, PAT was added to a traditional rather than a reform curriculum, and (2) the performance-based assessment was given to comparison classes as well as to experimental classes.

Method

Design. At University P, one comparison class (n = 16) studied a standard beginning college algebra curriculum, and one experimental class (n = 12), taught by the same instructor, studied the same standard curriculum but also used the PAT cognitive tutoring system. At University L, six comparison classes (n = 199) studied a standard curriculum, and three experimental classes (n = 88) used PAT along with the same curriculum. These nine classes were taught by a total of five different instructors, none of whom taught both an experimental and a comparison class.

In an ideal laboratory study, students would be randomly assigned to either the experimental or control condition. In field studies such as these, it is usually not administratively feasible to obtain random assignment. Typically, it is difficult enough to get course administrators to agree to having experimental and control conditions. Given the relative lack of design experiments that even meet pseudo-experimental standards (Cook and Campbell, 1979), the statistical inference risks of not having random assignment are relatively small. We have no reason

to suspect students were systematically self-selecting any of the course sections-- they were not aware of the manipulation at scheduling time-- and thus, this is likely to have been a natural randomization. Brown (1992) advocated such decisions which trade off ideal laboratory criteria in favor of the external validity and more general discovery potential of design experiments.

Student population. The two Universities both have a large developmental mathematics program. Reflecting the demographics of their respective cities, the dominant minority group in University P's student population is African American while University L's dominant minority group consists of Hispanic students. The proportion of minority students in the populations participating in this study was somewhat higher than the student body averages. The students' mean scores on the Mathematics portion of the SAT were statistically equivalent, though slightly higher for University P, averaging 420 (sd=68.2) for University P and 395 (sd=61.0) for University L ($F(1,267)=2.84$, $p = .09$).

Procedure. At University P the course met for 15 weeks, 4 days per week, 50 minutes per day. During 10 class sessions, students in experimental classes met in the computer lab and worked on 3 word problems with the tutor. At University L the course met for 15 weeks, 3 days per week, 75 minutes per day. During 7 class sessions, students in the experimental classes met in the computer lab and completed up to 16 problems with the tutor. University L students were required to come to the lab outside of class to finish any lessons not completed during class time. Students at University P were encouraged to use the tutor outside of class, but most did not do so. The cognitive tutoring software was stored on a server and accessed by students from Apple Macintosh machines.

Curriculum: lessons. The main tutor curriculum covered five lesson topics. The skills required for each lesson build upon the skills learned in previous lessons. (1) The Worksheet. In this lesson, students are introduced to word problems that involve a linear function of the form y=mx, y=mx+b or y=mx-b. Students read a problem statement and fill in a table of values. This involves identifying important quantities in the problem statement and using them as column labels; indicating units of measurement for each quantity; solving for values of y, given values of x; and

writing a formula for the relationship in the problem, by entering a variable for x and an arithmetic expression of y in terms of x. (2) Point-Plot Graphing. Here students continue to use the Worksheet and learn to map information from the table onto a graph. This includes labeling the axes of the graph; identifying upper and lower bounds for the graph based on the range of values in the Worksheet; setting a suitable scale for the graph, given the chosen bounds; creating points and placing them at the coordinates of the table values; and drawing a line to connect the points. (3) Finding unknowns with equations. As an elaboration on the use of the worksheet and graph, above, students are given an equation solving tool for finding values of x when given values of y. Students enter an equation and use a menu to select operations such as 'subtract from both sides', 'reduce fractions', or 'combine like terms'. (4) Slope-intercept graphing. A new graphical representation in this lesson requires students to write a function describing the problem situation and use the slope and y-intercept of the function to reason about rates of change and plot appropriate points on the graph. (5) Systems of equations. Students solve problems which involve reasoning about two equations of the form y=mx+b. They continue to use the worksheet and equation solving tools. In addition to solving for finding y, given x and solving for x, given y, students solve for x given that two equations, y and z, are equal to each other. Students use a new form of the grapher to graph their equations, focusing on the intersection of the two functions.

Curriculum customization. At University L, we worked with the head instructor to custom-define a set of approximately 16 problems for each of the five lessons. The tutor traced students' knowledge of each skill involved in a lesson, selecting and presenting additional problems to each student as needed. A lesson was complete when either all skills in a lesson had been mastered or no more problems were available to present. Our instructor collaborators at University P decided to implement PAT differently. The knowledge-tracing function of the tutor was turned off and all students were required to complete three problems for each lesson. The problems and curriculum sequence were custom-defined in conjunction with instructors and program administrators. Two additional lessons focused solely on solving linear equations with no context. At both universities these equation-solving lessons were presented using the

knowledge tracing function. Classroom instruction covered traditional algebraic symbol manipulation, e.g., simplifying equations, factoring, logarithms, and inequalities.

Instructor training. The instructor at University P oriented himself to the software over the course of one week and received some training on troubleshooting. The instructors at University L were trained in the use of the software over the course of two days. At both universities, the lead instructor worked with us to select lessons and decide when during the semester students would go to computer lab.

Performance-based assessment: "The Cellular Phone Problem". During the last week of class, all students were given approximately 20 minutes to work on the performance-based assessment (see Figure 2). Given pencil and paper (and at University P the optional use of a graphics calculator) the assessment asked students to do a "mathematical analysis" of cellular phone services, given a basic monthly charge and a rate per minute of usage for each service. At University L, the problem presented three cellular phone services. University P chose to use a simpler version of the problem with only the first two cellular phone services (Gold and Silver). The problem was contextualized as a requirement given by a "boss" to a "company employee". The goal was for the student, as the employee, to decide which cellular phone service would be appropriate for different officers of the company, based on the amount of time each officer used the cellular phone. The basic elements required for the mathematical analysis were clearly listed in the problem statement: (a) defining variables, (b) writing equations, (c) making tables, (d) constructing graphs, (e) finding slopes and intercepts, and (f) finding points of intersection. Additional problem-solving tasks were presented as criteria suggested by the boss, e.g., that the analysis should: (g) specify the amount of usage each service allows for a total cost of $100, and (h) state the range of usage for which each service is cheapest. The problem context stated that the employee would have to make decisions based on the mathematical analysis shortly after the boss provided the necessary information on each officer's normal phone usage. This implicitly encouraged students to produce a graph of the two functions from which values can be read off quickly.

Insert Figure 2 about here

Traditional final examination. All students took a final exam that covered traditional symbolic manipulation skills, including solving equations and inequalities, evaluating logarithmic expressions, simplifying equations, and graphing the equation of a line. Problems were presented as mathematical expressions with little context (no word problems), and no verbal responses were required. Typical problems included:

Solve: $2x^3 + 6x^2 - 20x = 0$

and

Simplify: $[(x^2 - 6x) - 8x\ (5 - 2x)] + 12x\ (x + 3)$

Such problems were not included in the PAT Equation Solving lesson, which addressed only linear equation solving.

Results and Discussion

Problem-solving ability. Performance on the Cellular Phone Problem was evaluated by breaking it down into eight component tasks corresponding to the eight requirements listed in the problem statement (a-h). A single researcher assigned each student's solution a score of 0, 0.5, or 1 point on each component, based on predetermined criteria. These component scores were analyzed separately and then averaged together to produce an overall mean score for each student that ranged between 0 and 1. Results showed that students in the experimental groups at both universities scored higher than students in the comparison groups overall (see Table 2, last row). At both sites, students overall scores on the Cellular Phone problem were related to their mathematical ability, as measured by their scores on the mathematics portion of the SAT test (MSAT). We performed an analysis of covariance (ANCOVA) with condition (experimental vs. comparison) as a between-subjects factor and MSAT as the covariate. At both universities, MSAT score significantly predicted Cellular Phone problem performance (University L: $F(1,237) = 53.51$,

$p < .001$; University P: $F(1,20) = 5.91$, $p < .05$). At University L, the difference between the experimental and comparison classes on their problem-solving performance was statistically significant in the ANCOVA ($F(1,237)=63.29$, $p <.001$). At University P, the difference between the experimental and comparison classes did not reach statistical significance in the ANCOVA ($F(1,20)=2.11$, $p =.16$) perhaps because of the small number of students in the two classes. Because the MSAT scores of 5 students were not available, the ANCOVA further reduced the number of students. A one factor analysis of variance (ANOVA) with condition as a between-subjects factor and including all students did reveal a statistically significant difference ($F(1,26)=4.49$, $p < .05$). The statistics reported in Table 2 are the ANOVA results.

Students at the two universities received comparable overall scores, despite the fact that University P presented a slightly less difficult problem (comparing only two linear functions instead of three) to students whose entering SAT scores were slightly higher than those at University L. On two of the eight component scores, Making a Table and Stating the Ranges of Usage, there were statistically significant differences between the experimental and comparison groups at both universities, with the experimental group outperforming the comparison group. On the remaining six component scores, Defining Variables, Solving for y = $100, Writing Equations, Finding Slopes and Intercepts, Constructing a Graph, and Finding Points of Intersection there were significant differences between the experimental and comparison classes at University L, but not University P. This is likely due to a lack of statistical power at University P, given the smaller number of students participating at that site.

Insert Table 2 about here

There was some evidence that students using the tutor had a better understanding of problem goals in that they were better able to state the range of values for which each cellular phone service was the cheapest. More students in the comparison classes seemed to believe that the problem goal had been met when they had solved the equations and found that one service

allowed a greater amount of phone time for a total cost of $100. However, the functions describing the total cost of each service intersect, creating two ranges of values. Basing a decision about the "best" cellular phone service on a single point does not take this fact into account.

Traditional algebraic skills. The overall scores of students in the experimental and comparison classes on the departmental final exams did not show a statistically significant difference at either university (University L: M(Experimental) = 68%, sd = 14.1%; M(Comparison) = 64%, sd = 16.3%, F(1, 285)= 2.67, p = .10; University P: M(Experimental) = 69%, sd = 18.1%; M(Comparison) = 75%, sd = 24.8%; F < 1). This suggests that the reduced lecture time students experienced in the experimental classes (because some lecture periods were spent in the computer lab instead) did not detract from those students' acquisition of traditional algebraic manipulation skills. An ANCOVA showed that final exam scores were also significantly predicted by MSAT scores at University L, but not at University P. At University L, the difference between the adjusted mean scores shows a trend towards an advantage for the experimental group (M(E) = 68.4%, sd = 15.6%; M(C) = 64.4%, sd = 15.5%; F(1, 239) = 3.52, p = .06).

Student evaluations. At the end of the semester at University L, the lead instructor gave his students the opportunity to write open-ended statements about what they liked best and least about the course. The instructor provided us with those evaluations in which students specifically mentioned the computer component. Of 12 such evaluations, 5 students described their work with PAT as a positive experience, 3 students described it positively, but with a qualification, and 4 students described it negatively. All the positive evaluations stated that the computer lab was "helpful". Examples of more specific comments include: "It guided the lessons, and helped me understand a lot," and "Computer lessons are very helpful, especially the word problems, because they make you think and that is very helpful for a student who is going to take higher math courses." The qualified positive evaluations all evaluated the computer lessons positively, but mentioned difficulties, particularly finding acceptable labels for values in the table. For example, "The computer lab was tedious to me because I got stuck on things like the right wording, not

math," and "The computer lessons are good, but they are a pain, especially when you can't label them correctly!" Most of the negative comments mentioned the amount of time required to complete the computer lessons, and perhaps indicated some unease about the non-traditional kinds of problems/activities in PAT, as in "Problems were too long," and "The class time we used for the lessons could have been used to learn something else about math."

Study 2: Spring 1996

We continued our investigation of the use of PAT for developmental algebra at the same two universities in the Spring semester of 1996. Both universities used a curriculum quite similar to the one they used in the Fall semester, though there was greater emphasis on applying the principles embodied in the tutor to authentic problem-solving in the classroom. Specifically, students took a series of quizzes which focused on the individual skills needed to solve complex real-world problems such as the Cellular Phone problem. The quizzes were created to help instructors better monitor students' development of problem solving and representation skills throughout the semester.

Another difference introduced during the Spring semester was the introduction of additional focus on word problems in the comparison class. At University L one comparison and one experimental class were taught by the same experienced instructor. This comparison class was given paper-and-pencil problems for homework that were taken from the set of problems the experimental class solved using PAT.

One of the major difficulties students experienced with PAT in the Fall was that of labeling the variables in each problem, as was evident in their student evaluations. This was remedied in the Spring when a new version of the tutor was put in place at both universities that included a much longer list of acceptable labels for the variables in each problem and implemented a spell-checker for the entered text. We did not collect student evaluations in the Spring, but teachers reported anecdotally that students were much more successful and less frustrated with labeling variables this time.

Despite the significant positive outcomes of the Fall study, the administration at University L made a commitment to use a commercial software environment for developmental mathematics starting in the Spring. This software environment is not based on a cognitive theory or empirical studies and the curriculum content is largely traditional symbol manipulation, not problem solving with multiple representations as in PAT. It was implemented with no instructor training, in fact, instructors were informed they would be using the system only 5 days before classes started. Some of these instructors had already participated in some of our PAT instructor training, as their sections had been targeted as comparison classes. Therefore we had fewer comparison classes than originally planned. We offered to include the courses using this commercial software system in our evaluation, however, this offer was refused by the administration. One reason given was that the new system would obviously not lead to positive results in the first semester of use.

Method

Design. At University L (n=138) there were three experimental classes taught by three different instructors (n=79). One of these experimental classes (n=27) was taught by an instructor who also taught a comparison class (n= 33) (the same-instructor comparison). Another comparison class (n=26) was taught by a different instructor. At University P, there was one experimental class (n=18) and no comparison class. This smaller set of students reflects the typical lower enrollment in the Spring sections of these courses.

Student population. At University P the student population included the same type of at-risk student as in Study 1, but also included students who were repeating the course and Seniors who had waited until their last semester to take a required mathematics course. The average MSAT score of these students was 428 (sd = 64.5).

Curriculum: lessons. The computer curriculum was the same as in Study 1. This time students at University P, like those at University L, were required to finish their computer lessons outside of class if they did not finish during the lab sessions. At University L, the students in the same-instructor comparison class were given paper-and-pencil versions of some of the problems

that the corresponding (same-instructor) experimental class completed on the computer. Thus the same-instructor comparison differed only in that students did not receive the cognitive tutor's assistance.

Curriculum: quizzes. All students at University P were given the three quizzes we constructed. At University L, all experimental classes and the same-instructor comparison class took the quizzes; the different-instructor comparison class did not. In the quizzes, the difficult skills needed to solve the Cellular Phone assessment, such as using the intersection of two lines to draw a conclusion about a relevant range of values, were introduced early in the semester as qualitative questions. For instance, in a problem comparing the costs of two rental car companies, one quiz contained the questions, "Which company costs less if we drive a very short distance? Will that company ever cost more than the other company?" Thus the quizzes served to prepare students for the complex final assessment.

Outcome measures. As in Study 1, students took the Cellular Phone performance-based assessment at the end of the semester in addition to their traditional final exam. University P again used the modified Cellular Phone Problem which included only two linear functions.

Results and Discussion

Problem-solving ability: Cellular Phone problem. The Cellular Phone problem was scored as in Study 1. Results are presented in Table 3. The experimental class at University P showed an increase in their overall scores (M=60%, sd = 19.1%) as compared to Study 1 (M = 41%, sd = 21.1%). An ANOVA shows this difference is a statistically significant increase over Study 1 ($F(1,26)=6.33$, $p < .05$).

Insert Table 3 about here

At University L, the two comparison classes averaged 45% on the performance assessment while the three experimental classes averaged 46%. However, these averages mask considerable

variability associated with the different instructor's classes, as shown in Figure 3. One experimental class (T4) averaged 36% overall on the Cellular Phone problem, lower than the two comparison classes (43% and 48%). The other two experimental classes (T2 and T3) averaged 55% and 47%, respectively. Several factors appear to have contributed to the relatively poor performance of instructor T4's experimental class. T4 had less experience using PAT and was less familiar with the curriculum changes than T2, and T4's class met only two days a week instead of three, as did T2 and T3. Thus students spent less time using PAT during class, and more time independently without instructor support. There is also the possibility that the classroom and lab practices of this instructor were substantially different from other instructors and these practices led to poorer student performance. As discussed below, T4's class also scored significantly lower on the traditional exam and on one of the quizzes.

Given the instructor variability, a better statistical comparison can be achieved by examining the experimental and comparison classes taught by the same instructor (T2 in Figure 3). We performed a one factor ANCOVA for this teacher alone with condition as the between subjects factor and MSAT as the covariate. The MSAT covariate has a positive relationship with Cellular Phone Problem scores ($\underline{F}(1,49) = 2.82$, $\underline{p} = .10$). T2's experimental class scored higher on the Cellular Phone Problem ($\underline{M}(E) = 55\%$, $\underline{sd} = 22.9\%$, adjusted $\underline{M}(E) = 57\%$, $\underline{sd} = 21.3\%$) than his comparison class ($\underline{M}(C)=48\%$, $\underline{sd} = 23.3\%$, adjusted $\underline{M}(C) = 46\%$, $\underline{sd} = 24.0\%$). The ANCOVA results indicate that this difference is near the .05 level of statistical significance ($\underline{F}(1,49) = 3.47$, $\underline{p}=.06$). This experimental effect is smaller than in Study 1 as a likely consequence of the use of PAT problems in the comparison classes.

T2's comparison class in Study 2 provides a tighter control on the nature of the effect of PAT than we had in Study 1. The greater impact of T2's experimental class over his comparison class replicates the same-teacher results in Study 1 at University P -- the improvement is not due to differences in instructors. Further, the improvement cannot be attributed to the inclusion of authentic problem solving situations since such problems were used in T2's comparison classes. The remaining difference is that students in the experimental class had the advantage of PAT's

intelligent assistance provided on demand and in the context of each student's individual problem solving and learning trajectory.

Insert Figure 3 about here

Problem-solving ability: Quizzes. In general, students performed well on the quizzes. At University P, students scores improved from 79% on the first quiz to 95% on the third despite the increasing difficulty of the quizzes. At University L, scores were initially higher (M(C)=86%, M(E)=87%, on quiz 1; M(C)=85%, M(E)=83%, on quiz 2), though the experimental classes showed a slight decrease on quiz 3 (M(C)=86%, M(E)=73%). This was due solely to a large decline in T4's class on quiz 3 (M(T4) = 48% on quiz 3)-- the other classes remained about the same on all three quizzes, despite their increasing difficulty.

Traditional algebraic skills. At University L, the same class that received very low scores on the Cellular Phone problem (T4 in Figure 3) also scored low on the final exam, bringing down the average of the experimental group, such that the comparison group outperformed the experimental group (M(E) = 62%, sd = 17.6%; M(C) = 69%, sd = 15.0%, F(1, 137) = 4.82, $p<.05$). When only the same-teacher (T2) comparison is considered, the difference favors the tutor class (M(E) = 73%, sd = 12.0%; M(C) = 69%, sd = 15.9%) but is not statistically significant (F(1, 59) < 1).

Study 3: Summer 1996

After two semesters in which the tutor was an adjunct to the traditional algebra curriculum, University P decided to reform their developmental algebra program to more fully integrate the technology and to align the course principles with the new standards for mathematics (AMATYC, 1995). The elements of the course, whose content was driven by the cognitive tutoring technology, included mathematical projects completed in small groups and reasoning with multiple representations. One objective was for students to see and understand mathematics from different

perspectives: verbal, symbolic, graphical, and numerical. Students learned to interpret mathematical models and consider alternative approaches to problem situations.

University L dropped out of the study because their administration decided to devote all of their developmental courses to the commercial software system mentioned earlier, despite having no evaluation results on student learning outcomes from the Spring semester. Further, the anecdotal reports we heard from the Spring instructors suggested that students did not benefit from this new course. However, the decision was made by the administration (not the instructors) to go forward, apparently in large part because a significant financial investment had been made to purchase the software.

Method

Design. In the summer of 1996, five experimental classes (n = 81) at University P participated. These classes were taught by four different instructors and used a reform curriculum. There were no comparison classes.

Student population. This course was attended by a different type of at-risk student than those in Studies 1 and 2. The summer students were generally from a higher SES background and had higher MSAT scores (447, sd = 59.1) than the Fall and Spring students, but were ranked lower in their high schools indicating poorer past achievement.

Curriculum. The 7-week reformed curriculum focused on quantitative literacy. The PAT curriculum from Studies 1 and 2 formed the heart of the curriculum, with the addition of a lesson on quadratic functions at the end. Students worked with the tutor for 1.5 hours, 3 times per week. The total amount of time spent using PAT was similar to Studies 1 and 2. Students also worked on small group projects that required two or more class periods to complete. The projects involved data, sometimes collected by the students themselves, that students would transform, model with a linear function, or otherwise analyze for use in making predictions and describing the relationships between variables. Groups of three students assumed the roles of problem presenter, analyzer and assembler, while all three carried out the problem operations. The problems presented in the

computer were mentioned during lecture periods and often used as examples. Homework from the textbook was de-emphasized.

Course structure. In previous summers, the course structure involved six 1.5 hour lectures a week, two on each of Monday, Wednesday, and Friday, and two 1.5 hour recitations on Tuesday and Thursday run by undergraduate teaching assistants. To facilitate greater learning-by-doing, including the group projects, this structure was modified so that lectures occurred on Tuesday and Thursday, group projects during one session on Monday, Wednesday, and Friday, and computer lab work on PAT during the other session on Monday, Wednesday, and Friday.

Outcome measures. Because there was no comparison class it was important to measure individual improvement. Therefore, we developed a second version of the authentic problem-solving test called the Storage Space test to serve as a pre-test. Its format exactly paralleled the Cellular Phone problem. Traditional algebraic skills were also measured both pre- and post-test, in that the final exam included 17 identical items from the mathematics placement test taken by all entering freshman.

Results and Discussion

Traditional algebraic skills. Overall, students scores nearly doubled over the course of the semester. As a whole, the classes averaged 37% for the 17 items on the placement test, and 70% for those same 17 items on the final exam. This shows that the new course format was successful in helping students to master traditional algebraic skills, despite the much greater emphasis on authentic problem solving.

Problem-solving ability: Storage Space and Cellular Phone problems. Students scores on the authentic problem-solving improved from 20% to 67% overall ($F(1,77) = 208.8$, $p < .001$; see Table 4). All component skills except solving for x, given y, were improved.

Insert Table 4 about here

General Discussion

Goals and Problem Identification

Our results address the features of design experiments listed in Table 1. The needs assessment (feature a in Table 1) uncovered a student profile for which the adaptive nature of cognitive tutors was a good fit. Further, we identified a large discrepancy in developmental mathematics curricula between the great amount of time spent on algebraic formalisms and the relatively little time spent on other representations (words, tables, graphs) and, more importantly, on applying such representations in analyzing situations and solving problems. The fact that less than 7% of the 215 students in the traditional classes in Study 1 were able to make a reasonable conclusion about the Cellular Phone Problem verified this aspect of our needs assessment (see Table 2). Our curriculum reform efforts were in response to this need. We encouraged greater focus on multiple representation and problem situation analysis in the curriculum materials, as well as adding the PAT software, and introduced assessments that better reflect students' future needs for algebraic representations and reasoning processes. When PAT was integrated into a reformed curriculum, we saw dramatic learning gains-- gains that are particularly relevant to student retention and success in college. By the end of Study 3, 60% of the students using PAT were able to draw a reasonable conclusion about the Cellular Phone problem (see Table 4).

In targeting the needs of the multiple stakeholders (feature b) in the problem of developmental mathematics, it is important to focus on appropriate objectives. We aimed for improvement of student learning on multiple assessments (feature c) over time as the goal for our design solution. Specifically, we did not want to trade an emphasis on application skills for a decrement in students' basic, formal algebra skills. The comparison of a traditional curriculum to one with the addition of PAT (Study 1, both sites) showed that, overall, students who used PAT were able to apply algebraic representations in the analysis of a real-world problem situation more

accurately than students in comparison classes. And in addition, allocating time to use PAT, time away from traditional lectures, did not negatively affect students' ability to perform traditional symbolic manipulation, as demonstrated by their final exams. In fact, at University L, where students spent more self-paced time with PAT, there appeared to be some beneficial effects of PAT on traditional algebra as well (in Study 1 and for the same-teacher comparison in Study 2).

Theory

We addressed the design problem with theoretical principles in hand including the general ACT theory of cognition and more specific theoretical principles regarding the nature of mathematical cognition (feature d).

The ACT cognitive architecture and cognitive tutor design. Previous research on cognitive tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995) has demonstrated the successful application of the ACT theory. Production rule models of problem solving in LISP programming and geometry theorem proving led to cognitive tutors with dramatic impact on student learning, accelerating learning three-fold with the LISP tutor, increasing post-test performance by a letter grade with two geometry tutors. The success of PAT described here and at the high school level in Koedinger, Anderson, Hadley, & Mark (in press) extends these results beyond symbol manipulation to situation analysis and problem solving with multiple representations. The cognitive tutor technique of model tracing implements the simplest and perhaps most important benefits that an individual human tutor, coach, or knowledgeable peer can provide: staying silent while you do productive work, warning you when your efforts may be unproductive, and providing advice when you need it. Simply put cognitive tutors increase student opportunities for success by providing as-needed support to maximize the time students are thinking within their own zone of proximal development (Vygotsky, 1978).

Mathematical cognition. The design guidance provided by the ACT architecture is a broad framework within which many degrees of freedom are still open. Most importantly, while ACT provides guidance for the form and grain size of production rules, it does not provide any guidance

on the content of productions that characterize student performance in the domain of interest. To further constrain the design process, theories of knowledge representation and content acquisition are needed that are more specific, in the current case, to mathematics learning. The ACT prescription of characterizing student knowledge in production rule units focuses the task analysis on performance knowledge rather than factual knowledge. Furthermore, it provides a discipline that helps the cognitive modeler do a task analysis at the right grain size -- one that is most likely to correspond with students' rate of knowledge acquisition. Traditional topic level analyses are too large and will lead to tutorial advice that is not specific enough to students' needs. A micro-feature analysis (e.g., to create a neural net) is too small and requires more work than is necessary to achieve effective instruction.

The ACT-R theory says that new thinking processes are learned via an analogy mechanism that relies on declarative encodings of problem states (cf., Blessing & Anderson, 1996). If these states are not visible to the learner, learning will be more difficult and error prone -- learners must generate these hidden states themselves. Two design principles follow: (1) create interfaces that reify, or make visible, hidden states like goals (Corbett & Anderson, 1995), plans (Koedinger & Anderson, 1993), or situation models (Nathan, Kintsch, & Young, 1992) and (2) provide encouragement and support for learners to generate these hidden states on their own, for instance, via verbal hints about goals (McKendree, 1990) or self-explanation prompts (Bielaczyc, Pirolli, & Brown, 1995; Chi, de Leeuw, Chiu, LaVancher, 1994).

Design Solution

PAT and corresponding curriculum reforms implement the above two principles to help students in acquiring the enigmatic skill of algebraic modeling. First, the PAT Worksheet and Pattern Finder reify, or make visible, algebraic modeling process by having students use their existing arithmetic skills to work through concrete instances of problems before abstracting to algebra (Koedinger & Anderson, in press). Second, PAT and the curriculum reforms embed algebraic instruction in situations, like the decisions required in the Cellular Phone Problem, and

representations, like words, tables, and graphs, that are more familiar for students. This makes it more likely that students will be able to correctly generate the hidden processes needed to successfully comprehend and produce algebraic symbols.

In addition to technology design, engineering the context of technological implementation is critical to success. The context includes integration of the technology into the curriculum, into the existing technological infrastructure (the hardware and network software), and into the broader administrative structure. It is particularly for such issues that Collins (1992) and Brown (1992) have emphasized the need for multiple expertise in design process (feature e), instructors as co-investigators (feature f), and multiple facets of innovation (feature g).

With regard to curriculum integration, one benefit of PAT is that it can be customized to fit alternative design constraints. During the first semester (Study 1), both of our university sites chose to use PAT as an added component to their algebra course. Traditional algebra skills were taught through lectures and textbook homework, and occasionally class would be held in the computer laboratory where students would use PAT to practice solving word problems and using multiple representations. In another example of flexibility to instructor needs (feature f), one of our sites chose not to use knowledge tracing. Instead they thought it was more fair to students if each were to complete the same number of problems in each lesson, even if they did not successfully master each component skill in the lesson. The second semester of the study, both universities integrated PAT more fully into their courses. For instance, students were given quizzes that emphasized the problem-solving skills that would be assessed at the end of the semester, and instructors participated in a training seminar that introduced strategies for integrating PAT with their traditional lessons and stressed the new role of the instructor as a facilitator of knowledge construction. Finally, in the third semester, after positive results for PAT had been demonstrated, the entire course at University P was redesigned (feature h) to be centered around the tutor technology and focus on the principles of authentic problem-solving and multiple representations.

Integration of the technology also includes system integration. This type of integration involved getting the tutor software to function at each site, with all the technological idiosyncrasies of each set-up. Designers must contend with each site's types of computers, amounts of memory, operating system software, and server connections. With a software product under development, as PAT was, this involves error diagnoses, adjustments, and reinstallations, sometimes mid-semester. The time and expertise of the designated computer operator may be limited, and instructors usually wish to concentrate on teaching with the technological tool, not becoming familiar with its technical side. Instructors who are willing to invest the time needed to learn about the "nitty-gritty" of the software may not have access to needed parameters within the university's computer system.

Successful integration of software into the classroom also involves helping it to mesh with the social milieu of the university. All parties involved, including instructors, computer systems operators, committees for human subjects research, mathematics departments, and campus learning skills centers, and designated staff (office and lab assistants) must come to agreements about timelines, division of labor, and costs and available resources. This is not trivial, and administrative policies must be respected.

Formative Evaluation

We monitored the systematic variations in our design (feature i) over an extended period of time (feature j), allowing multiple sources of evidence (feature c) for the success of our design solution. We made sure that our assessments were in alignment with both the objectives of the curriculum, and features of the technology. A design change like a reform curriculum may not show a desired effect if the assessments are not appropriately "aligned", measuring reform objectives. A partially reformed curriculum plus PAT was shown to be superior to the partial reforms alone, as evidenced by University L's same-teacher comparison in Study 2. Evidence that partial reforms alone were superior to a traditional curriculum is given by comparing University L's comparison classes in Study 2 to those in Study 1. Finally, we found that a fully reformed

curriculum plus PAT was superior to a traditional curriculum plus PAT, by comparing Study 3 to Study 1 for University P (see Figure 4). Thus this engineering design success reflects the effectiveness of the complete package, i.e., PAT, the redesigned curriculum, and changes in the role of the instructor, and the strategies for implementation (feature g).

Insert Figure 4 about here

Adding course reforms to the technology over three semesters led to a doubling of students mean scores on the problem-solving assessment from 30% in the control group at the end of the first semester, to 67% in the experimental group at the end of the third semester. We attribute this to the use of PAT and to the increasing integration of ACT principles and PAT features into the curriculum including learning-by-doing, using multiple representations, and problem-solving in authentic situations. This was accomplished through increased teacher training, the addition of PAT-like problem solving in comparison classes, and improvement of the software itself (feature k). The changes are summarized in Table 5.

Insert Table 5 about here

Conclusion

In an engineering design experiment, greater emphasis is put on finding a solution to a problem than on the scientific goal of understanding all the causal relations that exist or do not exist between variables (cf. Schauble, Klopfer, & Raghavan, 1991). Guided by theoretical principles, selective use of mathematical reform notions, and our past experience, we have attempted to create a package with the best chance of improving upon current practices. Given the low number of such studies showing significant positive learning changes in real settings relative to the huge

number of laboratory studies with unclear generalizability, we have chosen to first find a complete package that works -- PAT plus a reform curriculum works.

Design experiments in real settings require compromises with the ideals of laboratory experiments. For example, it is not always possible to control for teacher effects by having the same teacher teach both an experimental and comparison class. We achieved same-teacher controls at University P in Study 1 and University L in Study 2, but resource limitations, moral problems, and imitation effects prevented us from implementing such conditions at both sites every semester. In particular, instructor preparation time is typically quite limited and thus preparing two different versions of the same course is a substantial burden. Further, we observed that experience in teaching an experimental class may "contaminate" teachers such that their instruction in control classes begins to incorporate experimental principles in subtle and not-so-subtle ways (cf. "diffusion or imitation of treatment" in Cook & Campbell, 1979, p. 54). Such "contamination", as viewed from an experimental perspective, is a positive outcome from the engineering perspective. Given the difficulties in creating effective changes in teaching practices, the possibility that technology implementations, like the use of PAT, may lead to positive imitations outside the computer lab is particularly appealing.

Consistent with our experiences, Brown (1992) noted that, "Control groups were difficult to engineer because of resource limitations, even moral problems" (p. 166). Administrators and other stakeholders often perceive students in control conditions as being deprived of the benefits of educational innovation, and such perceptions often precede implementation and testing. At University P, we found positive results of PAT use in Study 1. As experimenters we felt it important to replicate these results and advocated control groups in Study 2, however, there were no classes available to serve as controls in Study 2. Then in Study 3, administrators objected on moral grounds because the summer students all shared living quarters and did not want control students to resent being deprived of the opportunity to use the computers.

We addressed the real difficulties in achieving within-semester controls through the addition of comparisons across semesters. This requires a commitment to the consistent use of the

same assessment tools across semesters. Getting such a commitment from instructors and administrators is not easy and certainly requires up-front planning. Unfortunately we did not fully achieve this goal, for instance we did not use the same exams every semester at each site in our measures of basic algebra skills.

Despite the difficulties in engineering control groups, design experiments should certainly strive for the ideals of laboratory experiments. Nevertheless, we cannot afford to dismiss or ignore studies that only partially achieve them. Many educational decisions are made without any empirical or theoretical input, as happened in the case of University L's decision to drop out of our study. We think it is better to apply results from imperfect studies than to make decisions by intuition alone. As Brown (1992) pointed out, "Given the systemic nature of the environment, however, simple controls can never be entirely satisfactory; but they can provide insights into the operation of some of the major variables." (p. 167).

The three studies presented here provide evidence both for the effectiveness of PAT and the effectiveness of increased incorporation of PAT principles in the regular classroom. Prior evaluations of PAT at the high-school level showed a large impact of the combined effect of PAT and a reformed curriculum over a traditional curriculum without technology (Koedinger, et al., in press). The current Study 1 eliminated the confound between PAT and the curriculum by controlling for the affect of curriculum. Comparing PAT use with a traditional curriculum to the traditional curriculum alone, we found a significant impact of PAT alone. In studies 2 and 3, we saw, in addition to improvements to PAT, increasing use of PAT problems and instructor facilitated learn-by-doing activities outside of the computer lab. As a consequence both experimental and control classes led to increased student outcomes over their corresponding conditions from Study 1.

The diffusion or imitation of treatment effect was certainly in part a consequence of dedicated instructors, one in particular at each site, seeking to innovate and improve student learning. But in addition, we believe these changes were facilitated by use of the PAT technology such that they would not have occurred as easily or as quickly through the usual avenues of

instructor workshops and materials development. We have also observed this diffusion effect in the dissemination of PAT at the high school level, where PAT is now in use at over 20 high schools. In reflecting on the reasons for PAT's affect on teacher change, we see two. First, by acting in effect as a teacher's aid, PAT reduces and changes some of instructors' usual teaching and management duties. PAT selects and delivers activities to each student and then is constantly available to address most of a student's questions. The instructor is freed to address individual student needs that go beyond PAT's capabilities. The instructor's role changes from information deliverer to facilitator, and more time is available for higher-quality interactions with students. Teachers often comment on this benefit of cognitive tutor use (e.g., Werthheimer, 1990) and a third-party evaluator documented such changes in a previous field student of a cognitive for high school geometry (Schofield, Evans-Rhodes & Huber, 1990). Teaching during lab classes is different but not necessarily easier than classroom instruction -- this was particularly true in the earlier semesters of the current experiment when technical and software problems were more frequent. However, preparation for lab days tends to be less work than preparing lectures. Thus, instructors may have had more time to reflect on their practices and consider implementing changes.

A second influence of PAT on teacher change is one of diffusion of principles and approaches. PAT provides an extensive and active example of a different way of teaching than is currently common in developmental math classes. Rather than the short term exposure that might occur in reviewing materials or viewing a lecture on mathematics reform, instructors had regular and repeated experience with PAT throughout the lab sessions in Study 1 and in the pilot implementations prior to Study 1. Initially, the PAT activities were most apparent to instructors and in Study 1 instructors at University L began to ask whether they might use some of the PAT problems in the regular classroom. The diffusion of PAT objectives was clearly apparent in the use of the PAT-like quizzes in Study 2. Both experimental and comparison classes began to address these objectives in their regular classroom sections. Further diffusion of PAT principles of focusing on learning by doing and teaching by facilitating in the context of this doing began to

emerge in Study 2 such that administrators at University P began to plan ambitious reforms for the structure and content of the course. These reforms, implemented in Study 3, increased learn-by-doing time by reversing the usual time allocation that emphasized lecture over supported problem solving in recitations. Further, group projects were created and scaffolded with structured roles, thus providing students with peer tutoring in the context of problem solving.

This design experiment is on-going. We expect even greater improvements as the consequence of consistent monitoring of desired reform objectives and continued software and course redesign to better achieve these objectives.

References

AMATYC: American Mathematical Association of Two-Year Colleges (1995). Crossroads in mathematics: Standards for introductory college mathematics. Memphis, TN: Author.

Anderson, J.R. (1993). Rules of the mind. Hillsdale, NJ: Erlbaum.

Anderson, J.R., Boyle, C.F., Corbett, A., and Lewis, M.W. (1990). Cognitive modeling and intelligent tutoring. Artificial Intelligence, 42, 7-49.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4, 167-207.

Blessing, S. & Anderson, J. R. (1996). How people learn to skip steps. Journal of Experimental Psychology: Learning, Memory and Cognition, 22, 3.

Bielaczyc, K., Pirolli, P. & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. Cognition and Instruction, 13 (2), 221-252

Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. Journal of the Learning Sciences, 2, 141-178.

Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18 (3), 439-477.

Collins, A. (1992). Toward a design science of education. In E. Scanlon and T. O'Shea (Eds.), New directions in educational technology. New York: Springer-Verlag.

Cook, T. D. and Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Boston, MA: Houghton-Mifflin.

Corbett, A.T. and Anderson, J.R. (1992). Student modeling and mastery learning in a computer-based programming tutor. In Proceedings of the Second International Conference on Intelligent Tutoring Systems. Montreal, Canada.

Corbett, A. T. & Anderson, J.R. (1995). Knowledge decomposition and subgoal reification in the ACT programming tutor. In Proceedings of the 7th World Conference on Artificial Intelligence

in Education, (pp. 469-476). Charlottesville, VA: Association for the Advancement of Computing in Education.

Kloosterman, P. and Stage, F.K. (1992). Measuring beliefs about mathematical problem solving. School Science and Mathematics, 92(3), 109-115.

Koedinger, K.R. and Anderson, J.R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. Cognitive Science, 14, 511-550.

Koedinger, K.R. and Anderson, J.R. (1993a). Reifying implicit planning in geometry: Guidelines for model-based intelligent tutoring system design. In Lajoie, S., and Derry, S. (Eds.) Computers as cognitive tools, (pp. 15-45). Hillsdale, NJ: Erlbaum.

Koedinger, K. R. and Anderson, J. R. (1993b). Effective use of intelligent software in high school math classrooms. In Proceedings of the World Conference on Artificial Intelligence in Education, (pp. 241-248). Charlottesville, VA: AACE.

Koedinger, K.R. and Anderson, J.R. (in press). Illustrating principled design: The early evolution of a cognitive tutor for algebra symbolization. Interactive Learning Environments.

Koedinger, K.R., Anderson, J.R., Hadley, W.H. and Mark, M.A. (in press). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8. Also appears in Proceedings of the 7th World Conference on Artificial Intelligence in Education, (pp. 421-428). Charlottesville, VA: Association for the Advancement of Computing in Education.

Kozma, R.B. (1991). Learning with media. Review of Educational Research, 61, 179-211.

MAA: Mathematics Association of America. (1996). Quantitative reasoning for college graduates: A complement to the standards [On-line]. Available: http://www.maa.org/past/ql/ql_toc.html.

Madison, B.L. and Hart, T.A. (1990). A challenge of numbers: People in the mathematical sciences. Washington D.C.: National Academy Press.

McKendree, J. E. (1990). Effective feedback content for tutoring complex skills. Human Computer Interaction, 5, 381-414.

Nathan, M. J., Kintsch, W., & Young, E. (1992). A theory of algebra word problem comprehension and its implications for the design of computer learning environments. Cognition and Instruction, 9, 329-389.

NCTM: National Council of Teachers of Mathematics (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: Author.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. Journal of Research in Science Teaching, 28(9), 859-882.

Schofield, J.W., Evans-Rhodes, D., and Huber, B.R. (1990). Artificial intelligence in the classroom: The impact of a computer-based tutor on teachers and students. Social Science Computer Review, 8, 24-41.

Vygotsky, L.S. (1978). Mind in society. Cambridge, MA: Harvard University Press

Wertheimer, R. (1990). The Geometry Proof Tutor: An "Intelligent " Computer-based tutor in the classroom. Mathematics Teacher, 308-313.

Acknowledgments

Support was provided by the Fund for the Improvement of Post-Secondary Education, U.S. Dept. of Education, grant number P116B41269 to Dr. Koedinger. This evaluation would not have been possible without the Practical Algebra Tutor software, for which John Anderson, Bill Hadley, Mary Mark, Ray Pelletier, and Steve Ritter have made major design, curriculum development, or programming contributions. We also wish to thank Bonnie Goins, Audra Charity, and Roxanne Eckenrode for their research assistance, Lora Shapiro and Mark Clark for their dedication as site coordinators, and all the instructors who participated in the experiment. Portions of these data appear in Koedinger, K. R. and Sueker, E. L. F. (1996), PAT goes to college: Evaluating a cognitive tutor for developmental mathematics, Proceedings of the International Conference on the Learning Sciences (pp.180-187).

Table 1. Components and features of design experiments, and their instantiation in this study.

Components	Features	Feature as Instantiated by this Design Experiment
1. Goals and problem identification	a. Needs assessment	Identification of developmental mathematics problem
	b. Multiple stakeholders+	Design guided by instructional objectives of students, instructors, and administrators
	c. Multiple assessments*+	Performance-based tests and basic skills exams
2. Theory	d. Theoretical basis+	ACT-R; problem-solving; learning in math domain
3. Design solution	e. Multiple expertise in design process*	Contributions from computer scientists, psychologists, instructors, administrators and instructor educators
	f. Instructors as co-investigators*	Instructors & administrators helped formulate research questions, customize tutor, design assessments
	g. Multiple facets of innovation*+	Curricular reform, performance technology, tutoring technology, instructor training
	h. Test changes most likely to succeed 1st*	Added cognitive tutor to existing courses before reforming a course
4. Formative evaluation	i. Systematic variation*	Classes taught by same instructor with & without tutor
	j. Monitoring	Multiple comparison conditions within and between sites and instructors, over extended periods of time
	k. Formative methods*	Flexible design revision mid-semester; improvement of tutor and course based on results
5. Summative evaluation & dissemination	l. Objective evaluation*	Assessments scored objectively with high interrater reliability, but third-party evaluation not achieved.
	m. Feasible & easily disseminated+	Tutor technology reifies approach and makes transition to new curriculum and teaching roles easier

* listed in Collins (1992); + addressed by Brown (1992)

Table 2. Mean Scores on Eight Components of the Cellular Phone Application Problem at Two Research Sites, Fall 1995.

	University L		University P	
Component	Experimental (n = 88)	Comparison (n = 199)	Experimental (n = 12)	Comparison (n = 16)
(a) Defining variables	.83 ***	.54	1.00	.84
(b) Writing equations	.72 ***	.51	.75	.69
(c) Making a table	.52 ***	.22	.67 **	.16
(d) Constructing a graph	.38 ***	.11	.42	.34
(e) Finding slopes and intercepts	.12 ***	.03	.02	.02
(f) Finding points of intersection	.16 ***	.02	.12	.06
(g) Solving for x, given y	.49 ***	.28	.54	.31
(h) Stating ranges	.17 ***	.07	.17 *	.00
Mean Overall	.42 ***	.22	.46*	.30

* $p < .05$; ** $p < .01$; *** $p < .001$. Reported p-values are based on a one factor analysis of variance with condition as the between-subjects factor.

Table 3. Mean Scores on Eight Components of the Cellular Phone Application Problem at Two Research Sites, Spring 1996.

Component	University L				University P
	Overall Experimental (n = 79)	Overall Comparison (n = 59)	Same-instructor Experimental (n = 27)	Same-instructor Comparison (n = 33)	Experimental (n = 18)
(a) Defining variables	.68	.70	.74	.67	1.00
(b) Writing equations	.92	.88	.94	.89	1.00
(c) Making a table	.70	.80	.70	.70	.94
(d) Constructing a graph	.44	.63**	.52	.65	.56
(e) Finding slopes and intercepts	.30	.22	.46*	.24	.17
(f) Finding points of intersection	.15	.12	.29	.17	.42
(g) Solving for x, given y	.34*	.15	.53**	.24	.61
(h) Stating ranges	.13	.17	.20	.24	.53
Mean Overall	.46	.46	.54	.48	.65

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 4. Mean Scores on Eight Components of the Storage Space and Cellular Phone Application Problems at One Research Site, Summer 1996.

	University P, Experimental Classes (n = 81)	
Component	Pre-Test (Storage Space Problem)	Post-Test (Cellular Phone Problem)
(a) Defining variables	.43	.89 ***
(b) Writing equations	.51	.95 ***
(c) Making a table	.31	.89 ***
(d) Constructing a graph	.07	.84 ***
(e) Finding slopes and intercepts	.02	.17 **
(f) Finding points of intersection	.00	.70 ***
(g) Solving for x when given y	.34	.34
(h) Stating ranges	.05	.60 ***
Mean Overall	.22	.67 ***

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 5. Changes to the developmental mathematics courses at universities P and L over the course of the monitored formative design experiment.

Semester	Site(s)	Change	Result
Fall 1995	P and L	Added PAT to traditional curriculum.	Better at applied problem-solving compared to control classes with no detrimental effect on traditional skills.
	L	Required students to complete PAT lessons outside of class, based on knowledge tracing.	Possible positive effect on traditional skills?
Spring 1996	P and L	Partially reformed curriculum and tutor: added quizzes; improved labels for PAT worksheet.	Improved scores on applied problem-solving compared to Fall 1995.
	L	Gave PAT problems on paper to comparison classes	Greater emphasis on multiple representations in the classroom.
	P	Required students to complete PAT lessons outside of class.	Slight positive effect of technology on traditional skills.
	P and L L	Increased instructor training. Included comparison instructors in training.	Large teacher variability; smaller differences between PAT and control classrooms.
Summer 1996	P	Reformed curriculum: added group projects requiring extended problem-solving, non-text-centered coursework, and PAT-like problems on exams.	More than 100% better on applied problem-solving than control students in Fall 1995.

Edit Windows Students Problem Grapher Equations 8:01:29

SOLVE
SUBSTITUTE

Problem Statement

PROBLEM

We need to rent a large truck for one week. Rental companies charge an initial cost plus an additional cost for each mile driven. One company, Hertz, will rent a 27 foot truck to us for $525 plus $0.21 per mile. Another company, Avis, will rent us the same size truck for $585 plus $0.13 per mile.

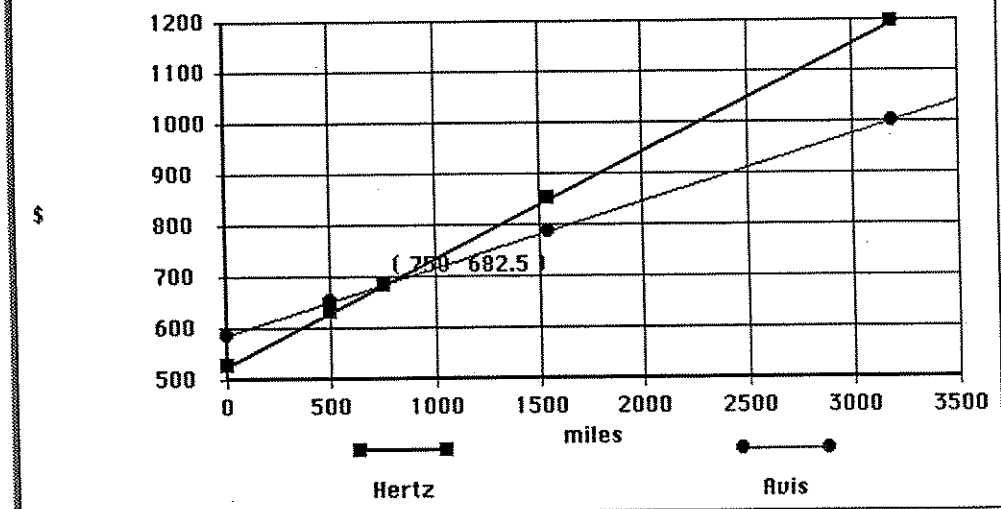
1. What is the initial cost, in dollars, of renting this model truck from Hertz? From Avis?
2. How much would it cost to rent the truck from Hertz if we will be driving it 500 miles? From Avis?
3. If we have budgeted a total of $850 to rent this truck, how many miles can we drive it if we rent it from Hertz?
4. If we have budgeted a total of $1000 to rent this truck, how many miles can we drive it if we rent it from Avis?
5. How many miles would we need to drive the truck so that the cost of rental from each company would be equal?

Worksheet

	Distance	Hertz	Avis
UNIT	miles	$	dollars
FORMULA	m	525+0.21m	585+0.13m
1	0	525.	585.
2	500	630.	650.
3	1547.619...	850	786.1905...
4	3192.3077...	1195.3846...	1000
5	750	682.5	682.5

Grapher

	Lower Bound	Upper Bound	Scale Size
Miles Settings	0	3500	500
$ Settings	500	1200	100



Messages

You have completed the worksheet. You have completed work with the grapher. You have completed this problem.

Equations

525+0.21m=585+0.13m

m=750.

Skillometer

- ✓ Labelling worksheet
- ✓ Identifying units
- ✓ Entering a given
- ✓ Solving for y
- Solving for x
- Defining a variable
- Writing an expressio
- Making a graph
- Labelling axes on th
- Labelling lines
- Drawing lines
- ✓ Changing axis bounds
- ✓ Changing axis interv

Figure 2.

Cellular Phone Problem

You are told that tomorrow you are to order cellular phone service for all the officers in your company. Your boss tells you that she will be providing you with the necessary information about the amount of "airtime" (number of minutes of phone time) per month that each officer will need. She also informs you that she will have this information for you about an hour before you must present your report and decision to the President of the company. Furthermore, she makes it very clear that your future with the company will depend on how well you perform this duty.

Knowing that you will need at least an hour just to put your report together, you contact the local cellular phone company. They give you the following information about their three available services.

- **Economy Service:** With this plan each person is charged $19.95 per month and $0.31 per minute of airtime.
- **Silver Service:** With this plan each person is charged $40.95 per month and $0.16 per minute of airtime.
- **Gold Service:** With this plan each person is charged $80.95 per month with no charge for airtime.

To prepare for tomorrow you must do a mathematical analysis of these three different plans. This analysis should include **defining variables, writing equations, making tables, constructing graphs, finding slopes** and **intercepts**, and **finding points of intersection.**
Your boss suggests that you look at these plans over a range of airtime from 0 to 500 minutes per month, and look at how much airtime you can get per month with each of these plans **for a total cost of $100.** She also makes it very clear that you must include **the range of airtime for which each plan is the cheapest.**

PLEASE SHOW ALL YOUR WORK.

Figure 3.

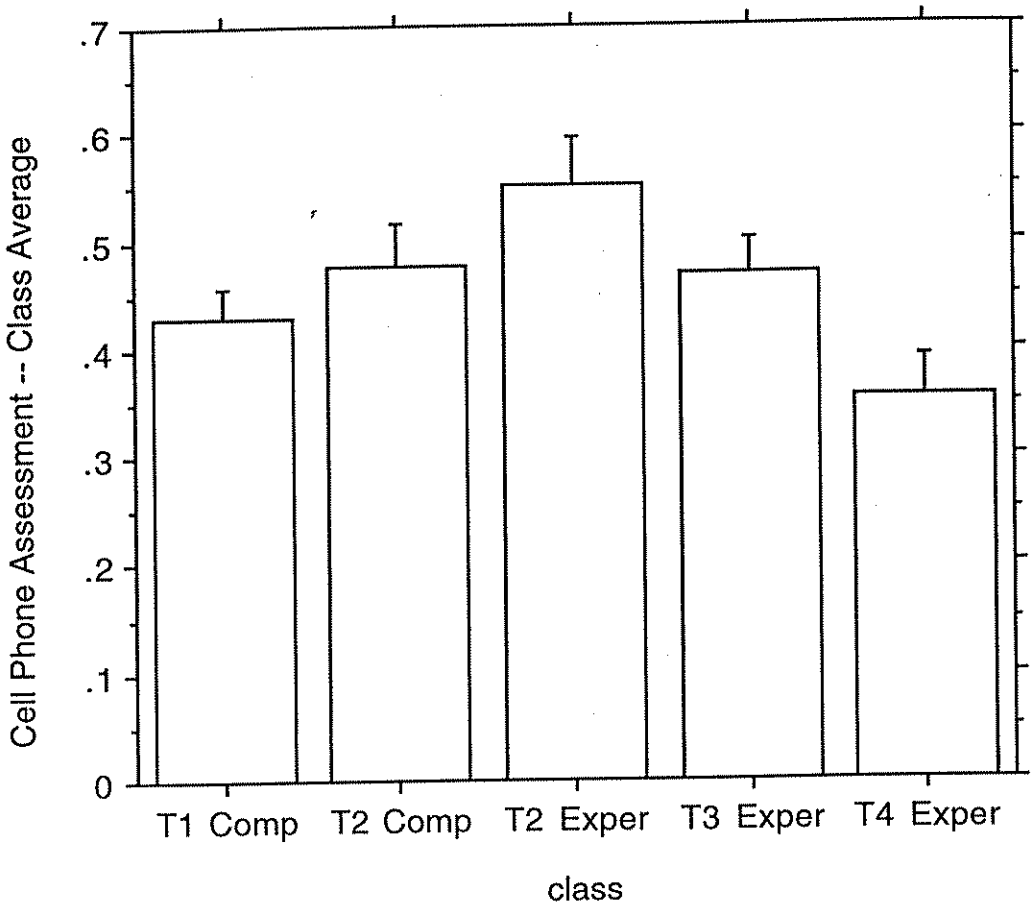

Cell Phone Assessment -- Class Average
.7
.6
.5
.4
.3
.2
.1
0
T1 Comp
T2 Comp
T2 Exper
T3 Exper
T4 Exper
class

Figure 4.

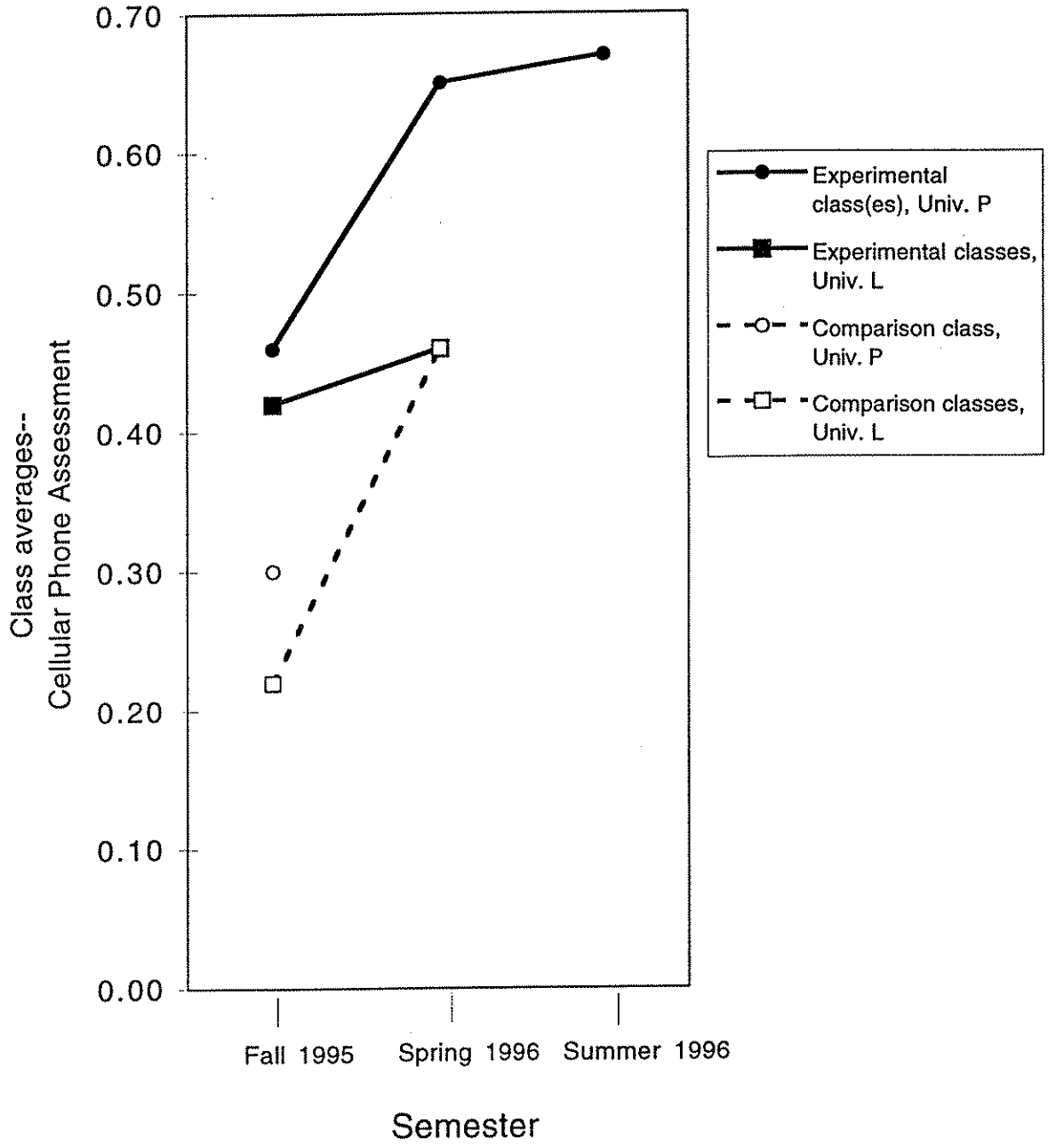

0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Class averages--
Cellular Phone Assessment
Fall 1995
Spring 1996
Summer 1996
Semester
Experimental class(es), Univ. P
Experimental classes, Univ. L
Comparison class, Univ. P
Comparison classes, Univ. L

Figure Captions

Figure 1. PAT (Practical Algebra Tutor) provides just-in-time intelligent support for the use of multiple algebraic representations in the analysis of real-world problem situations.

Figure 2. "The Cellular Phone Problem," the problem-solving assessment given to students at the end of each semester.

Figure 3. Teacher variability at University L in Study 2.

Figure 4. Improvement of the design solution over time at two research sites.