A Large-Scale Evaluation of an Intelligent Discovery World: Smithtown

Valerie J. Shute

Air Force Human Resources Laboratory
Brooks Air Force Base, TX

Robert Glaser

Learning Research and Development Center
University of Pittsburgh, PA

Abstract

Smithtown is an intelligent tutoring system designed to enhance an individual's scientific inquiry skills as well as to provide an environment for learning principles of basic microeconomics. It was hypothesized that computer instruction on applying effective interrogative skills (e.g., changing one variable at a time while holding all else constant) would ultimately lead to the acquisition of the specific subject matter. This paper presents an evaluation of Smithtown in two studies of individual differences in learning. Experiment 1, an exploratory study, demonstrated that Smithtown fared very well when compared to traditional instruction on economics and delineated the performance indicators which separated better from worse learners in this discovery environment. Experiment 2 extended the findings from the exploratory study using a large sample of subjects ($N = 530$) from a different population interacting with Smithtown and showed that the performance indicators relating to hypothesis generation and testing were the most predictive of successful learning in Smithtown, accounting for considerably more of the variance in our learning criterion than a measure of general intelligence. Overall, the system performed as expected. Tutoring on scientific inquiry skills resulted in increased knowledge of microeconomics. The differentiating behaviors between more and less successful subjects were in agreement with specific behaviors relating to individual differences found in general studies on problem solving and concept formation. From an instructional perspective, the behaviors we have denoted can serve as a focal point for relevant intervention studies. From a design perspective, findings from these studies suggest modifications to intelligent tutoring systems so they may be more like the individualized teaching systems they have the potential to be.

Smithtown is an intelligent tutoring system designed as a guided discovery world whose primary goal is to assist individuals in becoming more systematic and scientific in their discovery of laws for a given domain. A second goal of the system is to impart specific content knowledge in microeconomics, specifically the laws of supply and demand.

This paper presents a large-scale evaluation of *Smithtown* with regard to these two goals. The first study compared declarative knowledge acquisition between subjects interacting with this system, students enrolled in an introductory economics course, and a control group using both quantitative and qualitative measures. Focusing on the group interacting with *Smithtown*, we analyzed differences in the behaviors (or performance indicators) between those individuals that were successful in this type of discovery environment versus those less successful. "Success" was defined as a large gain score in performance from a pretest battery of economic con-

The authors wish to acknowledge the many persons whose contributions have been invaluable to this project: Bill Alley, Jeff Blais, Jeff Bonar, Ray Christal, Kathleen Katterman, Pat Kyllonen, Alan Lesgold, Kalyani Raghavan, Wes Regian, Paul Resnick, and Dan Woltz. A special note of gratitude must be extended to the creative and diligent programmers of Smithtown: Jamie Schultz and Audrey Peterson.

Support for the large-scale testing and analyses of Smithtown was provided by the Learning Abilities Measurement Program (LAMP), part of the Air Force Human Resources Laboratory. The Center for the Study of Learning is funded by the Office of Educational Research and Improvement of the U.S. Department of Education. The opinions expressed do not necessarily reflect the position or policy of either AFHRL or OERI and no official endorsement should be inferred.

Correspondence and requests for reprints should be sent to Valerie J. Shute, AFHRL/MOE, Brooks AFB, TX 78235

cepts to posttest battery scores. The second study analyzed data from a large group of subjects (N = 530) interacting with the system to see which performance indicators were correlated with a dependent measure of learning as well as a general intelligence measure. Of interest was whether these data replicated the findings from the first study that employed a smaller sample from a different population, as well as investigating the nature and range of the relationship between general intelligence and learning.

Scientific inquiry can be seen as a problem-solving activity involving both top–down and bottom–up processing of information (Greeno & Simon, 1988). We are particularly interested in training scientific inquiry skills which include: (a) generating and testing hypotheses, (b) observing, recording, and organizing data resulting from experimental tests, (c) modifying hypotheses in accordance with the results, and (d) inducing regularities and laws (Shute, Glaser, & Raghavan, 1989).

Generating and testing hypotheses using observations and empirical findings is important to scientific work, as well as to the acquisition of knowledge in general. When hypotheses are generated and new information is obtained, they serve as a basis for confirming or refuting perceived regularities and lawful relationships. There are two problems associated with induction and hypothesis testing. First, many learners can induce regularities or patterns, but do not treat them as hypotheses to be tested. Second, even when subjects realize that they should test a hypothesis, they may use faulty methods or procedures that do not guarantee that the inferences drawn on the data are reasonable and/or relevant to the world or system being observed.

Previous studies of induction have mainly focused on inducing a rule or classifying relatively abstract stimuli into categories on the basis of feedback about classification errors and other information (see Pellegrino & Glaser, 1980; Smith & Medin, 1981). Given our interest in exploratory environments, we see this large literature as relating mostly to passive induction where learners induce rules, make hypotheses, and classify and taxonomize observations on the basis of experimenter-controlled presentation of predetermined instances. However, a more active process is apparent when the learner can select variables, design instances, and interrogate his or her existing knowledge and memory for recent events. To study the latter form of induction, we apply a research paradigm that allows us to examine active experimentation in which learners explore and generate new data and test hypotheses with the data they have accumulated in the course of their investigations. Recent experimental technology and computer modeling have made this type of experimentation feasible (Klahr & Dunbar, 1988; Bonar, Cunningham, & Schultz, 1986; Michalski, 1986; Yazdani, 1986).

Facilitation of scientific inquiry skills has been investigated by White and Horowitz (1987) via their 'Thinker Tools' environment. Their approach was to first motivate students to want to learn by pointing out errors and inconsistencies in their current beliefs. Second, the students were guided through a series of microworlds, each one more complex than the preceding one with the objective of evolving more precise mental models of the subject matter (i.e., Newtonian mechanics). Third, the students had to formalize their developing mental models by evaluating a set of laws describing phenomena in the microworld. Finally, the students had to apply the selected law to see how it predicted real world phenomena.

A difference between White & Horowitz's approach and our approach is the degree of student control in the learning process. It is our belief that a more active process can be more facilitating to knowledge and skill acquisition, especially in conjunction with tutorial assistance on strategies related to testing generalizations. We believe that discovery learning can contribute to a rich understanding of domain information by enabling students to access and organize information themselves.

Thus, applying interrogative skills is the 'active process' that leads to learning in discovery situations. A proposition to be evaluated in this research is that effective interrogative skills are teachable or trainable if they can be articulated and practiced under circumstances which require their use.

Another hypothesis to be tested in this research is that intelligent tutorial guidance on effective inquiry skills, combined with a discovery world environment, can transform haphazard problem solving procedures into efficient, methodical learning procedures. Such a transformation arises from an individual's own actions and hypotheses. Thus, a second proposition to be evaluated in this research is that focusing on the tutoring of specific inquiry skills should consequently lead to learning the subject matter.

The remainder of this article will be organized as follows. First, we overview the system and present the two knowledge bases in *Smithtown*: Inductive inquiry skills and economic knowledge. Second, we show how individuals can maneuver within the *Smithtown* environment. Third, we include a section describing our exploratory investigation (Experiment 1) into individual differences in learning within this environment. Fourth, we present results from comparing the learning outcomes of subjects using *Smithtown* with another instructional treatment and no treatment. This section also includes the results from an analysis of effective performance characteristics of the subjects in the experimental group. Fifth, we discuss a large scale confirmatory analysis of data (Experiment 2) obtained from subjects using *Smithtown*. Finally, we conclude with a general discussion of the educational and scientific importance of these studies.

SMITHTOWN

The main goal of *Smithtown* is to enhance students' general problem solving and inductive learning skills. It does this in the context of microeconomics, providing an environment that fosters learning the laws of supply and demand. *Smithtown* is a highly interactive program, allowing students to pose questions and conduct experiments within the computer environment, testing, and enriching their knowledge bases of functional relationships by manipulating various economic factors.

Since *Smithtown* was designed to be a guided discovery environment, there is no fixed curriculum. Rather, the student generates his or her own hypotheses and problems, not the system. After generating a hypothesis (e.g., 'Does increasing the price of coffee affect the demand for Cremora?'), the student tests it by executing a series of actions, such as collecting baseline data on Cremora (e.g., the equilibrium price, the quantity demanded at that price), entering the coffee market and increasing the price of coffee, the returning to the Cremora market and observing the ensuing changes to relevant variables. To make this affect more salient, data may be plotted, such as superimposing two demand curves for Cremora, both before and after the price change was made to coffee. This series of actions for creating and executing a given "experiment" defines a student solution.

Smithtown has the instructional goal of teaching general problem solving skills. Instead of a curriculum-based instructional sequence, *Smithtown* relies on a process of constantly monitoring student actions, looking for evidence of good and poor behaviors, and then coaching students to become more effective problem solvers. The system keeps a detailed history list of all student actions, grouping them into (i.e., interpreting them as) behaviors and solutions. *Smithtown* diagnoses solution quality in two ways. It looks for overt errors by comparing students solutions with its "buggy critics" which are sets of actions or nonactions that constitute suboptimal behaviors. It also compares student solutions with its own "good critics" or expert solutions. Discrepancies between the two are collected into a list of potential problem areas and passed on to the coach for possible remediation.

Another area of 'intelligence' in *Smithtown* resides in its knowing about economic relationships among different variables. After a student conducts an experiment or series of experiments, collects data testing a hypothesis, and evaluates the results, he or she is in a position to make a generalization about an economic phenomenon. The student may state this generalization in the hypothesis window. The system compares the learner's input with known relationships, and if the student states a valid principle or law governing economic variables (e.g., As price increases, quantity demanded decreases), he or she is informed that their articulated hypothesis was correct (e.g., 'Congratulations! You have just discovered what economists refer to as the law of demand'), incorrect, or not understood by the system.

Figure 1 shows the flow of information and control in *Smithtown*; that is, the possible interactions one may have with the system as well as the modules of the program. The student takes some action in the environment. Particular sequences of actions constitute different inquiry behaviors that the system matches against known behaviors, both good and bad. The system's coach provides feedback to the student based on the type of error shown (i.e., an overt error: demonstrating a buggy behavior, or an error of omission: not doing something that was appropriate at the time), mediated by some pedagogical heuristics (e.g., first address buggy behaviors before addressing errors of omission, or if several behaviors are confirmed, address the more critical one with the higher weight first). If the student action is to specify a hypothesis, the system pattern matches the statement against known economic relationships and provides feedback appropriately.

Most intelligent instructional systems require some kind of knowledge base (de-

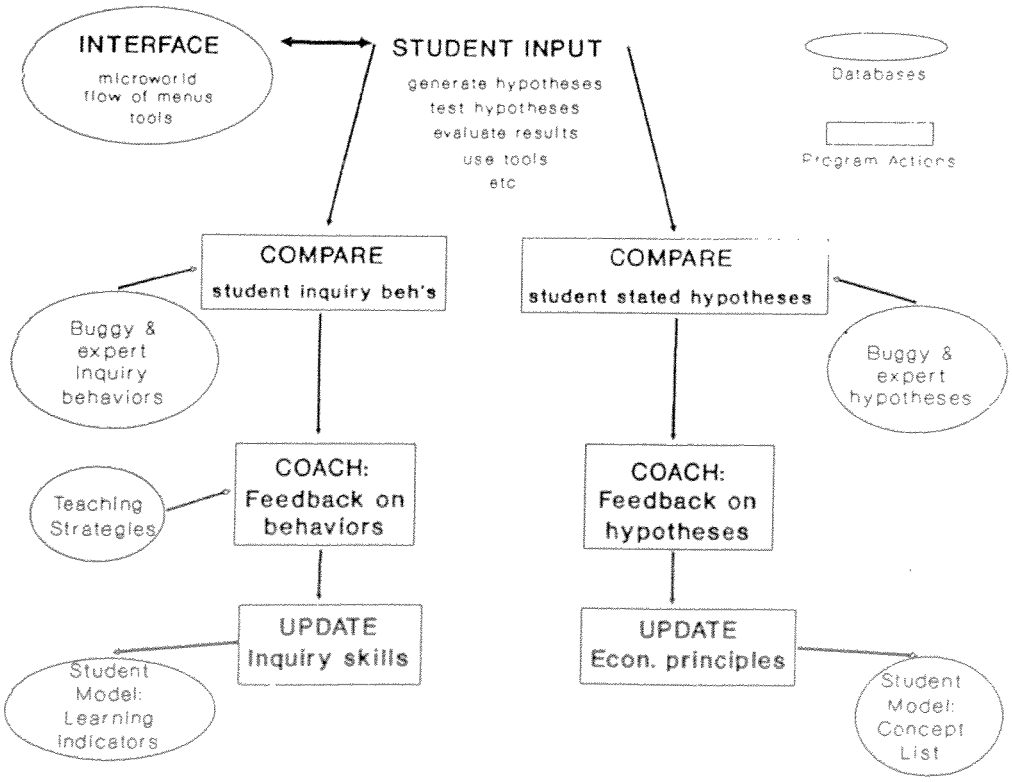


Figure 1. Flow chart of information and control in Smithtown

clarative or procedural) to be learned. We now detail *Smithtown*'s two-system knowledge bases: one for inductive inquiry skills (procedural knowledge, or knowing how to do X) and the other for economic concepts (declarative knowledge, or knowing about X).

Inductive Inquiry Skills

Scientific inquiry behaviors were delineated and categorized by an earlier study conducted with *Smithtown* yielding information about effective and ineffective behaviors for interrogating a new domain (see Shute & Glaser, in press). Some examples of "good" inquiry behaviors would be changing one variable at a time while holding everything else constant and conscientiously recording relevant data in the online notebook. These behaviors were coded into rules and the system monitors a learner's actual behaviors with respect to these rules. Thus the system recognizes sequences of good behaviors and also sequences of ineffective or buggy behaviors.

If a student is performing satisfactory in the environment (i.e., not repeatedly manifesting buggy behaviors and learning economic concepts at a reasonable rate), he or she will not be interrupted except to receive occasional congratulatory feedback when relevant. However, if the system determines that a student is floundering or demonstrating buggy behaviors, the coach will intervene and offer assistance on the specific problematic behavior(s). For instance, if a student changes many variables at one time without first looking at the baseline data, the following rule would be invoked (paraphrased):

If *The student changes more than two variables at a time prior to collecting baseline data for a given market, and it is early in the session where the experiment number is less than four,*
Then *Increment the 'Multiple Variable Changes' bug count by 1 and pass the list to the coach for possible assistance.*

If this rule count surpasses a threshold value (e.g., three times), then the coach would appear on the screen, informing the student, *"I see that you're changing several variables at the same time. A better strategy would be to enter a market, see what the data look like before any variables have been changed, then just change one variable while holding all the others constant."*

We also created a list of specific performance measures to determine the type of actions yielding differential performance in this environment. These performance or learning indicators were extracted from the student history list and arrayed by complexity, from low-level, simple counts of actions (e.g., total number of notebook entries made) to higher-level, complex behaviors (e.g., number of times a generalization of a concept was made across related goods). These indicators appear in Appendix A and serve as one data source in this study on individual differences in learning in *Smithtown*.

Economic Concepts

The second knowledge base concerns functional relationships among economic variables relating to supply and demand in a competitive market. The concepts were selected following discussions with an economics professor about relevant concepts for an introductory microeconomics course. Definitions can be seen below:

Demand: The buyer's side of the market is called demand. The law of demand says that the quantity of a product which consumers would be willing and able to buy during some period of time is inversely related to the price of the product. Graphing this relationship results in a demand curve showing how the quantity demanded of a good or service will change as the price of that good or service changes, holding all other factors constant.

Supply: The seller's side of the market is called supply. The law of supply is that

the quantity of a product which producers would be willing and able to produce and sell is related to the price of the product in a positive function. Graphing price and quantity supplied results in a supply curve.

Equilibrium: There are many factors that influence the price of a given product, but when a price is reached where the quantity supplied is equal to the quantity demanded, that market is at a point of equilibrium. Competitive markets always tend toward points of equilibrium.

Surplus: If the market price is higher than the equilibrium price, buyers will demand smaller quantities than sellers are supplying. This will create a surplus. Surpluses of unsold goods will tend to lower the price down toward the equilibrium level.

Shortage: If the market price is lower than the equilibrium price, buyers will demand larger quantities than sellers are supplying, thus creating a shortage. Shortages will lead to price increases, and the price will rise toward the equilibrium level.

Change in Demand: A change to variables other than price will cause the entire demand curve to shift, depending on which variable is changed and the magnitude of the adjustment. Some of the variables in *Smithtown* that can be manipulated and that shift the demand curve are: per capita income, population, interest rates, weather, consumer preferences, and the price of substitute and complementary goods.

Change in Supply: Again, changing certain variables other than price in *Smithtown* will cause the entire supply curve to shift, depending on the variable and the amount of change. The variables or 'town factors' that can be manipulated to effect a supply curve shift include: labor costs, number of suppliers, as well as some of the variables mentioned above (e.g., weather).

New Equilibrium Point: Equilibrium, once established, can be disturbed by changes in demand and/or supply. If demand and/or supply change, a surplus or shortage will result at the original price, and the price will move toward a new equilibrium. A shortage at the original price will cause the old price to rise to the new level and cause changes in the quantities supplied and demanded. A new equilibrium will be established at the second price and the second quantity.

Additional Concepts: Besides the above economic concepts, at least two more can be learned from the discovery world, although they are not explicitly recognized by the system: cross elasticity of demand and supply. Cross elasticity of demand indicates how a change in one market affects the demand in a related market while cross elasticity of supply indicates how a change in one market affects the supply in a related market.

To learn the concepts embedded in *Smithtown*, students are free to manipulate variables, observe the effects, and apply the online tools to organize their information in an effective way. Tools available for these activities include a notebook for collecting data from experiments (Figure 2), a table to organize data from the notebook (Figure 3), a graph utility to plot data (Figure 4), and a hypothesis menu to formulate relationships among variables (Figure 5). Three history windows allow the students to see a chronological listing of all actions, data, and concepts learned.

Students' experiments are independently created and executed, thus unique to each individual. The system recognizes two types of systematic investigations: (1) explorations—observing and obtaining information from *Smithtown* in order to generate hypotheses about the microeconomic concepts and laws, and (2) experiments—a series of student actions conducted to confirm or differentiate hypotheses (see Shrager, 1985, for a similar demarcation). Experiments are associated with a specific prediction from the 'Prediction' menu while explorations are not. Moreover, the system

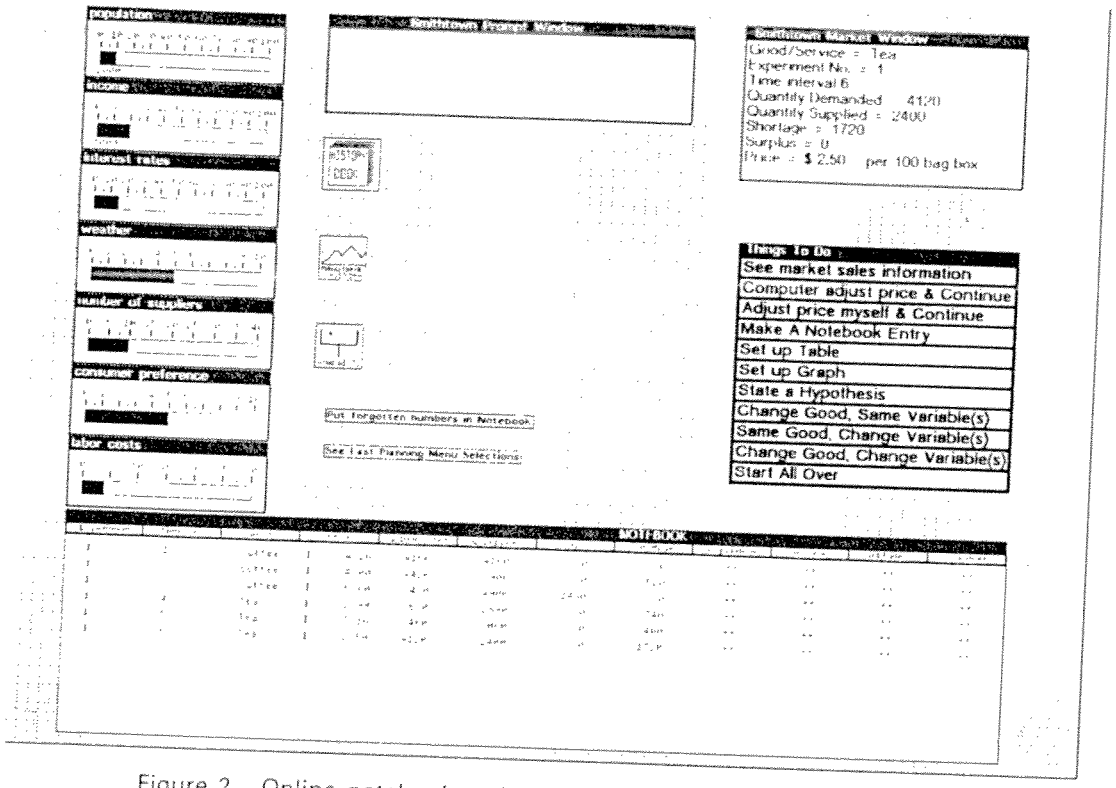


Figure 2. Online notebook and other screen features of Smithtown

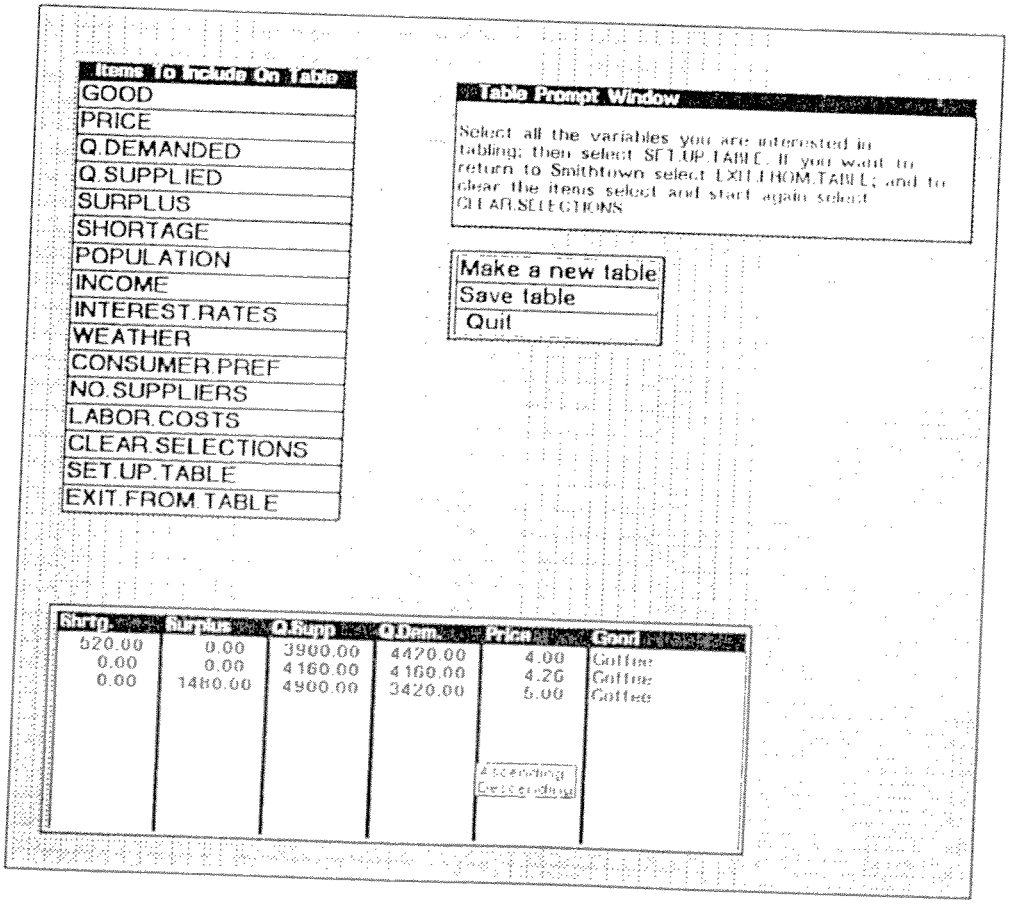


Figure 3. Table package for ordering data

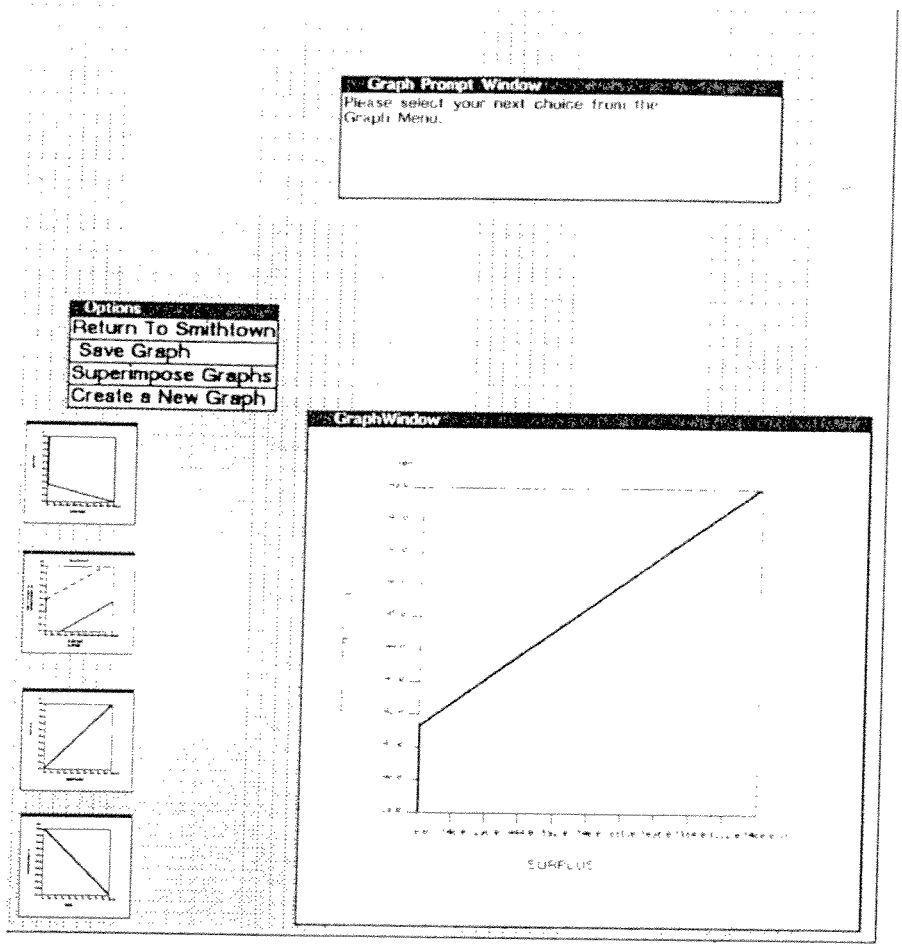


Figure 4. Graph package for plotting data

does not provide coaching while a learner is in 'exploratory mode'. Only when a person is classified as conducting an experiment does he or she receive feedback on behaviors.

The procession of events a person goes through in creating an experiment are fixed. First a student selects a market to investigate from the 'Goods Menu' (see Figure 6a). The selection of markets to include in the system was based on interesting relationships existing among different goods, such as complementary associations (e.g., Ground beef and hamburger buns), substitute goods (e.g., coffee and tea) as well as more complex relationships (e.g., large cars, compact cars, and gasoline). Next, the student informs the system of his or her experimental intentions by identifying variables of interest from the 'Planning Menu' (see Figure 6b). After choosing the focal variables for further experimentation, a student is free to make changes to any of the town factors (see Figure 6c). For each new experiment, the system asks the student if he or she would like to make a prediction regarding the planned experiment. If the student says 'No', the next menu to appear is the 'Things To Do Menu'. If the student replies 'Yes', a window appears where specific statements can be entered about predicted outcomes. For example, if a student wanted to investigate the relationship between income and the demand for large cars, and then proceeded to increase the

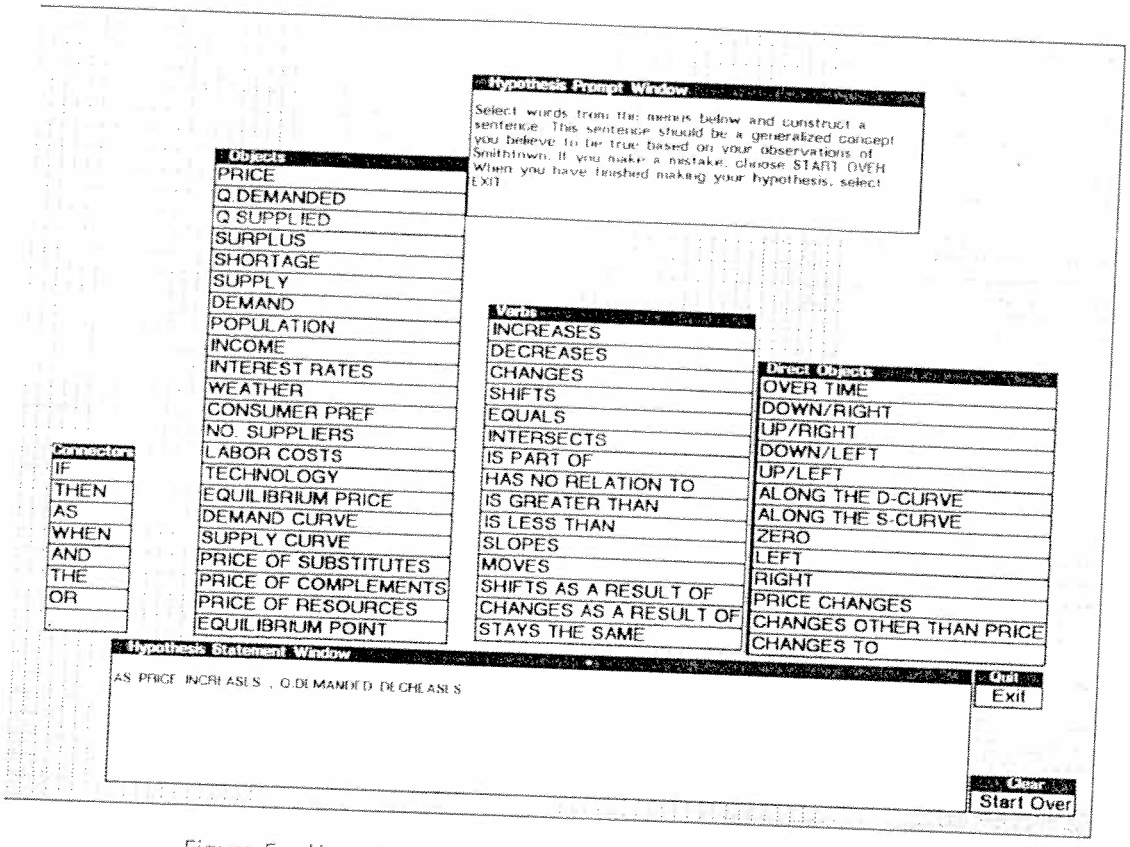


Figure 5. Hypothesis menu with the law of demand specified

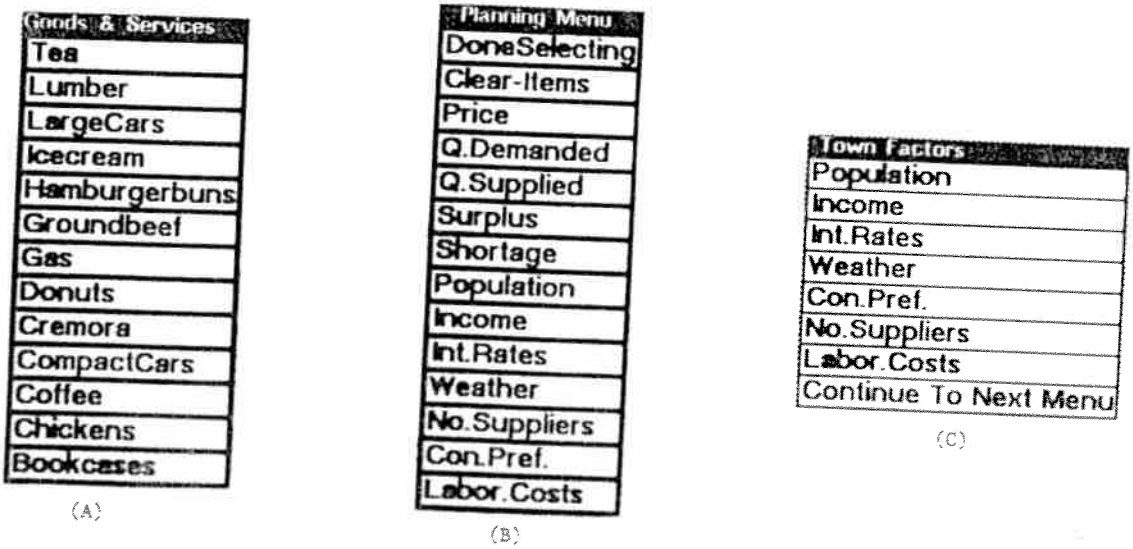


Figure 6a. Goods and services menu (Smithtown markets) Figure 6b. Planning menu items (focal variables for upcoming experiment) Figure 6c. Town factors menu

per capita income, a correct prediction would be, "Demand for large cars will increase."

Subsequent to setting up an experiment, the subject engages in activities from the 'Things To Do' menu. All formal experiments are implemented from this main menu where ten options are provided, outlined below.

1. *See market sales information.* This window displays the current information on the state of the market (see Figure 2, 'Smithtown Market Window').
2. *Computer adjust price.* The computer will increase or decrease the price, whichever brings the current market closer to equilibrium. This occurs in successive approximations rather than changing the state immediately into equilibrium.
3. *Self adjust price.* This option provides the student with an online calculator and allows the price of the particular good to be changed within a prescribed range of values, specific, and realistic to each good.
4. *Make a notebook entry.* The student selects variables to record and the current values are automatically put into the notebook.
5. *Set up table.* The table package allows the student to select variables of interest from the notebook, put them together in a table, and sort on any selected variable, by ascending or descending order.
6. *Set up graph.* The graph utility allows a student to plot data collected from his/her explorations and experiments. This provides an alternative way of viewing relations between variables.
7. *Make a hypothesis.* The hypothesis menu allows students to make inductions or generalizations from relationships in the data they have collected and organized. There are four connected menus of words and phrases comprising the hypothesis menu. First, the "connector menu" includes the items: if, then, as, when, and, and the. The "object menu" contains the economic variables used by the system. The "verb menu" describes the types of change, like decreases, increases, shifts as a result of, and so on. Finally, the "direct object menu" allows for more precise specification of concepts such as: over time, along the demand curve, changes other than price, etc.

8–10. *Experimental frameworks.* Three "experimental frameworks" provide the student with easy maneuvering within and between experiments. These include: Change Good, Same Variable(s); Same Good, Change Variable(s); and Change Good, Change Variable(s). They are used to change to a new market while holding the independent variables the same, change town factor(s) while holding the market constant, or change town factor(s) and the market, respectively.

Three history windows are also included in the system, accessible to both students and system. Histories are maintained in the Student History window of each action taken as students continue to perform different explorations and experiments. In addition, the Market History window keeps a record of all variables and associated values from every experiment conducted. Finally, there is the Goal History window, providing a representation of economic concepts the student has successfully learned as well as those yet to be learned.

In order to optimally induce the lawful regularities in this environment, a model sequence of iterative behaviors in *Smithtown* would involve: exploring the world (informally), developing a plan for investigation (more formally), choosing online tools or techniques for executing the plan, collecting and recording data from the experiment, organizing the results, seeing if the data confirm or negate prior beliefs,

constructing a problem representation, modifying the problem based on discrepant results, refining the problem based on additional information, recognizing discrepancies between the result and expectations, testing findings in additional realms, and finally, generalizing a principle or law. However, people differ in their application of these skills. Individual differences will be discussed in the next section.

EXPERIMENT 1: EXPLORATORY STUDY OF INDIVIDUAL DIFFERENCES AND LEARNING

The two main research questions underlying this investigation were: (1) Did individuals interacting with *Smithtown* acquire as many economic concepts as students from a traditional classroom environment? and (2) In terms of specific 'learning indicators', what are the characteristics of those individuals who were more successful in learning in this type of environment as compared to those less successful? The data source for the first question corresponded to scores on a battery of pretests and posttests of economic knowledge developed by an economics expert working on the project. The data sources corresponding to the second question were detailed computer history lists of all student actions as well as verbal protocols from each student about justifications for each action. These data were used in computing values for the learning indicators for each person across three two-hour sessions with the tutor.

This exploratory study of individual differences and learning took a problem solving perspective. Sternberg (1981) made a distinction between two forms of metacognition in problem solving: global planning and local planning. Global planning refers to a strategy that applies to a set of problems and does not focus on the characteristics of a particular problem. Local planning refers to a strategy that is sufficient for solving a particular problem within a given set. Sternberg finds that better reasoners spend more time in global planning of a strategy for problem solution and relatively less time in local planning. Similarly, in studies of writing, Hayes and Flower (1986) point out that experts attend more to global problems than do novices. Novices focus on the conventions and rules of writing while experts make more changes that affect the meaning of the text. In physics, differences in problem solving between novices and experts also relate to surface and deep problem representations (Larkin, McDermott, Simon, & Simon, 1980; Simon & Simon, 1978). Experts work in a more top–down manner indicating that a general solution plan is in place before they begin the manipulation of specific equations while novices approach problems in a more bottom–up manner, manipulating equations to solve the unknown.

These findings indicate that individual differences in inductive problem solving can be defined in terms of the global and local aspects of performance, or attention to specific versus more general features of the problem-solving task. In a discovery learning environment, following findings by Klahr and Dunbar (1988), we translate this distinction to date-driven performance in contrast to behavior which is more hypothesis-driven. In our environment/task, an individual obtains data (either self- or computer-generated). On the basis of these data, the individual then induces generalizations or hypotheses which drive further data collection, data organization, and experimentation. Based on the literature cited above, we anticipated that good reasoners might display hypothesis-driven performance earlier in their discovery activity, and use their hypotheses as performance goals in contrast to more sustained but indiscriminate data collection. For example, a good subject may plan to test the effect of changing the population of *Smithtown* on the demand for donuts, hypothesizing that if population increased, the demand for donuts would increase. He or she would record baseline data on donuts,

make the desired change to population, record the new data, and c mpare these data. A less planful subject may similarly change the population and look at the data within the donut market, but without a higher level goal or hypothesis in mind. Data-driven induction is not completely unacceptable since individuals come to the task (*Smithtown*) with preconceived notions of regularities in the world of economics and they manipulate data and experiment on the basis of their a priori hypotheses. So, the discovery process that we study here does involve some combination of data-drive induction and hypothesis-generated data which guide performance.

METHOD

Subjects

Thirty undergraduate students enrolled at the University of Pittsburgh participated in this study. None had any formal economics training or previous economics courses. The age range was from 18 to 25 years and there was about an equal distribution of males and females in this sample. Subjects comprising the *Smithtown* and control groups were obtained from responses to campus advertising and paid for their participation. Subjects enrolled in an introductory economics class volunteered to participate. All subjects were debriefed about the purpose of the experiment at its conclusion.

Procedure

Three groups of subjects were used in the study: (a) Students who received classroom instruction on introductory economics, (b) A control group which received no economics instruction, and (c) Students interacting with *Smithtown*. There were ten subjects per group. All subjects took a pretest battery of economic concepts, received their respective interventions, and then took the posttest battery. The elapsed time between test batteries was about equal for all groups (i.e., about two weeks). The economics classroom group had two and one half weeks of instruction on the issues of supply and demand, the control group simply returned in two weeks for the posttests (no intervention) and the *Smithtown* group spent, individually, about five hours interacting with the system, broken down into three sessions across two weeks.

The chapter covered by the economics class during the treatment phase corresponded to the identical material/curriculum covered by the group working with *Smithtown* (i.e., the same introductory economic principles involving the laws of supply and demand in a competitive market).

Prior to their first real session with the system, the group using *Smithtown* were given a *Guide to Smithtown*. This three-page booklet informed them of their goal (i.e., to discover principles and laws of economics) and how to best achieve that goal (i.e., to imagine themselves as scientists, gathering data and forming and testing hypotheses about emerging economic principles and laws). The Guide overviewed some of the online tools available in *Smithtown* with examples provided on how to use them. Finally, the Guide emphasized that the individual would probably make errors or get stuck, but to try to learn from the mistakes. A glossary of terms (e.g., mouse, menu) concluded the Guide and the students were free to take it home with them between sessions. The Guide did not contain any information about economics principles.

The test battery used in this study was developed by an economics instructor at the University of Pittsburgh. The battery consisted of two tests, multiple choice, and short answer, and parallel forms were constructed for pretests and posttests. The tests were pilot tested to ensure clarity of instructions, proper timing, and the appropriate level of difficulty. After test development, the batteries were reviewed by an

Table 1. Percent Correct on Pretests and Posttests (Multiple Choice = MC; Short Answer = SA; Combined MC and SA = AVG)

	Control			Classroom			Smithtown		
	MC	SA	AVG	MC	SA	AVG	MC	SA	AVG
Pretest									
M =	46.0	54.0	50.0	46.8	50.0	48.4	48.0	47.0	47.5
SD =	11.5	19.2		9.2	12.8		11.5	13.1	
	Control			Classroom			Smithtown		
	MC	SA	AVG	MC	SA	AVG	MC	SA	AVG
Posttest									
M =	54.8	59.7	57.3	64.0	85.7	74.8	61.0	84.3	72.7
SD =	14.6	17.2		9.4	8.5		11.6	6.3	

independent economics instructor for content validity (i.e., completeness and accuracy).

RESULTS

Group Comparisons

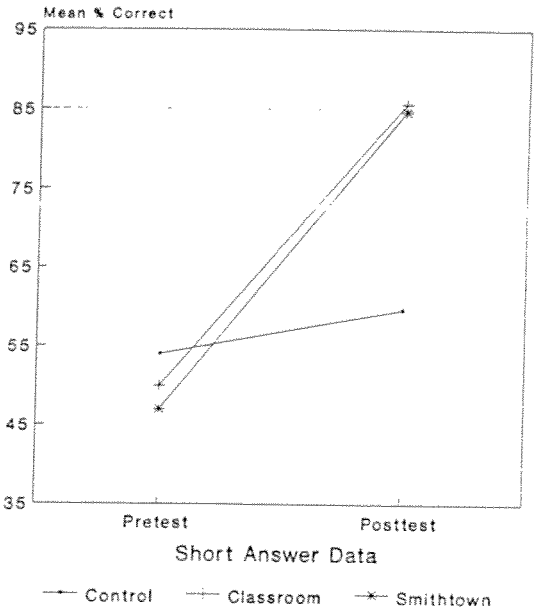
The group means by testing occasion (pretest, posttest) and test type (multiple choice, short answer) are presented in Table 1 and Figures 7a and 7b. First, note in Table 1 and Figure 7 that the three groups did not appear to differ on their pretest scores which assessed incoming economics knowledge (around 50% accuracy on both pretests). A post hoc MANOVA, computed on data from the two pretests, confirmed this observation: $F_{4,54} = 0.49$; $p = 0.74$.

The primary hypothesis of this study was that fostering the use of specific inquiry skills should facilitate the learning of specific domain knowledge; in this case, economics. This was tested by the interaction between testing occasion (i.e., pre vs. posttest for both the multiple choice and short answer tests) and the instructional treatment (i.e., control, classroom, and *Smithtown*). That is, did the three groups perform differently from pretest to posttest? This interaction was significant when a MANOVA was computed on these date ($F_{4,54} = 5.66$; $p < .001$).

Subsequent analyses were also conducted on planned comparisons between the different treatment groups. The first comparison between the experimental group (*Smithtown*) and the classroom group was not significant ($F_{2,26} = 0.36$; $p = .70$), implying that the two groups did not differ on relative pretest to posttest improvements. However, the comparison between treatment (classroom and *Smithtown*) and control was significant ($F_{2,26} = 16.86$; $p < .001$). Thus, the classroom and *Smithtown* groups showed equivalent improvements and greatly exceeded the performance of the control group.

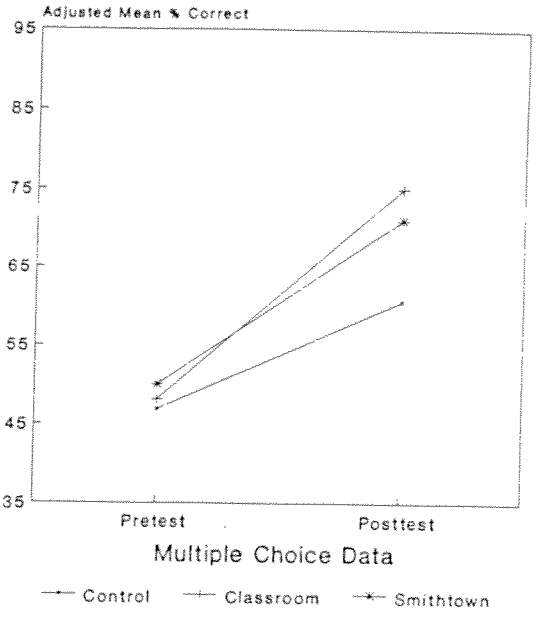
When the data were analyzed separately for each test type (multiple choice and short answer), the comparison of classroom and *Smithtown* groups versus the control group revealed no significant differences for the multiple choice test[1] ($F_{1,27} = 1.63$) but a significant difference for the short answer test ($F_{1,27} = 34.94$; $p < .001$). That is, the instructional treatments appar-

[1] In Figure 7b, the 'adjusted percent correct' was used for the multiple choice test data to adjust the mean score for guessing: Number right - (Number wrong/number of alternative choices - 1).



(A)

Figure 7a. Experiment 1: Pretest and posttest data from short answer test (by groups)



(B)

Figure 7b. Experiment 1: Pretest and posttest data from multiple choice test (by groups)

ently had their greatest effect on the cognitively complex task of recalling and articulating economic concepts (e.g., *List as many important factors as you can causing the demand curve for a good or service to shift to the left or right*) as opposed to the cognitively simpler task of simply choosing a correct response from alternatives.

Thus, the *Smithtown* group, with considerably less time on task, performed the same as students in the traditional classroom environment on tests of economic concepts. It is important to note that although the economics classroom group received almost twice as much instruction/exposure to the subject matter as did the *Smithtown* group (i.e., about 11 hours vs. 5 hours, respectively), the groups did not significantly differ on their posttest scores. Moreover, the *Smithtown* system did not tutor economic knowledge directly. Rather, the tutorial assistance was in terms of directing the subjects' scientific skills.

Analysis of Successful and Unsuccessful Learning Behaviors

Of particular interest to our research was the behavior within the *Smithtown* group. We wanted to know what the more successful subjects did differently than the less successful subjects in terms of specific learning behaviors. For example, maybe the successful subjects were simply more active in the environment, or recorded data into their notebooks more conscientiously, or perhaps the effective subjects generated more testable hypotheses compared to the less effective subjects.

We examined whether the low-level behavioral indicators relating to 'activity level' differentiated effective from ineffective learners, compared with the higher level indicators relating to 'data management' and 'thinking and planning' skills (see Appendix A for a listing of all learning indica-

tors). These learning process data (or learning indicators) thereby allowed us to capture learning in progress and then examine the precise behaviors that yielded more from less successful performance in this guided discovery environment. Our theoretical framework suggests a major role for the higher level indicators.

Table 2 presents the ten experimental subjects with their pretest and posttest scores. Of interest were those individuals scoring: (a) Low on the pretests but high on the posttests (more successful), and (b) Low on the pretests and low on the posttests (less successful). We are not interested in individuals who scored high on both the pre- and the posttests as that would imply some domain-related incoming knowledge. Only one individual (CR) fell into this category. As seen in Table 2, there are five subjects with large gain scores and four subjects with smaller gains on the economic test batteries. Having made this distinction between 'more' and 'less' successful learners in *Smithtown*, we can contrast these two groups by their performance indicator data.

Table 2. Smithtown Subjects' Scores on the Economic Tests (Combined Scores: Multiple Choice and Short Answer Tests)

Subjects	Pre	Post	Gain
Large Gain (above mean)			
CF	40.2	83.4	43.2
BW	53.7	89.9	36.2
JH	42.7	73.9	31.2
ML	54.2	84.4	30.2
CS	42.7	70.4	27.7
Small Gain (below mean)			
SS	53.3	75.4	22.1
JS	54.2	75.4	21.2
HT	43.1	56.8	13.7
OY	51.7	69.4	17.7
Constrained Gain (high on Pre and Posttests)			
CR	77.8	84.4	6.6
Overall *M*	51.0	76.0	25.0
SD	12.0	10.0	11.0

We collapsed data for each subject across their three sessions with *Smithtown* into a single index for each of 30 learning indicators. The indicators can be broken down into three rational categories: (1) General activity level indicators, (2) Data management skills, and (3) Thinking and planning behaviors. Each of these broad categories encompasses multiple individual indicators.

Two data sources were used to compute the performance indices: detailed computer history lists of all student actions, and verbal protocols from each student about justification for each action (i.e., what they expected to see after a particular action, and what their plans were for further experimentation). We then standardized each person's indicator scores.

As expected, more and less successful subjects differed mostly on performance measures relating to thinking and planning skills (i.e., the category representing the most complex learning indicators and reflecting effective experimental behaviors). There were fewer, but substantial differences on indicators from the data management skills category. Indicators from the activity level category did not discriminate between our two sets of subject. "Differences" in this context were defined as at least one standardized unit between the two groups per indicator. The particular indicators that best differentiated our subjects were as follows (ordered from most to least differentiating):

Generalization. The more effective subjects would test their developing economic beliefs in different markets to see if they were upheld, while the less effective subjects typically would not initiate experimentation across markets. These behaviors are represented by indicators 22 and 23 (in Appendix A) and involve both generalizing emerging principles to *related* markets

(e.g., investigating the effects of a manipulation on substitute or complementary goods) or testing beliefs out in *unrelated* markets to see the limits and extent of a particular concept. Since some of the town factors have global effects and some have only limited effects, it is good scientific practice to try out things in various markets. For instance, changing the prevailing interest rate in *Smithtown* would affect the demand for large cars but not for ice cream, while changing the population would impact the demand in relation to both markets.

Complexity of Experiments. Effective subjects also completed more actions within a given experiment and investigated fewer markets overall compared to the less effective subjects (indicators 29 and 4). These behaviors reflected the richness and tenacity of an individual's actions within an experiment. A thorough, systematic investigation of a concept was indicated by more connected actions (e.g., repeatedly changing the price of a good until the market reached equilibrium); aimless behavior was indicated by fewer connected actions. Furthermore, across the three sessions, the more successful subjects' number of actions per experiment increased showing that their experiments became more complex as they gained additional domain knowledge. This was not the case with the less successful group.

Systematic Variable Changes. Indicator 28 measured the number of variables manipulated per experiment. Given the freedom of the environment, it could be tempting to make changes to multiple variables concurrently; however, ensuing results would thereby be obscured as to what caused the state of market affairs. The biggest problem for the less successful subjects was that they persisted in changing multiple variables simultaneously. The more successful subjects changed fewer variables at a time, typically just single variables.

Adequate Data Collection. Another discriminating indicator from the thinking and planning category involved collecting sufficient amounts of data before making a generalized hypothesis regarding any of the economic concepts (indicator 24). Good scientific methodology involves generalizing a concept based on enough examples or instances of a phenomenon rather than on inadequate data which may include elements of chance or confounding variables. We set as our sufficiency criterion having at least three related rows of notebook entries before using the hypothesis menu. The more successful subjects did not attempt to make general hypotheses prior to collecting enough data on a given concept while the less successful subjects were content to make impulsive generalizations based on inadequate data.

Planning an Experiment. Higher-level planning behavior (indicator 20) was demonstrated by the more competent subjects. They tended to set up an experiment and execute it to completion. Actions corresponding to this indicator involved selecting variables from the planning menu, then utilizing those variables in subsequent controlled manipulations. This type of higher-level planning was rarely evidenced among the less successful subjects. This result is in accord with Sternberg's (1981) findings that persons scoring high on reasoning tests spent more time than low scoring persons on global planning, and less time on local planning. Similarly, Anderson (1987) investigated individual differences in students' solutions to Lisp programming problems and found that the poorer students tended to be less planful in their problem-solving activities compared with better subjects.

Predicting Experimental Outcomes. The proficient subjects in our study tended to make more outcome predictions for their experiments, while the less proficient subjects made considerably fewer predictions (indicator 16). To illustrate, for an experiment involving the increase of the price of

gasoline, a valid prediction could be rendered that 'Quantity demanded of gasoline will decrease'. The more effective subjects stated their predictions on more occasions than the less effective subjects. However, there was no way of determining whether the fewer predictions of the less effective subjects were due to knowledge deficiency or to a general lack of motivation.

Notebook Entries. In terms of data management skills, this study revealed that the successful subjects made more notebook entries, overall, compared with the less effective subjects (indicator 9). In addition, those entries tended to be more consistent with, and relevant to, the focus of their investigation (indicator 13). For example, notebook entries were typically made by the proficient subjects following any variable changes, and the entered variables usually had been selected beforehand in the planning menu as those of interest. In addition, proficient subjects collected baseline data into their notebooks (i.e., values of variables before being altered), sometimes the less proficient subjects generally failed to do, rendering later comparisons to changed data very difficult.

Relation of Successful Learning to Prior Scientific Training. In addition to comparisons between subjects based on their standardized indicator values, demographic information was obtained from all ten subjects. Two questions concerned previous scientific training: (a) what science courses the subject had taken since high school, and (b) what his or her major was. According to a hypothesis that different backgrounds caused the observed differences in scientific behaviors, we would have expected the less successful subjects to have taken fewer science courses. This was not the case. The less successful group had taken considerably more science courses since high school (total = 27) compared to the more successful group (total = 8). Moreover, the more and less successful groups had the same number of declared science majors (i.e., 3 per group). Thus, differential exposure to science training did not seem to determine who demonstrated better scientific behaviors.

EXPERIMENT 2: LARGE SCALE STUDY OF LEARNER DIFFERENCES IN SMITHTOWN

The previous study found some evidence for individual differences in learning and discovery strategies. We next addressed two main questions. First, what is the relationship between general intelligence and learning outcome (i.e., knowledge and skill acquisition from *Smithtown*), and second, do the findings from Experiment 1 generalize to a large sample from a different population? Experiment 1 tested the effectiveness of the system in comparison to economics learned in a traditional classroom environment and additionally found some areas differentiating more from less successful individuals in learning from the *Smithtown* system. In Experiment 2, we included a measure of general intelligence in our analyses. We were interested in seeing the nature and range of individual differences in learning. In particular, how much of these differences are attributable to general intelligence (or general ability)? Is it, simply, an individual's general intelligence that determines the nature and range of what they will learn, or is it something more, such as specific behaviors and strategies which are trainable (unlike general intelligence which is believed to be more fixed and inflexible)?

As part of the Learning Abilities Measurement Program (LAMP) at the Air Force Human Resources Laboratory, we tested a group of 530 subjects with a modified version of *Smithtown* which automatically tallied and summarized performance indicators at the end of a three and a half-hour session (instead of about five hours with *Smithtown* by subjects in the exploratory study).

Method

Subjects Subjects consisted of 530 enlisted Air Force recruits on their sixth day of basic training at Lackland Air Force Base, Texas. The gender distribution of subjects was approximately ¾ males and ¼ females. All subjects were between the ages of 17 and 27 years and had high school (or equivalent) educations.

Procedure Subjects were given a briefing prior to the tutor which informed them of their 'mission' (i.e., to manipulate the environment, acting as scientists, and to try to learn as many concepts as possible regarding basic laws of microeconomics). A short five minute game preceded *Smithtown*, designed to familiarize them with the mouse and menus. They next read an online Guide to *Smithtown*, saw a demonstration of a simple, online experiment, then entered the hypothetical marketplace on their own.

The number of concepts learned was our criterion measure (i.e., principles and laws correctly stated to the system via the hypothesis window). There were 12 concepts that could have been learned, and our subjects' criterion data ranged from 0 to 6. Since there was only 3.5 hours allotted for *Smithtown* interaction, and the first hour or so was typically spent familiarizing oneself with the environment, it was not surprising that the maximum number of concepts learned was only six.

A measure of general intelligence was available for each subject. The Armed Forces Qualification Test, or AFQT, is a composite score derived from the Armed Services Vocational Aptitude Battery consisting of the subtests: Arithmetic reasoning, word knowledge, paragraph comprehension and numerical operations.

RESULTS

Cluster Analysis on Performance Indicators

We computed a hierarchical cluster analysis based on the correlation matrix of the learning indicators to reduce the number of indicators, described in Experiment 1, to a more manageable set, and also to test alternative, objective groupings of the indicators (rather than the more subjective, rational categorization used in Experiment 1). This procedure was not employed in Experiment 1 because the small number of subjects would have made the results unreliable. We employed the ADDTREE/P clustering program (Corter, 1982) which implements Sattath and Tversky's (1977) Additive Similarity Tree model because ADDTREE/P has been shown to fit empirical data particularly well, especially discrete data such as our learning indicators. Applying the ADDTREE program to the 27 × 27 correlation matrix of indicators yielded a moderately good fit (r^2 = .71; stress = 0.06), but more importantly, yielded a highly interpretable solution. At the top level, the indicator data formed three main clusters: (a) Basic Activities; (b) Data Management; and (c) Scientific Behaviors. These three clusters further decomposed at the next level down. Basic Activities were subdivided into: (1) busy or gross-level activities, and (2) directed activities; Data Management subdivided into: (3) notebook usage, and (4) other tool applications; and Scientific Behaviors were subdivided into: (5) data-driven inquiry, (6) organizing experiments, and (7) hypothesis-driven inquiry behaviors. Note that this cluster analysis solution confirms the rational specification of learning behaviors posited in the design phase of the system (Shute, Glaser, & Raghavan, 1988). The cluster analysis solution is shown in Figure 8, and is characterized below by the contributing variables.

(1) Gross Activities: This cluster is defined by the number of variables changed at one time (VCPERTM), total number of independent variables changed (INDVAR), number of variables changed that were specified in the planning menu (PMVC), number of notebook entries of changed independent variables (ECIVAR), average number of variables changed per experiment (VARCHAN), and the average number of actions taken within a particular experiment (AVGACTS).

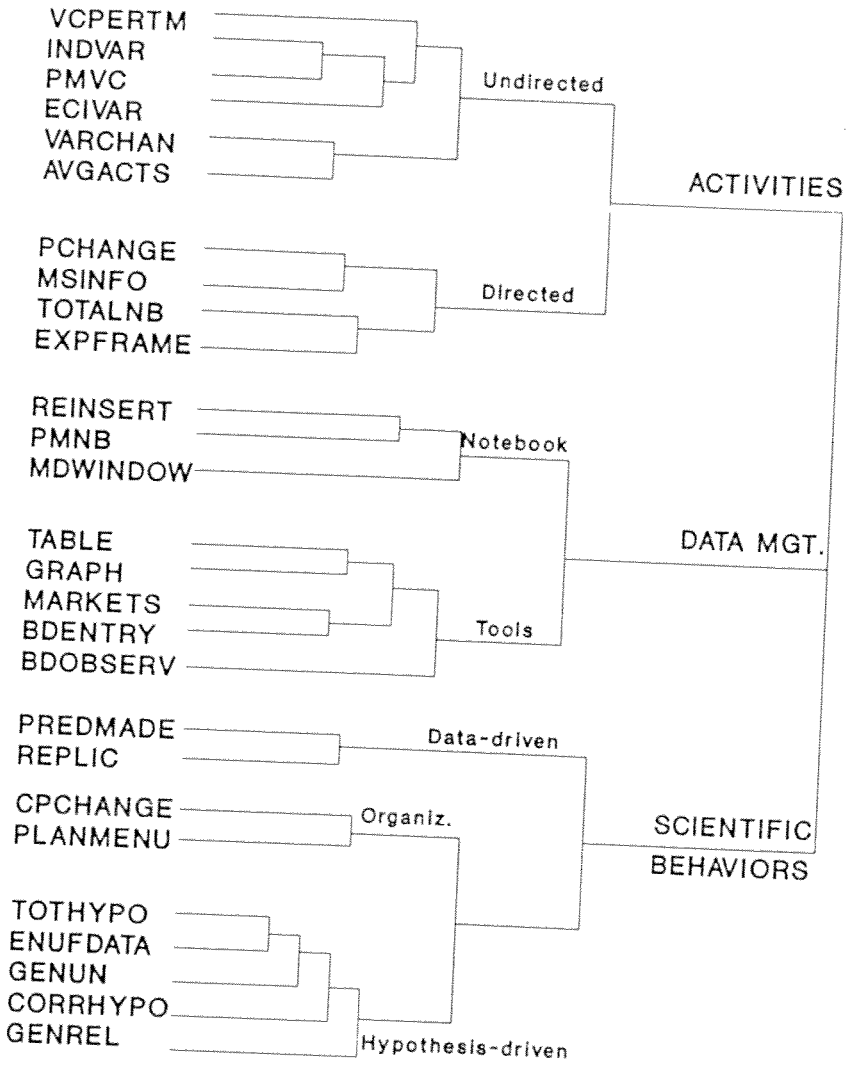


Figure 8. Hierarchical cluster analysis solution of learning indicators

(2) Directed Activities: This cluster is defined by the number of times price changes were made (PCHANGE), number of times the market sales information window was viewed (MSINFO), the total number of notebook entries made (TOTALNB), and the number of times the experimental frameworks were used to direct experiments (EXPFRAME).

(3) Notebook Usage: This cluster is defined by the number of times data from past experiments was inserted into the notebook (REINSERT), number of notebook entries of variables that had been specified in the planning menu (PMNB), and number of times the market data history window was viewed to see past variables and associated values (MDWINDOW).

(4) Tool Usage: This cluster is defined by the number of times the table package was applied (TABLE), number of times the graph package was used (GRAPH), number of markets investigated (MARKETS), number of times baseline data was entered into the notebook (BDENTRY), and number of times baseline data was observed (BDOBSERV).

(5) Data-driven Experiments: This cluster is defined by the number of specific pre-

dictions made of an experimental outcome (PREDMADE), and number of times an experiment was replicated (REPLIC).

(6) Organizing Experiments: This cluster is defined by the number of times the computer was requested to make price adjustments toward an equilibrium state (CPCHANGE), and number of times the planning menu was used to organize an experiment (PLANMENU).

(7) Hypothesis-Driven Experiments: This cluster is defined by the total number of hypotheses made (TOTHYPO), number of times sufficient data was recorded prior to rendering a hypothesis (ENUFDATA), number of times findings were generalized to unrelated markets (GENUN), ratio of the number of correct hypotheses made divided by the total number of hypotheses (CORRHYPO), and the number of times findings were generalized across related markets (GENREL).

Most of the clusters are readily interpretable. However, the distinction between cluster 5 (data-driven experimentation) and cluster 7 (hypothesis-driven experimentation) requires some elaboration. When a person conducts a **local** experiment (e.g., increases the number of compact car suppliers in *Smithtown*) and renders a specific prediction about the ramifications (e.g., the price of compact cars will go down), it is characterized as data-driven experimentation. This contrasts with a more **general** or hypothesis-driven experiment where an individual will attempt to generalize specific, local findings across different markets (e.g., investigating the relationship between price and quantity demanded in the gasoline, lumber, and ground beef markets), inducing a general principle operating in a competitive market.

Correlational and Regression Analyses

Seven composite scores, one for each major cluster category, were computed for each subject by summing standardized indicator scores within each cluster. The correlations of these variables and our criterion measure (number of concepts learned) can be seen in Table 3. From these data, it is apparent that the indicators relating to hypothesis-driven behaviors (i.e., the effective scientific inquiry skills) were the most highly correlated with successful learning. In addition, spending too much time managing the online notebook seemed to have a slightly negative effect on subsequent learning.

Regression analyses of these data were computed testing full and restricted models. First, all seven variables and two way interactions were tested (full model) predicting our criterion of number of concepts learned. This resulted in a multiple R = .70. Next, we computed a backward elimination of the interactions, and only three interactions remained in the equation (multiple R = .69). Finally, we computed a regression analysis with backward elimination of the main effects, and the results included the following main effects and interactions in the equation (multiple R = .69; $F_{7,522}$ = 66.81, $p < .001$): Gross level activities, Directed activities, Organization, Hypothesis-driven behaviors, Organization by Hypothesis-driven behaviors, Gross level activities by Hypothesis-driven behaviors, and Directed activities by Hypothesis-driven behaviors.

The three significant two-way interactions are characterized as follows. The interaction involving the variables: Organization and Hypothesis-driven behaviors (t =

Table 3. Correlation between Composite Indicators and Number of Concepts Learned

Composite Performance Indicator	Number of Concepts Learned
Gross Activities	−.08
Directed Activities	.05
Notebook Usage	−.12*
Tool Usage	−.08
Data-driven Behaviors	.03
Organization	.06
Hypothesis-driven Behaviors	.65**

N = 530; *$p < .01$; **$p < .001$

-4.3; $p < .001$) showed that if a person had a low value for hypothesis-driven behaviors, he or she would benefit (i.e., learn more concepts) from organizing and planning experiments. On the other hand, if a person had a high value for hypothesis-driven behaviors, less time spent planning and organizing, and more time spent actively and systematically conducting experiments was better as far as learning more concepts.The significant interaction involving Gross activities by Hypothesis-driven behaviors ($t = -4.9$; $p < .001$) showed a similar pattern where, for low values of hypothesis-driven experimentation, a person slightly benefitted from more activities in the environment, but for higher levels of hypothesis-driven behaviors, less busy (i.e., more focused) behaviors led to the acquisition of the subject matter. A different pattern is seen with the interaction of the variables: Directed activities and Hypothesis-driven behaviors ($t = 2.90$; $p < .01$). If a person did not act in a hypothesis-driven manner, engaging in more directed actions was not helpful in learning economic concepts. However, if a person was more hypothesis-driven, he or she would benefit from directed activities carried out in conjunction with scientific behaviors. Although these interactions are interesting, they only account for about 4% of the variance in our dependent measure while the majority of variance (42%) is explained by the single variable: Hypothesis-driven behaviors.

The simple correlation between our measure of general aptitude, the AFQT, and number of concepts learned ($r = .18$; $p < .01$) indicates that some amount of general intelligence is implicated in the learning outcome. However, when AFQT was included in a regression analysis involving the variables discussed above, the amount of *unique* variance accounted for by AFQT in predicting the number of concepts learned was less than 1%, compared to 38% of the unique variance attributable to hypothesis-driven behaviors (cluster 7). Thus, while general intelligence is certainly a component of learning, specific scientific behaviors account for considerably more variance in our criterion measure.

Another question we asked concerned the correlation between each of the composite variables and general intellectual ability. In other words, which behaviors did the subjects with higher AFQT scores engage in during *Smithtown* interactions? These correlations are shown in Table 4.

The pattern of correlations suggested that the high ability individuals engaged in directed, systematic activities, approaching the task in a manner concurrently bottom–up (data-driven) and top–down (hypothesis-driven). This was achieved by first conducting local experiments, then gradually expanding the scope of the findings across markets to test and refine developing hypotheses. It may have been that the subjects' high general ability enabled them to collect local data while having a goal state in mind.

Cluster Analysis on Cases

In contrast to a cluster analysis of variables, a cluster analysis on cases can detect consistent patterns or styles of interacting with *Smithtown*. A cluster analysis of cases addresses the different ways an individual may interact with *Smithtown*, and the effectiveness of these alternative approaches as far as ultimate knowledge acquisition. For example, someone may adopt the more obsessive approach of changing variables and conscientiously recording all values in the online notebook, regardless of relevance. This style contrasts with a more systematic approach of generating a hypothesis about some variable relationships, making the ap-

Table 4. Correlation between Composite Indicators and General Aptitude (AFQT)

Composite Performance Indicator	AFQT Score
Gross Activities	.06
Directed Activities	.27**
Notebook Usage	−.07
Tool Usage	−.10
Data-driven Behaviors	.13*
Organization	.07
Hypothesis-driven Behaviors	.24**

$N = 530$; $^*p < .01$; $^{**}p < .001$

Table 5. Cluster Solution of Composite Learning Behaviors

Cluster	1	2	3	4	5
Active	.50	−.35	−.31	1.15	.98
Data Mgt.	−.19	.33	−.23	1.85	−.17
Scientific	−.11	−.23	.32	−0.10	.99
N	170	183	153	11	13
N Concepts Learned	.21	.14	.78	0	1.38

propriate change(s), recording the data, and observing the results of the change.

A cluster analysis on cases (i.e., subjects as opposed to variables), allocates individual cases to clusters, classifying them based on the (squared) euclidean distances between cases and clusters. Each case is assigned to the cluster for which its distance to the classification center is smallest. We computed this analysis with respect to the three higher level composite variables: Activities, Data Management and Scientific Behaviors. The cluster analysis produced five distinct clusters of subjects, shown in Table 5.

We then compared the different clusters of subjects in terms of the criterion measure. An ANOVA was performed on the data with number of concepts learned[2] as the dependent variable, and the five cluster groups as the independent variables. Groups differed significantly: $F_{4,525} = 31.10$; $p < .001$. As seen in Table 5, the group learning the most concepts (i.e., cluster 5, $N = 13$), was characterized by having relatively high effective scientific behaviors and activities. The group learning the least concepts (i.e., cluster 4, $N = 11$) engaged in high activity and data management behaviors but fewer scientific behaviors.

[2] The mean number of concepts learned per group was fairly low. This may be due to several factors: the time on the system was very short (about 2 hours), the population was different (i.e., basic recruits as compared to university students), and the system did not always recognize some of the alternative representations of concepts thus did not tally the concept as being learned.

To test the hypothesis that engaging in only scientific behaviors is a sufficient condition for success in this type of environment, post hoc comparisons were computed testing the difference between cluster 3 (the group evidencing only scientific behaviors) and the other groups. In four contrasts (i.e., cluster 1 and 3, 2 and 3, 4 and 3, and 5 and 3), the subjects in cluster 3 learned significantly more concepts than all other groups, except those in cluster 5. Thus, the subjects learning the most from *Smithtown* were those that engaged in scientific behaviors and were, in general, active in their environment, albeit, in a directed manner. The less successful individuals in *Smithtown* (e.g., Cluster 4) spent most of their time managing data, busily occupied in a less directed manner, and not being very scientific during the learning process.

When the two groups learning most and least (Clusters 5 and 4, respectively) are compared on their profiles, we see that both groups have high loadings on the 'activities' variable. Since the three variables are orthogonal, an interpretation of this pattern is that when learners engage in any of the indicators tallied under the 'activities' variable, they will be successful in *Smithtown* only if they are goal or hypothesis-driven, conducting experiments that are systematically planned and executed. This stands in contrast to those being 'active' only with the goal state (local level) of having their data arranged neatly.

In summary, Experiment 2 found significant differences in knowledge outcome that were directly related to hypothesis-driven behaviors. When a measure of general intelligence was investigated in relation to the learning criterion, specific behaviors (i.e., those involved with goal or hypothesis-driven activities) were found to be much stronger predictors of successful learning in this type of environment than was the measure of general intelligence, which tends to be a more stable trait. These particular scientific behaviors are presumably trainable if they can be specified into rules, which is what we did in the 'inductive inquiry skills' knowledge base in *Smithtown*.

GENERAL DISCUSSION

In a computerized laboratory environment, students had the opportunity to engage in active, discovery learning of economic concepts by manipulating variables in a hypothetical town and seeing the repercussions. Overall, the system worked as we had hoped: Tutoring on the scientific inquiry skills resulted in learning the domain knowledge as a by-product, evidenced in Experiment 1 where performance on the posttest by *Smithtown* subjects was comparable to the performance by subjects from an introductory economics class.

In general, it appears that in the rather complex task involved in these two studies, many of the behaviors that differentiated successful and less successful subjects are similar to those identified in previous studies with both laboratory and more realistic tasks (e.g., Klahr & Dunbar, 1988; Shrager, 1985; Sternberg, 1985). Individual differences in performance from Experiment 1 were primarily a function of the hypothesis-driven behaviors applied by the subjects during *Smithtown* interaction. In particular, findings from Experiment 1 showed that the most effective learning behaviors were related to the category: Thinking and planning skills. Similarly, from Experiment 2, we showed a strong correlation ($r = .65$) between the composite variable: Hypothesis-driven experimentation and the dependent measure: Number of concepts learned. The cluster analysis conducted on cases confirmed this finding whereby the two groups of subjects who learned the most concepts (i.e., Clusters 3 and 5) were set apart from the other groups by virtue of their application of scientific behaviors. These subjects were interrogating the discovery world in a systematic manner, generating (top–down) and then testing (bottom–up) hypotheses about possible relationships among the economic variables. The less effective groups spent more time managing their data and doing other 'local' activities in the environment.

In summary, the successful individuals in both experiments employed more powerful heuristics compared to the less successful individuals. They manipulated fewer variables, holding variables constant while one variable was systematically explored. Less successful subjects did not seem to realize the power of the heuristic. Successful subjects took their time to generate sufficient evidence before coming to a conclusion while the less successful subjects were more impulsive and attempted to induce generalizations based on inadequate information. The more effective subjects tended to think in terms of generalizing their hypotheses and explorations beyond the specific experiment or market they were working on. They conceived of a lawful regularity as a general principle and as a description of a class of events rather than a local description. These subjects were also more sensitive to the existence of deeper explanatory principles in addition to local data descriptions; they appeared to realize that discovery was not only a function of data, but that they needed to generate some rule that could provide them with a goal for their actions. In this sense they tended to be more hypothesis-driven than the less successful subjects.

In regard to inductive problem solving, as Greeno and Simon (1988) state and as Klahr and Dunbar (1988) describe the interplay between rules and instances, the best learning strategy is a combination of bottom–up and top–down processing. In our subjects, this seemed to be the case: the better subjects would predict variable relationships and then test those hypotheses out, concurrently exploring and collecting data which led to further generalizations. Our less effective subjects seemed to be limited to a more data-driven (or bottom–up) approach, often falling short of grasping the larger picture. This is in accord with findings from investigations of novice–expert differences in problem solving (e.g., Larkin, McDermott, Simon, & Simon, 1980). Furthermore, the importance of higher level planning in this inductive discovery environment is in agreement with studies of individual differences in reasoning tasks (e.g., Sternberg, 1985). Suc-

cessful subjects consistently planned an experiment and then executed it to completion, according to plan, in sharp contrast to the more haphazard, less planful approach applied by less successful subjects in their experimental methodologies.

In Experiment 2, there was a significant correlation between the composite variable: hypothesis-driven behaviors and a general intelligence measure: AFQT score ($r = .28$; $p < .001$). This implies that the brighter individuals in our sample of 530 tended to be more systematic and controlled in their learning behaviors than those with lower AFQT scores. Furthermore, the correlation between AFQT score and our learning criterion was $r = .18$ implicating general intelligence in the final learning outcome. However, AFQT score only accounted for a small proportion (< 1%) of the learning outcome variance while the specific indicators, subsumed under the variable: hypothesis-driven behaviors, accounted for a much larger proportion of outcome variance (38%). The importance of these findings for instruction are that the particular scientific behaviors we have outlined (e.g., generalizing concepts across different markets, collecting sufficient instances of a phenomenon prior to stating a hypothesis, etc.) can be trained and hence, individuals can learn to be more methodical and scientific, thereby leading to the induction of general principles.

Learning from any complex environment is believed to represent a four-way interaction involving:

(a) the subject matter or curriculum,
(b) the instructional environment (e.g., discovery, didactic),
(c) the desired knowledge outcome (e.g., mental model, automatic skill), and
(d) learner style (e.g, passive vs. active, holistic vs. analytic processing) (see Kyllonen & Shute, 1989, for a complete discussion of this interaction).

In terms of these four dimensions, *Smithtown* may be characterized as follows:

(1) The subject mater is microeconomics as well as scientific inquiry skills.
(2) The instructional environment is a guided discovery environment where tutorial assistance is on the inquiry skills, not the economics knowledge.
(3) The desired knowledge outcome is a mental model of how the laws of supply and demand operate in a competitive market and also how to systematically conduct experiments to extract the various laws and relationships.
(4) Learner style was free to vary so that we could determine optimal and suboptimal behaviors in this environment.

For this type of environment, knowledge outcome, and subject matter, the most optimum learner behaviors we have found from the two experiments are systematic, hypothesis-driven activities. What about those subjects who are not characterized by these attributes? One way in which an ITS can increase its effectiveness is to adapt itself to an individual's strengths and weaknesses. In the case of *Smithtown*, this would take the form of providing more guidance for those less scientifically oriented on the particular skills determined to be important to learning from *Smithtown*. Since this system, as implemented in both experiments, was more discovery learning than guided[3], the more effective subjects were more self-directed and scientific. To optimize learning for all subjects, additional guidance, at least in the beginning sessions, is required for the less scientific persons. To make the program more flexible (i.e., to adapt its level of guidance based on

[3] For both experiments discussed in this paper, we set the threshold relating to the coach's intervention fairly high. That is, a subject needed to demonstrate 3 buggy behaviors or errors of omission before the coach would provide feedback. This threshold is modifiable and alternative environments may be created by manipulating the threshold value (e.g., turn it off completely for a discovery world, or set it to 1 to give immediate feedback).

subject behaviors) one additional rule could be incorporated into the 'teaching strategies' module. This rule could check the student model (i.e., the 'batting averages' per critic) for evidence of students' buggy or floundering behaviors, then intervene with immediate feedback until the behavior in question was no longer being demonstrated. Statistics are already maintained by the system (in the student model) on the frequency of unsystematic behaviors, thus the real-time adjustment of the current threshold of intervention would provide for more additional tutoring on those inquiry skills that were most difficult.

In conclusion, we have described two studies of individual differences in learning from an exploratory environment. Although in both studies the tutor only assisted on procedural problem areas (i.e., those related to various scientific inquiry behaviors), subjects did seem to extract domain knowledge during the course of their investigations and experimentations within *Smithtown*. Overall, the system worked as we had hoped: Tutoring on the scientific inquiry skills resulted in learning some principles and laws of microeconomics. Although we did not have information from the larger study about a subject's prior knowledge of economics to make a valid treatment effect statement, we did have that information with the smaller study (i.e., all subjects were selected on the basis of having no formal economics background). We have begun to delineate those skills and behaviors which are important to scientific discovery. The behaviors we have identified in this paper agree with the findings from related research (e.g., Klahr & Dunbar, 1988; Langley, Simon, Bradshaw, & Zytokow, 1987). In addition, these specific behaviors relate to individual differences found in general studies on problem solving, concept formation, and so on. From an instructional perspective, the behaviors we have denoted can consequently serve as a focal point for relevant intervention studies. From a design perspective, findings from these studies suggest modifications to intelligent tutoring systems, in general, so that they may be more like the individualized teaching systems they have the potential to be.

REFERENCES

Anderson, J.R. (1987, July). *Using intelligent tutoring systems to analyze data*. Paper presented at the Cognitive Science Society Conference, Seattle, WA.

Bonar, J.G.; Cunningham, R., & Schultz, J.N. (1987). *An object-oriented architecture for intelligent tutoring* systems. (Tech. Rep. No. LSP-3). Pittsburg, PA: Learning Research and Development Center. In Proceedings of the ACM Conference on Object Oriented Programming Systems, Languages, and Application.

Corter, J.E. (1982). ADDTREE/P: A PASCAL program for fitting Additive Trees based on Sattath & Tversky's ADDTREE Algorithm. *Behavioral Research Methods and Instrumentation, 14*, 353–354.

Greeno, J.G., & Simon, H.A. (in press). Problem solving and reasoning. In R.C. Atkinson, R. Herrnstein, G. Lindzey, & R.D. Luce (Eds.), *Stevens' handbook of experimental psychology* (Rev. ed.). New York: Wiley.

Hayes, J.R. & Flower, L.S. (1986). Writing research and the writer. *American Psychologist, 41*, 1106–1113.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1–48.

Kyllonen, P.C., & Shute, V.J. (1989). A taxonomy of learning skills. In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences*. San Francisco: Freeman.

Langley, P., Simon, H.A., Bradshaw, G.L., & Zytkow, J.M. (1987). *Scientific discovery: Computational explorations of the creative process*. Cambridge, MA: MIT Press.

Larkin, J., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Models of competence in solving physics problems. *Cognitive Science, 4*, 317–345.

Michalski, R.S. (1986). Understanding the nature of learning: Issues and research directions. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach, Volume II*. Los Altos, CA: Morgan Kaufmann.

Pellegrino, J.W., & Glaser, R. (1980). Components of inductive reasoning. In R. Snow, P. Federico & W. Montague (Eds.), *Aptitude, learning and instruction (Vol. I)* (pp. 177–218). Hillsdale, NJ: Erlbaum.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319–345.

Shrager, J.C. (1985). *Instructionless learning: Discovery of the mental model of a complex device*. Unpublished doctoral dissertation. Carnegie-Mellon University, Pittsburgh, PA.

Shute, V.J., & Glaser, R. (in press). An intelligent tutoring system for exploring principles of economics. In R. Snow & D. Wiley (Eds.), *Improving inquiry in*

social science: A volume in honor of Lee J. Cronbach. Hillsdale, NJ: Erlbaum.

Shute, V.J., Glaser, R., & Raghavan, K. (1989). Inference and discovery in an exploratory laboratory. In P.L. Ackerman, R.J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences*. San Francisco: Freeman.

Simon, D.P., & Simon, H.A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Erlbaum.

Smith, E.E., & Medin, D.L. (Eds.) (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. Cambridge, England: Cambridge University Press.

Sternberg, R.J. (1981). Intelligence and nonentrenchment. *Journal of Educational Psychology, 73*, 1–16.

White, B.Y., & Horowitz, P. (1987). *Thinker Tools: Enabling children to understand physical laws*. (Report No. 6470). Cambridge, MA: Bolt, Beranek, and Newman.

Yazdani, M. (1986). Intelligent tutoring systems survey. *Artificial Intelligence Review, 1*, 43–52.

APPENDIX A: ORIGINAL 30 LEARNING INDICATORS

I. General Activity Level

1. Total number of actions
2. Total number of experiments
3. Number of changes made to the price of the good
4. Number of markets investigated
5. Number of independent variables changed
6. Number of computer-adjusted prices
7. Number of times market sales information was viewed
8. Number of baseline data observations of market in equilibrium

II. Data Management Skills

9. Total number of notebook entries
10. Number of baseline data entries of market in equilibrium
11. Entry of changed independent variables
12. Number of reinsertions of changed independent variables to the online notebook
13. Number of "relevant" notebook entries divided by total number of notebook entries where "relevant" refers to those variables specified in the Planning Menu
14. Number of times the table package was used "correctly" divided by the total number of times the table was used, where "correctly" means less than 6 variables tabulated, and sorting was done on variables with differing values
15. Number of times the graph package was used "correctly" divided by the total number of times the graph was used, where "correctly" means plotting relevant variables, saving graphs, and superimposing graphs with a shared axis
16. Number of specific predictions made divided by the number of general hypotheses made
17. Number of correct hypotheses divided by the total number of hypotheses made

III. Thinking and Planning Skills

18. Number of notebook entries of Planning Menu items
19. Number of times notebook entries of Planning Menu items were made divided by the number of planning opportunities the subject had
20. Number of times variables were changed that had been specified beforehand in the Planning Menu.
21. Number of times an experiment was replicated
22. Number of times a concept was generalized across unrelated goods
23. Number of times a concept was generalized across related goods
24. Number of times the student had sufficient data for a generalization (i.e., at least 3 data points in the notebook before using the Hypothesis Menu)
25. Number of times a change to an independent variable was sufficiently large enough (i.e., greater than 10% of the possible range)
26. Number of times one of the experimental frames was selected (i.e., chose "same good, change variable," "change good, same variables" or "change good, change variable")

27. Number of times the Prediction Menu was used to specify a particular outcome to an event
28. Number of variables changed per experiment
29. Average number of actions per experiment
30. Number of economic concepts learned per session