8

Inference and Discovery in an Exploratory Laboratory

VALERIE J. SHUTE[1]
ROBERT GLASER
KALYANI RAGHAVAN *LDRC, University of Pittsburgh*

Formulating and testing hypotheses using observations and empirical findings is central not only to scientific work, but to the acquisition of knowledge in general. As individuals obtain new information they infer hypotheses which serve as a basis for confirming or refuting perceived regularities and lawful relationships. In the research described here we employ a computer laboratory, which we call an intelligent discovery world, to study the strategies students use to explore this environment. Our interest focuses on the study of individual differences in strategies of inference and discovery including comparative studies of successful and less successful learners. We are also investigating the impact of tutorial assistance on discovery skills.

Central to the process of induction and hypothesis formation is carrying out cognitive performances that ensure that inferences drawn are plausible and

[1]Currently at Air Force Human Resources Laboratory, Brooks Air Force Base, Texas.

relevant to the world or system being observed. One determines the plausibility of inductions and stated hypotheses by referring to knowledge obtained about the system. Thus the students' process of inference depends on the application of observation, experimentation, and data organization that enable the specification and testing of the knowledge obtained through experiments, hypotheses, and confirmations. As Holland et al. (1986) wrote, "The study of induction, then, is the study of how knowledge is modified through its use" (p. 5).

The kind of learning that we are considering has a long research history in experimental psychology, mostly in the context of laboratory and knowledge-lean tasks. In recent years, investigators have undertaken studies set in more complex situations, studies of machine learning, experimental studies, and computer simulation of problem solving and discovery tasks (Klahr and Dunbar, 1988; Kuhn and Phelps, 1982; Langley et al., 1987). Still, relatively few studies have investigated the domains taught in schools or other forms of formal education. Some exceptions are studies of microworlds in physics (Champagne and Klopfer, 1982; diSessa, 1982; White, 1984; White and Horowitz, 1987).

As we have indicated, the environment presents inductive problem-solving information, and the problem solver must attempt to find a general principle or structure that is consistent with this information. An important example is scientific induction. In medical and technical diagnosis a set of symptoms is presented from which one must induce the fault or cause. To paraphrase Greeno and Simon's (1984) description: Solving an induction problem can proceed in two ways, and in most tasks a combination of the methods is used. A top-down method involves generating hypotheses about the structure and evaluating them with information about the observed instances. A bottom-up method involves storing information about observations and events and making judgments about new events on the basis of similarity or analogy to the stored information. To perform the top-down method, the problem solver requires a procedure that generates or selects hypotheses, a procedure for evaluating hypotheses, and then a way of using the hypothesis generator to modify or replace hypotheses that are found to be incorrect. To use the bottom-up method, the problem solver needs a method of extrapolating from stored information, either by judging similarity of new stimuli to stimuli stored in memory or by forming analogical correspondences with stored information.

To a large extent, classic studies of induction have 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 and Glaser, 1980; Smith and Medin, 1981). Given our concern for exploratory environments, we perceive this large literature as pertaining, for the most part, to *passive* induction in which the learners induce rules, make hypotheses, and classify and taxonomize observations on the basis of sets of predetermined

instances designed by the experimenter. 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. In the latter form of induction, we need 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 (Bonar, Cunningham, and Schultz, 1986; Michalski, 1986; Yazdani, 1986).

In our research program we have been investigating the learning of topics in elementary physics, basic electronics, and economics. In this chapter, we report on the results of an empirical study conducted using a computer program called *Smithtown,* for learning economics. The environments that we design enable us to investigate a range of inductive or discovery learning, from learning in purely discovery environments to more guided discovery worlds. What we are learning from our work is that as students explore phenomena they can be guided and coached in the interrogation of a subject matter, analyzing their own understandings and misunderstandings, assessing progress toward their goals, and revising their problem-solving and learning strategies.

Our exploratory systems are designed to record, structure, and play back to students their own problem-solving processes. Such systems have been developed in algebra and geometry in which they provide a structured "trace" of problem solutions so that students can see the alternative paths that they have tried (Anderson et al. 1984; Brown, 1983). Previous papers report early work on the design and implementation of some intelligent discovery world environments (Reimann, 1986; Shute and Glaser, in press), and this paper describes an initial study of individual differences in exploration, data collection, and hypothesis formation in an exploratory world of microeconomic laws.

Smithtown is a computer program that provides a discovery environment for learning elementary microeconomics. An ideal sequence of iterative behaviors in *Smithtown* would include: exploring the world (informally), developing a plan for investigation (more formally), choosing on-line 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 out findings in additional realms, and, finally, generalizing a principle or law.

The focus of the study we are discussing is students' inductive inquiry skills, which in this context refers to the students' effectiveness in collecting, organizing, and understanding data, concepts, and relationships in a new domain. This system has been implemented on Xerox 1108/1186 LISP machines,

allowing self-paced, individualized, and interactive instruction in a rich data source (see Shute and Glaser, in press, for an overview of the system).

We hypothesize that discovery learning can contribute to a rich understanding of domain information by enabling the student to access and organize information. Furthermore, a proposition to be evaluated in this work is that effective interrogative skills are teachable if the particular skills involved can be articulated and practiced under circumstances which require them to be used.

Intelligent tutorial guidance, in conjunction with a discovery world environment, has the potential to transform a student's problem-solving performance into efficient learning procedures rooted in an individual's own actions and hypotheses. In such experiential learning, students are introduced to new subject matter and are given the opportunity to compare their observations with their current beliefs and theories, which they may reject, accept, modify, or replace (see Glaser, 1984). As they develop this knowledge, students ask questions, make predictions, make inferences, and generate hypotheses about why certain events occur with systematic regularity. Significant experience of this kind in discovering principles in a field of knowledge should affect the relationship learners perceive between themselves and the knowledge and their way of behaving when they forget a solution procedure or encounter an unprecedented problem (Cronbach, 1966).

KNOWLEDGE BASES IN *SMITHTOWN*

The primary purpose of the system is to help students become more methodical and scientific in learning a new domain. The first knowledge base—or "expert"—that we will discuss is concerned with efficacious inquiry skills.

The First Knowledge Base: Inductive Inquiry Skills

An earlier study conducted with *Smithtown* yielded information about more and less effective behaviors for interrogating and inducing information from a new domain (reported in Shute and Glaser, in press). This information was subsequently coded into rules that the system monitors in conjunction with a learner's actual behaviors. Thus, the system knows of sequences of good behaviors and also sequences of ineffective—"buggy"—behaviors.

The system leaves a student alone if she or he is performing adequately in the environment. However, if the system determines that a student is floundering or demonstrating buggy behaviors, the computer "coach" will intervene and offer assistance on the specific problematic behavior(s). For instance, if a student persists in changing many variables at one time without first collecting baseline data into the on-line notebook, the rule that would be invoked would look like the following.

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 does get fired and the number of times it has been invoked has surpassed some threshold value (for instance, four times), then the coach would appear and say, "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."

In addition to the rules monitored by the system, we developed a list of performance measures, called *learning indicators,* that enable us to determine what type of actions or behaviors yield better performance in this type of environment. We created a range of learning indicators, from low-level, simple counts of actions (for example, total number of activities taken within *Smithtown)* to higher-level, complex behaviors (for example, number of times a manipulation to an independent variable was made that showed an obvious change in the dependent variables). These indicators will be discussed in a later section and serve as one data source for our study on individual differences in learning in *Smithtown.*

The Second Knowledge Base: Economic Concepts in *Smithtown*

The second knowledge base, or expert, in the system knows about the functional relationships among economic variables which comprise valid economic concepts and laws. The system has a defined instructional domain which is decomposed into key concepts that are organized in a bottom-up manner (that is, from simpler to more complex ideas). An understanding of these concepts should result from the student's experiments in the microworld. The hierarchy of domain knowledge was developed by first reviewing six introductory microeconomics textbooks and determining the presentation order of information, and then discussing the optimal ordering of these concepts for student learning in the classroom with a college instructor of economics.

Although a student is not required to learn the concepts in any prescribed order, the hierarchy shown in Figure 8-1 provides the system with information about where the student is likely to be with regard to his or her knowledge acquisition. For example, the student can more readily understand the concept of equilibrium having first learned the laws of supply and demand. For the reader unfamiliar with this domain, we will now describe the basic concepts in microeconomics that can be learned using *Smithtown.*

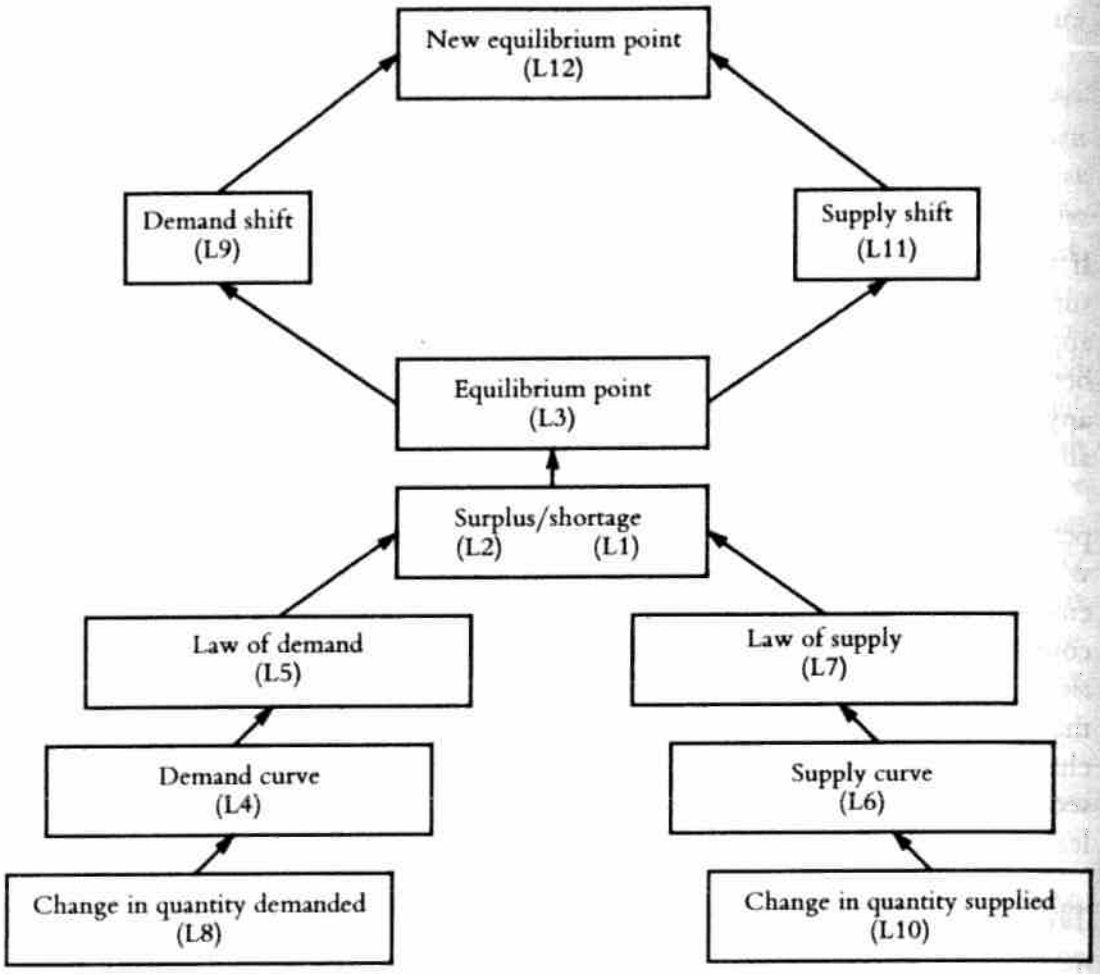


Figure 8-1 Hierarchy of supply and demand concepts.

Supply and Demand The buyer's side of the market is called *demand.* The law of demand states that the quantity of a product which consumers would be willing and able to purchase during some period of time is inversely related to the price of the product. If the price of gasoline goes up, consumers will demand a smaller quantity of gasoline; if the price goes down, consumers will demand larger quantities. If we graph this relationship, we get what is called a *demand curve* (see Figure 8-2) showing how the quantity demanded of a product will change as the price of that product changes, holding all other factors constant.

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 by a positive function. If the price of color televisions goes up, producers will tend to offer more television sets for sale. If the price of color television sets goes down, producers will reduce the number of television sets they put on the market. If we graph this relationship,

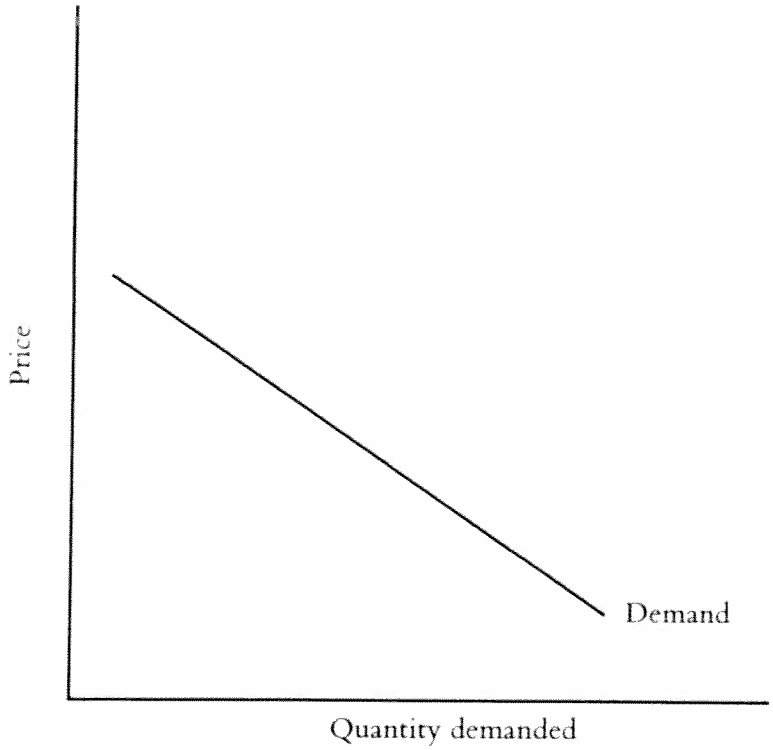


Figure 8-2 Graph of a typical demand curve.

we get what is called a *supply curve* (see Figure 8-3). A supply curve is a graph showing how the quantity supplied of some commodity will change as the price of that commodity changes, holding all other factors constant.

Equilibrium, Surplus, and Shortage There are many factors that influence the price of a given product, but when a price is reached at which the quantity that sellers want to sell is equal to the quantity that buyers want to buy, we say that the market is at a point of *equilibrium* (see Figure 8-4). Competitive markets always tend toward points of equilibrium. 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 convince sellers to lower their price down toward the equilibrium level. If, for some reason, 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.

Changes in Supply and Demand A change in the price of a good will influence the quantities demanded and supplied and cause movement along a fixed curve. A change to variables other than price will cause the entire curve (demand or supply) to shift depending on which variable is changed and the

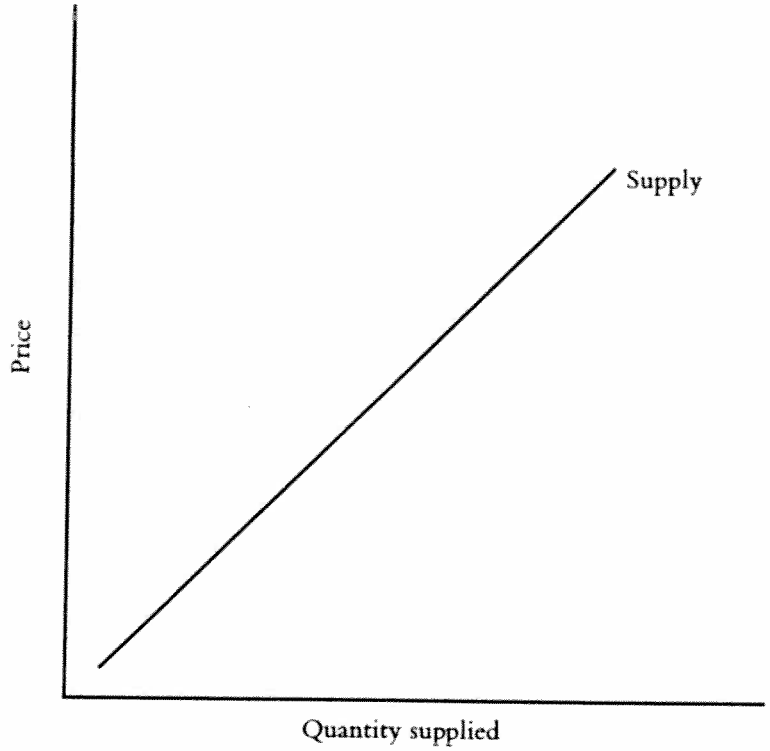


Figure 8-3 Graph of a typical supply curve.

magnitude of the adjustment. We refer to the variables in *Smithtown* that can be manipulated as *town factors;* they include per capita income, population, interest rates, weather, consumer preferences, labor costs, number of suppliers, and the price of substitute and complementary goods. For instance, if the population of *Smithtown* were increased from 10,000 to 25,000 persons, then the demand for automobiles would increase, resulting in a shift to the right of the demand curve for cars. Alternatively, if the number of suppliers of a particular good were to decrease, this would affect the supply curve for that commodity, resulting in a shift to the left. These shifts are depicted in Figures 8-5 and 8-6.

New Equilibrium Point Competitive markets tend to converge toward equilibrium points. Equilibrium, once established, can be disturbed by changes in demand or supply or both. If demand 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, and may be seen in Figure 8-7.

In addition to these economic concepts there are at least two more that can be extracted from the discovery world, although they are not explicitly recog-

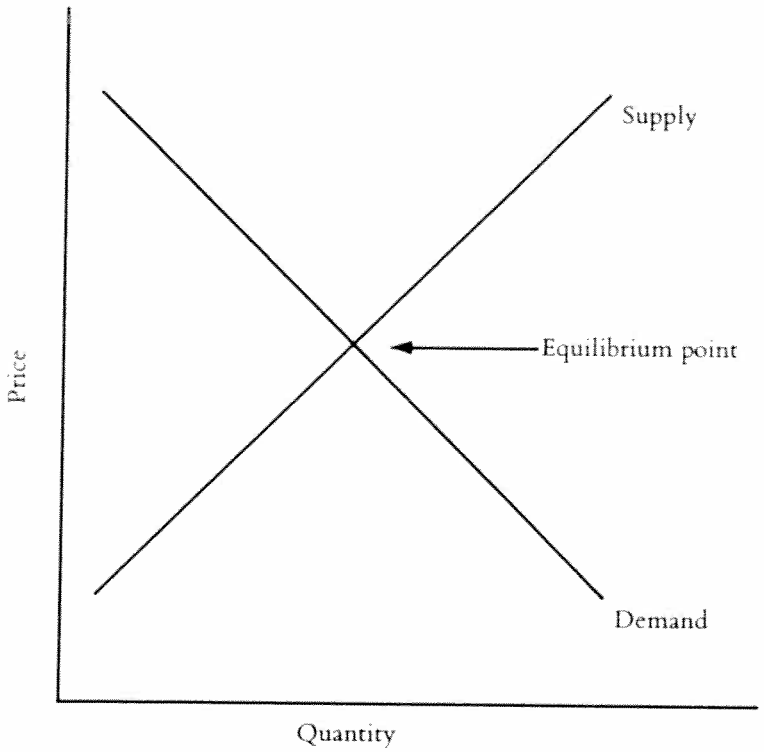


Figure 8-4 Graph of demand and supply curves intersecting to form an equilibrium point.

nized 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, whereas cross elasticity of supply indicates how a change in one market affects the supply in a related market.

MANEUVERING THROUGH *SMITHTOWN*

Students can discover regularities in the market by manipulating variables, observing effects, and using tools to organize the information in an effective way. The on-line tools for scientific investigations in *Smithtown* include a notebook for collecting data, a table to organize data from the notebook, a graph utility to plot data, and a hypothesis menu to formulate relationships among variables. Three history windows allow the students to see a chronological listing of actions, data, and concepts learned.

First, a student selects a market to investigate from the "Goods Menu" and informs the system of his or her experimental intentions by choosing variables she or he is interested in from the "Planning Menu." For each new experiment, the system asks the student if he or she would like to make a prediction

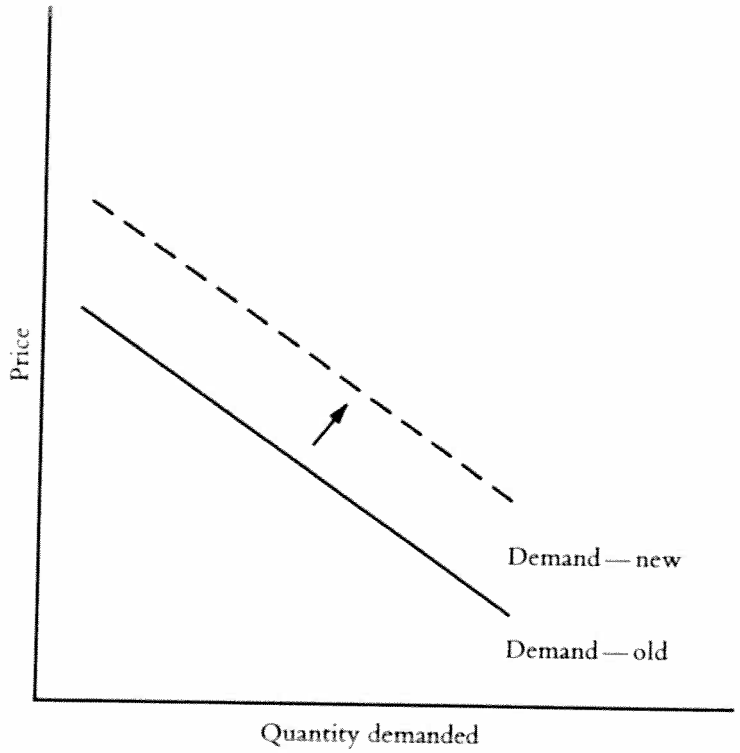


Figure 8-5 Graph of a demand curve shift (an increase).

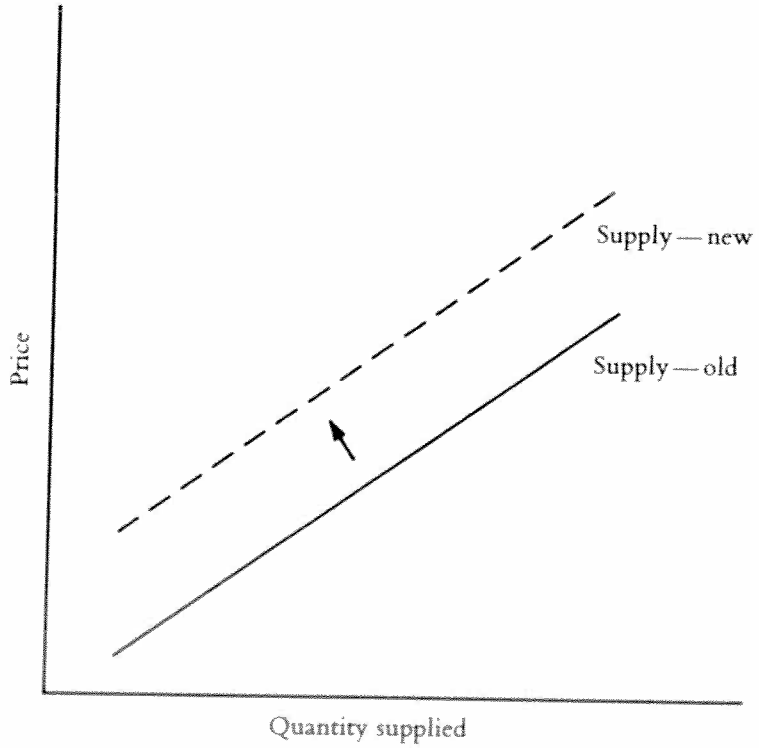


Figure 8-6 Graph of a supply curve shift (a decrease).

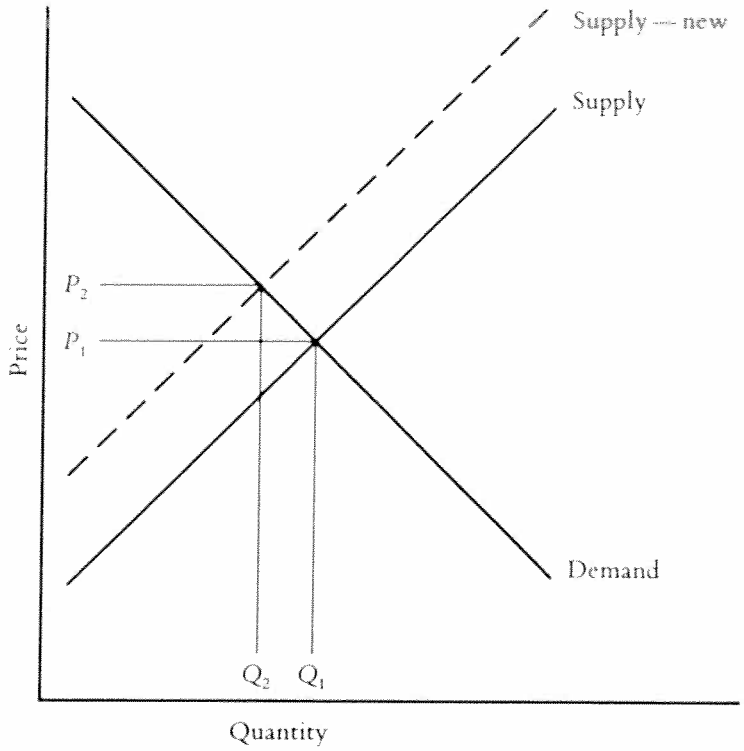


Figure 8-7 Graph of a supply curve shift causing a new equilibrium price.

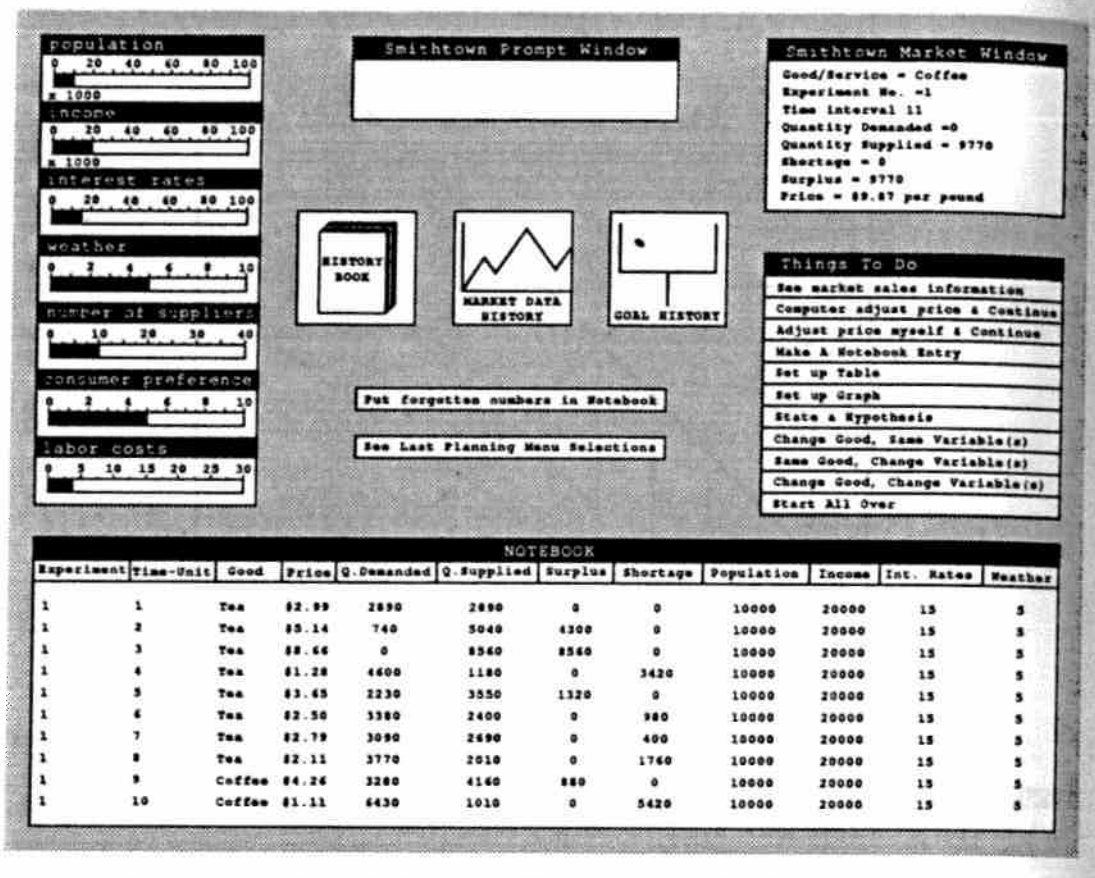
regarding the planned experiment. If the student chooses "No," the next menu of options is the "Things To Do Menu." If the student responds "Yes," a window appears where specific statements can be entered about predicted outcomes to a planned manipulation. For example, if the student's experiment was to increase the price of gasoline in order to see the repercussions in the market place, one prediction might be, *"The quantity demanded [of gasoline] will decrease."* Explorations and experiments are directed from the "Things To Do Menu" which provides ten options. Seven of the ten options are listed here. There are also three *experimental frameworks,* which are options 8 through 10.

1. *See market sales information.* This window displays information on the current state of the market.

2. *Computer adjust price.* The computer will increase or decrease the price, whichever brings the current market closer to equilibrium.

3. *Self adjust price.* This option provides the student with an on-line calculator and allows the price of the particular good to be changed.

4. *Make a notebook entry.* The student selects variables to record, and the current values are automatically put into the notebook (see Figure 8-8).



Experiment	Time-Unit	Good	Price	Q.Demanded	Q.Supplied	Surplus	Shortage	Population	Income	Int. Rates	Weather
1	1	Tea	$2.99	2890	2890	0	0	10000	20000	15	5
1	2	Tea	$5.14	740	5040	4300	0	10000	20000	15	5
1	3	Tea	$8.66	0	8560	8560	0	10000	20000	15	5
1	4	Tea	$1.28	4600	1180	0	3420	10000	20000	15	5
1	5	Tea	$3.65	2230	3550	1320	0	10000	20000	15	5
1	6	Tea	$2.50	3380	2400	0	980	10000	20000	15	5
1	7	Tea	$2.79	3090	2690	0	400	10000	20000	15	5
1	8	Tea	$2.11	3770	2010	0	1760	10000	20000	15	5
1	9	Coffee	$4.26	3280	4160	880	0	10000	20000	15	5
1	10	Coffee	$1.11	6430	1010	0	5420	10000	20000	15	5

Figure 8-8 Screen display of *Smithtown* with notebook entries made.

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 (see Figure 8-9).

6. *Set up graph.* The graph utility allows a student to plot data collected from his or her explorations and experiments. This provides an alternative way of viewing relations between variables (see Figure 8-10).

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 actually four interconnected menus of words and phrases comprising the hypothesis menu (see Figure 8-11). First, the "Connector Menu" includes the items "if," "then," "as," "when," "and," and "the." Next, the "Object Menu" contains the economic indicator 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." As students combine words or

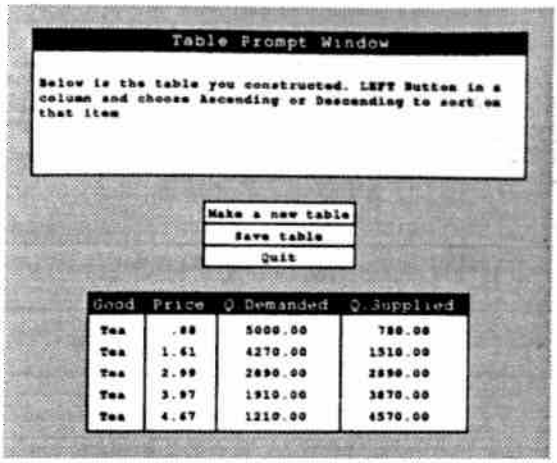


Figure 8-9 Screen display of the table package with four variables represented (ordered on price in ascending order).

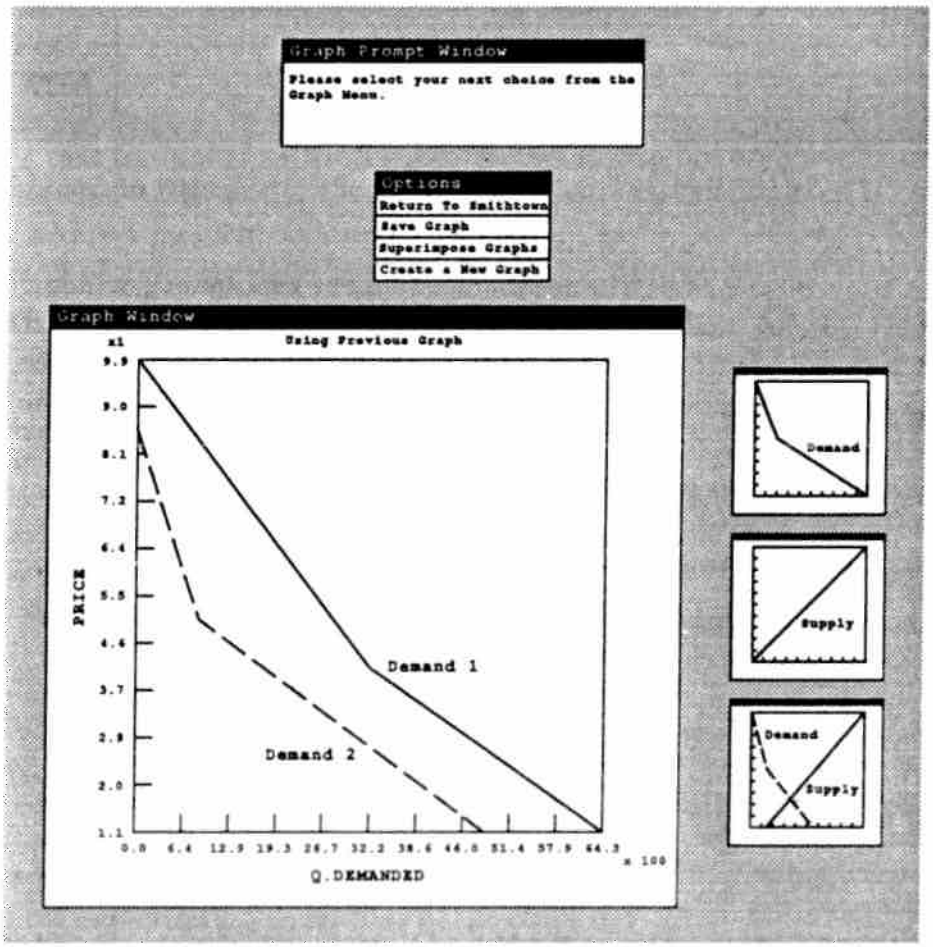


Figure 8-10 Screen display of the graph package with supply and demand curves superimposed.

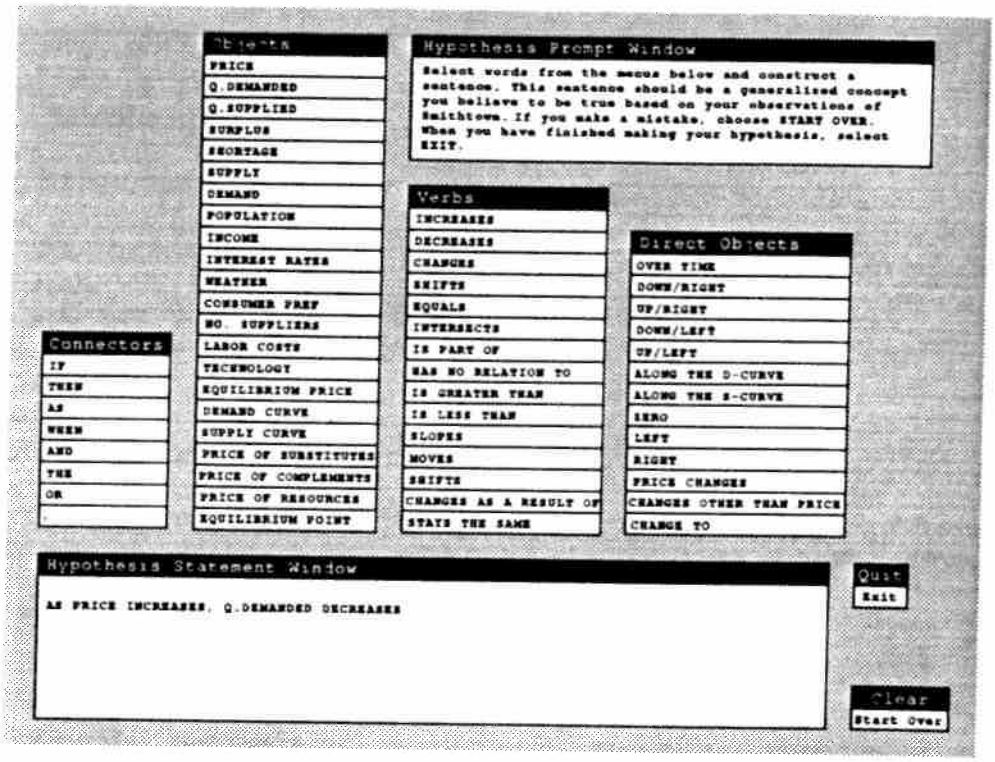


Figure 8-11 Screen display of the hypothesis menu with the law of demand stated.

phrases from these menus, the resultant statement appears in a window below. A pattern matcher analyzes key words from the input and checks whether this matches stored relationships for each targeted concept. For instance, if the student stated "As price increases, quantity demanded decreases," the system would match that to the law of demand which it understands to be the inverse relationship between price and quantity demanded.

Experimental Frameworks

The three experimental frameworks, options 8, 9, and 10, 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) within the current market, or to change the town factor(s) and the market.

History Windows

Three history windows are included in the system which are accessible by both the students and the system. As students continue to interact with *Smithtown,* histories accumulate that delineate the various actions resulting from different

explorations and experiments. This summary is maintained in the "Student History Window." The "Market History Window" keeps a record of all variables and associated values that the student has manipulated. Finally, there is the "Goal History Window." This provides a representation of what the student has successfully learned in terms of concepts targeted by the system.

LEARNING AND INDIVIDUAL DIFFERENCES

In this section we describe an exploratory study of learning and individual differences in performance in this intelligent discovery world environment. The system was able to categorize sequences of student actions as being more or less effective and intervened with a hint at times when the student was floundering.

This study was undertaken with two main goals in mind. One goal was to evaluate *Smithtown* to see if individuals interacting with it actually acquired any of the economic concepts embedded in the environment, such as the law of demand, equilibrium point, and so on. The second goal was to determine the performance characteristics of those individuals who were more successful in learning in this type of environment as compared to those less successful. Another implicit goal was to examine the computer architecture and interface features that facilitated or overly constrained an exploratory environment.

The kind of inference-discovery task that we are studying has been interpreted within a problem-solving framework by Klahr and Dunbar (1988) who conceived of the interplay between hypothesis formation and experimental design phases of the discovery process as a search between two problem spaces —a hypothesis space of rules and an experimental space of instances.

> This means that, first, we need to account for the identification of relevant attributes, for, unlike the conventional concept-formation studies, our situation does not present the subject with a highly constrained attribute space for hypotheses. Second, we need a more complex treatment of the instance generator, because in our context it consists of an experiment, its predicted outcome, and the observation of the actual outcome (p. 7).

Klahr and Dunbar placed their subjects in a discovery context by first teaching them how to use an electronic device — a computer-controlled robot tank called BigTrak — and then asked them to discover how a particular function works. They formulated a general model of scientific discovery as dual search that shows how search in the two problem spaces shapes hypothesis generation, experimental design, and the evaluation of hypotheses. Strategy differences among subjects were a consequence of the efficiency of search in the hypothesis space. Successful subjects were classified as theorists and those who abandoned hypothesis testing in order to search the experiment space were classified as experimenters.

We also take a problem-solving perspective in our investigation and are guided in our search for individual differences by certain general findings in problem-solving performance. For example, Sternberg (1981) makes a distinction between two forms of metacognitive performance: global planning and local planning. Global planning refers to a strategy that applies to a set of problems instead of focusing on the characteristics of a particular problem; global planning refers attention to the context or overall characteristics of the group of problems. Local planning refers to a strategy that is sufficient for solving a particular problem within a given set; local planning is less sensitive to general context and focuses more on the difficulty of carrying out the specific operations of a problem-solving task. Sternberg finds that better reasoners spend relatively more time in global planning of a strategy for problem solution and relatively less time in local planning. Such a distinction is also evident in studies of expert versus novice problem solving. In studies of writing, Hayes and Flower (1986) point out that experts attend more to global problems than do novices. Experts and novices attend to different aspects of a text. Novices focus on the conventions and rules of writing; experts make more changes that affect the text's meaning. The perceptions of the novices are more local or shallow whereas those of the expert are more global and have more overall meaning. The strategies used by novices are local strategies concerned with the deletion and addition of words and phrases whereas experienced writers are concerned more with strategies that involve changes in content and structure. In physics (Larkin et al., 1980; Simon and Simon, 1978), differences in problem solving between novices and experts also relate to surface and deep problem representations. The novice's representation of a problem results in a local form of problem solving in which they work with equations to solve the unknown. Experts, in contrast, work in a more top-down manner indicating that a general solution plan is in place before they begin the manipulation of specific equations.

The findings just outlined direct our attention to conceivable differences between good and poor inductive problem solvers in terms of the global and local aspects of their performance or their attention to specific versus more general features of the problem-solving task. In a discovery situation, taking a lead from Klahr and Dunbar, we translate this distinction to data-driven performance in contrast to behavior which is more rule or hypothesis driven. In our task, an individual starts out with attention to computer-generated observations or to subject-designed experiments. On the basis of these data, he or she induces generalizations or hypotheses which drive further data collection, data organization, and experimentation. Based on the studies on problem solving we have described, we can anticipate that good reasoners might display rule-driven performance earlier in their discovery activity and use rules as a performance

goal rather than more sustained attention to data collection, although the latter is necessary at certain points in the course of discovery.

Furthermore, in addition to behaviors at a general level, we must also look at more direct performance components. We refer to specific performance heuristics manifested by good reasoners that may not be available to others. A good example in discovery performance is the heuristic of identifying one variable as a dimension of examination and holding all other variables constant while the chosen one is varied systematically. Lawler (1982), in discussing computer-based microworlds that use LOGO language, refers to this as variable stepping. He points out that Piaget judged variable stepping to be an essential component of formal operational thought—a powerful idea because it is universally useful and crucial to the process of scientific investigation. In this regard we look for individual differences in our discovery worlds that relate to such performance procedures.

As a general caveat in the work reported here, we must point out that scientific discovery involves a whole array of processes including observing and gathering data, finding regularities that describe the data, formulating and testing the generalizability and limitations of these regularities, and formulating and testing explanatory theories. In this study we are primarily concerned with a subset of these processes, principally with discovery that starts with a data set that can be investigated and that derives descriptive rules, laws, or regularities from them. As has been pointed out (Bradshaw, Langley, and Simon, 1983), "the generation of data, and even the invention of instruments to produce new kinds of data, are also important aspects of scientific discovery. And in many cases, existing theory, as well as data, steer the course of discovery" (p. 971). We consider in this chapter the path from data to descriptive laws about data (not necessarily explanatory theories). This subset of scientific work is important in discovery and in our concern with individual differences in induction from data and the process by which inductive discovery is carried out. Also to be kept in mind is the fact that data-driven induction is not completely "pure." Individuals come with previous conceptions of regularities in the data, and they manipulate data and experiment on the basis of hypotheses they generate. So the discovery process that we study here will involve some combination of data-driven induction and hypotheses-generated data which guide performance.

Subjects

Three groups of subjects were involved in the experiment: (1) students who received traditional classroom instruction in introductory economics, (2) a control group which received no economics instruction, and (3) students interacting with *Smithtown*. There were ten subjects in each group. All subjects were

from the University of Pittsburgh, and none had any formal economics training or previous economics courses. The *economics group* were students who volunteered to participate in an experiment and who were enrolled in an introductory microeconomics course. About half of the *control group* consisted of psychology students who took the tests for class credit; the other half consisted of students selected from those who responded to ads placed around the campus for subjects who had no economics background. They took the tests and received payment for their time. The *experimental group* were individuals who similarly responded to ads placed around the University of Pittsburgh campus. They were paid for their participation. It should be noted that the chapters covered by the economics class during the testing interval corresponded to the identical material covered by *Smithtown* (that is, the same introductory economic principles involving the laws of supply and demand in a competitive market). All subjects were debriefed about the purpose of the experiment at its conclusion.

Test Materials

The test battery on microeconomics was developed by an economics instructor at the University of Pittsburgh. The tests were initially piloted by individuals who provided feedback about the tests as to the clarity of instructions, the timing of the tests, and the general level of difficulty. The battery consisted of two tests, multiple choice and short answer. After test development, the batteries were reviewed by an independent economics instructor for content validity —that is, completeness and accuracy.

Multiple-Choice Test Two alternate forms were created for the pretest and posttest. This involved knowledge of various concepts and principles of microeconomics. Subjects were to circle the best answer from the four alternatives given. An example of a pretest item from the test is:

> The supply curve of houses would probably shift to the left (decrease) if:
>
> a) construction workers' wages increased
> b) cheaper methods of prefabrication were developed
> c) the demand for houses showed a marked increase
> d) the population increased

A corresponding posttest item was constructed for each of the pretest items. The counterpart to the above question is:

> Which of the following is likely to move a supply curve for beef to the right (an increase)?
>
> a) a rise in the price of beef

b) a decrease in the price of cattle feed
c) an increase in the wages of farm laborers
d) a decrease in the price of raw hides

Short-answer Test This test involved the same concepts to be defined by the subject for both the pretests and the posttests. Elaborated knowledge was required to define different concepts, come up with instances of a given concept, or draw a curve on a labeled but empty grid. Two examples from the short answer test include:

1. What is market equilibrium?
2. List as many important factors as you can causing the demand curve for a good or service to shift over to the left or right.

Each answer on the short answer test was scored with reference to a list of necessary and sufficient elements.

Procedures

Subjects from the economics group were administered a pretest battery in their class prior to the lectures and readings on the laws of supply and demand. They received about two and one-half weeks of instruction on this part of the curriculum after which they were retested in the classroom with the posttest battery.

The control group completed the pretest battery and then returned in about two weeks for the posttests. This interval corresponded to the pretest and posttest intervals for the other two groups.

The experimental group took the pretest battery individually then signed up for three additional 2-hour sessions. This translated to a total of 5 hours on the computer (session 1 = pretest battery plus demonstration of the system; session 2 = 2 hours on the computer; session 3 = 2 hours on the computer; and session 4 = 1 hour on the computer and 1 hour for the posttest battery). The sessions were spread out over a two-week period to correspond to the same time frame as the economics group and the control group. Prior to the first real learning session with the system, students were given a *Guide to Smithtown* in session 1. This guide informed them of their goal—to discover principles and laws of economics—and told them that the best way to achieve that goal is to imagine themselves as scientists, gathering data and forming and testing hypotheses about emerging economic principles and laws. The guide overviewed some of the on-line 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 should try to learn from the mistakes. A

glossary of terms concluded the guide, and the students were free to take it home with them between sessions.

RESULTS

The first question addressed whether the three groups were *initially comparable* on their scores on the pretest battery of multiple-choice and short-answer problems. Table 8-1 shows the summary statistics for the raw data; the mean percentage scores for the pretest battery and for the posttest battery, collapsed across multiple choice and short answer, are plotted in Figure 8-12.

As shown in Table 8-1 and Figure 8-12, the three groups *are* initially comparable whereas on the posttest both the economics group and the experimental group surpass the control group. First we computed an ANOVA, a repeated measures design where the grouping factor was *treatment group* and the trial factors were *test type* and *pretest* versus *posttest* condition. The most important interaction that we were interested in was pretests and posttests by treatment group, collapsed across tests, $F_{2,27} = 2.99$, $p = .067$. This shows that the three groups did differ in terms of their pretest to posttest changes in scores. We then computed a Hotelling T^2 test, contrasting all three pairwise combinations of groups on the pretest battery, yielding the following nonsignificant T^2 values:

Economics by control group	$T^2 = .03$, $p = .77$
Control by experimental group	$T^2 = .11$, $p = .42$
Economics by experimental group	$T^3 = .03$, $p = .80$

Table 8-1 Summary Statistics: Means and Standard Deviations

	Control group		Economics group		Experimental group	
	MC	SA	MC	SA	MC	SA
Pretest						
M=	11.50	16.20	11.70	15.00	12.00	14.10
SD=	2.88	5.77	2.31	3.83	2.87	3.93
Posttest						
M=	13.70	17.90	16.00	25.70	15.20	25.30
SD=	3.65	5.17	2.36	2.54	2.90	1.89

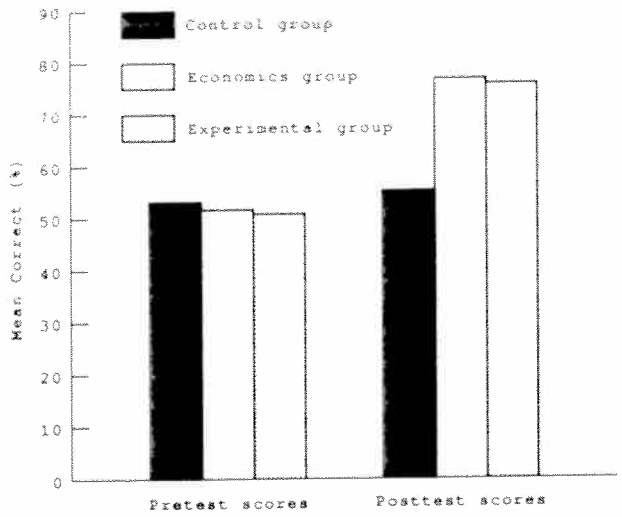


Figure 8-12 Pretest and posttest scores by treatment group (collapsed multiple choice and short answer test data).

After their respective interventions the groups differed; however the economics group and the experimental group ended up with equivalent posttest scores. It is important to note that students in the experimental group spent only five hours interacting with the discovery world compared to 2.5 weeks (or about 11 hours) of classroom lectures and recitation covering identical curricular information.

Hotelling's T^2 analysis allows us to see particular differences between independent groups on their test scores. The mean vectors for each group can be extracted from the summary statistics. First, we compare the posttest scores of economics students and the control group: $T^2 = 1.02$, $p = .003$. As expected, these two groups differed overall in their test scores. Individual t-tests on the data showed that the difference is primarily associated with the responses on the short answer posttest. The economics students had much more complete and articulate responses than the control group ($t = 4.28$, $p = .0005$). Second, the results from this analysis revealed that the economics group and the experimental group performed the same not only on their pretest scores but on their posttest scores as well: $T^2 = .031$, $p = .774$. The experimental group, with significantly less time on task, performed comparably with the students in the traditional classroom environment. No differences were found between any of the individual tests. Third, we compared the control and the experimental groups. We expected that there would be a difference between these two groups in their test composites given the experimental group's interaction with the system. This comparison also showed a significant difference between the posttests: $T^2 = 1.24$, $p = .001$. Individual t-tests were generated for each of the tests, and the short-answer posttest, again, was the major reason for the differ-

ences ($t = 4.25$, $p = .0005$). The experimental group had much more complete responses than the control group.

Individual Differences in the Experimental Group

The results from the between-group analyses suggest that, overall, *Smithtown was* effective in teaching a targeted set of microeconomic concepts comparable to a traditional classroom environment. We now further examined the experimental group data to see how differential interaction with this exploratory world affected subsequent learning. In other words, some individuals learned more than others from the system, and we wanted to know what it was that the more successful individuals did compared to the less successful persons in extracting and understanding new knowledge. "Successful," in this context, refers to someone who started out with a low pretest score on the battery of economics tests and, after interacting with the system, ended up with a high posttest score. Thus the two interesting comparisons are between those scoring (1) low on the pretest and low on the posttest, and (2) low on the pretest but high on the posttest. We were not interested in those who scored high on both the pretest and the posttests as they seemed to have started out with some domain-related knowledge. Table 8-2 lists each of the ten experimental subjects with their associated pretest and posttest scores (percent correct).

We are interested in comparing individuals who scored above the mean gain score with those below it. Thus there is a pool of five subjects having large

Table 8-2 Subjects' Scores on the Economic Tests (percent correct)

Subject	Pretest	Posttest	Gain
BW	53.7	89.9	36.2
JS	54.2	75.4	21.2
SS	53.3	75.4	22.1
ML	54.2	84.4	30.2
HT	43.1	56.8	13.7
CR	77.8	84.4	6.6
JH	42.7	73.9	31.2
CF	40.2	83.4	43.2
OY	51.7	69.4	17.7
CS	42.7	70.4	27.7
Mean	51	76	25
Standard deviation	12	10	11

gains and five subjects with small gains. These subjects will be discussed after the presentation of the learning indicators.

Table 8-3 lists the performance measures or learning indicators that were computed for each individual across sessions. For this exploratory study we collapsed data from the sessions into a single index for each indicator; we will look at changes over time later on. Two data sources were used in computing these values: (1) detailed computer history lists of all student actions, and (2) verbal protocols from each student about justifications for each action, what they expected to see after a particular action, and what their plans were for further experimentation.

Comparison of Subjects

Subjects BW, CF, HT, and OY all began the experiment at about the same level of knowledge as measured by pretest scores, but after the sessions with *Smithtown,* subjects BW and CF (more successful) greatly surpassed subjects HT and OY (less successful) on the posttest battery. In terms of gain scores (that is, posttest score minus pretest score), BW and CF scored over one standard deviation above the average gain score whereas HT and OY scored about one standard deviation below it.

	Pretest	Posttest
BW and CF	47.0	86.7
HT and OY	47.4	63.1

The question reduces to: What did BW and CF do, in terms of the indicators, that HT and OY did not do? Table 8-4 summarizes the standardized scores for these two pairs of subjects. The largest differences (ordered) between these two groups are for the following ten indicators: 22, 6, 24, 29, 9, 20, 16, 23, 28, and 13. The difference scores for all of these indicators exceeds .90 standardized units.

The first observation is that the majority of these indicators are from the most cognitively complex set of behaviors delineated, in other words, those in the thinking and planning category with six of the difference scores greater than .90. Next, there are three main differences between the two groups in the data management category. Finally, only one significant difference score is from the activity/exploration category. The progression of behaviors across these three categories goes from simply being active in the environment (activity/exploration) to being efficient (data management) to, finally, being effective (thinking and planning).

Table 8-3 Learning Indicators

Activity level

1. Total number of actions.
2. Total number of experiments.
3. Number of changes to the price of the good.

Exploratory behaviors

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.

Data recording

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.

Efficient tool usage

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 fewer than six 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.

Use of evidence

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.

Consistent behaviors

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.

Effective generalization

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.

Table 8-3 Learning Indicators (*continued*)

24. Number of times the student had sufficient data for a generalization (that is, at least three data points in the notebook before using the hypothesis menu).

Effective experimental behaviors

25. Number of times a change to an independent variable was sufficiently large enough (greater than 10 percent of the possible range).
26. Number of times one of the experimental frames was selected (that is, chose "same good, change variable," "change good, same variable," 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. (In the initial sessions, a low number indicates effectiveness, whereas in the later sessions a higher number indicates effectiveness since the domain knowledge increases and the student can deal with interrelationships among variables.)
29. Average number of actions per experiment. This should be an increasing function over sessions.
30. Number of economic concepts learned per session.

We will now discuss each of these ten indicators in their relation to individual differences in performing in this type of environment. We will illustrate the between-subjects' differences in each of the three relevant categories with excerpts from their verbal protocols and student procedure graphs developed to depict student solution paths.

Thinking and Planning Discriminating Indicators

This category represents the more complex learning indicators relating to experimental behaviors. First, the data show that the subcategory of effective generalizations was a very good discriminator between these subjects. Overall, BW and CF attempted to generalize findings across markets (indicators 22 and 23) to see if developing beliefs extended beyond the current market. This included both generalizing to related markets (for example, investigating the effects of a manipulation on substitute or complementary goods) or testing beliefs in unrelated markets to see the limits and extent of a particular concept. To illustrate, BW (more successful) was careful to try out his developing ideas in different markets to test his hypotheses. In the first session he was investigating the tea market, testing the idea that increasing the population caused an increase in the quantity demanded (it actually shifts the demand curve). BW increased the population and then said:

Table 8-4 Z-scores for Subjects on Each Indicator

Indicator/group	BW and CF	HT and OY	Difference
General activity/exploration levels			
Activity level			
1.	0.32	0.73	0.41
2.	−0.11	−0.29	0.18
3.	0.19	0.55	0.36
Exploratory behaviors			
4.	0.17	−0.32	0.49
5.	0.42	0.13	0.29
6.	1.09	−0.70	1.79*
7.	0.69	0.80	0.11
8.	0.22	−0.21	0.43
Data management skills			
Data recording			
9.	1.17	−0.01	1.18*
10.	0.21	−0.31	0.52
11.	0.48	0.90	0.42
12.	−0.37	0.33	0.70
Efficient tool use			
13.	0.70	−0.20	0.90*
14.	−0.67	0.10	0.77
15.	0.67	−0.06	0.73
Use of evidence			
16.	1.29	0.17	1.12*
17.	0.91	0.64	0.27
Thinking and planning skills			
Consistent behaviors			
18.	0.02	0.87	0.85
19.	0.88	0.00	0.88
20.	−0.13	1.01	1.14*
Effective generalizations			
21.	1.26	0.93	0.33
22.	1.88	−0.59	2.47*
23.	0.50	−0.50	1.00*
24.	1.52	−0.02	1.54*
Effective experimental behaviors			
25.	0.77	−0.01	0.78
26.	0.50	0.07	0.43

Table 8-4 Z-scores for Subjects on Each Indicator (*continued*)

Indicator/group	BW and CF	HT and OY	Difference
27.	0.66	1.13	0.47
28.	−1.13	−0.19	0.94*
29.	1.01	−0.35	1.36*
30.	0.86	0.59	0.27

*Indicates that the corresponding indicator differentiates the successful from less successful subjects by a standardized unit ≥ .90.

> Well, the quantity demanded did go up, it was 2,550 last time, although I would have thought it would have gone up more, twice as many people drinking tea [he had doubled the population]. So, quantity demanded did go up. There was a bit of a shortage. Well, I'd be pretty sure that it [shows the relationship between population and quantity demanded]. . . . I think it would, but since I haven't tested it out, I can't really say. I would change the good to take care of that problem.

Since some of the town factors have global effects and some have limited effects, it is a good strategy to try out things in different markets. After looking at the effects of interest rates on the compact car market, then switching to the doughnut market to see if interest rates affected anything there, BW concluded:

> OK, so I guess interest rates only influence expensive things like compact cars or big cars, but not doughnuts or hamburger buns. I bet there are things that influence everything, like income influences everything.

In contrast, subjects from the less successful group never generalized a concept across markets (related or unrelated goods). For any given market, they would make a hypothesis from the current data set and presume that it held across all goods without actually testing that notion out. In fact, because of the way the hypothesis menu was implemented in this version of *Smithtown,* it was possible to state a number of correct hypotheses from a single market; however that is not good scientific behavior.

The next indicator that differentiated the two groups had to do with using the planning menu to set up an experiment, specifying variables to investigate, and actually conducting an experiment based on those stated variable manipulations (indicator 20). Sternberg (1981, 1985) discusses two metacomponents,

global planning and local planning, isolated from a complex reasoning task. In a study of planning behavior in problem solving he found that more intelligent persons scoring high on reasoning tests tended to spend relatively more time than low-scoring persons on global (higher-order) planning and relatively less time on local (lower-order) planning. Poorer reasoners, however, seemed to emphasize local rather than global planning relative to the better reasoners. Similarly, Anderson (1987) investigated individual differences in students' solutions to LISP programming problems and found that the poorer students tended to plan less in their problem-solving activities. These findings are similar to those of our study: individuals who engage in planning an experiment are more successful (measured by our gain scores criterion) than those who do not. To illustrate, CF (more successful) decided to test the effects of weather on the demand for ice cream (where weather can range from 1—cold and wet—to 10—warm and dry). From the planning menu she chose the variables to investigate: price, quantity demanded, quantity supplied, surplus, shortage, and weather. After changing the weather index from a medium default value of 5 to a value of 10, she said, "OK, then that means, I think, there should be an increased demand for ice cream." She collected and recorded the data, observed that, indeed, the quantity demanded of ice cream went up, and chose the framework "same good, change independent variable" so that she could stay in the ice cream market and manipulate the weather variable further. From the new planning menu she selected the same variables as before, then changed the weather: "I'm gonna make the weather really bad. I'll put it at 1. . . . I think there'll be a surplus now, at the other extreme." This prediction was confirmed by her data. The other two subjects that were less successful evidenced much less front end (higher-order) planning of an experiment and typically only selected a few (or irrelevant) variables from the planning menu. Often changing an independent variable has effects on certain other variables, and a given experiment should focus on those. That is, a population increase could have an effect on the *demand* of a good and, in the long run, on the *price* of that good.

The next discriminating indicator (indicator 29) reflects the richness and tenacity of an individual's actions within an experiment as measured by the average number of actions taken per experimental episode. A thorough, systematic investigation of a concept is indicated by more connected actions within an experiment while more aimless behavior is seen by fewer connected actions. If a person were to move around randomly in this environment, making changes, moving on to new things, with little or no thread of consistency, then each experiment would have a small number of actions taken within a given market. Subjects BW and CF were not random movers. Their method of investigation was to choose a market and do many things within that market, always observing the effects of their manipulations and recording them in the on-line notebook. Thus, the average number of actions within their experiments was much

greater than for subjects HT and OY. In addition, across 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. The less successful subjects did not demonstrate a similar increase in complexity of experiments over time; rather, their average number of actions went up and down across sessions.

Relevant to these results, Sternberg and Davidson (1982; see also Sternberg, 1985) looked at individual differences in the solution to insight problems where individuals were free to spend as long as they liked in the solution process. They computed a correlation of .62 between the time spent and the score on the insight problems; thus, persistence and involvement in the problems was significantly correlated with success in solution. They argue that more intelligent persons do not give up, nor do they fall for the obvious, often incorrect, solutions.

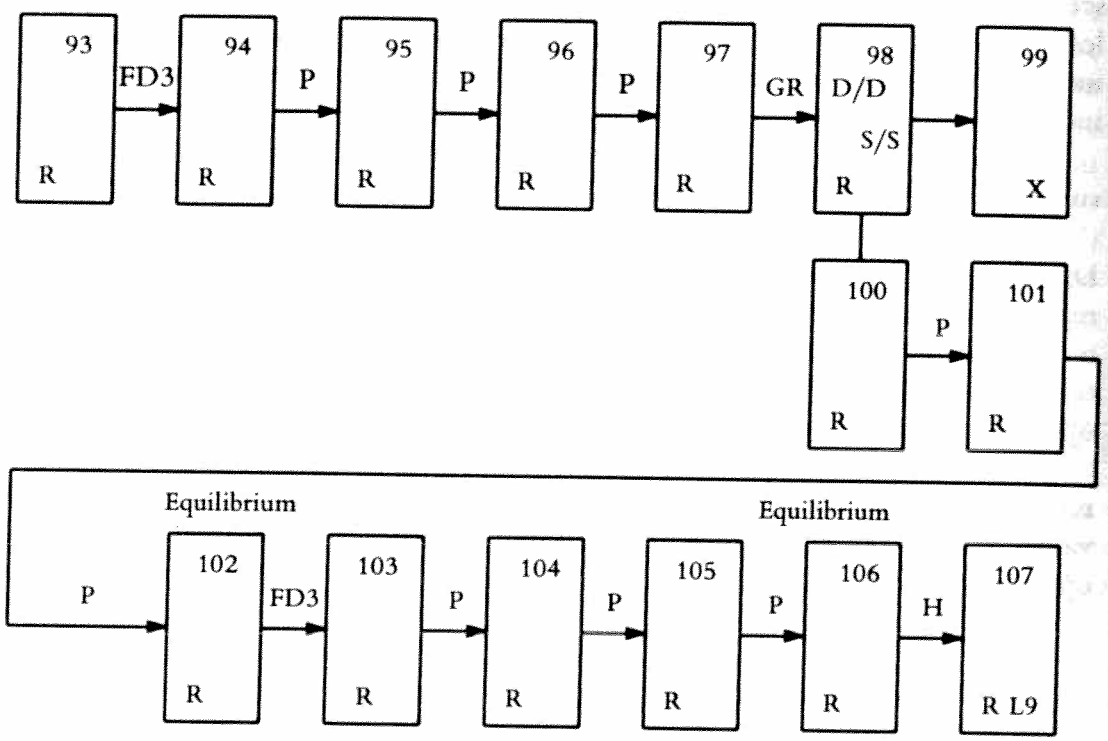
This activity is captured in *student procedure graphs;* we constructed these for subjects based on the idea of the problem behavior graphs of Newell and Simon (1972) showing student actions and the resulting state of knowledge. A state of knowledge is represented by a node, and the application of an operator is represented by an arrow pointing to the right.The result of the operation is the node at the head of the arrow. Vertical lines connecting nodes indicate a return to a previous state of knowledge because no new information was supplied. The operators and their symbols used for our purposes are listed below. Each operator is recorded above or below a horizontal arrow; an operator below the arrow indicates that the variable was changed back to its original default or baseline value. Most of the nodes (rectangles) contain symbols representing the resulting operation, also listed below.

	Operators and variables		Operations
P	Price	R	Notebook recording
G	Good	S	Supply curve
H	Hypothesis	D	Demand curve
FD	Town factor (demand shifts)	/	Superimposed curves (e.g., S/D)
FS	Town factor (supply shifts)		
GR	Graph	X	Error
T	Table		

Learning goals, meaning economic concepts that can be discovered, are indicated by symbols beginning with the letter "L" followed by a number — for example, the law of demand is L5. Their meaning can be seen in Figure 8-1.

Following are two examples of how the student procedure graphs are used to visually illustrate the flow of problem-solving activity in more and less efficient individuals (in relation to indicator 29).

Figures 8-13 and 8-14 exhibit obvious differences in experimental behavior using data from BW and another subject showing below average gain (subject SS) whose performance illustrates well the contrast between focused and fragmented search. The horizontal movement depicted in the graph of BW's performance (see Figure 8-13) shows much more focused and connected persistent behavior than the vertical, less relevant movement in SS's experimental



P	Price	X	Incorrect attempt at demand shift
R	Records data	L9	Demand shift
H	Hypothesis	D/D	Superimposes two demand curves
GR	Graph	S/S	Superimposes two supply curves
FD3	Income		

Figure 8-13 Student procedure graph of a more successful subject where horizontal movement of the graph indicates market investigation prior to the second hypothesis.

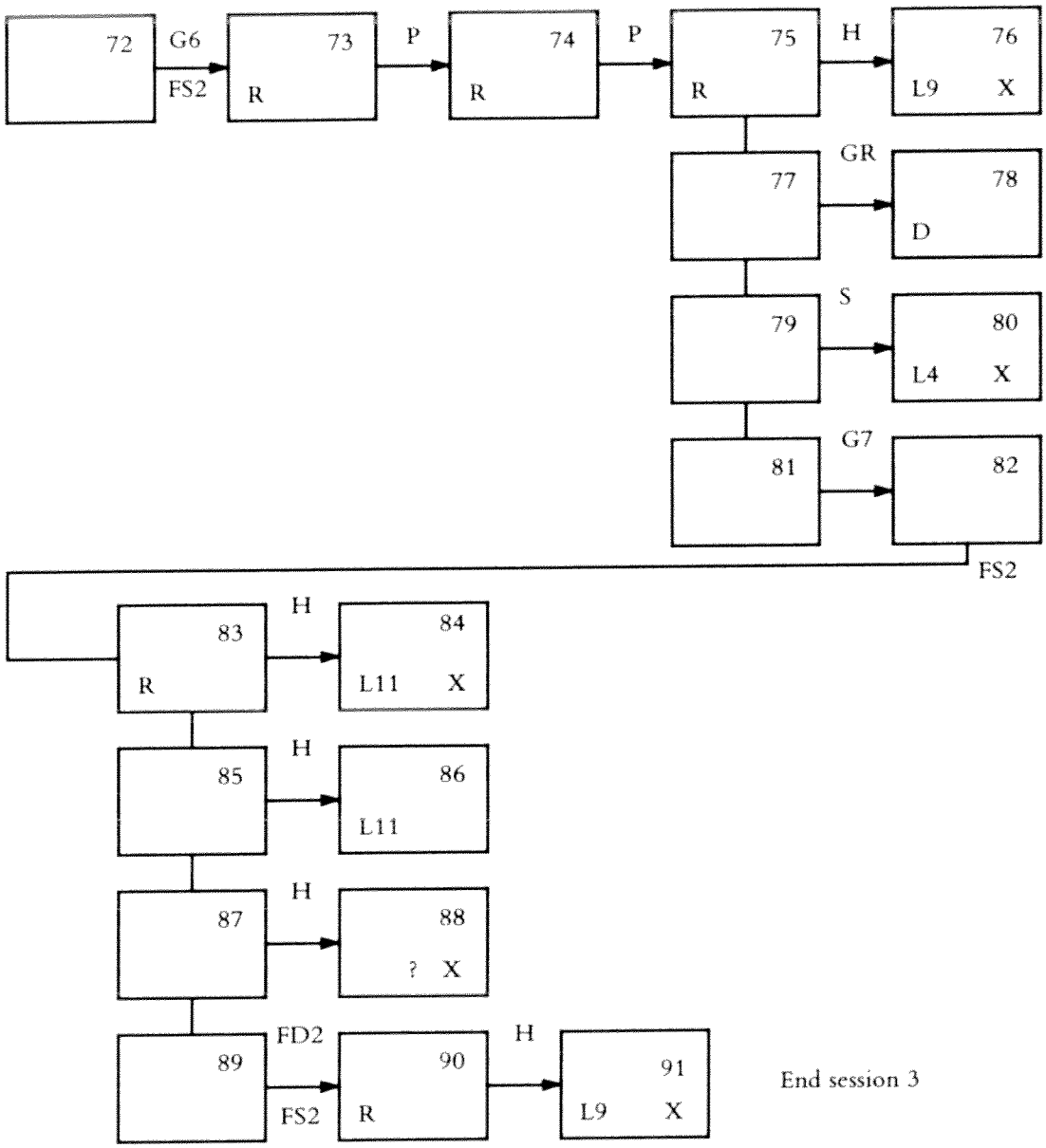


Figure 8-14 Student procedure graph of a less successful subject where vertical movement of the graph indicates a lack of experimentation.

behavior, as seen in Figure 8-14. In Figure 8-13, BW (more successful) began investigating the large car market by collecting data for the market when income was $20,000. At nodes 94 through 97 he changed the average income to $30,000 and collected additional data by changing price three times. Next, he plotted a demand curve (node 98) with income at $20,000 and then at $30,000, and at node 99 he made a hypothesis that when income increased quantity demanded increased and the demand curve would move to the right. During the period from nodes 100 to 102 he had the computer adjust the price back to equilibrium. From 103 to 106, he changed income to $40,000 and again had the computer adjust the price back to equilibrium. The subject said, "It's only if you change something other than price that you get a new demand curve." Finally, at node 107 he hypothesized correctly that demand curves shift as a result of changes other than price (that is, one characterization or description of what causes demand curves to shift).

In contrast to the systematic, persistent performance evidenced by BW, subject SS (less successful) spent a considerable amount of time generating hypotheses that were unrelated to the current experiment. Although both subjects were attempting to characterize a demand shift, Figure 8-14 and the following summary of actions clearly demonstrate an ineffective experimental procedure.

At node 73, SS entered the market for gasoline, and from nodes 74 to 75 she changed the price from $1.18 a gallon to $1.00 a gallon and then down to $0.75 a gallon. At node 76 she hypothesized that "as the price of complementary goods decrease, the quantity demanded increases." This was incorrect. She then tried to graph a demand curve (77 to 78) but was unsuccessful. During the period involving the nodes 79 to 80 she hypothesized that as price increases the demand curve shifts down and to the left. This was incorrect. The subject then entered the coffee market suddenly and without any apparent reason (82). At 83 she changed labor costs from $4.00 an hour to $20.00 an hour followed by three more incorrect hypotheses (84 to 88):

- As labor costs increase, the quantity supplied decreases and shortage increases.
- Quantity demanded has no relation to labor costs.
- Quantity demanded has no relation to the price of resources.

During 89 and 90 she decided to change the population from 10,000 to 50,000 and then returned the labor costs to $4.00 per hour. Finally, at 91, she again attempted a hypothesis that as population increases quantity demanded increases and quantity supplied increases. This, too, was not quite right.

A major difference in experimental behavior illustrated here seems to be one of staying with a problem until it is solved. Subject BW, when his initial hypothesis turned out to be incorrect, did more experimenting to understand more precisely the nature of the problem. In contrast, subject SS, who apparently was motivated by just getting a hypothesis correct, tried different hypotheses, some of which were wild guesses since there was no relation between the stated hypotheses and the experiments actually conducted.

The next indicator to discriminate between more or less effective performance was indicator 28: changing only a limited number of variables per experiment in which the fewer variables that are changed the better the subsequent performance. BW and CF (more successful) were very conscientious in changing only one variable at a time per experiment. Given the freedom of the environment, the temptation was often to make changes to multiple variables concurrently; however, the ensuing results were obscured as to what was actually responsible for the state of current market affairs. Subjects HT and OY (less successful) often fell prey to this temptation of making multiple changes. For example, while investigating the market for large cars, OY, when asked what he was going to do, responded, "I want to just go back and change some stuff." He then proceeded to change interest rates from 15 percent to 6.7 percent, number of suppliers (that is, large car dealerships) from ten to twenty, consumer preference (meaning the popularity of large cars) from 5 (medium) to 10 (very high) and then back to 5, per capita income from $20,000 to $25,000, and then interest rates from 6.7 percent to 9 percent. He did this all at once without collecting any data between the changes. When asked about what he would predict as a result of all of the changes, he said, "I think they'll still buy the cars, because the income is higher now . . . but the interest rates are higher . . . but since they're making more income, then I think they can afford it." OY's working memory capacity has obviously been overloaded at this point, and he fails to even consider the effects of the number of suppliers, consumer preference, or any of the potential interactions. Upon inspecting the market data, he sees that, in fact, there is an overall surplus of large cars. This he views as confirmation of his prediction, but it obviously is confounded by the fact that he had raised the number of large car dealerships in *Smithtown* as well as the per capita income. These last two actions actually have opposing effects: increasing the number of dealers would result in a surplus of cars but increasing the income would cause a shortage of cars.

The last indicator falling under the thinking and planning category involves collecting sufficient amounts of data before making a hypothesis of 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 inadequate data which may include elements of chance, confound-

ing variables, or other things. BW (more successful) investigated the concept of surplus and its relationship to price, quantity demanded, and quantity supplied. In the following protocol, it is apparent that he investigated the concept from many angles, collecting more than enough data before rendering a hypothesis. BW has just had the computer adjust the price of hamburger buns (raising the price):

> The price went up a lot, and there's a big surplus. . . . Well, as I found out before, as price goes up, the quantity demanded goes down, quantity supplied goes up. So, by now the quantity demanded has gone below the quantity supplied, and there's a surplus. So, the next time around I think the price should go back down 'cause there's a lot of hamburger buns around here. They'll go on sale.

He watched the price slowly converge on equilibrium, interrupting the computer adjustments with price adjustments himself until he found the equilibrium price where, "at $1.55 it came out right . . . no surplus and no shortage." BW speculated,

> OK, so I found out when there's no surplus and no shortage the price won't change. I could phrase that into a hypothesis also. When there's a surplus, price decreases, and when there's a shortage, price increases. . . . If surplus is greater than zero, then the price decreases.

When asked if he could characterize surplus any other way, he responded,

> Well, it's just quantity supplied minus quantity demanded. I can state that. *I've got enough examples!* There's a surplus when the quantity supplied is greater than the quantity demanded.

He then used the hypothesis menu and formalized the above statement into a successful specification of "surplus." Immediately afterwards, he used the same data and logic to characterize "shortage."

In contrast, HT (less successful) was content to make predictions and hypotheses based on single events and nonreplicated experiments. This was a poor strategy for this subject to follow since her data management skills were neither efficient nor consistent. Moreover, sometimes she forgot or misconstrued what the previous data were, not bothering to go back and retrieve the omitted data. For instance, after spending a long time in the final session trying to determine the influence of population changes on some of the dependent variables, she conducted an experiment which involved decreasing the population of *Smithtown* from 10,000 to 5,000. At that time, she was investigating the

doughnut market, and the experimenter asked what she expected to see as the result of this population decrease.

HT: So, less people will eat [doughnuts].

EXPERIMENTER: What about quantity supplied and price?

HT: When population decreases, demand . . . quantity demanded decreases, and quantity supplied decreases . . . price increases.

The market actually depicted the price and the quantity supplied remaining the same while the only change was in the quantity demanded, which changed as a function of the demand curve shift. Next, HT's actions centered around price changes to get an equilibrium price for the doughnut market in the smaller sized town. She did not replicate the experiment with the population change, and later, when attempting to articulate a hypothesis, she remembered erroneous results and showed little understanding of cause and effect among the variables:

HT: OK, so, I think when population decreased, the price decreased. That's why there is changes between quantity supplied and quantity demanded.

EXPERIMENTER: What was the first thing that happened?

HT: I think quantity demanded decreased . . . and when quantity demanded decreased, price decreased. Quantity supplied . . . let's see, population decreased . . . quantity supplied decreased.

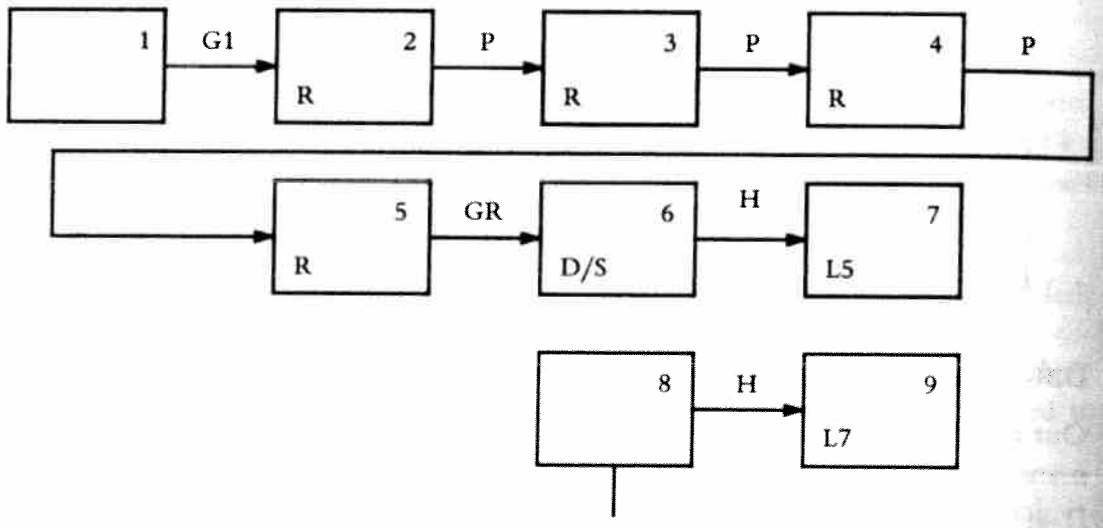
Data Management Discriminating Indicators

Our more successful subjects, BW and CF, generally exhibited very good data management skills, using their notebooks efficiently and consistently. They typically made notebook entries following variable changes and included variables in their notebooks that had been specified beforehand in the planning menu. In contrast, the less successful subjects (HT and OY) never became fully automatic in entering data to their notebooks. They continued to forget to record important information throughout the three sessions and had to rely on the history window to insert forgotten data. They also excluded variables whose values were changed or that were listed in the planning menu. In addition, they continued to omit baseline data. This latter omission was a major problem when attempting to attribute causes to market conditions.

Indicator 9 concerns the total number of notebook entries and indicator 13 the number of relevant notebook entries made. With regard to just the total number of notebook entries, the more entries, the better the performance. As to

the *type* of notebook entries, the more relevant notebook entries that were made overall, the better the performance, where relevant variables are those specified in the planning menu as the variables the subject was interested in exploring and collecting data on. This measure indicates whether the individual used the notebook efficiently for recording important information.

To illustrate the contrast in types of data recording skills, Figures 8-15 and 8-16 show examples of students with better and worse recording skills. In Figure 8-15, BW (more successful) entered the tea market and, prior to changing any variables, decided "to see what the initial conditions are." He followed the observation with a notebook entry of the baseline data, seen in nodes 1 and 2. At node 3 he increased the price of tea from $1.83 a box to $2.50 a box "to see if there's a relation between price and quantity demanded and quantity supplied." This price change was also duly recorded in the notebook. During nodes 4 and 5 BW continued to investigate this relationship by decreasing the price twice more, following each change with a notebook entry. He then



G1	Tea
R	Records data
P	Price change
GR	Graph
H	Hypothesis
L5	Law of demand
L7	Law of supply
D/S	Superimposed demand and supply curves

Figure 8-15 Student procedure graph of a more successful subject showing good data recording skills.

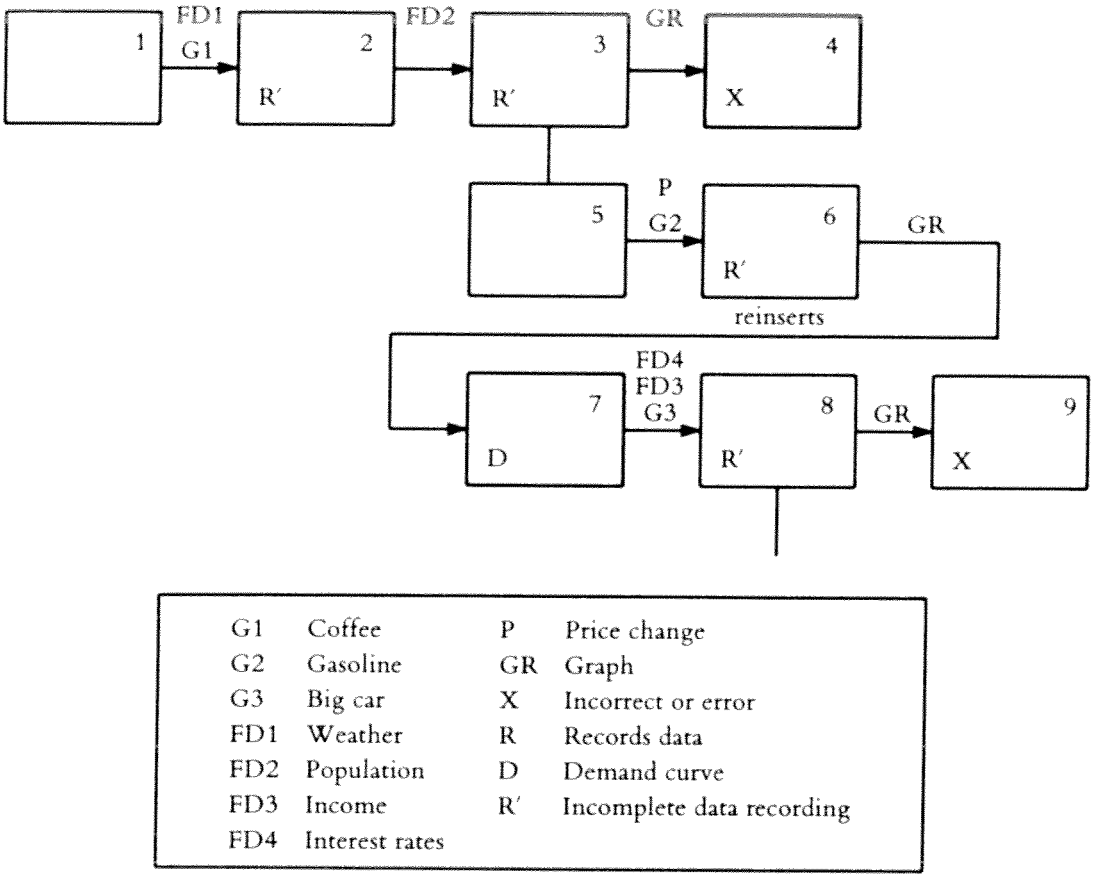


Figure 8-16 Student procedure graph of a less successful subject showing poor data recording skills.

graphed a demand curve (6) and successfully superimposed a supply curve, saving the graph for future reference. This systematic performance led to the correct induction of the laws of demand and supply (7 through 9):

- As price increases, the quantity demanded decreases.
- When price increases, quantity supplied increases.

Our less successful subject, HT, demonstrated inefficient data recording skills (see Figure 8-16). She rather haphazardly entered data into the notebook and failed to systematically record variable changes. Figure 8-16 illustrates some of the multiple variable changes and subsequent failure to record sufficient data into the on-line notebook. In Figure 8-16, the subject started out in the coffee market and changed the weather conditions from a mediocre value of 5 to a slightly less pleasant value of 3 to see if that would affect the demand for coffee, seen in nodes 1 and 2. She predicted that if the weather decreased (became

worse), then the price of coffee would increase. However, since she had failed to record any baseline data for the coffee market, she was unable to make the appropriate comparisons.

She then decided to ignore the weather influences, and at 3 changed the population from 10,000 to 4,000 persons. She predicted that if the population decreased, then a surplus would result. But, as with the above situation, she had failed to record the baseline data for when the population was 10,000, so she reinserted the necessary data from the past experiments. Next, HT tried to graph some data at node 4: price by quantity demanded and price by population. However, in each case there was only one data point per variable and thus no line could be drawn. She then switched to the market for gasoline (5 and 6) and raised the price from $1.50 a gallon to $4.00 a gallon. Since she again had failed to record the data from the market when gasoline was $1.50 a gallon, she had to reinsert this information into the notebook. With these additional data she tried to graph it again, and at 7 she successfully plotted a demand curve. At node 8 she entered the market for large cars. Her first action there was *not* to record the baseline data, but to change income from $20,000 to $30,000. She predicted that "if the income increases, people will have more money to buy large cars, so the price of the cars will increase, and the quantity demanded will increase, the quantity supplied will decrease, and there will be a shortage." When she saw the data resulting from her increase to the per capita income, however, price and quantity supplied actually stayed the same while the quantity demanded did, in fact, go up. Since she did not have the baseline data, she was unable to tell whether or not her prediction was confirmed since "increase" and "decrease" have to be interpreted relative to some other data. Finally, at node 9, she changed interest rates from 15 percent to 10 percent, predicting that if interest rates decrease then price will decrease and quantity demanded will increase and quantity supplied will decrease. She stated that she believed her prediction was confirmed, but later realized that price had stayed the same. Her last action in this segment shows how she tried to graph price against interest rates, but she did not have enough data.

Indicator 16 is the last major discriminating index in the data management category and deals with the number of specific predictions made by an individual in relation to the number of general hypotheses. In this case, the higher the ratio, the better the overall performance. Studies that have investigated individual differences between novices and experts in solving physics problems have found that an important distinction between the two groups is that whereas the experts used a "working-forward" strategy, the novices used a "working-backward" strategy (Simon and Simon, 1978; Larkin et al., 1980; Chi, Glaser, and Rees, 1982). These studies suggest that the novices may be more data driven whereas experts may be schema driven in the sense that their representation of a problem accesses a repertoire of solution methods. Thus, the novices' limita-

tions are derived from their inability to infer further knowledge from the literal cues in the problem statement. In contrast, these inferences necessarily are generated in the context of the relevant knowledge structures that experts possess. Predictions in *Smithtown* serve as a foundation or stepping stones to more general abstract principles and laws of economics. Our more successful subjects seemed to be able to work forward toward a goal—that is, they knew where they were going—in contrast to our less successful subjects who often got stuck at the more superficial or data level of investigation. To illustrate, CF (more successful) was interested in looking at the relationship between the coffee and tea markets, "because they are similar." First she increased the price of coffee and collected data on the resulting decreased quantity demanded and increased quantity supplied. Next, she chose the framework "change good, keep independent variables the same," changing to the tea market. Since the price of coffee had been increased, more people had shifted to drinking tea; thus the tea market came up with an initial shortage, confirming her initial prediction that "If the price of coffee increases, then the quantity demanded of tea will increase." She remained in these two markets and went on to investigate the concept of a new equilibrium point and demand shifts. She continued making predictions, observing the data, then proceeded on to successfully articulate the rules underlying the higher-level concepts. The less successful subjects skipped among markets, failing to make sufficient predictions in order to test developing hypotheses that would have led ultimately to the discovery of more economic concepts.

Activity/Exploration Discriminating Indicator

This last category concerns the number of times the subject had the computer make a price adjustment toward equilibrium (indicator 6). From the beginning sessions, our better subjects immediately grasped the utility of letting the computer make price adjustments while both the pattern of the changes it made and the effects on the market condition were observed. After his first computer change of the price, subject BW (more successful), when asked if the change was in accord with his expectations, said

> Well, yes, I thought so. Quantity demanded for hamburger buns was very high, and there were very few hamburger buns, so, it seems that suppliers would be able to get more for them. So, the price went up and there's still a shortage of hamburger buns. If I let the computer adjust the price again, the price will probably go up again.

He demonstrated an understanding that when the computer changed the price, it provided him an opportunity for observing systematic changes and relation-

ships. Although he did not have enough data to conceptualize "equilibrium point," he had started to understand that when shortages exist, prices go up. Our less successful subjects tried to use the option "computer adjust price," but they did not really grasp its purpose. It was revealed in the second session that subject HT (less successful) had no idea what was going on:

HT: Just now I had the "Computer Adjust Price."

EXPERIMENTER: Yes. Do you understand what's going on?

HT: No, I have no idea about that.

EXPERIMENTER: What happened when you chose that? How did it adjust the price?

HT: So, the price now increased from $1.70 to $1.90, and the quantity demanded decreased, decreased just a little bit. The quantity supplied increased a little bit too. No surplus, and the shortage is 6. Population is the same. So the price increased.

EXPERIMENTER: What would happen if you let the computer adjust the price again? Would it go up, down, or stay the same?

HT: I don't know much about the "Computer Adjust Price." It can increase or decrease, reduce the price.

The subject continued to have difficulty with this throughout the first two sessions (or 80 percent of the entire time with *Smithtown),* not realizing the benefits of observing the computer make price adjustments toward equilibrium.

Performance differences between our two groups were probably a function of the interaction of all of the aforementioned performance indicators. The behaviors that differentiated the subjects consisted of generalizing concepts across markets where the generalizations were a result of well thought out and executed plans, having sufficient data collected prior to the generalization, engaging in more complex experiments within a given market and not moving randomly among markets (that is, staying in an experiment long enough to extract valuable information), changing variables in a parsimonious and systematic fashion, recording important data in the notebook from different experiments, and generating and testing predictions that could lead to the induction of economic principles and laws.

GENERAL DISCUSSION

The comparisons between the economics classroom and the experimental group in their pretest and posttest results suggest that learning in the exploratory

world is at least as effective as traditional classroom learning. In fact, when learning time is compared, the students interacting with *Smithtown* spent less than half the amount of time formally learning economics compared to the length of time spent by the students in the economics classroom. It is possible that a group receiving classroom instruction *and* the intelligent discovery world could do even better. This remains an empirical question.

Our second, more compelling concern, was with the experimental group. In particular, we wanted to know how individuals learn or do not learn in this type of environment and on what measures the better and poorer learners differ. We illustrated that the contrasting pairs of subjects differed mostly on measures related to thinking and planning skills (that is, effective experimental behaviors) with fewer but significant differences in data management skills. The behaviors that differentiated the subjects were the following:

1. *Generalizing concepts across markets.* The better subjects would try out economic concepts in different markets to see if they were supported whereas the less effective subjects would not bother to extend an experiment across markets.

2. *Engaging in more complex experiments within a given market and not moving randomly among markets.* Typically, the better subjects had many more actions within a given experiment and investigated fewer markets overall compared to the less effective subjects.

3. *Changing only one variable at a time and holding all others constant.* The biggest problem for the poorer subjects was that they persisted in changing multiple variables simultaneously. The better subjects changed fewer variables at a time, typically just single variables.

4. *Basing generalizations on sufficient data.* We set as our criteria 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 whereas the less successful subjects were content to make careless and impulsive generalizations based on inadequate data.

5. *Conducting an experiment based on a planned manipulation or set of manipulations.* The planning and inferencing abilities of the better subjects allowed them to set up an experiment and execute it thoroughly, whereas the less successful subjects rarely evidenced advance (higher-level) planning throughout the experimental sessions.

6. *Generating and testing experimental predictions.* The better subjects tended to be more hypothesis or rule driven (working forward towards a goal), whereas the less efficient subjects were more data driven in experimentation. When

evidence does not confirm a hypothesis, further experimentation is required to modify the hypothesis. The better subjects generally recognized and implemented this approach, whereas others engaged in less systematic activities.

7. *Entering data into the on-line notebook.* Better subjects had more notebook entries overall compared to the less effective subjects. In addition, those entries tended to be more consistent with, and relevant to, the focus of their investigation.

8. *Using the computer to make price adjustments of a good towards equilibrium.*

We obtained demographic information from all subjects along with the pretest battery; two questions asked were (1) what science courses the subject had taken since high school, and (2) what the subject's major was. Subject BW (more effective) had taken just two science courses (physics I and II), and he was a sophomore, majoring in math. Subject CF (more effective) had taken three science courses (physics I, II, and III) and was also a sophomore, majoring in electrical engineering. In our less effective group, subject HT had five science courses (physics, two semesters of calculus, Fortran, and chemistry), and she was a freshman, majoring in pharmacy; subject OY had taken three science courses (chemistry, biology, and physics) and was a sophomore majoring in electrical engineering. These pairs could have differed in their scientific investigative behaviors as a function of past academic courses or variables relating to learning style differences. Thus, according to a hypothesis that different backgrounds were a cause of the observed differences, we would have expected the less scientific subjects to have taken fewer science courses. This was not the case. In fact, the less successful group had an average of four prior science courses while our more successful group had an average of 2.5 science courses since high school. In addition, each of the subjects was a science major. Of the original ten subjects in our experimental group, this same pattern was found. Dividing the subjects into two groups of five each based on their gain score, the two groups had the same number of declared science majors in each (three per group). However, the less successful group had taken considerably more science courses since high school (total of twenty-seven) compared to the more successful group (total of eight). Thus the idea of differential exposure to science training seems not to be a major factor in determining who will demonstrate better scientific behaviors.

Although this study focused on contrasting subjects in a descriptive and exploratory sense, the question arises as to whether the findings generalize to the population at large. As part of the Learning Abilities Measurement Program (LAMP) at the Air Force Human Resources Laboratory, the first author tested a large group of subjects—527 basic recruits at Lackland Air Force Base, Texas

—with a modified version of the system which included forty-four performance indicators automatically tallied in real time and summarized by the computer at the end of a 3.5-hour session. Using a measure of general intelligence as the dependent variable (the AFQT, a composite score derived from the Armed Services Vocational Aptitude Battery), we list the results of a correlational analysis of the indicators with AFQT score. Following are the indicators which significantly correlated with the AFQT score at $p < .001$:

1. Engaging in more complex experiments within a given market was tallied by the average number of actions per experiment. This indicator had a significant correlation with AFQT score; therefore the more connected actions taken in an experiment was associated with a higher AFQT score. Related to the *nature* of the experimentation, we also tallied the total number of markets investigated. A stepwise regression analysis was run on the data with AFQT score as the dependent variable. "Number of markets" was one of the five most predictive variables with an inverse relationship to AFQT. That is, the fewer the number of markets investigated, the more predictive of higher AFQT score.

2. The average number of independent variables changed at one time (that is, per experiment) had a significant negative correlation to AFQT score, implying that the fewer variables changed at a time, the better the performance.

3. Making hypotheses based on sufficient data was estimated by the indicator computing if the subject had at least three rows of related notebook entries before using the hypothesis menu. This correlated with AFQT score in our larger sample. Thus the better subjects relied on more data before formulating general principles and laws.

4. When a subject specifies his or her intentions for an experiment via a contrived manipulation on a variable or set of variables in the planning menu and actually conducts the experiment with those variables, this indicator is incremented. There was a significant correlation between planned performance and AFQT score. This implies that the more intelligent persons tended to engage in more higher-level, advance planning of an experiment.

5. Making and testing predictions of experimental outcomes and then observing the results for confirmation or negation of the prediction is effective interrogative behavior and tallied by this indicator. In the larger study, the overall number of predictions that a subject made was correlated with AFQT score.

6. The quantity and quality of on-line notebook entries were two significant indicators discriminating among subjects. First, the total number of entries in the notebook was significantly correlated with AFQT score; therefore the

higher AFQT scores were associated with more notebook entries overall. Variables entered into the notebook that had been specified in the planning menu was the second indicator, correlating with AFQT score. This implies that higher AFQT scores are associated with consistent behaviors, meaning formulating a planned set of variables to investigate and reliably entering those variables into the notebook.

Other indicators from the larger study that significantly correlated with AFQT score included the total number of actions taken in the experimental sessions, total number of economic concepts learned, and the number of experimental frameworks utilized by the subject. The total number of actions taken in the experimental sessions correlated with AFQT score, implying that the more intelligent persons were more active overall than the less intelligent persons. This must be viewed in light of the other indicators relating to the quality of performance, however, since it is not a matter of simply being "busy" in the environment but of being active in a connected, directed, systematic sense. For total number of economic concepts learned, the higher AFQT scores were associated with learning more concepts in the 3.5-hour session. Finally, the number of experimental frameworks utilized by the subjects correlated with AFQT score, implying that the experimental frameworks were employed more by the successful individuals as a planning procedure than the less successful persons.

Thus, the larger study seems to corroborate many of the findings from the descriptive analyses, extending and more precisely delineating individual differences in learning from this type of environment.

A limitation of the present study was the collapsing of data across sessions for this initial investigation. This can result in a loss of information that is valuable for examining individual differences and changes in knowledge and skills over time. Another limitation was that the use of difference scores on the economic tests as the measure of success was not ideal. That is because our primary focus for *Smithtown* was on the learning of good inquiry skills and only secondarily on the acquisition of economic knowledge. The ideal criterion (and data we plan to collect) should be the transfer of skills across domains; that is, how well students perform in a new environment with a similar structure or architecture but which differs in content from *Smithtown*. Currently there are several other systems being developed that fit these criteria, and further studies are planned which will investigate transfer of learning of these inquiry skills to new domains.

In general, it appears that in the rather complex task involved in this study 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. Individual differences in performance in our exploratory environment involved the following dimensions: generalization, goal setting and planning, more or less structured search, specific performance heuristics, and memory management.

Better 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. Better reasoners were 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 rule or hypothesis driven than the less successful subjects.

Better reasoners also engaged in more connected actions—structured search. They conceived of a particular market as a rich environment in which many actions needed to be taken in order to develop a structured understanding; disconnected probes did not assist them in their attempt at understanding. Less successful subjects, on the other hand, moved more frequently between markets. Their behavior was more fragmented and displayed a breadth of exploration, in contrast to a more in-depth search, in their attempt to establish meaning in a particular context.

Planning behaviors differentiated individuals whereby successful subjects planned their manipulations and experiments. Given the opportunity, successful subjects would structure a plan and then carry it out with specific information. The immediacy of carrying out some action was more desirable to the less successful subjects, comparable to jumping to solving equations in physics problems.

The successful individuals in our study 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 this heuristic, and for them it was a less desirable activity. Successful subjects took their time to generate sufficient evidence before coming to a conclusion, whereas the less successful subjects were more impulsive and attempted to induce generalizations based on inadequate information.

The necessity to manage memory was evident in the performance of the better subjects. They realized that they needed to store and display the information they had collected. Their data management performance was goal driven in the sense that the data collected were relevant to the current focus of their investigation. This contrasts with the poorer subjects' data management behaviors which were mostly inconsistent and often unrelated to an overall goal in their experimentation.

With regard to inductive problem solving, as Klahr and Dunbar (1988; see also Greeno and Simon, 1984) 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 while 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 investigating differences between novice and expert in problem solving (Larkin et al., 1980).

Furthermore, the importance of higher-level planning in this inductive discovery environment is in agreement with studies of individual differences in reasoning tasks (Sternberg, 1985). Better subjects consistently planned an experiment and then executed it to completion according to plan, in sharp contrast to the more haphazard, less planned approach applied by less successful subjects in their experimental methodologies.

This, then, is our initial study of individual differences in learning from an exploratory environment where students had the opportunity to engage in active discovery learning of economic concepts by manipulating variables in a hypothetical town and observing the repercussions. Overall, the system worked as we had hoped: tutoring on the scientific inquiry skills resulted in learning the domain knowledge as evidenced by the performance on the posttest battery.

We have begun to delineate those skills and behaviors which are important to scientific discovery. Although there is currently very little research being conducted in this area, the behaviors we have identified in this chapter fit in with the findings from related research (Klahr and Dunbar, 1988; Langley et al., 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.

ACKNOWLEDGMENTS

The authors wish to acknowledge the many persons whose contributions have been invaluable to this project: Jeff Blais, Jeff Bonar, Kathleen Katterman, Alan Lesgold, Paul Resnick, and Jamie Schultz.

The Center for the Study of Learning is funded by the Office of Educational Research and Improvement of the U.S. Department of Education. Support for computer equipment and software development was provided by the Cognitive Science Program of the Office of Naval Research. Additional support for the current large-scale testing and analyses was provided by the Air Force Human Resources Laboratory. The opinions expressed do not necessarily reflect

the position or policy of either NIE, ONR, or AFHRL, and no official endorsement should be inferred.

REFERENCES

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

Anderson, J., Boyle, C., Farrell, R., and Reiser, B. (1984). *Cognitive principles in the design of computer tutors.* Technical report, Advanced Computer Tutoring Project, Carnegie Mellon University, Pittsburgh.

Bonar, J. G., Cunningham, R., and Schultz, J. N. (1986, September). An object oriented architecture for intelligent tutoring. In *Proceedings of the ACM conference on object oriented programming systems, languages, and application.*

Bradshaw, G. F., Langley, P. W., and Simon, H. A. (1983). Studying scientific discovery by computer simulation. *Science,* 222(4627): 971–975.

Brown, J. S. (1983). Learning by doing revised for electronic learning environments. In M. A. White (Ed.), *The future of electronic learning.* Hillsdale, N.J.: Erlbaum.

Champagne, A., and Klopfer, L. (1982). *Laws of motion: Computer-simulated experiments in mechanics. Teachers Guide.* New Rochelle, N.Y.: Educational Materials and Equipment Co.

Chi, M. T. H., Glaser, R., and Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence,* Vol. 1. Hillsdale, N.J.: Erlbaum.

Cronbach, L. J. (1966). The logic of experiments on discovery. In L. S. Shulman and E. R. Keisler (Eds.), *Learning by discovery.* Chicago: Rand McNally.

diSessa, A. (1982). Unlearning Aristotelian physics; A study of knowledge-based learning. *Cognitive Science,* 6(1): 37–75.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist,* 39:93–104.

Greeno, J. G., and Simon, H. A. (1984). *Problem solving and reasoning* (Tech. Rep. No. UPITT/LRDC/ONR/APS-14). Pittsburgh: University of Pittsburgh, Learning Research and Development Center. To appear in R. C. Atkinson, R. Herrnstein, G. Lindzey, and R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (rev. ed.). New York: Wiley.

Hayes, J. R., and Flower, L. S. (1986). Writing research and the writer. *American Psychologist,* 41(10): 1106–1113.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, Mass.: MIT Press.

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

Kuhn, D., and Phelps, D. (1982). The development of problem-solving strategies. In H. W. Reese (Ed.), *Advances in child development and behavior* Vol. 17. New York: Academic Press.

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

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

Lawler, R. W. (1982). Designing computer-based microworlds. *Byte,* 7(8): 138–160.

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

Newell, A., and Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, N.J.: Prentice-Hall.

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

Reimann, P. (1986). *REFRACT: A microworld for geometrical optics.* Unpublished manuscript, LRDC, University of Pittsburgh.

Shute, V. J., and Glaser, R. (In press). An intelligent tutoring system for exploring principles of economics. In R. Snow and D. Wiley (Eds.), *Straight thinking.* San Francisco: Jossey Bass.

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

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

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

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

Sternberg, R. J., and Davidson, J. (1982). The mind of the puzzler. *Psychology Today,* 16 (June): 37–44.

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

White, B. Y. (1984). Designing computer activities to help physics students understand Newton's laws of motion. *Cognition and Instruction,* 1: 69–108.

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