In Klahr, D., & Kotovsky, K. (Eds.), (1989). Complex information processing: The impact of Herbert A. Simon. Hillsdale, NJ:Erlbaum.

4

Developmental Differences in Scientific Discovery Processes

Kevin Dunbar
David Klahr
Carnegie-Mellon University

ON THE ORIGINS OF DISCOVERY PROCESSES

Questions about the origins of scientific reasoning have been posed by developmental psychologists many times throughout the last 60 years (e.g., Karmiloff-Smith & Inhelder, 1974; Kuhn, Amsel, & O'Loughlin, 1988; Piaget, 1928; Vygotsky, 1934). The context of developmental questions about scientific reasoning can be expanded to include a number of broader questions—both descriptive and normative—about the nature of science and scientific reasoning. Within psychology, one approach to these questions has been to consider science a form of problem solving (e.g., Bartlett, 1958; Simon, 1977). The science-as-problem-solving view is stated most explicitly in Herbert Simon's characterization of scientific discovery as a form of search and in his elucidation of many of the principles that guide this search. For example, he has used the notion of search in a problem space to analyze what science is (Simon, 1977), how scientists reason (Langley, Zytkow, Simon, & Bradshaw, 1986; Kulkarni & Simon, 1988), and how scientists should reason (Simon, 1973). In this chapter, we follow a similar path, and apply the notion of search to the development of scientific reasoning strategies.

A contrasting view treats scientific reasoning as a form of concept formation. In the paradigmatic investigation of science-as-concept-formation, subjects are given examples or instances of a concept and are then asked to discover what the concept is (e.g., Bruner, Goodnow, & Austin, 1962). The extensive body of literature accumulated using this approach has revealed many differences between the reasoning processes used by adults and children when forming concepts. However, other than simply asserting that scientific reasoning is a type of concept formation, psychologists have not formally specified

how the cognitive processes involved in concept formation tasks are similar to those involved in scientific reasoning.

One way to specify this similarity is to build a model of the processes that are involved in both concept-formation tasks and problem solving. One model that has proved useful in this respect is Simon and Lea's (1974) Generalized Rule Inducer (GRI). Simon and Lea demonstrated how this single system encompasses both concept learning and problem solving. Within the GRI, concept learning requires search in two problem spaces: a space of instances and a space of rules. Instance selection requires search of an instance space, and rule generation requires search of a rule space. Simon and Lea's analysis also illustrates how information from each space guides search in the other. For example, information about previously generated rules may influence the generation of instances, and information about the classification of instances may determine the modification of rules.

A number of theorists (e.g., Cohen & Feigenbaum, 1983; Kulkarni & Simon, 1988; Lenat, 1977) have argued that the dual space search idea at the core of GRI can be extended to the domain of scientific reasoning, which takes place in a space of hypotheses and experiments. Using this idea, we developed a task that enables us to observe subjects' search paths in both spaces (cf. Klahr & Dunbar, 1988). Specifically, we studied the behavior of subjects who were attempting to extend their knowledge about a moderately complex device by proposing hypotheses about how it worked and then trying to determine whether or not the device behaved in accordance with their hypotheses. In this chapter, we use the task to investigate what components of the scientific reasoning process show a developmental course. Our goal is to understand how existing knowledge structures determine the initial hypotheses, experiments, and data analysis in a discovery task. Because we treat scientific reasoning as a search in two problem spaces, we explore the issue of whether there are developmental differences in how the two spaces are searched, and how search in one space affects search in the other.

Our subjects worked with a programmable, multifunctioned, computer-controlled robot whose basic functions they had previously mastered. We trained both adults and elementary-school children to the same criterion on basic knowledge in the domain before we asked them to extend that knowledge by experimentation. This training allowed us to analyze developmental differences among subjects who shared a common knowledge base with respect to the task domain. Our analysis focuses on their attempts to discover how a new function operates—that is, to extend their understanding about the device—without the benefit of any further instruction. In order to do this, our subjects had to formulate hypotheses and then design experiments to evaluate those hypotheses; the cycle ultimately terminated when they believed that they had discovered how to predict and control the behavior of the device.

The chapter is organized as follows. First, we briefly review some of the relevant literature on the development of scientific reasoning skills. Following this, we describe our task in detail, and then summarize two earlier studies using adult subjects.[1] These studies provide a context for the developmental questions. In the third study, we describe the performance of 8- to 11-year-old children on this task. On the basis of these three studies we propose a model for scientific reasoning, and then use it as a framwork for understanding the development of scientific reasoning strategies.

DEVELOPMENTAL ISSUES IN SCIENTIFIC REASONING

We have reviewed research on scientific reasoning in adults elsewhere (cf. Klahr & Dunbar, 1988), and in this section we concentrate on developmental issues. Research on scientific reasoning has typically treated different aspects of the overall process in isolation. In the developmental literature this approach has tended toward a polarization of views about the ontogenesis of scientific thought. One position is that improvements in scientific reasoning abilities are a consequence of a knowledge base that grows as the child develops (e.g., Carey, 1985; Keil, 1981). For example, Carey (1984) stated that

> the acquisition and reorganization of strictly domain-specific knowledge (e.g., of the physical, biological and social worlds) probably account for most of the cognitive differences between 3-year-olds and adults. I have argued that in many cases developmental changes that have been taken to support format-level changes, or changes due to the acquisition of some tool that crosscuts domains, are in fact due to the acquisition of domain-specific knowledge. (p. 62)

Under this extreme view, the actual processes that children use only *appear* to be qualitatively different from that of adults because children do not have the necessary knowledge to perform at adult levels.

The other view, exemplified by the work of Piaget (1952), purports that although there are obviously changes in the knowledge base as children grow older, they are not the primary source of the radical differences in the behavior of children and adults. Rather, children have qualitatively different representations of the world and strategies for reasoning about it (e.g., Inhelder & Piaget, 1958; Kuhn & Phelps, 1982). Research in this tradition has used tasks in which the role of knowledge has been minimized and the different developmental strategies are made transparent. With respect to the development of scientific reasoning strategies, this latter view makes very specific claims. Flavell (1977) succinctly described the difference between the reasoning strategies of adults and children as follows:

[1]Reported in Klahr & Dunbar, 1988.

> The formal-operational thinker inspects the problem data, *hypothesizes* that such and such a theory or explanation might be the correct one, deduces from it that so and so empirical phenomena ought logically to occur or not occur in reality, and then tests his theory by seeing if these predicted phenomena do in fact occur. . . . If you think you have just heard a description of textbook scientific reasoning, you are absolutely right. Because of its heavy trade in hypotheses and logical deduction from hypotheses, it is also called hypothetico-deductive reasoning, and it contrasts sharply with the much more nontheoretical and nonspeculative empirico-inductive reasoning of concrete-operational thinkers. (pp. 103–104).

Taken literally, this claim would lead to the conclusion that most adult subjects have not achieved the formal-operational level, because it has been well-established that adults find it extremely difficult to design experiments that provide a logical test of their hypothesis (e.g., Wason, 1968). Indeed, even well-trained scientists often draw invalid conclusions from the results of their experiments (e.g., Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Furthermore, the view of science as a hypothetico-deductive process is not consistent with recent descriptions of how scientists really work (cf. Harre, 1983; Kulkarni & Simon, 1988). Whether or not children's thinking is empirico-deductive is an open question. Although there has been a considerable amount of research on children's abilities to design experiments that test hypotheses, there has been little research that allows children to generate experimental results and then form hypotheses on the basis of these results. Therefore, one of the aims of our work with children was to discover what strategies they use in a scientific reasoning task, and how these strategies differ from those used by adults.

We believe that instead of framing the developmental question in terms of the dichotomy between a broadening of the knowledge base and a qualitative change in reasoning skills, it is more fruitful to provide a detailed characterization of the processes that are involved in scientific reasoning, and then to ask about the development of these processes. The specific approach in this chapter is based on the dual-space search idea introduced earlier, and our focus is on developmental differences in the search processes. By using the same task to investigate the types of hypotheses that subjects generate, and the types of experiments that they conduct, we avoid the problem of studying knowledge and strategies in isolation. This enables us to answer some more focused questions about the development of scientific reasoning skills.

Development of Experimental Strategies

Many developmental investigators have looked at the ability to design informative experiments. One common approach is to allow children to design (or select) simple experiments that will reveal the cause of an event (cf. Case, 1974; Inhelder & Piaget, 1958; Kuhn & Phelps, 1982; Siegler & Liebert, 1975; Tschirgi, 1980). For example, Kuhn and Phelps (1982) studied 10- to 11-year-old children attempting to isolate the critical ingredient in a mixture. They discovered that children's performance was severely impeded by "the power and persistence of invalid strategies," (i.e., experimental designs that were invalid, insufficient, or inefficient). Subjects commonly behaved as if their goal was not to find the cause of an effect, but rather to generate the effect. Tschirgi (1980) found that this tendency to generate a particular effect depends on whether the effect under investigation represents a good or a bad outcome. When the result of an experiment is undesirable (i.e., a bad outcome), subjects' tendency is to (correctly) vary *only* the hypothesized causal variable; in order to eliminate the bad outcome. However, for good outcomes, subjects tend to simultaneously vary *everything but* the hypothesized cause of the good outcome. Tschirgi found that adults were as likely to make this error as children.

Recent work on children's experimentation strategies by Kuhn and her co-researchers (Kuhn, Amsel, & O'Loughlin, 1988) showed some developmental changes in the ability to evaluate evidence. By presenting a large number of possible causes that might produce an effect and asking children to state what factor or combination of factors are the cause of the event, Kuhn and her colleagues discovered that children are more prone to ignore evidence that is inconsistent with their theory and are satisfied even when they know that their theory only accounts for some of the data. Furthermore, when children are asked to think of what data would be needed to disprove their theory, they have great difficulty. Taken as a whole, these studies suggest that children—and under some circumstances adults—frequently fail to distinguish between the goal of understanding a phenomenon and making it occur.

The approach to experimentation that we will take is one of discovering the strategies that subjects use to both design and evaluate the results of experiments. When experimentation is considered as a form of search it should be possible to delineate what types of cognitive processes govern the search of the experiment space and then specify the differences between adults and children with regard to these processes. In the following sections we describe the task and the type of hypothesis and experiment spaces that the subjects work in. This makes explicit the types of processes in which we expect to see developmental differences.

STUDYING THE DISCOVERY PROCESS: GENERAL PROCEDURE

The device we use is a computer-controlled robot tank (called "BigTrak") that is programmed using a LOGO-like language.[2] It is a six-wheeled, battery-powered vehicle, approximately 30 cm long, 20 cm wide, and 15 cm high. The device is used by pressing various command keys on the keypad on the top

[2]This same device was first used in a study of "instructionless learning" (Shrager, 1985; Shrager & Klahr, 1986).

of the device, which is illustrated in Fig. 4.1. BigTrak is programmed by first clearing the memory with the **CLR** key and then entering a series of up to sixteen instructions, each consisting of a function key (the command) and a one- or two-digit number (the argument). When the **GO** key is pressed, BigTrak executes the program.

The effect of the argument depends on which command it follows. For forward (↑) and backward (↓) motion, each unit corresponds to approximately one foot. For left (←) and right (→) turns, the unit is a 6° rotation (corresponding to 1 minute on a clock face. Thus, a 90° turn is 15 minutes). The **HOLD** unit is a delay (or pause) of 0.1 seconds, and the **FIRE** unit is one audiovisual event: the firing of the cannon (indicated by appropriate sound and light effects). The other keys shown in Fig. 4.1 are **CLS, CK,** and **RPT. CLS** Clears the Last Step (i.e., the most recently entered instruction), and **CK** Checks the most recently entered instruction by executing it in isolation. Using **CK** does not affect the contents of memory. We describe **RPT** later. The **GO, CLR, CLS,** and **CK** commands do not take an argument. To illustrate, one might press the following series of keys:

CLR ↑ 5 ← 7 ↑ 3 → 15 HOLD 50 FIRE 2 ↓ 8 GO

and BigTrak would do the following: move forward 5 feet, rotate counterclockwise 42 degrees, move forward 3 feet, rotate clockwise 90 degrees, pause for 5 seconds, fire twice, and backup 8 feet.

Certain combinations of keystrokes (e.g., a third numerical digit or two motion commands without an intervening numerical argument) are not permitted by the syntax of the programming language. With each syntactically legal key-stroke, BigTrak emits an immediate, confirmatory beep. Syntactically illegal key-strokes elicit no response, and they are not entered into program memory.

STUDY 1: ADULTS DISCOVERING A NEW FUNCTION

In this study (we use the term *study* to distinguish our procedures from our subjects' "experiments"), we established a common knowledge base about the device for all subjects, prior to the discovery phase. We instructed subjects about how to use all function keys and special keys, except for one. All subjects were trained to criterion on the basic commands. Then the discovery phase started. Subjects were told that there is a "repeat" key, that it takes a numerical parameter, and that there can be only one **RPT** in a program. Then they were asked to discover how **RPT** works. (It repeats the previous N instructions once.)

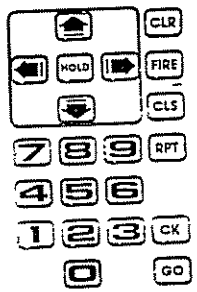


FIG. 4.1. Keypad from the BigTrak robot.

Procedure

Twenty Carnegie-Mellon undergraduates participated in the experiment. All subjects had prior programming experience in at least one language. The study consisted of three phases. First, subjects were given instruction and practice in how to generate a good verbal protocol. Next, the subjects learned how to use the BigTrak. All subjects mastered the device within about 20 minutes.

The third—and focal—phase began when the experimenter pointed out the **RPT** key and asked the subject to "find out how the repeat key works." Subjects were asked to speak aloud, to say what they were thinking and what keys they were pressing. All subject behavior during this phase, including all key-strokes, was videotaped. At the outset of this phase, subjects had to state their first hypothesis about how **RPT** worked before using it in any programs. When subjects claimed that they were absolutely certain how the repeat key worked, or when 45 minutes had elapsed, the phase was terminated.

Protocol Encoding

In this section we give a complete example of the kind of protocol that provides our basic source of data. (This listing, shown in Table 4.1, is one of our shortest, because it was generated by a subject who very rapidly discovered how **RPT** works.) At the outset, the subject (ML) forms the hypothesis that **RPT** N will repeat the entire program N times (003-004). (We call this kind of hypothesis *fully specified,* because both what will be repeated and the number of times it will be repeated are specified.) The prediction associated with the first "experiment" is that BigTrak will go forward 6 units (010-011). The prediction is consistent with the current hypothesis, but BigTrak does not behave as expected: it goes forward only 4 units, and the subject comments on the possibility of a failed prediction (013). This leads him to revise his hypothesis: **RPT** N repeats only the last step (019). At this point, we do not have sufficient information to determine whether ML thinks there will be one or N repetitions of the last step, and his next experiment (021) does not discriminate between the two possibilities. (We call this kind of hypothesis

partially specified, because of the ambiguity. In contrast, the initial hypothesis stated earlier (003-004) is *fully specified.*) However, his subsequent comments (024-025) clarify the issue. The experiment at (021) produces results consistent with the hypothesis that there will be *N* repetitions (BigTrak goes forward 2 units and turns left 60 units), and ML explicitly notes the confirming behavior (022). But the next experiment (026) disconfirms the hypothesis. Although he makes no explicit prediction, we infer from previous statements (023-025) that ML expected BigTrak to go forward 2 and turn left 120. Instead, it executes the entire **↑ 2 ← 30** sequence twice. ML finds this "strange" (028), and he repeats the experiment.

At this point, based on the results of only four distinct experiments, ML begins to formulate and verbalize the corect hypothesis—that **RPT** *N* causes BigTrak to execute one repetition of the *N instructions preceding the* **RPT** (030-034)—and he even correctly articulates the special case where *N* exceeds the program length, in which case the entire program is repeated once (035-037). Note that whereas the earlier hypotheses revisions maintained the role of *N* (it counted the number of times something was repeated), this final hypothesis gives *N* a new role: it determines what gets repeated. ML then does a series of experiments where he only varies *N* in order to be sure he is correct (038-046), and then he explores the issue of the order of execution of the repeated segment.

Aggregate Results

Overall Performance

Nineteen of the 20 subjects discovered how the **RPT** key works within the allotted 45 minutes. The mean time to solution (i.e., when the correct rule was finally stated) was 19.8 minutes. In the process of discovering how **RPT** worked, subjects generated, on average, 18.2 programs.

Of the 364 programs run by the 20 subjects, 304 were *experiments*; that is, they included a **RPT.** Another 51 programs were *control trials,* in which the subject wrote a program without a **RPT,** ran the program, then added **RPT,** and ran the program again. We label the initial program of the pair—as the one that does not include a **RPT**—as the control trial. Another seven programs we label as *calibration trials*: These were trials on which the subject attempted to determine (or remember) what physical unit is associated with **N** for a specific command (e.g., how far is ↑ 1). Only two programs that did not contain a **RPT** were unclassifiable.

We define a *common hypothesis* as a fully specified hypothesis that was proposed by at least two different subjects. Across all subjects, there were eight distinct common hypotheses. Protocols were encoded in terms of the fully specified hypotheses listed in Table 4.2. Subjects did not always express their hypotheses in exactly this form, but there was usually little ambiguity about the current hypothesis. We coded each experiment in terms of the

TABLE 4.1
Example of a Complete Protocol

```
002     EXP: SO HOW DO YOU THINK IT MIGHT WORK?
003     Uh... it would repeat all of the steps before it, however many times
004     I told it to repeat it.
005     EXP: WELL... SO START WORKING ON IT NOW THEN.
006     Ok.
007     EXP: AND TELL ME EVERYTHING THAT'S GOING ON IN YOUR MIND.
008     Ok.
009     EXP: NOW PRESS CLEAR.
010     Ok, um... I'm gonna make it go
011     forward two, and then I'm gonna make it repeat that twice.

012  00:30      ↑ 2   RPT 2
                                                          ↑ 4

013     Maybe not, uh....    ... repeat once

014  02:00      ↑ 1   ← 15   ↑ 1   HOLD 2   RPT 1
                                   ↑ 1   ← 15   ↑ 1   HOLD   4

015     Hmm.. guess that was it.
016     EXP: SO WHAT ARE YOU THINKING?
017     Um.. actually I have no idea now.
018     I'm trying to figure out what it is.
019     Um.. maybe it repeats the last step.
020     Ok, I'm gonna try that.  ..... repeat once.

021  03:30      ↑ 2   ← 30   RPT 1
                                                  ↑ 2   ← 60

022     All right, that backs up my theory.
023     Let me see if I can somehow make sure that that's what it does
024     is repeats the last step however many times that I tell it to,
025     so I'm gonna ... repeat it four times...

026  04:00    ↑ 2   ← 30   RPT 4
                                   ↑ 2   ← 30   ↑ 2   ← 30

027
028     That was strange, hmm... um... let me see that again.

029  04:30      ↑ 2   ← 30   RPT 4
                                   ↑ 2   ← 30   ↑ 2   ← 30

030     Ok, maybe it means repeat the last number...
031     however many steps before it that I put in,
032     that'll be the  number after the repeat.  For instance,
033     if I put repeat two, it'll repeat the last two steps.
034     If I put repeat five,it'll repeat the last five steps,
035     and if there's too many...
036     if the five is more than the number of steps in the program,
037     it'll just end it at whatever number of steps in the program,
038     so . . . repeat one, no, repeat two.
```

Note: CLR and GO commands have been deleted. BigTrak's behavior is shown in boldface type.

(Continued)

TABLE 4.1
(Continued)

```
039
040  06:00     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 2
                                  ↑ 2   ← 15   ↑ 2   FIRE 3   ↑ 2   FIRE 3

041     All right, I think I might have gotten it.
042
043  06:30     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 3
                        ↑ 2   ← 15   ↑ 2   FIRE 3   ← 15   ↑ 2 FIRE 3

044     Ok, I think I've gotten it.  I'm gonna make it repeat four times.
045        . . wanna repeat four...

046  07:30     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 4
                  ↑ 2   ← 15   ↑ 2   FIRE 3   ↑ 2   ← 15   ↑ 2   FIRE 3

047     Ok, now I'm trying to figure out which order the repeat step goes.
048     If it does the first part of the program or if it does...if it starts
049     from the last part of the program, where repeat...
050     if I say repeat one, does it repeat the first step in the program,
051     or does it repeat the last step I pressed in?  Um...repeat that
052     step...
053
054  09:00     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 1
                                  ↑ 2   ← 15   ↑ 2   FIRE 6

055
056     It goes from the last step,
057     and I don't understand why it doesn't go backwards.
058     Maybe it counts back two steps.
059     If I put repeat two, it would count back two steps,
060     starting from there and go until the last step.  Alright,
061     ...um...the last two steps were forward two and fire three,
062     so let me try and repeat that again.

063  10:00     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 2
                                  ↑ 2   ← 15   ↑ 2   FIRE 3   ↑ 2   FIRE 3

064     All right, now if I ... repeat five...
065     so if I put repeat four, it should do the whole program over again.
066  11:00     ↑ 2   ← 15   ↑ 2   FIRE 3   RPT 4
                  ↑ 2   ← 15   ↑ 2   FIRE 3   ↑ 2   ← 15   ↑ 2   FIRE 3

067     Well, I think I figured out what it does.
068     EXP: SO HOW DOES IT WORK?
069     Ok, when you press the repeat key and then the number,
070     it comes back that many steps and then starts from there
071     and goes up to, uh...it proceeds up to the end of the program
072     and then it hits the repeat function again.
073     It can't go through it twice.
074     ......
075     EXP: GREAT.
```

Note: CLR and GO commands have been deleted. BigTrak's behavior is shown in boldface type.

TABLE 4.2
Common Hypotheses and Percentage of Experiments Conducted Under Each

HYPOTHESIS[4]	% EXPERIMENTS UNDER EACH HYPOTHESIS	
	Adults	Children
HS1: One repeat of last N instructions.	02	00
HS2: One repeat of first N instructions.	04	00
HS3: One repeat of the Nth instruction.	03	01
HN1: One repeat of entire program.	06	03
HN2: One repeat of the last instruction.	04	05
HC1: N repeats of entire program.	14	21
HC2: N repeats of the last instruction.	20	08
HC3: N repeats of subsequent steps.	02	00
HC4: N-1, N/2 or +N repeats.	00	17
HC5: N repeats of last 2 steps.	00	07
Partially specified	03	27
Idiosyncratic	14	01
No Hypothesis	28	10
	100	100

[4]Hypotheses are labeled according to the role of N: HS - selector; HN - nil; HC - counter.

hypothesis held by the subject at the time of the experiment, and Table 4.2 shows the proportion of all experiments that were run in study 1 while an hypothesis was held.[3] (The final column in Table 4.2 refers to the children's performance in study 3, to be described in a later section.)

On average, subjects proposed 4.6 different hypotheses (including the correct one). Fifty-five percent of the experiments were conducted under one of the eight common hypotheses listed in Table 4.2. Partially specified hypotheses, which account for 3% of the experiments, are defined as those in which only some attributes of the common hypotheses were stated by the subject (e.g., "it will repeat it *N* times."). An idiosyncratic hypothesis is defined as one that

[3]As noted earlier, HS1 in Table 4.2 is the way that BigTrak actually operates.

was generated by only one subject. Such hypotheses are not listed separately in Table 4.2. There were no stated hypotheses for 28% of the experiments.

The Hypothesis Space

The eight common hypotheses—which account for over half of the experiments—can be described in terms of four attributes: The role of *N*, the type of element to be repeated, the boundaries of the repeated element, and the number of repetitions. The resulting *hypothesis space* is shown in Table 4.3, together with an abstract test program and an indication (in the rightmost column) of how BigTrak would execute the test program, if it operated according to the hypothesis in question.

This space can be represented in terms of *frames* (cf. Minsky, 1975). The basic frame for discovering how **RPT** works is depicted at the top of Fig. 4.2. It consists of four slots, corresponding to the four attributes listed: *n*-role, unit of repetition, number-of-repetitions, and boundaries-of-segment. A fully instantiated frame corresponds to a fully specified hypothesis, several of which are shown in Fig. 4.2. There are two principle subsidiary frames for **RPT**, *N*-role:*counter* and *N*-role:*selector*. Within each of these frames, hypotheses differing along only a single attribute are shown with arrows between them. All other pairs of hypotheses differ by more than one attribute. Note that the hypotheses are clustered according to the *N*-role frame in which they fall.

Recall that subjects were asked to state their hypothesis about **RPT** *before*

TABLE 4.3
Attribute-Value Representation of Fully-Specified Common Hypotheses[5]

Rule	N-role	Rep-type	Bounds	# of reps	Prediction
HS1	selector	segment	last N	i	abcdCDef
HS2	selector	segment	first N	i	abcdABef
HS3	selector	instruction	Nth fm start	i	abcdBef
HN1*	nil	segment	all	i	abcdABCDef
HN2*	nil	instruction	prior	i	abcdDef
HC1	counter	segment	all	N	abcdABCDABCDef
HC2	counter	instruction	prior	N	abcdDDef
HC3	counter	segment	all following	N	abcdefEFEF

Test Program: abcdRPT2ef

[5]1) * rules do not use N; 2) Uppercase letters in predictions show executions under control of RPT2; 3) Underlined letters reflect ambiguity in "repeat twice."

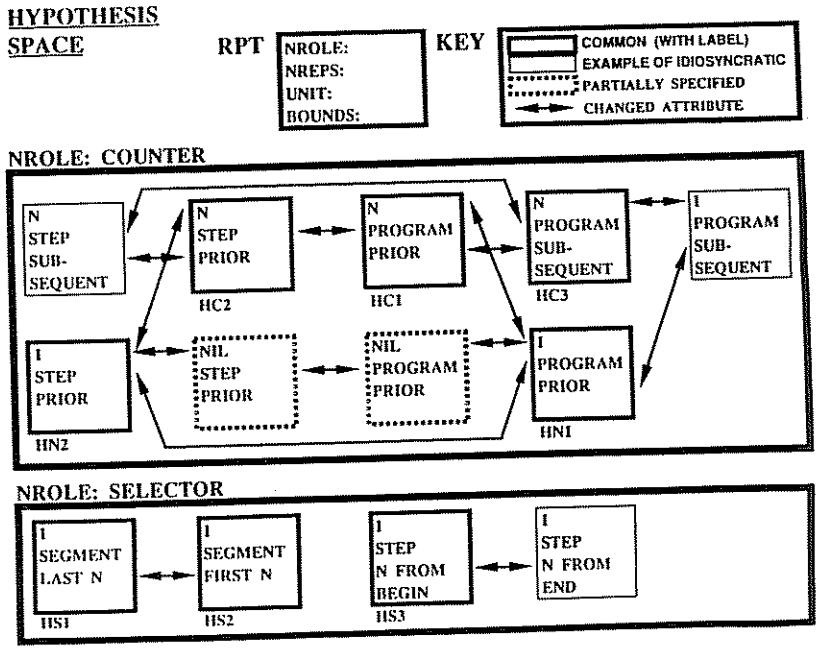


FIG. 4.2. Frames for hypotheses about how **RPT N** works. Heavy borders corresond to common hypotheses from Table 4.2; dashed borders correspond to partially specified hypotheses; arrows indicate a change in the value of a single attribute. (All possible hypotheses are not shown.)

actually using it in an experiment. This procedure enabled us to determine what frame is constructed by searching memory for relevant knowledge. No subject started off with the correct frame. Seventeen of the 20 subjects started with the *N*-role:*counter* frame. That is, subjects initially assume that the role of *N* is to specify the number of repetitions, and their initial hypotheses differed only in whether the repeated unit was the entire program or the single instruction preceding **RPT** (HC1 and HC2). This suggests that subjects drew their initial hypotheses by analogy from the regular command keys, where *N* determines the number of times that a command is executed.

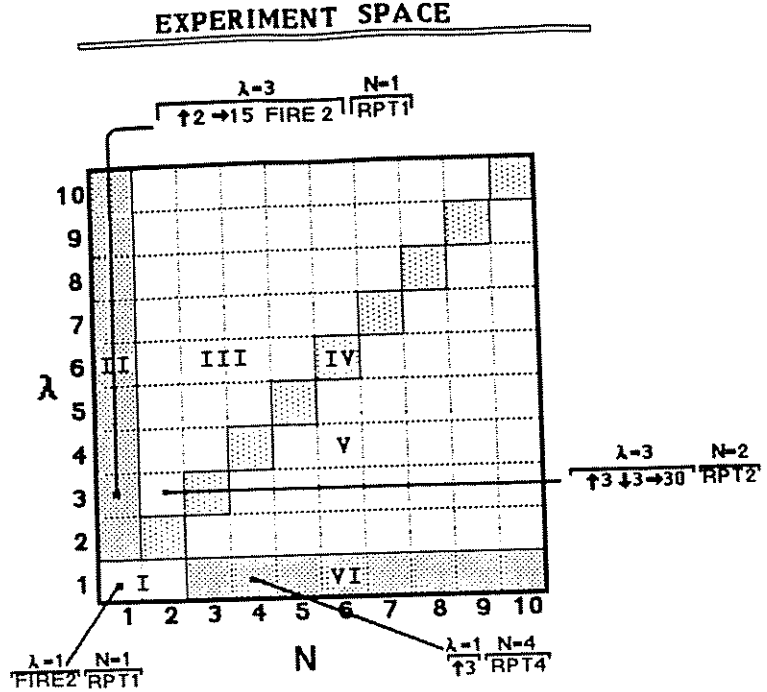
Having proposed their initial hypotheses, subjects then began to revise them on the basis of experimental evidence. Subjects eventually changed from an N-role:*counter* frame to the *N*-role:*selector* frame. Fifteen of the subjects made only one frame change, and four of the remaining five made three or more frame changes. This suggests that subjects were following very different strategies for searching the hypothesis space. We discuss strategic variation later in this chapter.

The Experiment Space

Subjects test their hypotheses by conducting experiments; by writing programs that include **RPT** and observing BigTrak's behavior. But it is not immediately obvious what constitutes a good or informative experiment. In constructing experiments, subjects are faced with a problem-solving task that parallels their effort to discover the correct hypotheses, except that in this case search is not in a space of hypotheses, but in a space of experiments.

A useful characterization of the experiment space is one that abstracts over the specific content of programs and refers to only two dimensions of their experiments. The first is the value of N—the argument that repeat takes. The second is λ—the length of the program preceding the **RPT.** Within the N-λ space, we identify six distinct regions according to the relative value of N and λ and their limiting values. The regions are depicted in Fig. 4.3, together with illustrative programs. At the bottom of the figure, we indicate which of the common hypotheses would be confirmed by experiments in each region. Here we define the regions and indicate the general consequences of running experiments in each.

- Region I. One-step programs $N = 1$ or 2, (e.g., **↑ 1 RPT 1,** or **↑ 1 RPT 2**). although an incrementalist strategy would suggest that this is a good starting place for exploring the experiment space, such exepriments are totally undiscriminating: as shown in Fig. 4.3, they produce behavior consistent with all but HC3 in Table 4.2. Furthermore, the ambiguous distinction between "repeat once" and "repeat twice," mentioned earlier, is exacerbated with a one-step program. Regardless of whether the value of N is 1 or 2, the command will be executed twice.
- Region II. Multistep programs with $N = 1$ (e.g., **↑ 1 FIRE 1 → 15 RPT 1**). Experiments in this region are consistent with hypotheses of the form "it repeats the previous step," such as HC2 and HN2. They rule out hypotheses that the entire program is repeated once (HN1) or N times (HC1).
- Region III. Programs with at least three instructions and a value of N less than λ and greater than 1 (e.g., **↑ 1 Fire 1 → 15 RPT 2**). As long as no two adjacent instructions are identical, programs in this region are consistent only with HS1 (the correct hypothesis). For example, the program [**↑ 2 → 15 FIRE 4 ← 30 RPT 3**] is inconsistent with every common hypothesis except HS1.
- Region IV. Here, $\lambda = N$ (e.g., **↑ 1 FIRE 1 → 15 RPT 3**). In addition to HS1, these experiments are consistent with hypotheses that **RPT** causes a repetition of the entire program (HN1), as well as with HS2 (Repeat first N steps once).
- Region V. In this region, N is greater than λ (e.g., **↑ 1 FIRE 1 → 15 RPT 5**). In this situation, BigTrak effectively sets N equal to λ, so experiments



		Region					
		I	*II*	*III*	*IV*	*V*	*VI*
Hypothesis	*HS1*	+	+	+	+	+	+
	HS2	+	–	–	+	+	+
	HS3	+	–	–	–	–	+
	HN1	+	–	–	+	+	+
	HN2	+	+	–	–	–	+
	HC1	+	–	–	–	–	–
	HC2	+	+	–	–	–	–
	HC3	–	–	–	–	–	–

FIG. 4.3. Regions of the Experiment Space, showing illustrative programs and confirmation/disconfirmation for each common hypothesis. (Shown here is only the 10 × 10 subspace for the ful 15 × 16 space.)

in this region tend to support the hypothesis that N is irrelevant and that HN1 is the correct hypothesis.

- Region VI. Experiments in this region have one-instruction programs with values of N greater than 2 (e.g., **FIRE 1 RPT 6**). This region is similar to Region V and also serves as the testing ground for hypotheses that N corresponds to the number of repetitions (HC1-HC3). These hypotheses are disconfirmed in this region, but some subjects persevere here nevertheless.

Other formulations are possible, but we will use the N–λ space in our analysis. We do not claim that subjects have this elaborated representation of the experiment space. Instead, it enables us to classify experiments according to the kinds of conclusions that they support.

Strategic Variation in Scientific Discovery: Theorists and Experimenters

As noted earlier, subjects started with the wrong frame; thinking that N functions as a counter. The most significant representational change occurred when subjects switched from the N-role:*counter* frame to the N-role:*selector* frame. Once subjects made this change, they quickly discovered how the **RPT** key works. Subjects used two different strategies to switch frames. Thirteen subjects were classified as experiment space searchers because they induced the correct frame from the result of an experiment in region III of the experiment space. For convenience, we refer to them as "Experimenters." The remaining seven subjects searched the hypothesis space for information to construct a frame that was consistent with the experimental data that they had observed. We call them "Theorists." Theorists did not have to conduct an experiment in region III of the experiment space to generate the correct frame.

Experimenters: General Strategy

Experimenters went through two phases. During the first, they explicitly stated the hypothesis under consideration, and conducted experiments to evaluate it. They proposed a number of hypotheses within the N-role:*counter* frame, however they eventually realized that the N-role:*counter* frame was inadequate and they switched to a search of the experiment space. In this second phase, Experimenters conducted experiments without explicit statement of an hypothesis. Prior to the discovery of how the **RPT** works, Experimenters conducted, on average, six experiments without statement of an hypothesis. Furthermore, these experiments were usually accompanied by statements about what would happen if N or λ were changed. By pursuing the approach of changing N and λ, Experimenters eventually conducted an experiment in region III of the experiment space. When the subjects conducted an experiment in this region, they noticed that the last N steps were repeated and proposed HS1—the correct rule.

Theorists: General Strategy

The strategy used by Theorists was to construct an initial frame, N-role: *counter*, and then to conduct experiments that tested the values of the frame. When they had gathered enough evidence to reject an hypothesis, Theorists switched to a new value of a slot in the frame. For example, a subject might switch from saying that the prior step is repeated N times to saying that the prior program is repeated N times. When a new hypothesis was proposed, it was always in the same frame, and it usually involved a change in only one attribute. These subjects eventually accumulated enough evidence to reject the N-role:*counter* frame entirely. Knowing that sometimes the previous step and sometimes the previous program was repeated, Theorists could infer that the unit of repetition was variable and that this ruled out all hypotheses in the N-role:*counter* frame—because those hypotheses all require a fixed unit of repetition. This realization enabled Theorists to constrain their search to an N-role that has a variable unit of repetition. As is shown in study 2, subjects can construct an N-role:*selector* frame without further experimentation. Following memory search, Theorists constructed the N-role: *selector* frame and proposed one of the hypotheses within it. They usually selected the correct one, but if they did not, they soon discovered it by changing one attribute of the frame as soon as their initial N-role:*selector* hypothesis was disproven.

Performance differences between Theorists and Experimenters are summarized in Table 4.4. The most important one is that Experimenters conduct more experiments than Theorists and that this extra experimentation is conducted without an explicit hypothesis statement. We have argued that this extra experimentation is indicative of searching the experiment space, and we have shown that Experimenters do indeed use more N–λ combinations than the Theorists. Furthermore, we have argued that instead of conducting a search of the experiment space, Theorists search the hypothesis space for an appropriate role for N. This is an important claim for which there was no direct evidence in the protocols. Our second study tests the hypothesis that it is possible to think of an N-role:*selector* hypothesis without exploration of the experiment space.

STUDY 2: HYPOTHESIS-SPACE SEARCH AND EXPERIMENTATION BY ADULTS

Our interpretation of subjects' behavior in Study 1 generated two related hypotheses: First, it should be possible for subjects to propose the correct rule without the benefit of any experimental outcomes. In Study 2, we tested this hypothesis by asking subjects to state not just one, but *several*, different ways that **RPT** might work, *before* doing any experiments. If subjects can think of the correct rule without any experimentation, then this can only be attributed to hypothesis space search because there is no experimental input. Second, if hypothesis-space search is unsuccessful, then subjects switch to a search of the experiment space. This hypothesis predicts that subjects who are unable to generate the correct rule in the hypothesis-space search phase will behave like the Experimenters of Study 1 and will discover the correct rule only after conducting an experiment in region III of the experiment space.

TABLE 4.4
Performance Summary of Experimenters and Theorists in Study 1

	Experimenters	Theorists	Combined
N	13	7	20
Time (minutes)	24.46	11.40	19.40
Experiments	18.38	9.29	15.20
Experiments with hypotheses	12.30	8.57	11.00
Experiments without hypotheses	6.08	0.76	4.2
Different hypotheses	4.92	3.86	4.55
Hypothesis switches	4.76	3.00	4.15
Experiment space verbalizations	5.85	0.86	4.10
/Vλ combinations used	9.9	5.7	8.45

Method

Ten Carnegie-Mellon undergraduates participated in this study. The familiarization part of Study 2 was the same as described for Study 1; subjects learned how to use all the keys except the **RPT** key. Familiarization was followed by two phases: hypothesis-space search and experimentation.

The hypothesis-space search phase began when the subjects were asked to think of various ways that the **RPT** key might work. In an attempt to get a wide range of possible hypotheses from the subjects, we used three probes in the same fixed order:

1. How do you think the **RPT** key might work?
2. We've done this experiment with many people, and they've proposed a wide variety of hypotheses for how it might work. What do you think they may have proposed?
3. When BigTrak was being designed, the designers thought of many different ways it could be made to work. What ways do you think they may have considered?

After each question, the subject responded with as many hypotheses as could be generated. Then the next probe was used. Once the subjects had generated all the hypotheses that they could think of, the experimental phase began: The subjects were allowed to conduct experiments while attempting to discover how the **RPT** key works.

Results and Discussion

Subjects proposed, on average, 4.2 different hypotheses. All but 2 subjects began with the *N*-role:*counter* frame, and 7 of the 10 subjects switched to the N-role:*selector* frame during Phase 1. The correct rule (HS1) was proposed by 5 of the 10 subjects. In the experimental phase all subjects were able to figure out how the **RPT** key works. Mean time to solution was 6.2 minutes, and subjects generated, on average, 5.7 experiments and proposed 2.4 different hypotheses.

The results of the hypothesis-space search phase of Study 2 show that it is possible for subjects to generate the correct hypothesis (among others) without conducting any experiments. This result is consistent with the view that the Theorists in Study 1 think of the correct rule by a search of the hypothesis space. The results of the experimental phase of Study 2 further support our interpretation of Study 1. All of the subjects who failed to generate the correct rule in the hypothesis-space search phase behaved like Experimenters in the experimental phase. They discovered the correct rule only after exploring region III of the experiment space. This finding is consistent with the view that when hypothesis-space search fails, subjects must turn to a search of the experiment space.

This study and the previous one have provided some initial answers to the question of how adults reason scientifically. The adults' performance provides a standard against which we can compare children's performance on the same task as was used in Study 1. Thus, in Study 3, children were given some initial training on how to use the BigTrak, and were then asked to find out how the **RPT** key works.

STUDY 3: SCIENTIFIC REASONING IN CHILDREN

As a result of our work with adults we can now pose some more specific questions than those outlined earlier. One set of questions deals with searching the hypothesis space. First, given the same training experience as adults, will children think of the same initial hypotheses as adults? If they do, then this would suggest that the processes used to construct an initial frame are similar in both adults and children. Second, when children's initial hypotheses are disconfirmed will the children assign the same values to slots as the adults? That is, are the processes that are used to search the hypothesis space similar in both adults and children? Finally, will children be able to change frames or will they remain in the same frame? Given that some adults—Theorists—were able to construct frames from a search of memory, will children be able to do so too? Failing that, will they be able to switch their strategy to a search of the experiment space—as did the experimenters, or will they stay within their initial frame?

Another set of questions concerns children's search of the experiment space. Children may search different areas of the experiment space than the adults, or they may even construct a different type of experiment space. Such a finding would suggest that the strategies used to go from an hypothesis to a specific experiment are different in adults and children. Another possibility is that children may evaluate the results of experiments in a different way from adults. Kuhn and her colleagues' work suggests that the ability to evaluate experimental evidence is one of the major differences in reasoning strategies

between adults and children. However, in her tasks, the opportunity for an interaction between data and theory is not present because the children cannot continually cycle from hypotheses to experiments.

Method

Subjects

Twenty-two third to sixth graders from a local private school participated in the study. All of the children had 45 hours of LOGO instruction prior to participating in this study. We selected this group partly as a matter of convenience, because they were participating in another study on the acquisition and transfer of debugging skills (Carver, 1986; Klahr & Carver, 1988). More importantly, because we will be contrasting the children's performance with adult subjects—all of whom had some programming experience—our subjects' experience provided at least a rough control for prior exposure to programming instruction. Furthermore, the subjects' age range (8;2 to 11;8) spans the putative period of the emergency of formal operational reasoning skills, the hallmark of which is, as noted earlier, the ability to "reason scientifically." Also, in a pilot study, we discovered that children with no programming experience had great difficulty understanding what was expected of them on the task.

Procedure

As in study 1, the subjects were taught how to use the BigTrak and were then asked to discover how the **RPT** key works. The session ended when the child stated that he or she was satisfied that he or she had discovered how the **RPT** key works, or could not figure out how it worked. Two procedural modifications facilitated working with the children. First, if the children did not spontaneously state what they were thinking about, the experimenter asked them how they thought the **RPT** key worked. Second, if a subject persisted with the same incorrect hypothesis and did exactly the same type of experiment (i.e., λ and N were not changed) four times in a row, the experimenter asked the child what the purpose of the number with the **RPT** key was.

Results

In this section, we first discuss the overall results. Then we describe the types of hypotheses and experiments that the children proposed. We also point to some of the more important differences between the strategies used by the children and the adults.

Only 2 of the 22 children discovered the correct rule. Fourteen children (including the 2 who were correct) asserted that they were absolutely certain that they had discovered how **RPT** works. Four gave up in confusion, and

4 thought that it worked in a particular way some of the time. The children spent, on average, 20 minutes trying to determine how the **RPT** key works. They generated an average of 13 programs. Of the 285 programs run by the subjects, 240 were experiments, 23 were control experiments, 1 was a calibration, and 21 were unclassifiable. Children proposed 3.3 different hypotheses during the course of a session. This is only about 1 less than the mean number of hypotheses proposed by adults; but as shown in the second columnn of Table 4.2, the relative frequency of experiments run under different hypotheses was very different. The following paragraphs discuss these differences.

Partial Hypotheses

Nearly 30% of the children's experiments were conducted under partial hypotheses, whereas adults specified all but 3% of their experiments fully (see Table 4.2). Of those experiments children conducted under partial hypotheses, 51% did not mention the unit of repetition (i.e., whether it was a step, a program, or a segment), and 49% did not mention the number of repetitions that should occur. This statement of partial hypotheses could be the result of differences in the children's ability to articulate fully specified hypotheses, or it could result from the fact that the children often did not regard the attributes of number of repetitions and the unit of repetition as being salient attributes of the **RPT** key. With respect to the number of repetitions, the latter interpretation is supported by the finding that the children often failed to type in a number after pressing the **RPT** key, indicating that they did not see a number as being a necessary part of the **RPT** command. With respect to the segments, the issue is unclear. In any event, by not stating the unit of repetition or the number of repetitions, the children are indicating that they consider these attributes of the hypothesis to be secondary.

Exploring Only One Frame

All of the 20 children who failed to discover how **RPT** works proposed hypotheses that were solely in the *N*-role:*counter* frame. Even though the children observed many experimental outcomes that were consistent with the *N*-role:*selector* frame and not with their current frame, none of the children were able to induce the selector frame. This suggests two things: First, the children did not have sufficient knowledge available to generate the *N*-role:*selector* frame by searching the hypothesis space. Second, the children did not use experiment-space search to induce a new frame. Instead, they used it to induce new slot values for their current frame. As a result, the children generated a number of hypotheses within the *N*-role:*counter* frame that were not generated by the adults.

Many of the children who originally had an hypothesis with *N*-role:*counter* abandoned it in favor of a nil role for *N* or invented a new number of repetitions to account for the data. Seventeen percent of their experiments were

conducted using one of these hypotheses (HC4 in Table 4.2). These hypotheses were generated when the children were trying to account for the finding that **RPT** 2 only repeats the prior program once, not twice. These children either said that *N* had no role, or tried to accommodate the number of repetitions slot to fit the data. The children stated that the program was repeated *N*-1 times, *N*/2 times, or stated that the value of *N* replaced the value that was bound to the previous command (e.g., FIRE 3 RPT 8 will do a FIRE 3 FIRE 8). No adult generated such hypotheses.

Another type of hypothesis that appeared only in the children's data was that the last two steps of the program were repeated *N* times. Three of the 22 children proposed this type of hypothesis after conducting an experiment in region III with $N = 2$. Thus, the children proposed an hypothesis that was within the *N*-role:*counter* frame, yet was consistent with the observation that the last two steps of a program were repeated.

Each of these hypotheses is a way of staying within the *N*-role:*counter* frame while accounting for the finding that there were not *N* repetitions of a command or a program. These hypotheses were generated even though there was a large amount of evidence available that could disconfirm both the individual hypotheses and the frame itself. However, the children were content with hypotheses that could account for the results of the most recent outcome. That is, local consistency was sufficient, and global inconsistency was ignored.

Search of the Experiment Space

One question that we raised earlier was whether children's search in the experiment space would be different from that of the adults. As can be seen from Table 4.5, the adults and children differed in the number of experiments run in regions I and V ($\chi^2_5 = 31.4$ $p < .05$). Children ran twice as many experiments as the adults in region I and about one third as many as the adults in region V. Experiments in region I confirm any hypothesis and merely show that something is repeated, without providing any information about number of repetitions or what is repeated. Experiments in region V suggest that *N* is irrelevant, because they repeat the entire program once, whatever the value of *N*.

Although two thirds of the adult experiments were distributed over the experiment space in exactly the same way as the children's experiments, the hypotheses that they induced from these experiments were quite different. In particular, both adults and children conducted 17% of their programs in the (potentially) highly informative region III of the experiment space. Adults were able to induce the correct rule from experiments in this region, whereas children were not. Adults and children also conducted the same amount of experiments in region II of the experiment space yet reached different conclusions. Adults induced the hypothesis that the previous step was repeated,

TABLE 4.5
Percentage of Programs in Each Area of the Experiment Space for Adults (Study 1) and Children (Study 3)

	I	II	III	IV	V	VI
Adults	15	25	17	10	20	13
Children	30	21	17	11	7	14

whereas the children did not; they maintained the hypothesis that it is the program that is repeated. In the following paragraphs we will explore these interactions between search of the Experiment and Hypothesis spaces in more detail.

Differences in Search Strategies

Only two children generated the *N*-role:*selector* frame, so it is difficult to classify the other 20 children as either Experimenters or Theorists according to the same criteria used in Study 1. The earlier classification was based on how subjects switched from one frame to another. Clearly, when subjects only use one frame it is impossible to make this categorization. However, even without this criterion we can see that all 20 of the children who failed to generate the correct hypothesis can be classified as a type of Experimenter. The children were within the *N*-role:*counter* frame and their search of the hypothesis space consisted of changing the values of the slots within the *N*-role:*counter* frame. This was achieved by searching the experiment space to find values for the number of repetitions slot within the frame.

While the children were searching the experiment space to induce new hypotheses, their search was different from the adults: The adults searched the experiment space once they had abandoned the *N*-role:*counter* frame and the goal of their search was to induce a new frame. In contrast, the children used experiments to find new slot values within a frame that they were reluctant to abandon. Some experiments, because they were in uninformative regions of the experiment space, did confirm their incorrect hypotheses. Others did not, but children responded to disconfirmation either by misobservation or by ignoring the results and running yet another experiment that they were sure would confirm their prediction. This indicates that while the children were exploring both the Hypothesis and the experiment space, their search of the Hypothesis space was limited; their search of the Hypothesis space was constrained to staying within one frame—the *N*-role:*counter* frame.

Summary

There were three main differences between adults and children. First, children proposed hypotheses that were different from adults. Furthermore, these different hypotheses were induced from the same type of data as the

adult's hypotheses. Second, the children did not abandon their current frame and search the Hypothesis space for a new frame, or use the results of experiment space search to induce a new frame. Third, the children did not attempt to check whether their hypotheses were consistent with prior data. Even when children knew that there was earlier evidence against their current hypothesis, they said that the device *usually* worked according to their theory.

The analysis of the children's search strategies, as well as the earlier analysis of the adult group, have begun to yield a complex picture of the different ways that subjects can use experiments. In order to fully interpret these differences, it is necessary to introduce a theoretical framework that further explicates the distinction between the hypothesis space and the experiment space as well as the coordination of search in the two spaces. In the next section, we turn to that theoretical extension. Following that, we return to the comparative interpretation of our findings in terms of the framework.

A DUAL-SEARCH MODEL OF SCIENTIFIC DISCOVERY

Our model of scientific reasoning is based on Simon and Lea's (1974) Generalized Rule Inducer (GRI). As noted earlier, in the GRI, concept formation tasks involve search in two problem spaces—a space of rules and a space of instances. Simon and his colleagues extended this original idea to the analysis of several important scientific discoveries (Kulkarni & Simon, 1988; Langley, Zytkow, Simon & Bradshaw, 1986), and we extended it to provide a framework for the interpretation of results from experimental studies of scientific reasoning in the laboratory. In this section, we describe our model of Scientific Discovery as Dual Search (SDDS), and in the following section we use SDDS as a basis for further discussion of developmental issues.

SDDS: Summary[6]

The fundamental assumption is that scientific reasoning requires search in two related problem spaces: an hypothesis space, consisting of the hypotheses generated during the discovery process, and an experiment space, consisting of all possible experiments that could be conducted. Search in the hypothesis space is guided both by prior knowledge and by experimental results. Search in the experiment space may be guided by the current hypothesis, and it may be used to generate information to formulate hypotheses.

SDDS consists of a set of basic components that guide search within and between the two problem spaces. Initial hypotheses are constructed by a series of operations that result in the instantiation of a frame (cf. Minsky, 1975) with default values. Subsequent hypotheses within that frame are generated

[6]See Klahr and Dunbar (1988) for more detail.

by changes in values of particular slots, and changes to new frames are achieved either by a search of memory or by generalizing from experimental outcomes. Three main components control the entire process from the initial formulation of hypotheses, through their experimental evaluation, to the decision that there is sufficient evidence to accept an hypothesis. The three components, shown at the top of the hierarchy in Fig. 4.4 are SEARCH HYPOTHESIS SPACE, TEST HYPOTHESIS, AND EVALUATE EVIDENCE.

SEARCH HYPOTHESIS SPACE

The goal of this process is to form a fully specified hypothesis, which provides the input to TEST HYPOTHESIS. This can be achieved in two ways. The first is by searching memory for a frame that could be used to generate an hypothesis (EVOKE FRAME). The second is by conducting experiments and inducing a new frame from the results of these experiments (INDUCE FRAME). Once a frame has been instantiated, the subject must assign specific values to the slots so that a specific hypothesis can be generated. Again, there are two ways that this can occur. One is by conducting further experiments to determine what the slot values should be (USE EXPERIMENTAL OUTCOMES), and the other is to fill in the slots with their default values (USE PRIOR KNOWLEDGE).

TEST HYPOTHESIS

TEST HYPOTHESIS generates an experiment appropriate to the current hypothesis, makes a prediction, then runs and observes the result of the experiment. Experiments are designed in the E-SPACE MOVE process. This process consists of selecting a central focus for the experiment and then setting values for this focus. Once this is set the values of the other aspects of the experiment can be assigned. The output of TEST HYPOTHESIS is a description of evidence for or against the current hypothesis, based on the match between the prediction derived from the current hypothesis and the actual experimental result.

EVALUATE EVIDENCE

EVALUATE EVIDENCE decides whether the cumulative evidence—as well as other considerations—warrants acceptance, rejection, or continued consideration of the current hypothesis.

GENERATE OUTCOME

This process consists of an E-SPACE MOVE, which produces an experiment, RUNNING the experiment and OBSERVING the result. As we noted earlier the E-SPACE MOVE also occurs as a subprocess within SEARCH HYPOTHESIS SPACE.

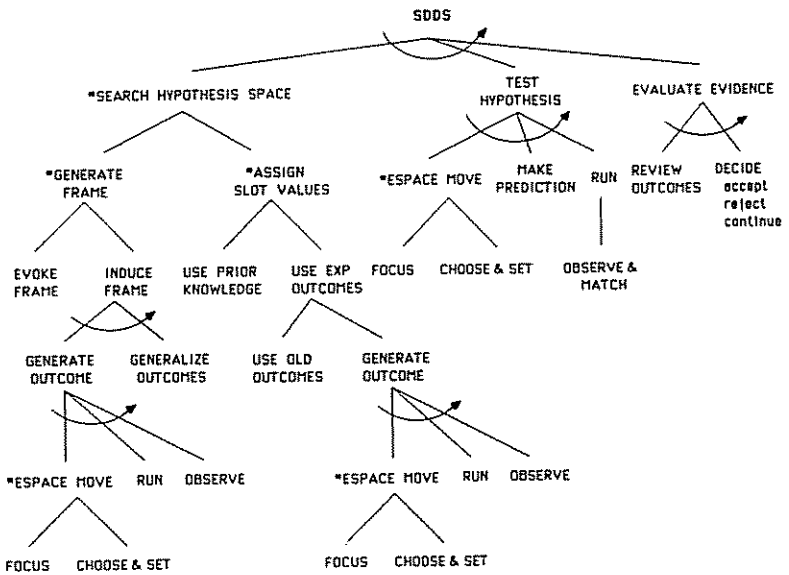


FIG. 4.4. Process hierarchy for SDDS. All subprocesses connected by an arrow are executed in a sequential conjunctive fashion. All process names preceded by an asterisk include conditional tests for which subprocess to execute.

E-SPACE MOVE

Experiments are designed by E-SPACE MOVE. The most imoprtant step is to FOCUS on some aspect of the current situation that the experiment is intended to illuminate. "Current situation" is not just a circumlocution for "current hypothesis," because there may be situations in which there is no current hypothesis, but in which E-SPACE MOVE must function nevertheless. (The multiple role played by experimentation is an important feature of the model, and is discussed further later.) If there is an hypothesis, then FOCUS determines that some aspect of it is the primary reason for the experiment. If there is a frame with open slot values, then FOCUS will select one of those slots as the most important thing to be resolved. If there is neither a frame nor an hypothesis—that is, if E-SPACE MOVE is being called by INDUCE FRAME—then FOCUS makes an arbitrary decision to focus on one aspect of the current situation.

Once the focal value has been determined, CHOOSE sets a value in the Experiment Space that will provide information relevant to it, and SET determines the values of the remaining, but less important, values necessary to produce a complete experiment.

Memory Requirements

A variety of memory requirements are implicit in our description of SDDS and must, by implication, play an important role in the discovery process. Here we provide a brief indication of the kinds of information about experiments, outcomes, hypotheses, and discrepancies that SDDS must store and retrieve.

- Recall that GENERATE OUTCOME operates in two contexts. Under INDUCE FRAME, it is called when there is no active hypothesis and when the system is attempting to produce a set of behaviors that can then be analyzed by GENERALIZE OUTCOMES in order to produce a frame. Therefore, SDDS must be able to represent and store one or more experimental outcomes each time it executes INDUCE FRAME.
- Another type of memory demand comes from EVALUATE EVIDENCE. In order to be able to weight the cumulative evidence about the current hypothesis, REVIEW OUTCOMES must have access to the results produced by MATCH in TEST HYPOTHESIS. This evidence would include selected features of experiments, hypotheses, predictions, and outcomes.
- Similar information is accessed whenever ASSIGN SLOT VALUES calls on USE PRIOR KNOWLEDGE or USE OLD OUTCOMES to fill in unassigned slots in a frame.

At this point in the model's development, the precise role of memory remains an area for future research.

The Multiple Roles of Experimentation in SDDS

Examination of the relations along all these processes and subprocesses, depicted in Fig. 4.4, reveals both the conventional and unconventional characteristics of the model. At the top level, the discovery process is characterized as a simple repeating cycle of generating hypotheses, testing hypotheses, and reviewing the outcomes of the test. However, below that level is a potentially complex interaction among the subprocesses. Of particular importance is the way in which E-SPACE MOVE occurs in three different places in the hierarchy:

1. As a subprocess deep with GENERATE FRAME, where the goal is to generate experimental evidence over which a frame can be induced. All of the Experimenters in study 1, and one of the children in study 3 used experiments for this purpose.
2. As a subprocess of ASSIGN SLOT VALUES where the purpose of the experiment is simply to resolve the unassigned slots in the current frame.

Both adults and children used this process, though it was used more extensively by children than by adults.

3. As a component of TEST HYPOTHESIS, where the experiment is designed to play its conventional role of generating an instance (usually positive) of the current hypothesis. This strategy was widely used by adults and children.

Note that the implication of the first two uses of E-SPACE MOVE is that in the absence of hypotheses, experiments can be used to generate hypotheses. Thus, experiments can be used for purposes other than the testing of hypotheses.

SDDS also elaborates the details of what can happen during the EVALUATE EVIDENCE process. Recall that three general outcomes are possible: the current hypothesis can be accepted, it can be rejected, or it can be considered further.

- In the first case, when there is sufficient evidence in favor of an hypothesis, the discovery process simply stops, and asserts that the current hypothesis is the true state of nature.
- In the second case, when an hypothesis has been rejected, the system returns to H-SPACE SEARCH, to either construct a new frame, or to fill in slot values of the currently active frame. If the entire *frame* has been rejected by EVALUATE EVIDENCE, then the model must attempt to generate a new frame using EVOKE FRAME. If the system cannot construct a new frame—as with the Experimenters and the children—then it will attempt to induce a new frame by running experiments. Having induced a new frame (which most of the children were unable to do), or having returned from EVALUATE EVIDENCE with a frame needing new slot values (i.e., a rejection of the hypothesis but not the frame), SDDS executes ASSIGN SLOT VALUES. Here too, if prior knowledge is inadequate to make slot assignments, the system may wind up making moves in the experiment space in an attempt to make the assignments. In both of these cases, the behavior would be the running of experiments without fully specified hypotheses. This was precisely what we saw in the second phase of the adult Experimenters' performance and for most of the children.
- In the third case, when there is not sufficient evidence to either accept or reject an hypothesis, SDDS returns to TEST HYPOTHESIS in order to further consider the current hypothesis. The experiments run in this context correspond to the conventional view of the role of experimentation. During MOVE IN E-SPACE, FOCUS selects particular aspects of the current hypothesis and designs an experiment to generate information about it.

DISCUSSION

As outlined earlier, one of the major goals in theories of cognitive development has been to tease apart the relation between the development of the knowledge base and the strategies that are applied to this knowledge base. In this chapter, we have recast these questions in terms of scientific reasoning as a search in two problem spaces. This approach allows us to make some initial observations about the components of the processes that show developmental trends. Our model shows that if the prior knowledge is not available, then subjects will resort to searching the experiment space (Study 2). Because children do not have the requisite knowledge that would enable them to construct the correct frame by searching the Hypothesis space, they, like the adults, must switch to a search of the experiment space. But when children search the Experiment space, their strategies are different from those used by the adults. Although the children conduct experiments that are similar to the adults, they induce different types of hypotheses and also evaluate evidence in different ways.

Different Experimental Strategies

Testing Hypotheses

Our model incorporates a goal that is central to the scientific process: testing hypotheses. The subjects also saw this as their goal. Over 70% of the experiments conducted by both the adults and the children were concerned with testing hypotheses. There were, however, some important differences in the hypothesis-testing strategies used by adults and children. Children often conducted a single experiment and then said that they had discovered how the device works, whereas adults conducted a number of experiments before they were convinced that an hypothesis was correct. Clearly, the criteria the children use for accepting hypotheses are very different from those used by adults.

The way children use disconfirming evidence differed substantially from that of adults. When an experiment produced disconfirming evidence, children attempted to conduct some new experiment that would confirm their hypothesis. Their goal was to generate some consistent outcomes, and their conclusion was that the device usually works the same way as their hypothesis. Thus, many of their experiments were designed to find evidence consistent with their hypothesis rather than to discover the correct hypothesis. Adults tended to be more sensitive to disconfirming evidence. Although adults did not abandon their hypothesis on the basis of a single disconfirming instance, they did attempt to understand inconsistencies. Children simply ignored them.

These findings are very similar to those reported by Kuhn et al. (1987). They found that when children have to judge what attributes of a ball make it produce a "good serve," they often proposed hypotheses that did not account

for all of the data and were content with saying that the attribute sometimes makes a difference. Kuhn, Amsel, and O'Loughlin (1988) also discovered that children found it difficult to determine what evidence was sufficient to reject their current hypothesis. They argued that one of the reasons that children find it difficult to evaluate hypotheses is that they do not have the ability to reflect upon a theory in the abstract. What their results and ours suggest is that in the EVALUATE EVIDENCE processes there are a number of subprocesses that bias interpretation toward the currently favored hypothesis. This may be due to an inability to remember previous outcomes or to the use of different subprocesses by adults and children.

Generating New Hypotheses

As our model indicates, another goal of experimentation is to generate new hypotheses when old ones have been disconfirmed. Again, there were many differences in how the children and adults did this. The adults tended to try only one or two hypotheses within a frame before abandoning the frame and switching to a search of the experiment space or searching memory for new frames. In contrast, all but two of the children stayed with the *N*-role:*counter* frame. These children proposed a number of hypotheses different from the adults as they attempted to reconcile experimental results with their hypotheses. They proposed a new hypothesis after only one experiment, they did not check to see if the results of the previous experiments were consistent with their hypothesis, and they were content with hypotheses that, from an adult's perspective, were highly implausible.

In terms of our model, these results suggest that the children's GENERATE OUTCOMES and GENERALIZE OUTCOMES processes do not include components specifying that a number of outcomes need to be generated and that the new hypothesis should be consistent with prior outcomes. Therefore, because of limitations in children's ability to GENERALIZE OUTCOMES, they tended to extract only the most local information from experiments. On the positive side, these results indicate that given a particular piece of experimental evidence, children are able to induce a rule that is consistent with the immediate result. Furthermore, children usually state the rule in a sufficiently abstract form so that it could account for a number of results. That is, they could state hypotheses in terms of any value of *N*, rather than in terms of the specific value that had been observed. However, although children of this age can induce new hypotheses from experimental data, the ability to correctly apply this inductive skill does not appear to be present.

Generating New Frames

Our adult Experimenters spent a considerable amount of time conducting experiments without an hypothesis in an effort to generate a new frame. The notable features about this strategy were that subjects usually conducted

three or four experiments before an hypothesis was proposed and that subjects proposed an hypothesis that was consistent with the results of the previous few experiments. Finally, the hypotheses that they proposed were plausible. Children rarely used this strategy. Recall that only 2 of the 22 children managed to evoke the correct frame from prior knowledge or induce it from experimental outcomes. It is clear that children rarely took the first main branch of SEARCH HYPOTHESIS SPACE once they had generated their initial frame.

Children's failure to propose more than one frame (*N*-role:*counter*), indicates that one of the major differences between adults and children is in the way that the results of previous experiments are used to evaluate evidence and to make new inductions. First, children did not use the information available to them to abandon their current frame. Rather, they spent much of their time using experimental results to ASSIGN SLOT VALUES to the *N*-role:*counter* frame. This suggests that either the children did not have the prior knowledge available to construct a new frame, or they could not deduce that the experimental evidence available disproved that the role of *N* was a counter, thereby allowing them to abandon that frame. A second major difference was that the types of inductions that the children generated from the data were not constrained by the results of prior experiments. Even those children who did discover that a segment of the program is repeated persisted in stating that the segment is repeated *N* times. The children either were unable to abandon their current frame, or did not have the knowledge available to construct a new frame that would be consistent with their results.

One of the central components of the previous analysis has been the idea that subjects search for information to construct frames. This search for new frames could occur in two ways. One way that subjects might construct a new frame is to search memory for information that allows them to construct a frame. This search process would be constrained by the problem specification, and by the results of prior experiments. A second possible way is to make some minor modification to a preexisting frame that already meets the task specifications. In the domain of machine learning, this idea has been used by Shrager (1985, 1987), and Falkenhainer (1987). Our model does not distinguish between these two possible ways of constructing frames, and subjects may have used either. Furthermore, it is possible that adults, having more knowledge available, may be able to import frames from other domains more readily than children.

Scientific Reasoning Skills: What Develops?

It depends. The developmental story that is beginning to emerge has several layers. At the level of subjects' global behavior on this task, there is little difference between the children and the adults. Both groups clearly understand the nature of the task and realize that they can only discover how the device works by making it behave, observing that behavior, and generating a summary statement that captures the behavior in a universal and general fashion.

That is, both the children and the adults know what the scientific reasoning process is supposed to look like. However, viewed at the level of overall success rates, there are profound differences in the consequences of how this general orientation toward discovery is implemented. The adults had a 95% success rate, whereas 90% of the children failed. These differences do not lie in the ability to generate informative experiments, for, as we saw earlier, there were few differences in the regions of the E-space that were visited by children and adults. There appears to be a crucial difference in the *reason* that those experiments were generated and in the *inductions* that are made from the results of those experiments. In terms of the model, children tended to move in the E-space in order to generate some data to patch a faulty hypothesis or to produce a desired effect, whereas adults used E-space search to generate a data pattern over which they could induce a new frame. With respect to inductive differences, we discovered that although all the children could induce new hypotheses from experiments, none of them were able to use an experimental result to induce a new frame. Inductions were local rather than global.

Another possible reason for these differences is knowledge about how to evaluate hypotheses. More specifically, children tend to have much less stringent criteria for evaluating evidence than adults. Two consequences of these lax criteria are that children accept hypotheses on the basis of incomplete evidence and that they maintain them in the face of much inconsistency. As we argued earlier, successful performance on this task depends on memory for previous experimental results. Children appear to lack the knowledge that the results of earlier experiments must be considered when evaluating an hypothesis. Research on designing factorial experiments (Siegler & Liebert, 1975) has shown that many children do not spontaneously realize that they must keep track of the results of experiments. Kuhn et al. (1988) also argued that children do not have the metacognitive skills available to properly evaluate evidence. Thus, children's ability to test hypotheses will not be the same as adults until they are able to utilize such information.

Conclusion

We have proposed that scientific reasoning requires search in two problem spaces and that the different strategies that we observed in children and in adults are caused by different patterns of search in these two problem spaces. We proposed SDDS as both a framework for interpreting these results and as a general model of scientific reasoning. Clearly, there are many aspects of the scientific reasoning process that we still do not fully understand, but we believe that SDDS offers a potentially fruitful framework for further exploration.

POSTSCRIPT: ACKNOWLEDGMENTS TO HERBERT SIMON

We are pleased to include this work in a Simon Festschrift, because his influence is evident in nearly every important aspect: in the focus on scientific discovery (Langley et al., 1986), in the methodology of verbal protocol analysis (Ericsson & Simon 1984), in the conceptualization of scientific discovery as search in two spaces (Simon & Lea 1974; Simon, 1977), and most fundamentally, in the assumption that the scientific discovery process is subject to systematic investigation. As Simon (1986) recently commented on his own research program:

> The hypothesis that drives this research is that scientific discovery is a problem-solving activity like other problem-solving activities that human beings engage in, using the same basic information-processing mechanisms that have been identified in those other processes. This hypothesis rests, in turn, on the belief that the scientist does not stand outside the lawful scheme of Nature; he is part of that scheme, and it is an important goal of scientific research to undestand his mental processes, just as it is to understand the processes of a star, an atom, or a cell. (p. 168)

Indeed, Simon's pervasive influence is disquieting, for it threatens our need to believe that we have made our own modest but unique contribution to the area. We are at least gratified to know that Herb has stayed away from developmental studies in this area. And, for a while, we felt that we were unique in initiating experimental studies within the "scientific discovery as search" view, because Simon's work on computational models of the discovery process was confined to the analysis of the historical record of practicing scientists. However, he has recently extended his work on scientific discovery to the experimental laboratory as well! It is difficult enough to stand on the shoulders of giants, but when they persistently expand the frontiers of knowledge, it is a daunting task to keep ones eyes fixed on new discoveries. For such a challenge, we are deeply indebted to Simon.

ACKNOWLEDGMENTS

This study was supported in part by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under contract No. N00014-86K-0349. Reproduction in whole or in part is permitted for any purpose of the United States Government.

REFERENCES

Bartlett, F. C. (1958). *Thinking.* New York: Basic Books.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1962). *A study of thinking.* (science ed.). New York.

Carey, S. (1984). Cognitive development: The descriptive problem. In M. S. Gazzaniga (Ed.), *Handbook of cognitive neurology.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.
Carver, S. M. (1986). *Transfer of LOGO debugging skill: Analysis, instruction and assessment.* Unpublished doctoral dissertation, Carnegie-Mellon University Pittsburgh.
Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology, 6,* 544–573.
Cohen, P. R., & Feigenbaum, E. A. (Eds.). (1983). *Handbook of artificial intelligence* (Vol. 3). Los Altos, CA: W. Kaufman, Inc.
Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.
Falkenhainer, B. (1987). Scientific theory formation through analogical inference. In *Proceedings of the 4th International Workshop on Machine Learning.* Los Altos, CA: Morgan Kaufmann.
Flavell, J. H. (1977). *Cognitive development.* Englewood Cliffs, NJ: Prentice-Hall.
Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*(2), 216–229.
Harre, R. (1983). *Great scientific experiments: Twenty experiments that changed our view of the world.* New York: Oxford University Press.
Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.
Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition, 3,* 195–212.
Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88,* 197–227.
Klahr, D., & Carver, S. M. (in press). Cognitive objectives in a LOGO debugging curriculum: Instruction, learning, and transfer. *Cognitive Psychology.*
Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*(1), 1–55.
Kuhn, D., & Phelps, E. (1982). The development of problem solving strategies. In H. W. Reese (Ed.), *Advances in child development and behavior.* New York: Academic Press.
Kuhn, D., Amsel, E., & O'Loughlin, M. (1988) *The development of scientific thinking skills.* Orlando, FL: Academic Press.
Kulkarni, D., & Simon, H. A. (in press). The processes of scientific discovery: The strategy of experimentation. *Cognitive Science.*
Langley, P., Bradshaw, G. L., & Simon, H. A. (1983). Rediscovering chemistry with the BACON system. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach.* Palo Alto, CA: Tioga.
Langley, P., Zytkow, J. M., Simon, H. A., & Bradshaw, G. L. (1986). The search for regularity: Four aspects of scientific discovery. In R. S. Michaliski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (Vol. 2). Los Altos, CA: Morgan Kaufmann.
Lenat, D. (1977). On automated scientific theory formation: A case study using the AM program. In J. E. Hayes, D. Michie, & L. Mikulich (Eds.), *Machine Intelligence 9.* New York: Halsted.
Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision.* New York: McGraw-Hill.
Piaget, J. (1928). *The child's conception of the world.* Boston: Routledge & Kegan Paul.
Piaget, J. (1952). *The origins of intelligence in children.* New York: International University Press.
Shrager, J. (1985). *Instructionless learning: Discovery of the mental model of a complex device.* Doctoral dissertation, Department of Psychology, Carnegie-Mellon University.
Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning, 2,* 247–276.
Shrager, J., & Klahr, D. (1986). Instructionless learning about a complex device. *International Journal of Man-Machine Studies, 25,* 153–189.

Siegler, R. S., & Liebert, R. M. (1975). Acquisition of formal scientific reasoning by 10- and 13-year-olds: Designing a factorial experiment. *Developmental Psychology, 11,* 401–402.
Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of Science, 40*(4), 471–480.
Simon, H. A. (1977). *Models of discovery.* Dordrecht-Holland: D. Reidel Publishing.
Simon, H. A. (1986). Understanding the processes of sciences: The psychology of scientific discovery. In T. Gamelius (Ed.), *Progress in sciences and its social conditions.* Oxford, England: Pergamon Press.
Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In L. W. Gregg (Ed.), *Knowledge and cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1–10.
Vygotsky, L. (1934). *Thought and language.* New York: Wiley.
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20,* 273–281.