Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114, 285–310.

Tsai, Y. (1983). Movements of attention across the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 523–530.

Ward, R., & McClelland, J. L. (1989). Conjunctive search for one and two identical targets. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 664–672.

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided Search: An alternative to the Feature Integration Model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15, 419–433.

(Accepted May 11, 1992)

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

COGNITIVE PSYCHOLOGY

Published quarterly by Academic Press, Inc., 6277 Sea Harbor Drive, Orlando, FL 32887-4900. Number of issues published annually: 4. Editor: Dr. Douglas L. Medin, Department of Psychology, Northwestern University, Swift Hall, 2029 Sheridan Road, Evanston, IL 60208.
Owned by Academic Press, Inc., 1250 Sixth Avenue, San Diego, CA 92101. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.
Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securities of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.
Total no. copies printed: average no. copies each issue during preceding 12 months: 2321; single issue nearest to filing date: 2355. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 1858; single issue nearest to filing date: 1869. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 0; single issue nearest to filing date: 0. Free distribution by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 52; single issue nearest to filing date: 52. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 1910; single issue nearest to filing date: 1921.

(Signed) Evelyn Sasmor, Director

COGNITIVE PSYCHOLOGY 25, 111–146 (1993)

Heuristics for Scientific Experimentation: A Developmental Study

DAVID KLAHR AND ANNE L. FAY

Carnegie Mellon University

AND

KEVIN DUNBAR

McGill University, Montreal, Quebec, Canada

Scientific discovery involves search in a space of hypotheses and a space of experiments. We describe an investigation of developmental differences in the search constraint heuristics used in scientific reasoning. Sixty-four subjects (technically trained college students, community college students with little technical training, 6th graders, and 3rd graders) were taught how to use a programmable robot. Then they were presented with a new operation, provided with a hypothesis about how it might work, and asked to conduct experiments to discover how the new operation really did work. The suggested hypothesis was always incorrect, as subjects could discover if they wrote informative experiments, and it was either plausible or implausible. The rule for how the unknown operation actually worked was either very similar or very dissimilar to the given hypothesis. Children focused primarily on plausible hypotheses, conducted a limited set of experiments, designed experiments that were difficult to interpret, and were unable to induce implausible (but correct) hypotheses from data. Adults were much better than children in discovering implausible rules. The performance deficits we found were not simply the result of children's inadequate encoding or mnemonic skills. Instead, the adults appear to use domain-general skills that go beyond the logic of confirmation and disconfirmation and deal with the coordination of search in two spaces: a space of hypotheses and a space of experiments.

We thank the children, parents, teachers, and administrators of Community Day School in Pittsburgh as well as the late Alan Spencer of the Community College of Allegheny County for their enthusiastic cooperation with this project. Thanks to Kate McGilly, Tom Ward, Doug Medin, and an anonymous reviewer for comments on earlier drafts and to Susan Glotfelty for meticulous assistance with protocol transcription and data analysis. Funds supporting this research were provided to the first author in part by the National Institute of Child Health and Human Development (R01-HD25211) and the A. W. Mellon Foundation, to the second author by a Post Doctoral Fellowship from the James S. McDonnell Foundation, and to the third author by Grant OGP0037356 from the Natural Sciences and Engineering Research Council of Canada. Correspondence and reprint requests should be addressed to David Klahr, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213.

0010-0285/93 $9.00

INTRODUCTION

Scientific discovery requires the integration of a complex set of cognitive skills, including the search for hypotheses via induction or analogy, the design and execution of experiments, the interpretation of experimental outcomes, and the revision of hypotheses (Klahr & Dunbar, 1988). There are two long-standing disputes about the developmental course of these skills: (a) the "child-as-scientist" debate asks whether or not it makes sense to describe the young child as a scientist; (b) the "domain-specific or domain-general" debate revolves around the appropriate attribution for whatever differences in children and adults may exist. The first issue is controversial because, although there is considerable evidence that young children possess some rudiments of scientific reasoning (Brewer & Samarapungavan, 1991; Karmiloff-Smith, 1988), there appears to be a long and erratic course of development, instruction, and experience before the component skills of the scientific method are mastered, integrated, and applied reliably to a wide range of situations (Fay, Klahr, & Dunbar, 1990; Kuhn, Amsel, & O'Loughlin, 1988; Mitroff, 1974; Kern, Mirels, & Hinshaw, 1983; Siegler & Liebert, 1975).

The second issue derives from a lack of consensus about the extent to which developmental differences in performance on scientific reasoning tasks results from domain-specific or domain-general acquisitions. This issue is analogous to questions in the memory development literature about the relative roles of content knowledge and broader mnemonic skills in accounting for age-related improvements in memory performance (Chi & Ceci, 1987). On the one hand, acquisition of domain-specific knowledge influences not only the substantive structural knowledge in the domain (by definition) but also the processes used to generate and evaluate new hypotheses in that domain (Carey, 1985; Keil, 1981; Wiser, 1989). On the other hand, in highly constrained discovery contexts, young children correctly reason about hypotheses and select appropriate experiments to evaluate them, even when the context is far removed from any domain-specific knowledge, (Sodian, Zaitchik, & Carey, 1991).

The two principal ways (aside from direct instruction) in which children acquire such domain-specific knowledge are observation and experimentation. Analysis of children's performance as *observational* scientists is exemplified by Vosniadou and Brewer's (in press) investigations of children's mental models of the earth. Such studies involve assessments of children's attempts to integrate their personal observations (e.g., the earth looks flat) with theoretical assertions conveyed to them by adults and teachers (e.g., the earth is a sphere). Similarly, children's understanding of illness concepts (see Hergenrather & Rabinowitz, 1991) is based primarily on their observations in the domain, rather than on their experiments. Issues of experimental design do not arise in this context. *Experimental* science adds to the demands of observational science the burden of formulating informative experiments. Studies investigating young children's ability to design factorial experiments (Case, 1974; Siegler & Liebert, 1975) focus on experimental aspects of science, as do studies of children's performance in experimental microworlds (e.g., Schauble, 1990).

We have approached the study of developmental differences in scientific reasoning by attempting to disentangle these different aspects of scientific discovery, while using a context that provides a plausible laboratory microcosm of real-world scientific discovery. We view scientific discovery as a type of problem solving (Klahr & Dunbar, 1988; Simon, 1977) in which domain-general heuristics for constraining search in a problem space play a central role. In this paper, we describe a study that illustrates some important developmental differences in subjects' use of several domain-general search heuristics. We compare the ability of children and adults to reason in a context designed to simulate some of the key problems faced by an *experimental* scientist. In our task, subjects' domain-specific knowledge biases them to view some hypotheses as plausible and others as implausible. However, they must rely on domain-general heuristics to guide them in designing experiments. In summary, our focus is on developmental differences in domain-general heuristics for experimental design, in a context where domain-specific knowledge influences the plausibility of different hypotheses.

Components of Scientific Reasoning

We view scientific discovery as a problem-solving process involving search in two distinct, but related, problem spaces. Our work is based on Klahr and Dunbar's (1988) SDDS framework (*S*cientific *D*iscovery as *D*ual *S*earch), which elucidates a set of interdependent processes for coordinating search in a space of experiments and a space of hypotheses. The three main processes are

1. Searching the hypothesis space. SDDS characterizes the process of generating new hypotheses as a type of problem-solving search, in which the initial state consists of some knowledge about a domain, and the goal state is a hypothesis that can account for some or all of that knowledge in a more concise or universal form. Several mechanisms have been proposed to account for the way in which initial hypotheses are generated. These include memory search, analogical mapping, remindings, and discovery of effective representations (Dunbar & Schunn, 1990; Gentner, 1983; Gick & Holyoak, 1983; Kaplan & Simon, 1990; Klahr & Dunbar, 1988; Ross, 1984; Shrager, 1987). Each of these mechanisms emphasizes a different aspect of the way in which search in the hypothesis space is initiated and constrained.

Once generated, hypotheses are evaluated for their initial plausibility. Expertise plays a role here, as subjects' familiarity with a domain tends to give them strong biases about what is plausible in the domain. Plausibility, in turn, affects the order in which hypotheses are evaluated: highly likely hypotheses tend to be tested before unlikely hypotheses (Klayman & Ha, 1987; Wason, 1968). Furthermore, subjects may adopt different experimental strategies for evaluating plausible and implausible hypotheses.

2. *Searching the experiment space.* Hypotheses are evaluated through experimentation. But it is not immediately obvious what constitutes a "good" or "informative" experiment. In constructing experiments, subjects are faced with a problem-solving task paralleling their search for hypotheses. However, in this case search is in a space of experiments rather than in a space of hypotheses. Ideally, experiments should discriminate among rival hypotheses. Subjects must be able to plan ahead by making predictions about which experimental results could support or reject various hypotheses. This involves search in a space of experiments that is only partially defined at the outset. Constraints on the search must be added during the problem-solving process.

One of the most important constraints is to produce experiments that will yield interpretable outcomes. This, in turn, requires domain-general knowledge about one's own information-processing limitations, as well as domain-specific knowledge about the pragmatic constraints of the particular discovery context. Furthermore, utilization of this knowledge to design experiments capable of producing interpretable outcomes requires a mapping from hypotheses to experiments and an ability to predict what results might occur.

3. *Evaluating evidence.* This involves a comparison of the predictions derived from a hypothesis with the results obtained from the experiment. Compared to the binary feedback provided to subjects in the typical psychology experiment, real-world evidence evaluation is not so straightforward. Relevant features must first be extracted, potential noise must be suppressed or corrected, and the resulting internal representation must be compared with earlier predictions. Theoretical biases influence not only the strength with which hypotheses are held in the first place—and hence the amount of disconfirming evidence necessary to refute them—but also the features in the evidence that will be attended to and encoded (Wisniewski & Medin, 1991).

Each of the components listed above is a potential source of developmental change, and most investigators have studied them in isolation. For example, classic concept learning studies (Bruner, Oliver, & Greenfield, 1966) focus on hypothesis formation and evaluation, but do not require subjects to design experiments. In contrast, studies of children's ability to design factorial experiments (Siegler & Liebert, 1975) do not require them to formulate and evaluate hypotheses. Finally, studies of children's ability to decide which of several hypotheses is supported by evidence focus on evidence evaluation, while suppressing both hypothesis formation and experimental design (i.e., Shacklee & Paszek, 1985). We have approached the study of scientific reasoning by using tasks that require coordinated search in *both* the experiment space and the hypothesis space, as well as the evaluation of evidence produced by subject-generated experiments. Rather than eliminating search in either space, we have focused on the coordination of both, because we believe that it is an essential aspect of scientific reasoning.

ASSESSING DIFFERENCES IN DOMAIN-GENERAL EXPERIMENTAL DESIGN HEURISTICS

In this paper, we focus on developmental differences in the heuristics used to constrain search in the experiment space. We were interested in the extent to which such heuristics would vary according to age, amount of formal scientific training, and the plausibility of the hypotheses under investigation. Although most studies demonstrate that subjects tend to attempt to confirm, rather than disconfirm, their hypotheses (cf. Klayman & Ha, 1987), such studies typically use hypotheses about which subjects have no strong prior beliefs about plausibility or implausibility. In contrast, we used a context in which plausibility played an important role.

Results from earlier investigations (Dunbar & Klahr, 1989) suggested that, in the domain in which we planned to test them, subjects at all ages and technical levels would be likely to share *domain-specific* knowledge that would bias them in the same direction with respect to the relative plausibility of different hypotheses. This allowed us to determine how search in the experiment space was influenced by the hypothesis plausibility. We expected the effects of age and scientific training to reveal differences in the *domain-general* heuristics used to constrain search in the experiment space. Such domain-general heuristics might include rules for effecting normative approaches to hypothesis testing as well as pragmatic rules for dealing with processing limitations in encoding, interpreting, and remembering experimental outcomes.

Subjects

Four subject groups participated: 12 Carnegie Mellon (CM) undergraduates, 20 community college (CC) students, 17 "sixth" graders (a mixed class of fifth to seventh graders, mean age 11 years) and 15 third graders (mean age, 9 years). The adult groups were selected to contrast subjects with respect to technical and scientific training. Sixth graders were selected because they represent the age at which many of the components of "formal reasoning" are purported to be available, and the third graders were chosen because pilot work had indicated they were the youngest group who could perform reliably in our task. In addition, the two younger groups match the ages of children studied in many other investigations of children's scientific reasoning skills (e.g., Kuhn et al., 1988).

CMs were mainly science or engineering majors who received partial course credit for

participation. They reported having taken about two programming courses, and they rated themselves between average and above average on technical and scientific skills. All CCs were nonscience majors (General Studies, Para-legal, Communications, Pre-nursing, etc.). They were recruited by posted advertisements and were paid for their participation. CCs had little training in mathematics or physical sciences beyond high school, and less than half of them had taken a college course in Biology or Chemistry. While 70% of them had used computer-based word processors and 45% had used spreadsheets, only 3 of the 20 had ever taken a programming course.

Children were volunteers from an urban private school and came primarily from academic and professional families. They were selected to be young "equivalents" of the CMs with respect to both the likelihood of ultimately attending college and age-appropriate computer experience. All sixth graders had at least 6 months of Logo experience, and most had more than a year of experience. All but one of the third graders had at least 1 month of Logo, with the majority having 6 months to a year of experience. Note that CCs had less programming experience than the third graders.

The BT Microworld

We used a computer microworld—called BT[1]—in which subjects enter a sequence of commands to a "spaceship" which then responds by carrying out various maneuvers. The discovery context was established by first instructing subjects about all of BT's basic features and then asking them to extend that knowledge by discovering how a new—and uninstructed—function works in the microworld. Subjects proposed hypotheses and evaluated them by experimenting, i.e., by writing programs to test their hypotheses.

The BT interface is shown in Fig. 1. The spaceship moves around in the left-hand panel according to instructions that are entered in its memory when subjects "press" (point and click) a sequence of keys on the keypad displayed on the right. The basic execution cycle involves first clearing the memory and returning BT to "base" with the CLR/HOME key and then entering a series of up to 16 instructions, each consisting of a function key (the command) and a 1- or 2-digit number (the argument). The five command keys are: ↑, move forward; ↓, move backward; ←, turn left; →, turn right; and FIRE. When the GO key is pressed BT executes the program. For example, one might press the following series of keys:

CLR ↑5 ←7 ↑3 →15 FIRE 2 ↓8 GO.

When the GO key was pressed, BT would move forward 5 units, rotate counterclockwise 42° (corresponding to 7 min on an ordinary clock face), move forward 3 units, rotate clockwise 90°, fire (its "laser cannon") twice, and backup 8 units. The time to enter and execute a program depends on the number of instructions and the value of their parameters. The program listed above would take approximately 20 s to enter and about 40 s to execute.

Procedure

The study had three phases. In the first, subjects were introduced to BT and instructed on the use of each basic command. During this phase, the display did not include the RPT key shown in Fig. 1. Subjects were trained to criterion on how to write a series of commands to accomplish a specified maneuver. In the second phase, subjects were shown the RPT key.

[1] BT is the simulated version of the "BigTrak" toy robot initially used by Shrager & Klahr (1986). One version of BT, written in LISP and run on a Xerox Dandelion workstation, was used by the CMs and the children, and the other, written in cT for the Apple MacII, was used by the CCs. On the Dandelion the display shown in Fig. 1 was approximately 11 × 14 in.; on the MacII, it was 7 × 9 in.

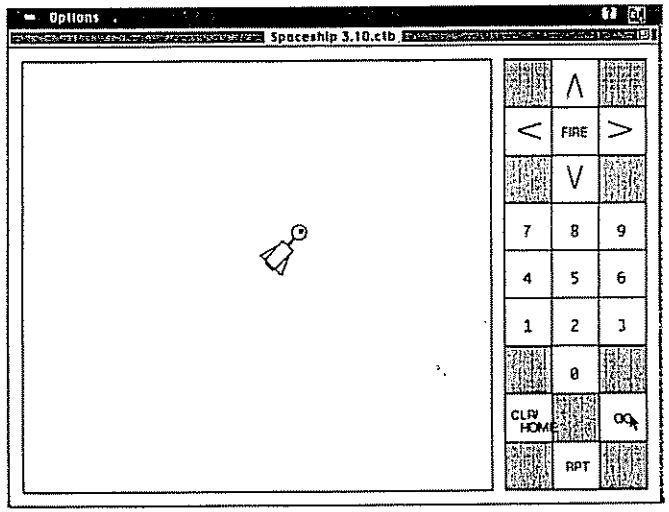

FIG. 1. The BT interface. Subjects enter commands by pressing the keypad on the right, and the BT "spaceship" maneuvers in the panel on the left.

They were told that it required a numeric parameter (N) and that there could be only one RPT N in a program. They were told that their task was to find out how RPT worked by writing at least three programs and observing the results. At this point, the Experimenter suggested a specific hypothesis about how RPT might work.

> One way that RPT might work is: (one of the four hypotheses described in the next section). Write down three good programs that will allow you to see if the repeat key really does work this way. Think carefully about your program and then write the program down on the sheet of paper. . . . Once you have written your program down, I will type it in for you and then I will run it. You can observe what happens, and then you can write down your next program. So you write down a program, then I will type it in, and then you will watch what the program does. I want you to write three programs in this way.

Next, the third—and focal—phase began. Subjects wrote programs (experiments) to evaluate the given hypothesis. After each program had been written, but before it was run, subjects were asked to predict the behavior of BT. Subjects had access to a record of the programs they had written (but not to a record of BTs behavior).

Subjects were instructed to give verbal protocols. This gave us a record of (a) what they thought about the kinds of programs they were writing while testing their hypotheses, (b) what they observed and inferred from the device's behavior, and (c) what their hypotheses were about how RPT actually worked. When subjects had written, run, and evaluated three experiments, they were given the option of either terminating or writing additional experiments if they were still uncertain about how RPT worked. The entire session lasted approximately 45 min.

Task Analysis

The BT hypothesis space. In previous studies with adults and grade school children (Klahr & Dunbar, 1988), we found that there were two very "popular" hypotheses about the effect of RPT N in a program:

A: Repeat the entire program N times,
B: Repeat the last step N times.

Subjects devoted a large proportion of their effort to exploring these two hypotheses. In contrast, there were two hypotheses that subjects were unlikely to propose at the outset:

C: Repeat the Nth step once,
D: Repeat the last N steps once.

The preference for A and B and the disinclination to propose C and D was found at all ages.

These four hypotheses about RPT N (as well as many others) can be represented in a space of "frames" (Minsky, 1975). The basic frame consists of four slots, corresponding to four key attributes: (1) the role of N; does it *count* a number of repetitions (as in A and B) or does it *select* some segment of the program to be repeated (as in C and D)? We call A and B *Counter* hypotheses and C and D *Selector* hypotheses. (2) The unit of repetition; is it a step (as in B and C), the entire program (as in A), or a group of steps (as in D)? (3) Number of repetitions; 1, N, some other function of N, or none? (4) Boundaries of repeated segment; beginning of program, end of program, Nth step from beginning, or end? Of the four slots, N-role is the most important, because a change in N-role from *Counter* to *Selector* mandates a change in several other attributes. For example, if N-role is *Counter*, the number of repetitions is N, whereas, if N-role is *Selector*, then number of repetitions is 1.

The BT experiment space. Subjects could test their hypotheses by conducting experiments, i.e., by writing programs that included RPT and observing BT's behavior. The BT experiment space can be characterized in many ways: the total number of commands in a program, the location of RPT in a program, value of N, the specific commands in a program, the numerical arguments of specific commands, and so on. (For example, counting only commands, but not their numerical arguments, as distinct, there are over 30 billion distinct programs [5^{15}] that subjects could choose from for each experiment. Even if we consider only programs with 4 or fewer steps, there are nearly 800 different experiments to choose from [$5^4 + 5^3 + 5^2 + 5$].) In this paper, we characterize the experiment space in terms of just two parameters. The first is λ—the length of the program preceding the RPT. The second is the value of N—the argument that RPT takes. Because both parameters must have values less than 16, there are 225 "cells" in the λ–N space. Within that space, we identify three distinct regions. Region 1 includes all programs with $N = 1$. Region 2 includes all programs in which $1 < N < \lambda$. Region 3 includes all programs in which $N \geq \lambda$. The regions are depicted in Fig. 2, together with illustrative programs, from the (4,1) cell in Region 1, the (3,2) cell in Region 2, and the (1,4) cell in Region 3.

Programs from different regions of the experiment space vary widely in how effective they are in supporting or refuting different hypotheses. (A complete analysis of the interaction between experiment space regions and hypotheses is given in Klahr, Dunbar, & Fay (1990). Here we summarize the major differences between the regions.)

1. Region 1 programs have poor discriminating power. For example, the Region 1 program shown in Fig. 1 would execute the final LT 5 command twice under both Rule B (Repeat the last step N times) and Rule D (Repeat the last N steps once).

2. Region 2 programs provide maximal information about all of the common hypotheses, because they can distinguish between Counters and Selectors, and they can distinguish *which* Selector or Counter is operative. Region 2 produces different behavior under all four rules for any program in the region, and varying N in a series of experiments in this region always produces different outcomes.

3. Region 3 experiments may yield confusing outcomes. For rules C (Repeat the Nth step once) and D (Repeat the last N steps once), programs in this region are executed under the subtle feature that values of N greater than λ are truncated to $N = \lambda$. Therefore, varying N

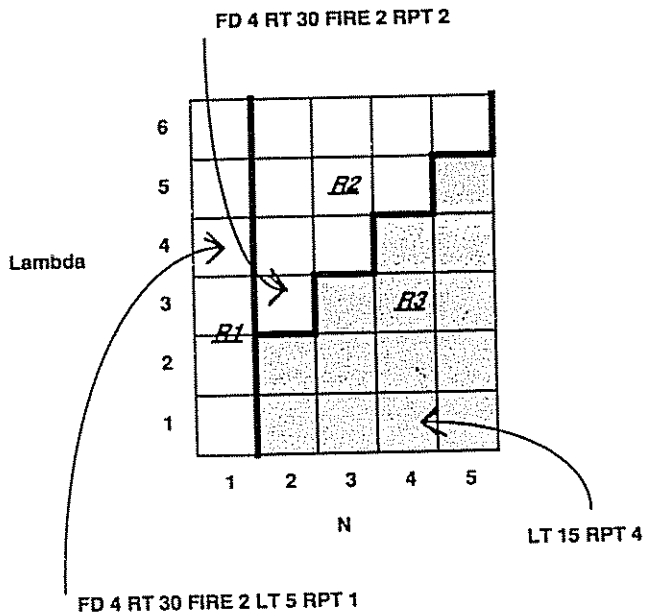

FIG. 2. Regions of the experiment space, showing illustrative programs (Shown here is only the 6 × 5 subspace of the full 15 × 15 space).

from one experiment to the next may give the impression that N has no effect. For example, Rule D would generate the same behavior for ↑4 Fire 2 RPT 3 and ↑4 Fire 2 RPT 4. Some of the programs in this region are discriminating, but others either don't discriminate at all, or they depend on the truncation assumption to be fully understood.

Design

One consequence of domain-specific knowledge is that some hypotheses about the domain are more plausible than others. In this study we explored the effect of domain-specific knowledge by manipulating the role of plausible and implausible hypotheses. Our goal was to investigate the extent to which prior knowledge—as manifested in hypothesis plausibility—affected the types of experiments designed and the interpretation of results.

We provided each subject with an initial hypothesis about how RPT might work. The suggested hypothesis was always wrong. However, depending on the condition, subjects regarded it as either plausible or implausible (recall that both children and adults in earlier studies regarded Counter hypotheses as highly plausible and Selector hypotheses as implausible). In some conditions the suggested hypothesis was only "somewhat" wrong, in that it was from the same frame as the way that RPT actually worked. In others, it was "very" wrong, in that it came from a different frame than the actual rule.

The BT simulator was programmed so that each subject worked with a RPT command obeying one of the two "Counter" rules or two "Selector" rules described above. We used a between-subjects design, depicted in Table 1. The Given hypothesis is the one that was suggested by the experimenter, and the Actual Rule is the way that BT was programmed to work for a particular condition. The key feature is that RPT *never worked in the way that*

TABLE 1
Design of Given-Actual Conditions

	Actual rule	
	Counter	Selector
Given hypothesis	B: Repeat last step N times	A: Repeat entire program N times
Counter	↓	↓
	A: Repeat entire program N times	D: Repeat the last N steps once
	D: Repeat the last N steps once	C: Repeat step N once
Selector	↓	↓
	A: Repeat entire program N times	D: Repeat the last N steps once

was suggested. In each Given-Actual condition, there were three CMs,[2] 5 CCs, four sixth graders, and four third graders (except for the Counter-Counter condition, which had three third graders and five sixth graders).

Changing from a hypothesis within a frame to another hypothesis from the *same* frame (e.g., from one Counter to another Counter) corresponds to *theory refinement*. However, as noted earlier, a change in N-role requires a simultaneous change in more than one attribute, because the values of some attributes are linked to the values of others. Changing from a hypothesis from one frame to an hypothesis from a *different* frame (e.g., from a Counter to a Selector) corresponds to *theory replacement*.

REPRESENTATIVE SUBJECT PROTOCOL

The raw protocols provided the basis for all performance measures. They are comprised of subjects' written programs as well as transcriptions of subjects' verbalizations during the experimental phase. Before presenting the quantitative analysis of subjects' behavior, we examine the verbal protocol of a single subject in order to illustrate a variety of interesting qualitative aspects of subjects' behavior. (The full protocol is listed in the Appendix.) Our goal is to convey a general sense of subject's approach to the task and to illustrate how we encoded and interpreted the protocols. In subsequent sections, we provide a detailed analysis based on the full set of protocols.

DP was a male CM subject in the Counter → Selector condition, and he was given Rule A: *Repeat entire program N times*. The actual rule was Rule C: *Repeat Nth step once*.[3] DP discovered the correct rule after five experiments. Two characteristics of DP's protocol make it interesting (but not atypical). First, even before the first experiment, DP proposed an alternative to the Given hypothesis (2: "I want to test to see if RPT repeats the statements before it"). Second, throughout the experimental phase, DP made many explicit comments about the attributes of the experiment space. He clearly attended to the properties of a "good" experiment.

DP's goal in his first experiment is unambiguous (2–9): to determine whether RPT acts on instructions before or after the RPT command. To resolve this question DP conducted an experiment with easily distinguished commands before and after the RPT key. (This ability

[2] These subjects were a subset of a larger group studied in a related experiment reported in Klahr, Dunbar, & Fay, 1990. That study used an extended set of problems that included two *different* Given-Actual pairs in each cell (e.g., for Counter-Counter, A → B *and* B → A).

[3] DP was in one of the conditions from the extended set. See previous footnote. However, his protocol is typical of other adult protocols.

to write programs that contain useful "markers" is an important feature of our subjects' behavior, and we will return to it later). This experiment allowed DP to discriminate between these two rival hypotheses. However, with respect to discriminating between the Given hypothesis (A), the Current hypothesis (B) and the Actual hypothesis (C), the program yielded ambiguous results. DP extracted from the first experiment the information he sought (17–18: "it appears that the repeat doesn't have any effect on any statements that come after it.")

For the second experiment DP returned to the question of whether the Given hypothesis (A), or the Current hypothesis (B) was correct, and he increased λ from 1 to 2. He also included one step following the RPT "just to check" that RPT had no effect on instructions that follow it (22–23). Thus, DP was in fact testing three hypotheses; A, B, and "after." Once again, he used commands that could be easily discriminated. He wrote another program from Region 3 of the experiment space ($\lambda = 2$, $N = 2$). DP observed that there were two executions of the ↑2 instruction, and he concluded (29–30) that "it only repeats the statement immediately in front of it." While this conclusion is consistent with the data that DP had collected so far, the hypothesis (B) was not in fact how the RPT key worked.

For the third experiment, DP continued to put commands after RPT just to be sure they were not affected. However, given that his current hypothesis had been confirmed in the previous experiment he next wrote a program that further increased the length of the program. This was his first experiment in Region 2. The goal of this experiment was to "see what statements are repeated" (33). He realized that the outcome of this experiment was inconsistent with his Current hypothesis (B), while the outcome of the previous experiment was consistent with B (47: ". . . it seemed to act differently in number two and number three"). The unexpected result led DP to abandon Hypothesis B and to continue beyond the mandatory three experiments.

For the fourth experiment, DP used a different value of N (53–54: ". . . repeat three instead of a repeat two, and see if that has anything to do with it.") Here too, DP demonstrated another important characteristic of many of our subjects' approach to experimentation. He used a very conservative incremental strategy, similar to the VOTAT (vary one thing at a time) strategies described by Tschirgi (1980) and the Conservative Focusing strategy described by Bruner, Goodnow, and Austin (1956). This approach still led him to put commands after the RPT, even though he was confident that RPT has no effect on them, and even though they placed greater demands on his observational and recall processes. (At the λ–N level, DP executed VOTAT consistently throughout his series of five experiments. The λ–N pairs were: 1-2, 2-2, 3-2, 3-3, 3-1. For the last three experiments, even the specific commands and their parameters remained the same, and only N varied.) This moved him from region 2 into region 3, and while analyzing the results of this experiment (59–69) in conjunction with earlier results, DP changed from the Counter frame to the Selector frame. First he noticed that "the number three" statement (i.e., the ↓1) was repeated twice in this case but that "the turning statement" was repeated (i.e., executed) only once (59–61). The implied comparison was with the previous experiment in which the turning statement (i.e., "the right 15 command" [43]) was the command that got repeated.

The next sentence is of particular interest: ". . . because when I change the number not only did it change . . . it didn't change the uh . . . the number that it repeated but it changed the uh . . . the actual instruction" (64–67). We believe that DP was attempting to articulate a change from the Counter frame to the Selector frame, as the following paraphrase of his comments indicates: "When I changed the value of N, it didn't change the *number* of repetitions, but it did change *which* commands got repeated."

DP went on to clearly state two instantiated versions of the correct rule by referring to previous results with $N = 2$ and $N = 3$, and he designed his fifth experiment to test his prediction with $N = 1$. The outcome of this final experiment, from Region 1, in conjunction with earlier results was sufficient to convince him that he had discovered how RPT worked.

RESULTS

Having presented an illustrative example of the performance of one subject, we now turn to a detailed analysis of the full set of protocols. Group differences are investigated with respect to four questions. First, how successful were subjects in discovering how RPT actually worked in the different experimental conditions? Second, how did subjects' interpretation of the task affect their goals and implicit constraints? Third, how did they search the experiment space? Fourth, how did their search of the experiment space affect the hypothesis that they finally stated?

In cases where the data for the two adult groups were not significantly different and the two children groups were not significantly different, the data were collapsed into adult (CM and CC) vs children (grades 6 and 3). For other analyses, the data for the two adult groups and the sixth graders all revealed the same pattern but the third graders showed an opposite pattern. In these situations, the groups were collapsed into Older (CM, CC, and 6) vs Youngest (grade 3). The χ^2 tests were used for all analyses involving response category contingency tables, except for cases of very small *N*s, in which Fisher Exact tests were used instead.

Success Rates

Domain-specific knowledge—as manifested in subjects' expectations about what "repeat" might mean in this context—played an important role in subjects' ability to discover the Actual rule. (See Fig. 3.) Regardless of what the Given hypothesis was, subjects found it easier to discover Counters (81%) than Selectors (35%), $\chi^2(1,N = 64) = 12.6, p = .0004$.[4] There was also a main effect for group: the correct rule was discovered by 83% of the CMs, 65% of the CCs, 53% of the sixth graders, and 33% of the third graders, $\chi^2(3,N = 64) = 7.48, p = .058$. This group effect is attributable to the Actual = Selector conditions, in which 56% of the adults but only 13% of the children were successful, $\chi^2(1,N = 32) = 4.99, p = .03$. For Counters, adults and children were roughly equal in their success rates (88% versus 75%), $\chi^2(1,N = 32) = .82, p = .65$.

The main effect for plausibility can also be attributed primarily to the children's performance in the Actual = Selector condition. Whereas 75% of the children discovered the rule when it was a Counter, only 13% discovered the rule when it was a Selector, $\chi^2(1,N = 32) = 12.7, p = .001$. Adults were also better at discovering Counters than Selectors (88% vs 56%), although the effect was not as strong as for children ($\chi^2(1,N = 32) = 3.86, p = .11$) due to the surprisingly poor performance by the CC subjects in the Counter-Counter condition.

[4] All χ^2 results with 1 *df* are reported with Yates correction for continuity.

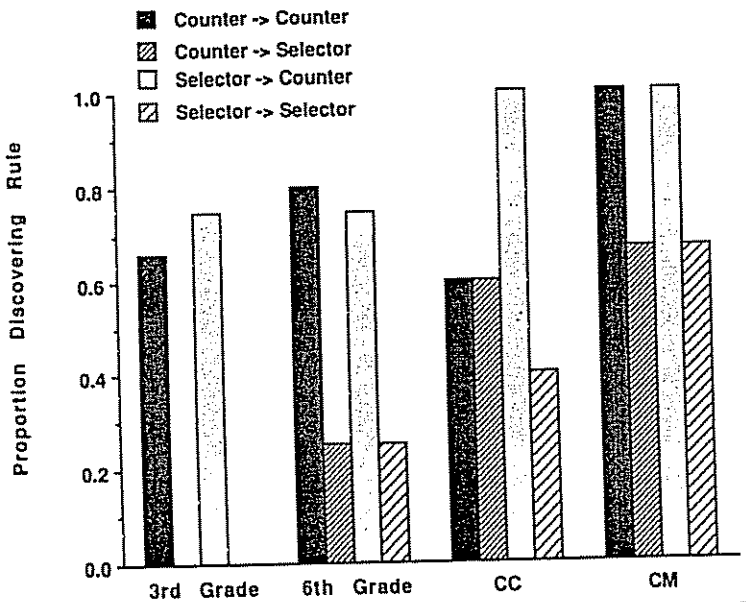

FIG. 3. Proportion of subjects in each group discovering correct rule for each Given–Actual condition.

In order to determine whether success rates differed between specific groups, we calculated the Fisher exact probability of success or failure among all six pairings of the four subject groups. The analysis yielded three low *p* values: CM versus sixth ($p = .08$), CM versus third ($p = .01$), and CC versus third ($p = .05$). Thus, overall success rate was affected only when grade level differences were extreme (both adult groups versus third graders) or when they were combined with training differences (CMs versus sixth graders).

Hypothesis Interpretation: Initiating Search in the Hypothesis Space

The purpose of presenting subjects with a Given hypothesis was to determine the extent to which search in the hypothesis space was influenced by the plausibility of the hypothesis being considered. This is one of the points at which domain-specific knowledge (which determines plausibility) might affect domain-general knowledge about experimental strategies, such as attempts to disconfirm, discriminating between rival hypotheses, and so on.

Prior to running the first experiment, subjects were asked to predict what would happen. Their predictions indicated the extent to which they understood and/or accepted the Given hypotheses. Each subject's re-

sponse to the Given hypothesis was assigned to one of three categories: I, accept the Given hypothesis; II, accept the Given, but also propose an alternative (see the protocol for Subject DP, presented earlier); and III, reject the Given, and propose an alternative. The number of subjects in each category is shown as a function of grade level and type of Given hypothesis in Table 2.

There was a main effect of Given hypothesis (Counter versus Selector) on type of response, $\chi^2(2,N = 60) = 5.47$, $p = .065$. This effect was attributable entirely to the third graders, who almost always accepted Counters and rejected Selectors (Fisher, $p = .029$). There was also a main effect for group, $\chi^2(6,N = 60) = 23.16$, $p = .0007$. This effect remained when analyzed separately for both Given = Counter, $\chi^2(6,N = 32) = 12.5$, $p = .05$, and Given = Selector, $\chi^2(6,N = 28) = 16.98$, $p = .009$.

In both conditions, the two adult groups always accepted the Given hypothesis, either on its own (Category I) or in conjunction with a proposed alternative (Category II) (the difference between CMs and CCs in their response patterns was not significant [$\chi^2(1,N = 29) = 2.18$, $p = 0.28$]). In contrast, no third grader and only two sixth graders ever proposed an alternative to compare to the Given (Category II). Children were approximately evenly divided between accepting the Given (Category I) or rejecting it (Category III). (The difference between third and sixth graders in their pattern of responses was not significant [$\chi^2(2,N = 29) = 2.06$, $p = .36$]). Overall, adults were more likely to consider multiple alternatives than children: 10 of 29 adults in category II, versus 2 of 31 children, $\chi^2(1,N = 60) = 5.71$, $p = .017$.

Of the 25 subjects who proposed alternatives to the Given hypothesis, 3 proposed alternatives that could not be coded as either Counters or Selectors. For the remaining 22, there was a strong effect of the type of Given hypothesis on the type of alternative proposed. Table 3 shows the

TABLE 2
Subjects' Responses When Given either Counter or Selector Hypothesis

		Group									
		CM		CC[a]		Sixth		Third[a]		Total	
Response category	Given:	C	S	C	S	C	S	C	S	C	S
I. Accept Given		3	3	9	4	5	4	6	1	23	12
II. Accept Given and propose alternative		3	3	1	3	1	1	0	0	5	7
III. Reject Given and propose alternative		0	0	0	0	3	3	1	6	4	9

[a] One third grader and three CCs did not respond to the "what will it do?" question.

TABLE 3
Type of Alternative Generated (for Categories II and II)

		Group							
		CM		CC		Sixth[a]		Third[a]	
Type of alternative	Given:	C	S	C	S	C	S	C	S
Same frame as given		3	1	1	0	3	1	1	0
Different frame from Given		0	2	0	3	0	3	0	4

[a] One sixth grader and two third graders generated alternatives that were unclassifiable.

number of subjects proposing alternatives from the same or different frame as the Given, as a function of group and type of Given. In each group, Given = Counter subjects who proposed alternatives always proposed another Counter, whereas, across all four groups, only 2 of the Given = Selector alternatives were from the Selector frame, $\chi^2(1,N = 22) = 11.8$, $p = .001$.

In summary, when responding to the Given hypothesis, adults were able to consider more than a single hypothesis, whereas children were not. When subjects did propose alternatives, they tended to propose plausible rather than implausible alternatives (i.e., Counters rather than Selectors). As we shall see in the next section, this propensity to consider multiple vs single hypotheses can affect the type of experimental goals set by the subjects, which in turn can be used to impose constraints on search in the experiment space.

Search in the Experiment Space

How did subjects solve the problem of designing a "good experiment" to discover how the RPT key worked? We address this question by analyzing the kinds of experiments that subjects designed. We start with an analysis of how domain-general knowledge about their own cognitive limitations was used by subjects to impose pragmatic constraints on the complexity of their experiments. Next, we do a static analysis of the distribution of experiments in the experiment space, and then we look at the *dynamics* of experiment space search by examining transitions from one experiment to the next. Finally, we examine the *interaction* between the experiment space and the hypothesis space by analyzing the ability of subjects to extract useful information from the outcomes of experiments and to use that information to evaluate their hypotheses.

Constraints derived from domain-general knowledge. Subjects' use of domain-general knowledge to constrain search in the Experiment Space can be investigated by analyzing (a) what they say about experiments and (b) the features of the experiments that they actually write. Each of these

knowledge sources, in turn, can be analyzed at the λ–N level, or at a finer grain of analysis that looks at the details of program content. In this section, we first summarize the results from the verbal protocols, and then we look at the features of their programs.

Subjects' verbal protocols contain many statements indicating both explicit understanding of the experiment space dimensions and what might be called a general notion of "good instrumentation:" designing interpretable programs containing easily identifiable markers. Subjects made explicit statements about both kinds of knowledge. The following statements by different adult subjects are typical: (a) "I don't want to have two of the same move in there yet, *I might not be able to tell if it was repeating the first one or if it was doing the next part of my sequence;*" (b) "I'm just going to make up some random but different directions *so that I'll know which ones get executed;*" (c) "I'm going to use a series of commands that will . . . *that are easily distinguished from one another*, and won't run it off the screen;" (d) "so I'm going to pick two [commands] that are the direct opposite of each other, to see if they don't really have to be direct opposites but I'm just going to write a program that consists of two steps, *that I could see easily*." (Emphasis added.)

Sixth graders were somewhat less articulate, but still showed a concern for both experiment space dimensions and program interpretability. Typical comments were (a) "I should have done FIRE because that was something *more standing out;*" (b) "Can I write one that has less steps so I can see if I can figure it out that way easier?" (c) "This time I'm not going to make it so long so it'll be easier." (d) "Maybe I should not go all the way [to the screen boundary] so I can tell if it does it again." Third graders rarely made such comments. We quantified subjects' appreciation of the dimensions of the experiment space by tabulating the frequency with which they made comments about "using longer programs," "using a different value of N," and so on. Eighty-three percent of the CMs made at least one such comment, compared to 60% of the CCs, 53% of the sixth graders, and 20% of the third graders, $\chi^2(3,N = 64) = 11.4, p = .01$.

In addition to these verbal statements, subjects demonstrated the effects of constraint imposition in their limited exploration of the experiment space. Permissible values for both λ and N range from 1 to 15. However, all subjects tended to constrain both the length of programs they ran and the value of N. Although the $\lambda \leq 4$ by $N \leq 3$ region of the experiment space represents only 5% of the full space, 50% of the CMs' 44 experiments were within it, as were 63% of the CCs' 117 experiments, 31% of the sixth graders' 68 experiments, and 31% of the third graders' 55 experiments. That is, subjects at all grade levels clustered their experiments in a small region of the experiment space, although adults did it more than children, $\chi^2(1,N = 284) = 21.96, p = .0001$.

At a finer level of detail, good instrumentation was assessed by the extent to which subjects observed three pragmatic constraints: (a) using small numeric arguments (values < 5) on move commands, so that the actions of BT are not distorted by having it hit the boundaries of the screen; (b) using standard units of rotation, such as 15 or 30 "minutes" (90 and 180°), for turn commands; and (c) using distinct commands in a program where possible.[5] Programs constrained in these ways produce behavior that is easier to observe, encode, and remember.

Subjects were scored as observing a constraint if they violated it on no more than one of their experiments. For both turns and moves, there was a main effect of group. On turn commands, 92% of the CMs, 95% of the CCs, 71% of the sixth graders, and 53% of the third graders observed the constraint, $\chi^2(3,N = 64) = 10.6, p = .01$). On move commands, 92% of the CMs, 85% of the CCs, 65% of the sixth graders, and 47% of the third graders observed the constraint, $\chi^2(3,N = 64) = 9.18, p = .03$. In contrast, there was no significant difference in the proportion of subjects in each group who observed the distinct command constraint, although for all groups the proportion was much lower than for the other two constraints: 42% of the CMs, 45% of the CCs, 24% of the sixth graders, and 53% of the third graders observed this constraint, $\chi^2(3,N = 64) = 3.23$, $p = .36$. The main effect for the first three analyses is not simply a training effect, because even when the CM subjects are eliminated, it remains.

It is possible that group differences on explicit statements about the experiment space are a consequence of older subjects' general superiority at verbalization. However, none of the other three measures depend on verbalization ability. Thus, both what subjects *said* and what they *did* support the conclusion that older subjects—even those with weak technical backgrounds—were better able than children to constrain their search in the experiment space and to design interpretable experiments.

Constraints derived from domain-specific knowledge. As noted earlier, subjects establish different goals in response to hypotheses that their domain-specific knowledge leads them to interpret as plausible or implausible. These different goals, in turn, should lead to different types of search constraints in the experiment space. More specifically, if the goal is to identify which of the program steps are repeated for Selector hypotheses, or to discriminate between Selectors and Counters, then subjects should write programs having more than N steps (i.e., with $\lambda > N$). In programs where λ is several steps greater than N, it is easy to distinguish among repeats of all steps, first step, last step, and N steps. On the

[5] But note that for any program with $\lambda > 5$, *some* command must be repeated, as there are only five distinct commands.

other hand, if the goal is to demonstrate the effect of a Counter, then subjects should use larger values of N and (for pragmatic reasons) relatively short programs (i.e., programs with $\lambda \leq N$). Both of these effects should be strongest for the first experiment, before subjects have any direct evidence for the actual rule.

Both of the adult groups' responses and sixth graders' responses were consistent with the normative account given above: Combining the three groups, only 10 of the 25 Older subjects given Counters wrote $\lambda > N$ programs, while 20 of the 24 given Selectors did so, $\chi^2(1,N = 49) = 7.95$, $p = .005$. Third graders showed the opposite pattern: Six of the seven given Counters but only 2 of the 8 given Selectors had first programs with $\lambda > N$. (Fisher exact, $p = .041$, two-tailed). Figure 4 shows the proportion of subjects in each condition whose first programs had $\lambda > N$.

These results show that while subjects at all grade levels used both domain-general and domain-specific knowledge to constrain their search, the groups differed in how they used the two kinds of knowledge. Adults and sixth graders, when given a plausible Counter, produced short programs with large values of N, suggesting a focus on the number of repetitions, rather than on what was to be repeated. On the other hand, when given an implausible Selector, Older subjects were likely to start with longer programs with smaller N's (i.e., programs with $\lambda > N$), suggesting a focus on what was to be repeated rather than the number of repetitions, which would allow them to discriminate between Counters and Selectors.

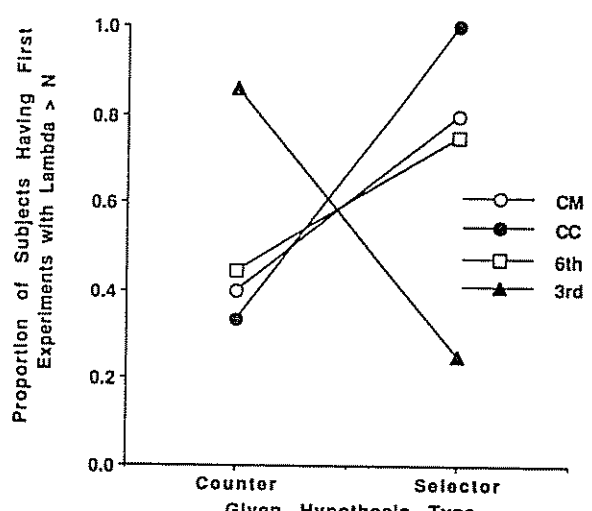

FIG. 4. Proportion of subjects in each group with $\lambda > N$ on first experiment.

We interpret the third graders' reversal of this pattern as evidence for their apparent inability to even *consider* a Selector hypothesis. When given one, they wrote relatively short programs with relatively large values of N, a strategy that is consistent with a goal of trying to convincingly demonstrate a Counter. When given a Counter, third graders used less extreme values of N, perhaps because they had less motivation to "prove a point."

Different experimental strategies can also be inferred by classifying experiments in terms of experiment space regions (see Fig. 1). As noted earlier, Region 2 is the most informative region, and adults appear to have understood its potential informativeness better than the children. Eleven of 12 CMs, 15 of 20 CCs, 10 of 17 sixth graders, and 8 of 15 third graders wrote at least one Region 2 experiment, $\chi^2(1,N = 64) = 3.56$, $p = .059$. Another way to extract useful information from the E-space is to write experiments from more than a single region. Adults were more likely to sample different regions than were children. Ninety-one percent of the adults (100% of CMs and 85% of the CCs) wrote experiments from at least two different regions of the experiment space. In contrast, only 29% of the sixth graders and 60% of the third graders sampled from more than one region, $\chi^2(1,N = 64) = 22.19$, $p = .0001$. Staying in one region of the experiment space is only detrimental if the region fails to discriminate between hypotheses (e.g., Region 1 for hypotheses B versus D) or if it fails to adequately demonstrate the correct hypothesis (e.g., Region 3 for hypothesis D). All of the third graders in Actual = Selector conditions who stayed in one region were in either Region 1 or 3. For the sixth graders in Actual = Selector conditions, 75% who stayed in one region were in Region 3. Thus, for the children, the failure to run experiments from different regions of the experiment space severely limited their ability to extract useful information from the outcomes of their experiments.

The common pattern here is that there is little or no difference between the CM and CC subjects, who, when combined, tend to differ from the two children's groups. For some measures, the sixth graders cluster with the adult subjects. Taken as a whole, this pattern suggests a developmental effect, rather than a training effect, for subjects' sensitivity to the potential informativeness of different types of experiments as a function of the Given hypothesis. Moreover, by some of our measures, this effect appears between third and sixth grades.

Dynamics of search in the experiment space. The analysis of the experiment space in terms of λ–N combinations or experiment space regions gives a picture of the properties of experiments aggregated over the entire search process, but it does not describe the *dynamics* of search in the experiment space. In this section we look at the search dynamics at two levels: (a) changes in N and λ from one experiment to the next, indepen-

dent of program-content change, and (b) program-content change across experiments, independent of changes in λ and *N*.

One type of domain-general knowledge that subjects might bring to this task is the "vary one thing at a time" (VOTAT) strategy mentioned earlier. When applied to the λ and *N* dimensions of the experiment space, VOTAT produces conservative moves: changes from one experiment to the next that do not vary both λ and *N* at the same time (including moves that vary neither). Overall, subjects were conservative on 56% of their moves. The proportion of conservative transitions (following the second experiment) at the λ–*N* level was calculated for each subject and analyzed using a 2 (Actual hypothesis) × 4 (Group) ANOVA. The analysis yielded no effect for group, or Actual hypothesis, $F < 1$, $p > .5$.

However, if we look at the next level of detail—at the specific program content—then we can define a conservative move as one that keeps program content constant (except for additions or deletions) so that the effects of changes in λ and *N* would be easy to detect. Under this definition, the CM adults are more likely to make conservative moves than the CC adults and the children. The mean proportion of conservative moves at the program-content level was .48 for CMs, but only .13, .12, and .09 for CCs, sixth, and third graders, respectively ($F[3,60] = 5.52$, $p = .003$). Using Scheffe *F* tests for pairwise comparisons, CMs were significantly different from all other groups ($p < .05$), but the CCs, sixth, and third graders were not significantly different from each other ($p > .5$).

Encoding experimental outcomes. The biases induced by domain-specific knowledge creates expectations about BT's behavior. These expectations might affect subjects' ability to accurately encode outcomes at variance with those expectations. Subjects' descriptions of experimental outcomes were scored as misencodings if they contained unambiguous errors in reporting BT's behavior during program execution (i.e., if they contained explicit descriptions of events that did not actually occur, or if they described a sequence of events, but omitted an important one.)

There was a main effect of Actual condition on misencoding: 63% of the subjects in the Actual = Selector but only 35% of the subjects in the Actual = Counters conditions misencoded at least one experiment, $\chi^2(1,N = 64) = 4$, $p = .045$. Within-group analyses showed that third graders misencoded more programs when the Actual Rule was a Selector than when it was a Counter, $\chi^2(1,N = 50) = 4.32$, $p = .04$. The other three groups were as likely to misencode programs when the Actual rule was a Selector as when it was a Counter. Even though the third graders' misencodings were more prevalent when the rule was implausible, they were not systematically distorted in the direction of their current hypotheses. For all the groups, when misencodings did occur, distortions of the actual results were as likely to occur in the direction of confirming the current hypotheses as disconfirming it.

There was also a main effect of group on misencoding: Twenty-five percent of the CMs, 40% of the CCs, 53% of the sixth graders, and 73% of the third graders misencoded at least one experimental outcome, $\chi^2(3,N = 63) = 7.3$, $p = .062$. To determine whether the tendency to misencode experimental outcomes differed between specific groups, we calculated Fisher Exact tests among all six pairings of the four subject groups. The analysis yielded three low *p* values: CM versus sixth ($p = .08$), CM versus third ($p = .02$), and CC versus third ($p = .05$).

These results are almost identical to the pattern of between-group differences in overall success rates reported earlier, and they suggest the possibility that third-graders' inability to discover Selectors is a consequence of their well-known mnemonic and encoding deficiencies. However, the interaction of BT rules and regions produced encoding and mnemonic demands that were *more* difficult in Counter conditions, where the third-graders did very well, than in the Selector conditions, where they all failed to discover the correct rule. In general, subjects in Actual = Counter conditions, who are always working with Rule A, tend to have much more complex behaviors to observe, encode, and remember than do subjects in Actual = Selector conditions. For example, the Region 2 program shown in Fig. 2 would execute nine instructions under Rule A, but only five under Rule D.[6]

Encoding demands for subjects working under Counter versus Selector rules were estimated by calculating the mean number of executed instructions for each subject's correctly and incorrectly encoded program outcomes.[7] A 2 (Actual Rule: Selector vs. Counter) × 2 (Encoding: Correct vs Incorrect) ANOVA revealed that subjects in the Counter condition had to observe and encode significantly more instruction executions than subjects in the Selector condition ($M = 25$ executions versus $M = 6.5$ executions), $F(1,17) = 16.68$, $p = .0008$, but there was no difference in program length for correctly versus incorrectly encoded programs, $F(1,17) = 1.04$, $p = .32$, and no interaction effect, $F(1,17) = .26$, $p = .62$. Furthermore, whereas children in the Actual = Counter conditions could

[6] More generally, the number of instructions executed for a given λ–*N* combination depends on the Actual rule. For Rule A (Repeat entire program *N* times), the number of instructions executed is 2 * λ in Region 1 and (N + 1) * λ in Regions 2 and 3. For Rule D (Repeat last *N* steps once), the number of executed instructions is λ + 1, λ + *N*, and 2 * λ, in Regions 1, 2, and 3, respectively.

[7] Some subjects contribute mean values for both the correctly and incorrectly encoded programs, whereas subjects who either misencoded or correctly encoded all programs only contribute one mean to the analysis.

correctly encode programs with a mean length of 22 executions, children in the Actual = Selector conditions misencoded programs that had only 7 executed instructions.

Thus, children's failures in the Selector condition cannot be attributed to general encoding and observation deficits. A more plausible explanation is that their Counter bias led them to focus on the verification that N repetitions (of something) had occurred. This attention to *number* of repetitions rather than *what* was being repeated caused them to misencode the outcome. However, it did not lead them to distort their encodings to fit a Counter rule. In only 4 of the 32 Actual = Selector programs where children stated their hypotheses was N given a Counter role, and in 3 of these cases the value of N was 1, and the encoding was correct.

Inducing Hypotheses from Experimental Outcomes

We began this analysis with a presentation of overall success rates, followed by a detailed analysis of the statics and dynamics of search in the experiment space. Throughout, we have pursued the theme that domain-specific knowledge about "repeat" determined the initial plausibility of hypotheses and that, in turn, hypothesis plausibility influenced the kind of domain-general knowledge that subjects brought to bear in imposing constraints on their search of the experiment space. One remaining question is whether or not, when their E-space search did lead them into Region 2, subjects were able to make use of its maximally informative results. In other words, how successful were subjects at coordinating the searches in the hypothesis space and the experiment space? Although Region 2 provides direct and discriminating evidence for all hypotheses, it is most useful for discovering Selector rules. Therefore, we expected that Selector subjects who had observed the outcome from one or more Region 2 experiments would be more likely to discover the correct rule than those who never entered Region 2.

In order to determine the effect of experiment space region on overall success rate, we calculated the probability of discovering the correct rule as a function of the Regions actually visited. When success rates are aggregated over all grades and conditions, there appears to be no benefit from having been in Region 2. Sixty-four percent of the 44 subjects who had one or more Region 2 experiments were successful, while 45% of the 20 who never entered Region 2 were successful, $\chi^2(1,N = 64) = 1.27, p = .26$. However, as predicted, a closer analysis reveals a clear effect of Region 2's utility for discovering Selectors. Table 4 shows the number of subjects in each Actual condition and grade level who were successful or unsuccessful according to whether or not they ever went into Region 2. As just noted, most subjects in the Actual = Counter conditions are successful, regardless of whether or not they entered Region 2. However,

TABLE 4
Number of Successful and Unsuccessful Subjects Who Did and Did Not Have at Least One Experiment in Region 2 for Different Actual Conditions

	Actual conditions			
	Counter		Selector	
	S	U	S	U
CM				
In R2	6	0	3	2
Not in R2	0	0	1	0
CC				
In R2	5	1	5	4
Not in R2	3	1	0	1
Sixth				
In R2	4	1	2	3
Not in R2	3	1	0	3
Third				
In R2	3	1	0	4
Not in R2	2	1	0	4

Note. S, Successful; U, Unsuccessful; R2, Region 2.

for all but one subject in the Actual = Selector conditions, having at least one experiment in Region 2 is a necessary but not sufficient condition for success.

DISCUSSION

We have approached the investigation of developmental differences in scientific reasoning in terms of dual search in a space of experiments and hypotheses. In the studies described here, we manipulated subjects' initial location in the hypothesis space in order to examine their search in the experiment space. Our discussion will focus primarily on developmental differences in experiment space search. However, it is important to recall that in all conditions—including Counter-Counter and Selector-Selector conditions—the Given hypotheses was always wrong. *Subjects always had to search for the correct hypothesis.* Therefore, we include a brief discussion of hypothesis space search, and then we conclude with a discussion of the broader implications of our findings.

Experiment Generation Heuristics

Compared to the unbounded size of the experiment space faced by a scientist in a laboratory, the BT domain may appear to be an unrealisti-

cally simple context in which to investigate experimentation skills. In fact, the potential size of the BT experiment space is surprisingly large. When command-level differences between experiments are taken into account, then the number of distinct experiments is in the billions (See the earlier description of the BT Experiment Space). Adult subjects quickly realized that although the specific commands in a program might be used as useful markers of BT's behavior, the essential attributes of the experiment space were the values of λ and N. Nevertheless, even the full λ–N space has 225 cells and subjects had to decide which of them to explore.

Both CM and CC adults were effective at drastically pruning the experiment space. Over half of their experiments occurred within the $\lambda \leq 4$, $N \leq 3$ area of the experiment space, which represents only 5% of the full space. In contrast, less than one-third of the children's experiments were so constrained. Furthermore, the pattern of results described in the previous section revealed a developmental trend in the overall systematicity and effectiveness with which subjects searched the experiment space. Our interpretation of this pattern is that it is a consequence of developmental differences in the application of a set of domain-general heuristics for searching the experiment space. The four principle heuristics are

1. Use the plausibility of a hypothesis to choose experimental strategy. As noted earlier, one of the most robust findings in the literature on scientific reasoning in adults is that subjects attempt to confirm, rather than disconfirm, their current hypothesis (Gorman, in press; Klayman & Ha, 1987). Similarly, developmental studies show that even when explicitly instructed to generate evidence that could potentially falsify a rule, children at the 6th-grade level or below perform very poorly (Kuhn, 1989; Ward & Overton, 1990). However, in this study, we found a more flexible kind of response. That is, both children and adults varied their approach to confirmation and disconfirmation according to the plausibility of the currently held hypothesis.

More specifically, subjects chose λ–N combinations that could either demonstrate or discriminate hypotheses according to their plausibility. When hypotheses were plausible, subjects at all levels tended to set an experimental goal of demonstrating key features of the given hypothesis, rather than conducting experiments that could discriminate between rival hypotheses. (However, when given Counters, the third graders did not emphasize the value of N to the extent that the other three groups did.)

For implausible hypotheses, adults and young children used different strategies. Adults' response to implausibility was to propose hypotheses from frames other than the Given frame and to conduct experiments that could discriminate between them. Our youngest children's response was to propose a hypothesis from a different, but plausible frame, and then to ignore the initial, and implausible, hypothesis while attempting to demonstrate the correctness of the plausible one. Third graders were particularly susceptible to this strategy, but by 6th grade, subjects appeared to understand the type of experiments that will be informative.

2. Focus on one dimension of an experiment or hypothesis. Experiments and hypotheses are both complex entities having many aspects on which one could focus. In this study, experiments could vary at the λ–N level, at the command level, or even at the level of arguments for commands. Similarly, for hypotheses, there are auxiliary hypotheses, ancillary hypotheses, and additional assumptions that are never directly tested (cf. Lakatos & Musgrave, 1970). An incremental, conservative approach has been found to be effective in both concept attainment (Bruner et al's "conservative focusing") and hypothesis testing (Tschirgi's, 1980, VOTAT strategy). This suggests that in moving from one experiment or hypothesis to the next or in moving between experiments and hypotheses, one should decide upon the most important features of each and focus on just those features.

Use of this focusing heuristic was manifested in different ways with respect to hypotheses and experiments. For hypotheses, it led all groups except the third graders to focus initially on the number of times something was repeated when given Counters, and what was repeated when given Selectors. This produced the λ–N pattern depicted in Fig. 4. For experiments, it led to a characteristic pattern of between-experiment moves that minimized changes at the command level. Here, the CM adults stood apart from the other three groups. They were much more likely than any of the three other groups to make conservative moves—that is, to minimize differences in program content between one program and the next. Although there are few sequential dependencies in the informativeness of experiment space regions, CM adults may have used this heuristic to reduce the cognitive load imposed when comparing the outcomes of two programs.

Interestingly, only the third graders failed to use this heuristic when searching the hypothesis space, whereas only the CM adults used it effectively when searching the experiment space. It is possible that, because the hypothesis search aspect of the discovery task is so familiar, all but the third graders were able to use the focusing heuristic. In contrast, when confronted with the relatively novel experimental design aspect of the task, even adults, if untrained in science, remained unaware of the utility of a conservative change strategy.

3. Maintain observability. As BT moves along the screen from one location to another, it leaves no permanent record of its behavior. Subjects must remember what BT actually did. Thus, one heuristic is to write short programs in order to make it easy to remember what happened and to compare the results to those predicted by the Current hypotheses. At

the level of individual commands, this heuristic produces small arguments for the ↑ and ↓ commands, so that BT does not go off the screen. There were clear differences in the use of this heuristic. Adults almost always used it, whereas the youngest children often wrote programs that were very difficult to encode. This heuristic depends upon knowledge of one's own information processing limitations as well as a knowledge of the device. Our finding that the third graders did not attempt to maintain observability, whereas the sixth graders and adults did, may be a manifestation, in the realm of experimental design, of the more general findings about the development of self-awareness of cognitive limitations (Brown, Bransford, Ferrara, & Campione, 1983; Wellman, 1983).

4. *Design experiments giving characteristic results.* Physicians look for "markers" for diseases, and physicists design experiments in which suspected particles will leave "signatures." In the BT domain, this heuristic is instantiated as "use many distinct commands." This heuristic maximizes the interpretability of experimental outcomes. It is extremely difficult to isolate the cause of a particular piece of BT behavior when many of the commands in a program are the same. All four groups were roughly equivalent in their use of this heuristic; on average, about half of all programs did not contain any repeated commands.

Overall, adults and children differed widely in their use of these heuristics. Adults not only appeared to use each of them but also appeared to be able to deal with their inherent contradictions. No subject ever used the 1,1 cell, even though it would yield the easiest to observe behavior, because it is so uninformative with respect to discriminating among rival hypotheses. Additionally, in a related study (Klahr, Dunbar, & Fay, 1990), adults' experiments were significantly *over*represented in the $\lambda = 3$, $N = 2$ cell of the experiment space. This cell represents the shortest possible Region 2 experiment, and its overrepresentation suggests a compromise between informativeness and simplicity. Adults' tendency to cluster their experiments in the 4×3 experiment space in the present study represents a similar compromise among competing heuristics.

In contrast, children either failed to use these heuristics at all or they let one of them dominate. For example, one approximation to the "characteristic result" heuristic would be to write long experiments that could generate unique behavior, although that would violate the "maintain observability" heuristic. Even on their first experiments, adults tended to write relatively short programs. Only one-third of them wrote first programs with $\lambda > 3$, whereas 80% of the children wrote programs with $\lambda > 3$.

Search in the Hypothesis Space

In this study, as in all our previous studies, subjects at each grade level found Counter hypotheses more plausible than Selector hypotheses. Indeed, the relative plausibility of Selectors was so low for third graders, that when Selectors were given they were disregarded and replaced by Counters. Klahr & Dunbar (1988, pp. 10, 11) suggest several potential sources of domain-specific knowledge about what "repeat" might mean that could account for this bias, including linguistic knowledge, programming knowledge, and the particulars of the BT command syntax.

Although we found no age-related difference in the bias toward Counters, there was a developmental difference in the implications of this bias, which was manifested in subjects' ability to consider multiple hypotheses. One-third of the adults but almost none of the children began by considering more than one hypothesis. Adults' tendency to test multiple hypotheses was unexpected, because results from previous studies using this domain (Klahr & Dunbar, 1988) as well as others (e.g., Mynatt, Doherty, & Tweney, 1977) indicate that subjects generally avoid testing multiple hypotheses. One possible explanation for this difference is that in the earlier studies, subjects had to generate their own initial hypotheses, whereas in the studies described here, subjects were given hypotheses to test. This procedural difference may have induced in subjects a mild skepticism for a hypothesis not of their own making, even when it was from a plausible frame. For adults, this resulted in the testing of multiple hypotheses, while for most children, it led to replacement of an implausible Given with a different hypothesis.

These results suggest that one developmental difference in hypothesis space search is in how the strength of belief in a hypothesis determines whether multiple or single hypotheses will be considered. The cognitive demands imposed by the need to search the hypothesis space for a plausible hypothesis to oppose an implausible one, and then to search the experiment space for a discriminating test, may have exceeded the capacity of our youngest subjects. This led them to avoid considering multiple hypotheses. If the initial hypothesis was plausible, then they accepted it. If it was implausible, their only recourse was to abandon it and replace it with one more plausible. In contrast, when adults were given a hypothesis, whether plausible or implausible, they had sufficient resources to contrast it with another one.

A Framework for Investigating the Development of Scientific Reasoning Skills

We opened this paper by alluding to two long-standing disputes about developmental differences in scientific reasoning skills. In this concluding section, we will briefly summarize the two dichotomies and then show how our results, when interpreted within the SDDS framework, enable us to address both issues in a productive and informative way.

The child as scientist. The positive view of the child-as-scientist is exemplified by the assertion that "Clearly, children go about their task as true scientists do, building theories about the physical, social and linguistic worlds, rather than reasoning as inductive logicians" (Karmiloff-Smith, 1988, p. 193). This position is stated even more forcefully by Brewer & Samarapungavan (1991), who conclude their review of children's knowledge about astronomy by stating that "the child can be thought of as a novice scientist, who adopts a rational approach to dealing with the physical world, but lacks the knowledge of the physical world and experimental methodology accumulated by the institution of science." In stark contrast is the claim that "the process in terms of which mental models, or theories, are coordinated with new evidence is significantly different in the child, the lay adult, and the scientist. In some very basic respects, children (and many adults) do not behave like scientists" (Kuhn, 1989, p. 687).

Unfortunately, one can find empirical support for each of these incompatible claims. On the one hand, results of formal studies, as well as abundant everyday experience, provide evidence that trained scientists, and even untrained lay adults, commonly outperform children on a variety scientific reasoning tasks. Indeed, the vast literature on differences between formal operations and concrete operations supports the position that there are substantial changes in scientific reasoning skills between the ages of 6 and 12. On the other hand, many studies show that adults demonstrate systematic and serious flaws in their reasoning (Kuhn et al., 1988; Schauble & Glaser, 1990), while very young children are capable of surprisingly competent reasoning about hypotheses testing and experimentation. For example, Sodian et al. (1991) report that 1st-grade children, when given a pair of mutually exclusive and exhaustive alternative hypotheses and a choice between two unambiguous experiments, can distinguish between conclusive and inconclusive tests of the hypotheses. Furthermore, in this constrained context, first graders can distinguish between testing a hypothesis and generating an effect.

Domain-general and domain-specific scientific reasoning skills. In this debate it is assumed that there are developmental differences in scientific reasoning skills and the question becomes "what causes them?" By one account, they are simply a side effect of the acquisition of domain-specific knowledge (Chi & Ceci, 1987). Support for this view comes from demonstrations that when children do have sufficient domain-specific knowledge, they often outperform adults (e.g., Chi, 1978). The domain-general view, in contrast, attributes the differences to both the development of basic information-processing components (such as encoding rates, scanning skills, retrieval speed (Kail, 1991), metamemorial skills (Wellman & Somerville, 1984), and problem-solving skills (Klahr & Robinson, 1981) as well as differences in the logical skills required to carry out the full set of processes involved in scientific reasoning.

Toward a resolution of both debates. We believe that the results of this study, when interpreted within the SDDS framework summarized earlier, contribute toward a resolution of both of these debates. Our approach has been to formulate an explicit characterization of the scientific discovery process, and to examine the developmental trajectory of its components. We described a framework (SDDS) in which scientific reasoning was conceptualized as a problem-solving process involving search in two spaces, and we focused on heuristics for searching the experiment space. We argued that the search constraint heuristics were domain-general, because they were applied in a context far removed from the situations in which they were acquired. However, the plausibility of specific hypotheses, which influenced search in both the hypothesis space and the experiment space, is based on domain-specific knowledge. In the study reported here, it results from subjects' strong biases about what "repeat" might mean.

Our study yielded a picture of both similarities and differences in the way that children and adults formulate hypotheses and design experiments to evaluate them. At the top level of the cycle of scientific reasoning—that is, the level of hypothesis formation, experimentation and outcome interpretation—older elementary school children approached the discovery task in an appropriate way. Most sixth graders and some third graders understood that their task was to produce evidence to be used in support of an argument about a hypothesis. Contrary to Kuhn et al. (1988), they were able to distinguish between theory (hypotheses) and evidence. However, when placed in a context requiring the coordination of search in two spaces, children's performances were markedly inferior to adults (both with and without technical and scientific training).

An examination of the fine structure of the subjects' sequences of experiments and hypotheses revealed that their overall performance differences could be attributed to characteristic differences in how they searched both the hypothesis space and the experiment space. The most important difference in hypothesis space search was in the way that adults and children responded to plausible and implausible hypotheses. When adults were given an implausible hypothesis, they established a goal of designing an experiment that could discriminate between the given implausible hypothesis and a plausible hypothesis of their own creation (usually one of the standard Counters).

When children were given hypotheses to evaluate, they were not insensitive to whether they were plausible or implausible, but they responded by generating a different goal than the adults'. In the implausible case, rather than simultaneously considering two alternative hypotheses,

children focused only on a plausible one of their own making (a Counter), and attempted to generate what they believed would be extremely convincing evidence for it. This was not an unreasonable goal, but it produced uninformative experiments. More specifically, in order to generate a convincing case for a Counter hypothesis, third graders chose large values of *N*, so that the effect of the number of repetitions would be unambiguous. Because their goal was demonstration, inconsistencies were interpreted not as disconfirmations, but rather as either errors or temporary failures to demonstrate the desired effect. When subsequent efforts to demonstrate the Counter hypothesis were successful, they were accepted as sufficient. Because the third graders did not seek global consistency, they extracted only local information from experimental outcomes. Analogous results with respect to lack of global consistency have been reported by Markman (1979). She demonstrated that children between 8 and 11 years old have difficulty noticing internal contradictions in relatively brief text passages. Markman suggested that children focus on the reasonableness of individual statements, rather than their collective consistency. Similarly, our youngest children selectively focused on specific experimental outcomes, rather than seeking a hypothesis that could account for all of them.

The BT context elicited behavior in our third graders that is characteristic of younger children in simpler contexts. Resistance to disconfirming evidence has been observed in studies of discrimination learning (Tumblin & Gholson, 1981), but it has been limited to much younger children. For example, Gholson, Levine, & Phillips (1972) found that kindergarten children maintained disconfirmed hypotheses on about half of the negative feedback trials, while by 2nd grade the rate dropped to 10%. The complexity of the discovery context, in conjunction with strong plausibility biases, may have caused our third graders to function like kindergarten children in the simpler discrimination learning task.

With respect to search heuristics in the experiment space, children were less able than adults to constrain their search, they tended not to consider pragmatic constraints, and they were unsystematic in the way that they designed experiments. These findings indicate that one of the problems for the younger children is to apply effective search constraints on their experiments. This viewpoint is consistent with research on the effects of constraints on problem solving in younger children. When presented with "standard" puzzles (involving search in a single problem space), young children perform much better when the order of subgoals is constrained by the structure of the materials than when they have to decide for themselves what to do first (Klahr, 1985; Klahr & Robinson, 1981). Here too, we find the third graders in our dual-search situation behaving analogously to younger children in single-search contexts. That

is, in our study, when given a task in which they had to impose multiple constraints on hypotheses and experimental design, children did not conduct appropriate experiments. However, in studies where both hypotheses and experimental choices are highly constrained, young children can select appropriate experiments (Sodian et al., 1991).

Overall, the SDDS framework has helped us to begin to answer some enduring questions about the development of scientific discovery skills. The results of our analysis, when combined with the related work from other laboratories, clarify the conditions under which children's domain-general reasoning skills are adequate to successfully coordinate search for hypotheses and experiments: (a) hypotheses must be easily accessible (such as the highly plausible Counters in our study) or few in number (as in the two-alternative situations used by Sodian et al.), (b) the experimental alternatives must also be few in number (also as in Sodian et al.), and (c) the domain must provide feedback relevant to discriminating among plausible hypotheses (as in Region 2 experiments in BT studies). It is important to reiterate the point that the performance deficits we found were not simply the result of children's inadequate encoding or mnemonic skills. As shown earlier, when experimental outcomes were consistent with children's expectations, they were correctly encoded, even though they were three times as long as those incorrectly encoded, but discrepant from children's expectations. Instead, the adult superiority appears to come from a set of domain-general skills that go beyond the logic of confirmation and disconfirmation and deal with the coordination of search in two spaces.

APPENDIX: A COMPLETE PROTOCOL FROM A CM (ADULT) SUBJECT (DP)

Note. CLR and GO commands have been deleted. BT's behavior is shown at the right in boldface type.

Given: A—Repeat entire program *N* times Actual: C—Repeat *N*th step once

1. What I want to do is um . . .
2. I want to test to see if repeat repeats the statements before it,
3. so I'll write just a little program that does something,
4. say turns left ten, and then put the repeat statement in
5. and we'll just repeat it twice 'cause it's not really important
6. for the first program.
7. And then just to check I'll put in a statement after the repeat
8. to see if it has any effect on that,
9. so it's just left ten, repeat it twice and then right ten.
10. EXP: OK, SO WHAT ARE THE COMMANDS?

11. Left ten, repeat two, and right ten.
12. *Experiment 1:* ←10 RPT 2 →10
13. **←20 →10**

14. Ok, it did what you would expect it would do,
15. it did repeat the left ten twice so it turned left twenty
16. and then back to the right ten,
17. so it appears that the repeat doesn't have any effect on any statements
18. that come after it,
19. so I guess now we'll try putting more than one statement
20. in front of the repeat and see what happens
21. we'll do left ten and forward two, repeat twice,
22. and just to check it again at the end we'll just have it fire once
23. and see if it repeats it or not
24. It's left ten, forward two, repeat twice, fire one.
25. *Experiment 2:* ←10 ↑2 RPT 2 FIRE 1
26. **←10 ↑4 FIRE 1**

27. Ok, it only turned left once,
28. but it went forward four,
29. so it looks like it only repeats the statement immediately in front
30. of it. Oh . . . um . . .
31. I guess for the last test we'll just try a more complicated program
32. with the repeat in the middle of it
33. and see what statements are repeated. Ahh, forward one and . . .
34. So this is just a long program with a repeat in the middle of it
35. to see what statements are repeated.
36. EXP: OK, SO WHAT ARE THE COMMANDS?
37. Forward one, right fifteen, back one, repeat two, left ten,
38. forward one.
39. Um, that's the old program. I don't think that you cleared. Ok, forward one, right fifteen,
40. back one, repeat two, left ten, forward one.
41. *Experiment 3:* ↑1 →15 ↓1 RPT 2 ←10 ↑1
42. **↑1 →15 ↓1 →15 ←10 ↑1**

43. Ok, that time it repeated the uh . . . the right fifteen command
44. EXP: SO DO YOU KNOW HOW IT WORKS,
45. OR WOULD YOU LIKE TO WRITE MORE PROGRAMS?
46. I'm not really sure how it works,
47. because it seemed to act differently in number two, and number three.
48. It would be best to write more.

49. EXP: OK.
50. Should I just go ahead?
51. EXP: YES JUST WRITE NUMBER FOUR.
52. Um . . . Let's just try the same program as before,
53. except we'll put a three in, repeat three instead of a repeat two,
54. and see if that has anything to do with it.
55. Forward one, right fifteen, back one, repeat three, left ten,
56. forward one, and go.
57. *Experiment 4:* ↑1 →15 ↓1 RPT 3 ←10 ↑1
58. **↑1 →15 ↓2 ←10 ↑1**

59. That time it repeated the statement here but it repeated it twice,
60. the number four . . . the number three . . .
61. it repeated the turning statement once . . . Um . . .
62. EXP: WOULD YOU LIKE TO TRY ANOTHER PROGRAM?
63. Um . . . I guess, I don't really have any idea of what it's doing
64. because when I change the number not only did it change . . .
65. it didn't change the uh . . .
66. the number that it repeated
67. but it changed the uh . . . the actual instruction . . . Um, I'm going to try . . .
68. I guess my conjecture is, right now, that it says repeat two
69. so it repeats the second instruction,
70. and here it repeats three and it repeats the third instruction.
71. So we'll try the same thing with repeat one,
72. and see if it repeats the first instruction . . .
73. Forward one, right fifteen, back one, repeat one, left ten, forward one, go.
74. *Experiment 5:* ↑1 →15 ↓1 RPT 1 ←10 ↑1
75. **↑1 →15 ↓1 ↑1 ←10 ↑1**

76. Ok, ok, I think I know what it does now.
77. EXP: OK . . .
78. When it hits the repeat statement . . .
79. when it says repeat one it means at this point repeat statement
80. number one
81. and in this case because it went forward and it turned and it went back
82. and then it came forward again, which is the first statement.
83. and it did something similar, I mean it went forward one, turned right
84. went back, and it hit repeat three and this is the third statement
85. so it went back again
86. EXP: OK, SO HOW, IN GENERAL, HOW DOES THE REPEAT KEY WORK?
87. If you type, it looks, when it hits the repeat statement,

88. if you look through the program and it's like repeat six
89. it takes the sixth statement and does that,
90. then when it hits the repeat statement it'll repeat the sixth statement once.
91. EXP: OK, GREAT.

REFERENCES

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Brewer, W. F., & Samarapungavan, A. (1991). Child theories versus scientific theories: Differences in reasoning or differences in knowledge? In R. R. Hoffman & D. S. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives*. Hillsdale, NJ: Erlbaum.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering, and understanding. In P. H. Mussen (Ed.), *Handbook of child psychology: Cognitive development, Vol. III*. New York: Wiley.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: NY Science Editions.

Bruner, J. S., Olver, R. R., & Greenfield, P. M. (1966). *Studies in Cognitive Growth*. New York: Wiley.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Case, R. (1974). Structures and strictures: Some functional limitations on the course of cognitive growth. *Cognitive Psychology*, 6, 544–573.

Chi, M. T. H. (1978). Knowledge structures and memory development. In R. S. Siegler (Ed.), *Children's thinking: What develops?* Hillsdale, NJ: Erlbaum.

Chi, M. T. H., & Ceci, S. J. (1987). Content Knowledge: Its Role, Representation, and Restructuring in Memory Development. In H. W. Reese (Ed.), *Advances in Child Development and Behavior*. (pp. 93–141). New York: Academic Press.

Dunbar, K. (1989). Scientific reasoning strategies in a simulated molecular genetics environment. In *Proceedings of the Eleventh Annual Meeting of the Cognitive Science Society* (pp. 426–433). Hillsdale, NJ: Erlbaum.

Dunbar, K., & Klahr, D. (1989). Developmental differences in scientific discovery strategies. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon*. Hillsdale, NJ: Erlbaum.

Dunbar, K., & Schunn, C. D. (1990). The temporal nature of scientific discovery: The roles of priming and analogy. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 93–100). Hillsdale, NJ: Erlbaum.

Fay, A. L., Klahr, D., & Dunbar, K. (1990). Are there developmental milestones in scientific reasoning? In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 333–339). Hillsdale, NJ: Erlbaum.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.

Gholson, B., Levine, M., & Phillips, S. (1972). Hypotheses, strategies, and stereotypes in discrimination learning. *Journal of Experimental Child Psychology*, 13, 423–446.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogic transfer. *Cognitive Psychology*, 15, 1–38.

Gorman, M. E. (in press). Using experiments to determine the heuristic value of falsification. In M. Keane & Gilhooly (Eds.), *Advances in the Psychology of Thinking, Vol. I*. Hemel Hempstead, Hertfordshire: Harvester Wheatsheaf.

Hergenrather, J. R., & Rabinowitz, M. (1991). Age-Related Differences in the Organization of Children's Knowledge of Illness. *Developmental Psychology*, 27, 952–959.

Holland, J., Holyoak, K., Nisbett, R. E., & Thagard, P. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Kail, R. (1991). Processing time declines exponentially during childhood and adolescence. *Developmental Psychology*, 27, 259–266.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, 22, 374–419.

Karmiloff-Smith, A. (1988). A child is a theoretician, not an inductivist. *Mind and Language*, 3, 183–195.

Keil, F. C. (1981). Constraints on knowledge and cognitive development. *Psychological Review*, 88, 197–227.

Kern, L. H., Mirels, H. L., & Hinshaw, V. G. (1983). Scientists' understanding of propositional logic: An experimental investigation. *Social Studies of Science*, 13, 131–146.

Klahr, D. (1985). Solving problems with ambiguous subgoal ordering: Preschoolers' performance. *Child Development*, 56, 940–952.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science*, 12, 1–55.

Klahr, D., & Robinson, M. (1981). Formal assessment of problem solving and planning processes in preschool children. *Cognitive Psychology*, 13, 113–148.

Klahr, D., Dunbar, K., & Fay, A. L. (1990). Designing good experiments to test 'bad' hypotheses. In J. Shrager & P. Langley (Eds.), *Computational models of discovery and theory formation*. San Mateo, CA: Morgan-Kaufman.

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation and information in hypothesis testing. *Psychological Review*, 94, 211–228.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96, 674–689.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. New York: Academic Press.

Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge*. New York: Cambridge University Press.

Markman, E. M., (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development*, 50, 643–655.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

Mitroff, I. I. (1974). *The Subjective Side of Science*. New York: Elsevier.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology*, 29, 85–95.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Piaget, J. (1952). *The origins of intelligence in children*. New York: International University Press.

Ross, B. H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology*, 16, 371–416.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology*, 49, 31–57.

Schauble, L., & Glaser, R. (1990). Scientific thinking in children and adults. In D. Kuhn (Ed.), Developmental perspectives on teaching and learning thinking skills. *Contributions to Human Development*, 21, 9–26.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' Transition from an Engineering Model to a Science Model of Experimentation. *Journal of Research in Science Teaching*, 28, 859–882.

Shaklee, H., & Paszek, D. (1985). Covariation Judgment: Systematic Rule Use in Middle Childhood. *Child Development*, 56, 1229–1240.

Shrager, J. (1987). Theory change via view application 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*, 10, 401–402.

Simon, H. A. (1977). *Models of discovery*. Dordrecht-Holland: D. Reidel Publishing Co.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, 62, 753–766.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1–10.

Tumblin, A., & Gholson, B. (1981). Hypothesis Theory and the Development of Conceptual Learning. *Psychological Bulletin*, 90, 102–124.

Vosniadou, S., & Brewer, W. F. (in press). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24.

Ward, S. L., & Overton, W. F. (1990). Semantic familiarity, relevance, and the development of deductive reasoning. *Developmental Psychology*, 26(3), 488–493.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology*, 20, 273–281.

Wellman, H. M. (1983). Metamemory revisited. In M. T. Chi (Ed.), *Trends in memory development research*. New York: Karger.

Wellman, H. M., & Somerville, S. C. (1984). The development of human search ability. In M. E. Lamb & A. L. Brown (Eds.), *Advances in developmental psychology*. Hillsdale, NJ: Erlbaum.

Wisniewski, E. J., & Medin, D. L. (1991). Harpoons and long sticks: the interaction of theory and similarity in rule induction. In D. H. Fisher, Jr., M. J. Pazzani, & P. Langley (Eds.) *Concept formation: Knowledge and experience in unsupervised learning*. San Mateo, CA: Morgan Kaufmann.

Wiser, M. (1989). Does learning science involve theory change? Paper presented at the Biannual Meeting of the Society for Research in Child Development, Kansas City, April 30, 1989.

(Accepted May 12, 1992)

NOTICE TO CONTRIBUTORS:

The publishers wish to call your attention to the following instructions for preparing manuscripts for *Cognitive Psychology*: Format and style of manuscript should conform to the conventions specified in the "Publication Manual of the American Psychological Association" (1200 Seventeenth Street, N.W., Washington, D.C. 20036; 1983 Revision), with the exceptions listed below. Please note that it is the responsibility of the author that manuscripts for *Cognitive Psychology* conform to the requirements of this journal.

INFORMATION FOR AUTHORS

Cognitive Psychology publishes original empirical, theoretical, and tutorial papers, methodological articles, and critical reviews dealing with memory, language processing, perception, problem solving, and thinking. This journal emphasizes work on human cognition. Papers dealing with relevant problems in such related areas as social psychology, developmental psychology, linguistics, artificial intelligence, and neurophysiology also are welcomed provided that they are of direct interest to cognitive psychologists and are written so as to be understandable by such readers. There are no maximum or minimum length restrictions for journal articles. Minor or very specialized studies are seldom accepted.

All manuscripts should be submitted to: Dr. Douglas L. Medin, Department of Psychology, Northwestern University, Swift Hall, 2029 Sheridan Road, Evanston, Illinois 60208.

Original papers only will be considered. Manuscripts are accepted for review with the understanding that the same work has not been published, that it is not under consideration for publication elsewhere, and that its submission for publication has been approved by all of the authors and by the institution where the work was carried out; further, that any person cited as a source of personal communications has approved such citation. Written authorization may be required at the Editor's discretion. Articles and any other material published in *Cognitive Psychology* represent the opinions of the author(s) and should not be construed to reflect the opinions of the Editor(s) and the Publisher.

Authors submitting a manuscript do so on the understanding that if it is accepted for publication, copyright in the article, including the right to reproduce the article in all forms and media, shall be assigned exclusively to the Publisher. The Publisher will not refuse any reasonable request by the author for permission to reproduce any of his or her contributions to the journal.

A manuscript submitted for publication is judged by three main criteria: (a) appropriateness of the subject matter for this journal; (b) significance of its contribution to knowledge; and (c) clarity and conciseness of writing. No changes in a manuscript may be made once it has been accepted and is in press.

Form. Type at least double-spaced throughout, including tables, footnotes, references, and figure captions, with 1 inch margins on all sides. Submit four complete copies. Each copy *must* include all figures and tables.

Number the pages consecutively. *Page 1* should contain the article title, author(s) name(s), and affiliation; at the bottom of the page type a short title, not exceeding 35 characters and spaces, and the name and complete mailing address (including zip code) of the person to whom proofs should be sent. The address of correspondence should be given as a footnote to the appropriate author's name. *Page 2* should contain a short abstract, approximately 100 to 150 words in length.

Headings. The organization of the paper must be clearly indicated by appropriate headings and subheadings.

Abbreviations. Do not use final periods with units of measure that are abbreviated (cm, s, kg, etc.) in text or in tables, except for "in." (inch).

Symbols. Underline letters that represent mathematical symbols; these will be set in *italic* type.

Equations. Number displayed equations consecutively, with the number placed in parentheses to the extreme right of the equation. Refer to numbered equations as Equation 1 or say "the first equation." Punctuate equations to conform to their place in the syntax of the sentence.