COGNITIVE SCIENCE Vol 23 (3) 1999, pp. 337–370 ISSN 0364-0213

The Generality/Specificity of Expertise in Scientific Reasoning

CHRISTIAN D. SCHUNN
George Mason University

JOHN R. ANDERSON
Carnegie Mellon University

Previous research on scientific reasoning has shown that it involves a diverse set of skills. Yet, little is known about generality or domain specificity of those skills, an important issue in theories of expertise and in attempts to automate scientific reasoning skills. We present a study designed to test what kinds of skills psychologists actually use in designing and interpreting experiments and contrast expertise within a particular research area with general expertise at designing and interpreting experiments. The results suggest that psychologists use many domain-general skills in their experimentation and that bright and motivated Carnegie Mellon undergraduates are missing many of these skills. We believe that these results have important implications for psychological and artificial intelligence models of expertise, as well as educational implications in the domain of science instruction.

I. INTRODUCTION

What are the reasoning skills required for making scientific discoveries? Research in cognitive psychology has infrequently, although consistently, addressed variations of this question since the advent of the cognitive revolution in the 1950s (e.g., Bruner, Goodnow, & Austin, 1956; Dunbar, 1993; Klahr & Dunbar, 1988; Langley, Simon, Bradshaw, & Zyktow, 1987; Mynatt, Doherty, & Tweney, 1977; Wason, 1960; Wason & Johnson-Laird, 1972). One of the reasons that scientific reasoning has been studied infrequently is its complexity—scientific reasoning involves a diverse collection of cognitive activities rather than a single cognitive process. By contrast, one of the reasons for the long-lived

Direct all correspondence to: Christian D. Schunn, Department of Psychology, MSN 3F5, George Mason University, Fairfax, VA 22030-4444; E-Mail: cschunn@gmu.edu; John Anderson, Departments of Psychology and Computer Science, Carnegie Mellon University.

interest in studying scientific reasoning is the importance of the process. Scientific discoveries are often considered among the pinnacle achievements of humanity.

Although much of the validity in studying scientific reasoning processes lies in the real-world nature of the task, it is an interesting fact that the great majority of psychological research on scientific reasoning has studied neither actual scientists nor used actual scientific tasks. Instead much of the research has used undergraduates working on artificial problems (e.g., Bruner et al., 1956; Klahr & Dunbar, 1988; Mynatt et al., 1977; Schunn & Klahr, 1992; Wason, 1960), although a few studies have investigated real scientists working on real scientific tasks (e.g., Dunbar, 1994; Kulkarni & Simon, 1988; Mitroff, 1974). Of course, there are many advantages to using undergraduates and artificial tasks, and there are clear disadvantages to using actual scientists and actual scientific tasks. Two important factors favoring the use of undergraduates and artificial tasks are the issues of control and statistical power. On the issue of control, scientists are not typically available to serve as subjects in the psychology lab, and their real-world environment is difficult, if not impossible, to experimentally manipulate. Moreover, the knowledge and skills of the scientist are not subject to control and frequently confound over-all intelligence with domain-specific knowledge. Relating to the issue of statistical power, if one tries to study scientists in their real-world setting, one almost inevitably winds up studying research on a specific problem in a specific lab. It is difficult to know how to generalize from such specific instances. Thus, studies of scientific reasoning have almost exclusively used numerous undergraduates in the psychology lab (e.g., Dunbar, 1993; Klahr & Dunbar, 1988; Qin & Simon, 1990; Schunn & Klahr, 1995), or small *N*, historical (e.g., Kulkarni & Simon, 1988) or observational (e.g., Dunbar, 1994; Mitroff, 1974) studies of scientists in their own environments.

Comparing results across these different study types could provide important information about the generality of the findings. For example, from results within only one study type, it is unclear which findings are general to all inductive reasoning tasks or all problem solving tasks, which are specific to scientific reasoning, which are specific to scientific reasoning in a particular scientific domain, and which are specific to only a few brilliant researchers. However, it is difficult to know how to bridge across these study types because they simultaneously differ on two important dimensions: the nature of the task (e.g., its difficulty and familiarity) and the nature of the participant populations (e.g., their general reasoning abilities and experience with research methods).

One approach to bridging these research traditions is to use a quasi-experimental design that involves just one representative scientific reasoning task and systematically manipulates various forms of expertise. This approach has two clear exemplars. The first such study was Voss, Tyler, and Yengo's (1983) study of scientific reasoning in political science. They gave four groups of participants a political science problem relating to agriculture in the Soviet Union: "Assume you are the head of the Soviet Ministry of Agriculture and assume crop productivity has been low over the past several years. You now have the responsibility of increasing crop production. How would you go about doing this?" The groups differed along several dimensions of expertise. The first group of participants were political science faculty whose domain of expertise was the Soviet

Union. The second group consisted of political science faculty whose domain of expertise was some other area (e.g., Latin America). The third group consisted of faculty from a chemistry department. The final group consisted of undergraduates taking a course on Soviet domestic policy.

Voss et al. (1983) found group differences at three different levels. First, they found that both groups of political scientists spent a great deal of time developing a representation of the problem, whereas the undergraduates and chemists spent little time developing representations. Second, they found that both groups of political scientists proposed only a few abstract solutions, whereas the undergraduates and chemists proposed a large number of more concrete solutions. Finally, they found that the domain-expert political scientists provided long argument chains in support of their solutions, whereas all other groups provided only short argument chains. Thus, they found that ability to produce long reasoning chains depends upon familiarity with the domain and that the skills of elaborating problem representations and generating abstract solutions depend upon familiarity with the task. It appears from these results that these skills are not general to all areas of science: the chemists performed no differently than the undergraduates. However, it is possible that the chemists possessed these skills, but simply did not have the supporting knowledge to use the skills in this task.

A second, related study, conducted by Shraagen (1993), which used a problem from sensory psychology. In this task, participants were told that the manufacturer of Coca Cola wanted them to design a study to determine what people taste when they drink Coca Cola, making comparisons to Pepsi Cola and another house brand. Four groups of participants were used: domain-experts, who were psychology faculty working within the domain of sensory psychology; design-experts, psychology faculty working in other areas of psychology; intermediates, who were graduate students from other areas of psychology; and beginners, who were undergraduates studying psychology.

Shraagen (1993) found that both groups of experts worked at the problem longer and produced more solutions than did the other groups. More interesting, design-experts, like domain-experts, used a much more systematic method for designing experiments than did the intermediates or the beginners. In particular, from analyses of the verbal protocols, Shraagen (1993) found that the experts were more likely to follow the general control structure of moving from understanding the problem to selecting a research paradigm to pursuing that paradigm to selecting design principles. Shraagen (1993) also found that the two groups of experts shared two heuristics in common: mental simulation and progressive deepening of the design. Yet, the domain-experts did differ from the design-experts in one important way: the domain-experts produced much better solutions, as rated by other domain-experts. In particular, it appeared that the design-experts produced solutions that, while perfectly logical, were unlikely to work for various practical reasons.

Thus, from Shraagen's (1993) study, one might conclude that domain expertise consists merely of practical knowledge of what will work in the given domain and that there is great generality to the skills of designing experiments. Interestingly, the heuristics of mental simulation and progressive deepening design have also been found in program-

ming expertise (Adelson, 1990; Adelson & Soloway, 1985), and thus it appears that they are general to expertise in many areas other than science.

While the Voss et al. (1983) and the Shraagen (1993) studies have provided some answers about the nature and generality of scientific reasoning skills, many questions remain. First, those two studies identified only a few of the skills required for scientific reasoning, all within the design process. Presumably a complex task, such as making scientific discoveries, requires many more skills within the design process and within other aspects. For example, there are also the processes of deciding how to measure and plot the outcomes of experiments, interpreting the experimental outcomes, generating hypotheses, and comparing results to hypotheses (cf. Schunn & Klahr, 1995). It would also be interesting to know how general these skills are.

Second, the tasks used in the Voss et al. (1983) and Shraagen (1993) studies were not typical scientific tasks, which usually have goals of obtaining new knowledge or theories or testing existing scientific theories. Instead, they were more like engineering tasks geared at obtaining practical outcomes. This issue raises the question of whether their results would generalize to more typical scientific tasks. For example, previous studies have found that people are less likely to conduct systematic experiments and to try to identify the causal relevance of individual variables when they have an engineering goal than when they have a science goal (e.g., Schauble, Klopfer, & Raghavan, 1991; Tschirgi, 1980). Thus, the lack of systematicity in the novices in the Shraagen (1993) study may be due to those participants mistakenly interpreting the task as an engineering task.

Third, the Voss et al. (1983) and Shraagen (1993) studies did not provide any opportunities for the participants to make use of feedback. Scientists in the real world rely on the ability to test their ideas empirically and to iteratively attack a problem (Tweney, 1990)—scientific questions are rarely answered in one experiment and especially not the first one. With a small amount of feedback (e.g., being presented with the outcomes of experiments), the task-experts in Shraagen's (1993) study may have developed much better experiments. Alternatively, with feedback, the domain-experts in the Voss et al. (1983) study may have demonstrated some procedural differences from the task-experts.

The current study was designed to address these issues. As with the Voss et al. (1983) and Shraagen (1993) studies, our study contrasts domain-experts with task-experts and task-novices. However, in contrast to the two previous studies, the current study had three new features. First, it uses a more typical scientific task of trying to distinguish among two general theories regarding an empirical phenomenon. Of course, no task can represent all sciences equally, and our task will be most representative of the scientific methodology used in psychology. Second, it makes use of a new computer interface that allows subjects to see the results of their experiments and to design multiple experiments based on the feedback they receive. Third, this study investigates the processes by which the participants examined the experimental outcomes.

Three general questions were at issue in our study and will shape the analyses and discussion of the results. First, is there a general set of skills that scientists use in designing and interpreting experiments? It may be that there are relatively few general skills that hold true across domains, or even across scientists. Such a paucity would explain why

previous studies have not described a greater number of general skills. That is, it may be that the majority of skills required for scientific reasoning are domain specific. For example, outcome interpretation skills may consist primarily of recognizing familiar patterns found within that domain.

Second, of the domain-general skills that do exist, are these general skills unique to scientists, or would any intelligent, motivated individual possess them? The abundance of studies that investigate scientific reasoning skills using undergraduate populations suggest that many psychologists believe that scientific reasoning skills can be found in non-scientists as well. Consider, for example, the skill of systematically investigating the causal relevance of different variables, a skill at which most scientists are likely to be proficient. While some adults and children may not have mastered this skill, it may be that many well-educated adults are proficient at systematic, causal investigation. For example, Kuhn (1991) found that students at college-track schools had much better basic scientific reasoning skills than did students at vocational-track schools, and that age (14−60 years) had little impact. One goal of our study is to examine the empirical generality of these basic reasoning skills, and thus provide important information about the nature and origins of these skills.

Third, assuming there are both domain-general skills and domain-specific skills, is the empirical generality of these skills easily predicted by a task analysis (or simple symbolic model)? According to the Singley and Anderson (1989) Identical Elements theory of transfer, skills should transfer across domains and tasks to the extent to which the domains and tasks share those skills. However, it may be that skills that logically should generalize do not transfer because they require supporting domain-specific knowledge. For example, the Investigate-Surprising-Phenomena heuristic proposed by Kulkarni and Simon (1988) requires domain-specific expectations to elicit a surprise response. As another example, Shoenfeld (1985) found that success in applying Polya's general problem solving heuristics depended on knowledge about the mathematical domain. Additionally, many recent theories of learning and transfer have proposed that transfer of knowledge and skills is a non-trivial matter (e.g., Clancey, 1993; Greeno, 1988; Lave, 1988; Suchman, 1987). For example, task-experts may not be able to transfer their general skills because the current task is not situated in their familiar domains.

Alternatively, it may be that there are domain-specific skills that are very easily acquired. That is, rather than requiring years of domain-specific experience, many of the domain-specific skills in scientific reasoning may be logically deduced or readily acquired with minimal feedback. Such a finding would be consistent with a general reasoning ability view of expertise and inconsistent with current theories of expertise, which assume that expertise consists primarily of domain-specific knowledge that requires years of experience to master (Ericsson, Krampe, & Tesch-Römer, 1993; Gobet & Simon, 1996).

The design of the current study was to contrast domain-experts with task-experts and task-novices on a novel problem from within the domain-experts' domain. We selected a problem that simultaneously satisfied three constraints: 1) the solution must be unknown to the domain-experts because science involves the discovery of previously unknown solutions; 2) the problem must be free of domain-specific jargon and easily understandable

to even task-novices; and 3) the solution must be obtainable through experimentation. We found a problem from within the domain of cognitive psychology that seemed to meet these three constraints. It particular, the problem that we gave the participants was to find the cause of the spacing effect in memory—that items with longer intervening intervals tend to be better remembered.

The structure of our study was to give domain-experts, task-experts, and task-novices a step-by-step introduction to the spacing effect, as well as two theories about the cause of the spacing effect. The participants' goal was to develop experiments that could determine which explanation of the spacing effect was correct. The participants used a computer interface to design the experiments. The interface provided a set of variables that could be manipulated, as well as the facility to easily conduct factorial-design experiments. The use of the computer interface allowed the participants to observe the outcomes of their experiments by embedding a mathematical model of memory consistent with existing research into the interface. Additionally, the computer interface produced a rich keystroke protocol that provided detailed information about how the participants designed and interpreted their experiments.

Distinctions of Expertise

Before turning to the experiment, a few words regarding the types of expertise that were being considered are in order. Both the Voss et al. (1983) and Shraagen (1993) research distinguished between expertise at the task of doing science and expertise in a particular scientific domain (e.g., political science, chemistry, or sensory psychology). Although this distinction has intuitive appeal, its underlying theoretical distinction requires further specification, especially if testable predictions are to be made. Moreover, it is yet unspecified how this distinction could be captured mechanistically or computationally.

One distinction from the philosophy, psychology, computer science, and neuroscience literature that provides further theoretical illumination to this issue is the procedural/declarative distinction (Anderson, 1976; Cohen & Squire, 1980; Ryle, 1949). Declarative knowledge is characterized by knowledge that people can report or describe (i.e., knowing that), whereas procedural knowledge is characterized by knowledge that appears in performance but cannot be reported (i.e., knowing how). There are many other behavior differences associated with this distinction (Anderson, 1976, 1983, 1993), and some researchers have argued for neurophysiological differences in their underlying representations (e.g., Cohen & Squire, 1980). Computationally, this distinction can be captured within a production system such as ACT-R (Anderson, 1993; Anderson & Lebiere, 1998). Specifically, procedural knowledge is embodied in productions, and declarative knowledge is embodied in declarative memory chunks.

One can conceive of the process of scientific reasoning as applying two types of reasoning procedures to two types of declarative knowledge. One type of declarative knowledge represents the specific scientific problem being attacked, its specification, and the results of specific experiments. If the problem is a novel problem, the problem-specific declarative knowledge will inherently be new to everyone, domain-novice or domain-

expert. The other kind of declarative knowledge is background knowledge that is privy to experts in the domain (e.g., chemical formulas, results of similar experiments, etc.). The procedures will also be of two types: some of the scientific reasoning procedures will be domain general (e.g., how to read a table), whereas others may be domain specific (e.g., knowing how to do chemical titration). It is important to note that domain-general procedures will fail to be applied across domains if they depend upon domain-specific knowledge.

The goal of our research was to determine whether there are domain-general procedures that expert scientists from different domains share, but are not found in all educated adults (i.e., are acquired as a result of detailed training and practice in a domain, but are still shared across domains). This was determined by comparing performance across domain-experts, task-experts, and undergraduates (task- and domain-novices). The assumption was that the domain- and task-experts would differ only in terms of domain-specific knowledge and domain-specific procedures, whereas the task-experts will differ from the undergraduates in terms of domain-general procedures. Predictions for particular skills will be made by considering the type of knowledge and procedures that they require.

An important caveat is that expertise is not dichotomous. Although for experimental reasons expertise is typically treated as a dichotomous variable, expertise in science, as in most domains, is a continuum of gradually accumulated knowledge and skill (e.g., child to adolescent to undergraduate student to graduate student to postdoc to faculty member). Moreover, the generality of the skills and knowledge are also more of a continuum than a dichotomy. For example, some task skills, such as knowing how to calibrate a particular instrument, are specific to a very narrow subdomain, whereas other task skills, such as multiple regression statistics, are general to a number of domains, and others still, such as knowing to avoid confounded variables, are general to all areas of science. Similarly, some knowledge, such as the size of a particular star, is useful in a narrow subdomain, whereas other knowledge, such as the chemical composition of water, is useful in many domains.

In our experiment, we focused on the extremes of generality: skills that are expected to be general to most if not all science versus skills and knowledge that are expected to be specific to a particular subarea within cognitive psychology. This focus minimizes the ambiguity of predictions as wells as maximizes the generality of the skills that are found.

II. METHODS

Participants

There were three categories of participants: cognitive psychology faculty who have done much of their research in the domain of memory (Domain-Experts; $n = 4$), social and developmental psychology faculty who have done very little or no research related to memory (Task-Experts; $n = 6$), and undergraduates ($n = 30$) from a variety of engineering, physical science, social science, and arts backgrounds.[1] The undergraduates were all from Carnegie Mellon University and participated for course credit. The Domain-Experts

and Task-Experts were a mix of senior and junior faculty with prolific publication records at strong research universities. The Domain-Experts and Task-Experts were well matched on a variety of background measures: the two groups did not differ significantly in the number of years since earning a PhD (17.5 versus 13.0, $t(8) = 0.8$, $p > .45$), number of journal publications (41.0 versus 36.0, $t(8) = 0.3$, $p > .7$), number of non-journal publications (27.5 versus 21.5, $t(8) = 0.6$, $p > .5$), and number of publications in the year preceding this study (3.0 versus 3.0, $t(8) = 0$, $p > .99$). However, there were large differences in the number of publications on the topic of memory (20.0 versus 1.5, $t(8) = 5.8$, $p < .0005$).

We expected the Domain-Experts alone to display skills specific to the task of studying memory and only the Domain-Experts and Task-Experts to display skills general to the task of conducting scientific experimentation. However, it is possible that some of the skills may be so general such that all groups would display the domain-general skills. Moreover, it is possible the Domain-Experts, Task-Experts, and only the brighter undergraduates would display the domain-general skills. Thus, as will be explained in further detail later, the undergraduates were divided into two groups according to ability.

Materials and Procedure

At the beginning of the experiment, the undergraduate participants were given a background questionnaire regarding 1) their current and previous psychology courses; 2) the number of courses, including high school, in physics, biology, chemistry, mathematics, computer science, philosophy, social and decision sciences, and architecture and engineering; 3) the number of lab courses; 4) their math and verbal SAT scores; and 5) whether or not they had taken part in a research project previously.

Next, all the participants were given a step-by-step introduction into the main task on the computer. The participants read the instructions aloud to familiarize themselves with talking aloud and to insure that they read all the instructions. The instructions began with an introduction to the simple form of the spacing effect—that spaced practice produces better memory performance than massed practice.[2] The participants were then told that their main task was to design experiments to discover the cause of the spacing effect. The spacing effect was chosen as the target problem because it is 1) simple to describe, 2) easy to understand, and 3) the cause yet unknown even to experts in the field of memory.

There are two common theories for the cause of the spacing effect. These theories were simplified and presented to the participants as possible causes of the spacing effect. The first theory was the shifting context theory, which stated that memories were associated with the context under study and that context gradually shifted with time. Under this theory, the spacing effect occurs because spaced practiced produces associations to more divergent contexts, which in turn are more likely to overlap with the test context. This theory is consistent with associative models of memory such as SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) and Landauer's random walk model (Landauer, 1975).

The second theory was the frequency regularity theory, which stated that the mind estimates how long memories will be needed based on regularities in the environment and,

in particular, adjusts forgetting rates according to the spacing between items. Under this theory, items learned with short intervening spaces are forgotten quickly because they need not be remembered for very long, whereas items learned with long intervening spaces are forgotten more slowly because otherwise they would be forgotten before they were needed again. In other words, it is assumed that when items do not appear after a long delay since their last appearance, it is less likely that they will ever appear again, if they previously appeared with short intervening delays than if they previously appeared with long intervening delays, and it is assumed that the mind is adjusting forgetting rates to reflect these expectations about information being needed again. This theory is consistent with rational models of memory (e.g., Anderson & Schooler, 1991).

After the introduction to the spacing effect and the two theories for its cause, the goal of the main task was given: to discover whether either, neither, or both of the two theories about the cause of the spacing effect are correct. To achieve this goal, the participants were asked to design experiments. Because we were interested in the process by which people interpreted data in addition to how they designed experiments, we built a computer interface, called the Simulated Psychology Lab, that produced simulated experimental outcomes and allowed participants to iterate through the process of designing and interpreting. The interface supported factorial experimental designs because that was the most common design generated by Domain-Experts and graduates students in pilot experiments.

Within the interface, participants designed experiments by selecting values for six independent variables. Participants could simultaneously manipulate up to four of the independent variables in any one experiment. The participants were told that the computer had been given the results of many actual experiments, and that it would show them the results of whatever experiment they generated. The six independent variables were divided into three variables of the source or study task and three variables of the test situation. All the manipulations would be between groups, and the basic task consisted of studying lists of words.[3]

The source task variables that the participants could manipulate were 1) repetitions, the number of times that the list of words was studied (2, 3, 4, or 5 times); 2) spacing, the amount of time spent between repetitions (from 1 min to 20 days); and 3) source context, whether the participants were in the same context for each repetition or whether they changed contexts on each repetition (either by changing rooms or by having their moods manipulated). The test variables included 1) the test task, free recall, recognition, or stem completion; 2) delay, the amount of time from the last study repetition until the test was given; and 3) test context, whether the participants were in the same context or a different context at test relative to study (either by changing rooms or having moods manipulated).[4] Figure 1 shows the six variables that could be manipulated and their current settings as they appeared on the screen. In this example, the experiment is only partially specified, with only the repetitions and spacing variables determined. Repetitions was not manipulated (it was held constant at 3) and spacing was manipulated (5 min versus 20 min). An experiment was not complete until values for all six variables were specified.

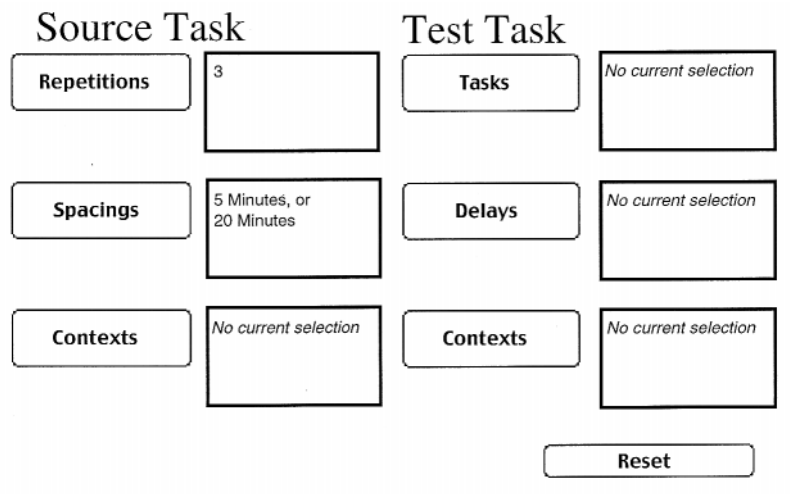

Figure 1. The interface used to display the variables that could be manipulated and their current settings in the experiment being designed.

The six variables that the participants could manipulate were ones that were produced in a pilot task involving the free generation of experiments on paper. Although the mapping of the two theories for the spacing effect onto these six variables is not simple, this relationship between theory and operational variable is typical of most psychological theories and experiments. Of interest in this study was how the participants would make the mapping from the theories onto these variables.

Participants selected variable settings by clicking on sliders and scrolling lists by using a mouse. Figure 2 provides an example of how values were selected for the source context variable. The slider on the left was used to select whether the variable was held constant across subjects (by setting it to 1; Figure 2A) or whether the variable was manipulated (by setting it to 2 or 3; Figure 2B). Values for repetition were selected using a slider that varied from 2 to 5. Values for context and test task were selected from a scrolling list of three options. Values for spacing and delay were selected by using a slider (which varied from 1 to 20) and a scrolling list of units (minutes, hours, or days). Thus, almost 400,000 unique experiment settings could be generated. Values had to be selected for all six variables, including the variables that were held constant in the given experiments; no default values were used. There was no restriction on the order of variable selection, and participants could go back to change their selections for any of the variables at any point in time until they selected to run the experiment.

Participants could only vary up to four variables simultaneously in any given experiment. A schematic of the size of the current design was displayed, and this was replaced with a "Too many factors!" warning if the participants tried to manipulate more than four variables at once. The restriction to only four variables was a practical consideration of being able to plot experiment outcomes; however, we were interested in whether the participants would seek to design experiments with more than four manipulated variables, and thus we did not inform the participants of the restriction in advance.

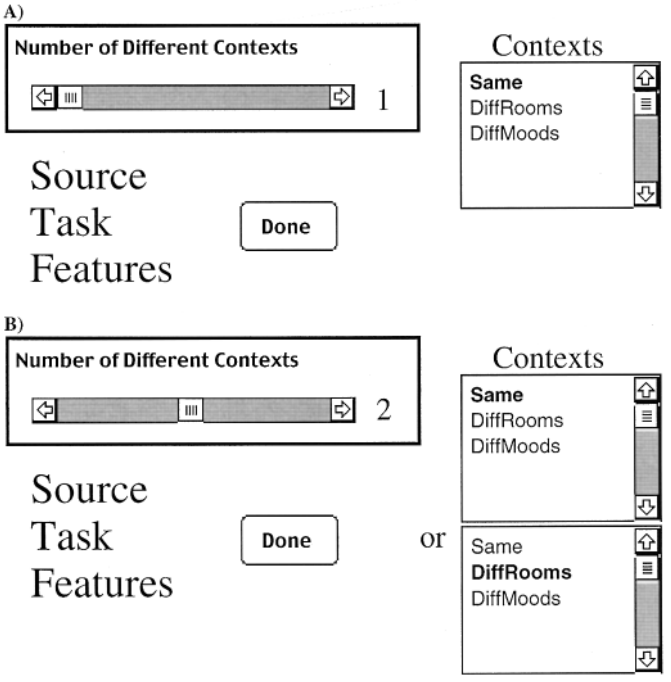

Figure 2. The interface used for selecting how many levels of a independent variable and the values for each level. A) Only one level is selected (i.e., this variable is held constant), and the constant value is "Same". B) Two levels are selected (i.e., this variables is manipulated), and the values are "Same" and "DifferentRooms".

The participants made predictions and were given outcomes in a table format, with all cells being shown at once. Tables rather than graphs were used because we thought that tables were easier for undergraduates to understand and manipulate. Before being given the table, participants had to select on which dimension each manipulated variable would be plotted (i.e., rows, columns, across tables vertically, or across tables horizontally). This choice was presented by using a schematic table with the correct number of dimensions for the current experiment and a button associated with each dimension. Participants made a selection for each dimension by clicking on the associated button and selecting from a provided list of the manipulated dimensions.

After choosing the table structure, participants predicted the mean percent correct for each cell in the design. Although this prediction task is more stringent than the prediction task psychologists typically give themselves (i.e., directional predictions at best, and rarely for all dimensions and interactions), we used this particular form of a prediction task because 1) assessing directional predictions proved a difficult task to automate; 2) numerical predictions could be made without explicit thought about the influence of each variable and possible interactions, and thus we thought it was less intrusive; 3) it provided further data about the participants' theories and beliefs about each of the variables; and 4)

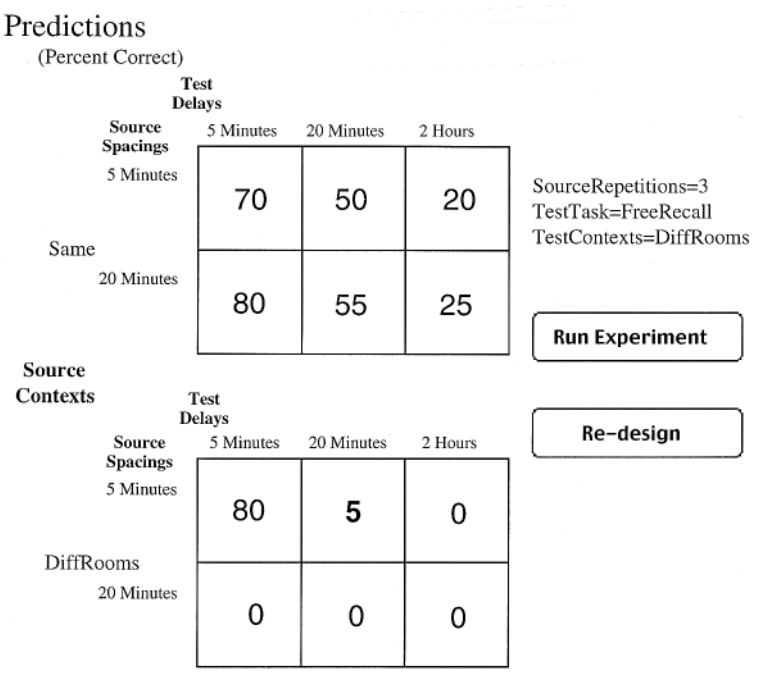

Figure 3. The interface used for making predictions. In this example, predictions have been entered for the majority of the cells (zero is the default value). Test Delays are in the columns, Source Spacings in the rows, and Source Contexts across tables. The value currently being entered is in bold type.

it provided some cost to large experimental designs (i.e., many more predictions to make) to simulate the increasing real-world cost of larger experimental designs.[5] During the prediction phase, participants could choose to redesign the experiment (or re-assign variables to the table dimensions). To the right of the prediction table, the values for the variables held constant were displayed as well (see Figure 3).

After completing their predictions, the participants were shown the results of their experiment. The same table format was used, including the display of the variables held constant and their values (see Figure 4). In addition, the outcome tables also displayed the participant's predictions in italics for each cell. To facilitate comparison across rows, columns, and tables, the row, column, and table marginals were also provided. To provide a rough evaluation of the quality of the predictions, the participants were also shown the Pearson correlation between the predictions and outcomes. Figure 4 also illustrates the key results in this task: the effects of spacing and delay, and their interaction (the drop-off with increasing delay is faster at smaller spacings), and the lack of an effect of source context.

To be roughly consistent with results from research on memory and the spacing effect, the following underlying model of memory was used to produced the experiment outcomes. Memory performance decreased as a power-function of Delay, in which the decay constant was an inverse power-function of Spacing, i.e., the spacing effect was roughly

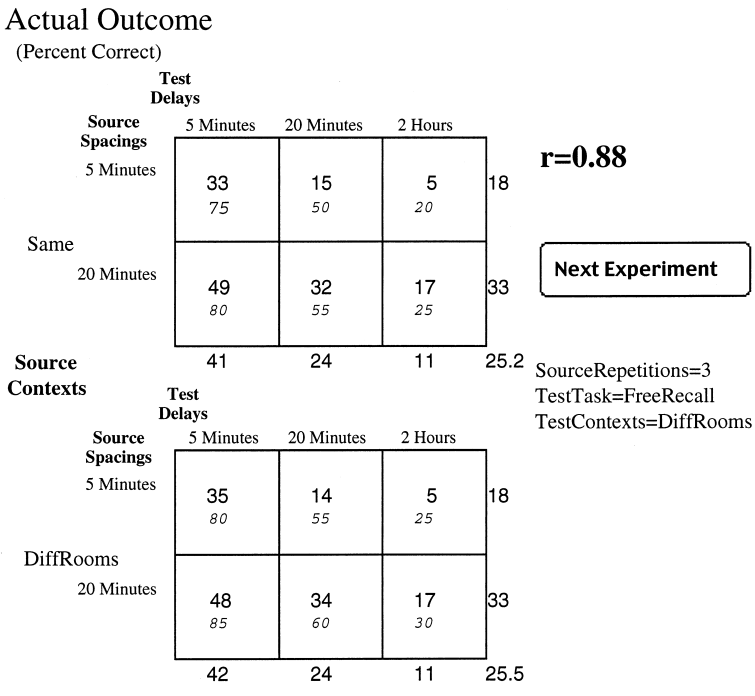

Figure 4. The interface used for displaying the outcomes of experiments. Actual outcomes are the main entry in each cell. Predicted outcomes are in italics. The *r*-value is the Pearson correlation between the predictions and actual outcomes.

consistent with the frequency regularity theory. Repetitions produced a very small, additive increase. Source Context had no effect on performance (ruling out the shifting context theory), whereas Test Context had a moderate additive effect (same context produced better performance than did different contexts, with different moods producing the worst performance). Performance for the recognition and recall test tasks was nearly identical, with recall performance merely being .9 of recognition performance. Stem completion, being an implicit test of memory, was not influenced by spacing. In sum, five of the six variables had an effect, and there were two significant interactions (Spacing $\times$ Delay and Spacing $\times$ Test Task). To simulate noise in the data, between 0% and 2% were added to or subtracted from each cell, with 0% and 1% noise being twice as likely as 2%.

While designing experiments, the participants also had the option of viewing the outcomes of previous experiments. Previous experiments were accessed by their experiment number and were displayed with all the same information of the original outcomes tables minus the predictions. However, as an additional memory aid, participants were also given paper and pen. They were instructed that the computer would keep track of all the experiments and outcomes, but that they were free to use the paper and pen as they wished.

Participants worked at the task-iterating through the process of experiment design, choosing a table structure, making predictions, and viewing outcomes until they felt that

they had found out what the cause of the spacing effect was or 40 min had elapsed. The primary data gathered in this experiment were keystroke data as the participants generated experiments, chose the table structures, and interpreted experiments. However, the participants were also asked to give a think-aloud verbal protocol throughout the task (Ericsson & Simon, 1993). Moreover, at the end of the task, participants were asked to verbally report their conclusions about the spacing effect, i.e., whether the Shifting Context theory, Frequency Regularity theory, both theories, or neither theory explained the spacing effect. The participants were also asked to give conclusions about the effects of each of the six variables.

III. RESULTS AND DISCUSSION

Overview

The goal of these analyses was to examine the presence and absence of domain-specific skills and knowledge (specific to the domain of the cognitive psychology of memory) and domain-general skills (general to most sciences). The skills were divided into three general classes: designing an experiment, making predictions, and interpreting outcomes. Skills were also divided into domain-general (skills expected to be useful at least generally in psychology) and domain-specific (skills expected to be useful only in memory experiments). Table 1 presents a full list of the skills. As it turned out, all the domain-specific skills happened to be experiment design skills, not because others do not exist, but simply because those proved to be the only ones that could be reliably assessed within this task. The differential representation of skills from the design, predict, and interpret categories (and the lack of skills associate with choosing a table structure) similarly reflects the degree to which such skills could be reliably assessed in the given task rather than the actual number of such skills associated with those types of activities in general.

To make predictions for each skill as to whether it was likely to be domain-general or domain-specific, productions implementing the skill were developed. If the production seemed applicable to scientific domains other than psychology and did not retrieve (i.e., require matching to) declarative knowledge specific to the cognitive psychology of memory, then the corresponding skill was classified as general. By contrast, if the production had to be written in such a way that it seemed applicable to only the cognitive psychology of memory or retrieved declarative knowledge specific to that domain[6], then the corresponding skill was classified as specific. English forms of the productions are listed in Table 1. A full production system model of this task that implements these skills and the many others required to do this task has been implemented in ACT-R. A detailed description of this model can be found in Schunn and Anderson (1998). In this detailed model, the only feature of importance to the current discussion is that some of the English productions listed in Table 1 actually correspond to many productions in the model (e.g., encoding main effects and interactions).

To test these predictions of generality (i.e., examine the empirical generality of the skills), we structured each of the analyses by contrasting performance for each of the

TABLE 1
List of Skills Examined by Skill Type and Skill Level, Along With English Form of Productions That Implement Them

	Skill	Production
Design		
General	Design experiments to test the given theories	If given theories to test, then set goal to test some aspect of theory.
General	Keep experiments simple	If variable is not relevant to hypotheses under test, then hold variable constant.
General	Use sensible variable values	
	(sufficient range)	If have multiple values to choose, then pick values that are far apart.
	(non-decreasing intervals)	If choosing third value of a variable, then choose equal increment as between first and second values.
	(minimal complexity)	If manipulating a variable, then choose simple, canonical manipulations.
General	Keep general settings constant across experiments	If not varying a variable, then pick the same value as used in the previous experiment.
Specific	Avoiding floor and ceiling effects	If choosing a variable value and know that value will produce very high or very low performance, then do not choose that value.
Specific	Knowledge of variables likely to interact	If testing a memory phenomenon, then consider varying test task.
Specific	Choose variable values useful in the given domain	If selecting a value for delay, then pick a value in the same range as the value selected for spacing.
Predict		
General	Make caricature predictions of interactions	If predicting an interaction, then consider only canonical qualitative interactions.
Interpret		
General	Make conclusions in light of theories under test	If finished encoding the results of an experiment, then relate results to theories under test.
General	Use evidence in support of conclusions about theories	If making a conclusion about a theory, then present supporting evidence from empirical results.
General	Encode main effects	If mean outcome for level 1 is significantly different than mean outcome for level 2, then conclude there is a main effect.
General	Encode interaction outcomes	If the effect of variable X is significantly different at different levels of Y, than conclude there is an interaction.
General	Ignore small noise levels in data	If an effect or interaction is very small, then ignore it.

participant groups. We predicted that the Domain-Experts alone would display the domain-specific skills and that both the Domain-Experts and Task-Experts would display the domain-general skills. However, it is possible that some of the skills may be so general that all groups would display the domain-general skills, especially because all of the undergraduates have had at least some science instruction.

The comparisons between the Task-Experts and undergraduates is likely to confound two factors: task expertise and general reasoning ability. To examine the influence of

reasoning ability skills, the undergraduates were divided into two groups by using a median-split on Math SAT, a factor found to be predictive of performance in this domain and other scientific reasoning tasks (e.g., Schunn & Klahr, 1993). If the differences between the undergraduates and the Task-Experts were due only to task expertise, then there should be no differences between the two groups of undergraduates. Those undergraduates above the median (660) were called High-Ability undergraduates ($n = 14$, mean = 728) and those below were called Mid-Ability undergraduates ($n = 16$, mean = 586). Because our undergraduates all had Math SATs above the US national median, we used the label Mid rather than Low to aid future comparison to studies conducted with other populations. The two groups did not differ significantly in terms of mean Verbal SAT (613 versus 620) or number of psychology courses (0.4 versus 0.4), science courses (7.2 versus 6.0), or lab courses (1.4 versus 2.8). But, the Math SAT difference was associated with a difference in major subject of study when using the four divisions of Computer Science, Engineering, or Math (6 High versus 1 Mid); Arts (2 versus 3); Humanities (3 versus 7); and Science (3 versus 3).

Along these lines, the results will be presented in figures in which the x axis lists the participant groups in the order Domain-Experts, Task-Experts, High-Ability undergraduates, and then Mid-Ability undergraduates. The statistical analyses focused on three comparisons: Domain-Experts versus Task-Experts, Domain- and Task-Experts versus High-Ability undergraduates[7], and High-Ability versus Mid-Ability undergraduates. We expected that 1) all the experts would be near the ceiling on all the domain-general skills; 2) the two groups of experts would differ only on domain-specific skills; and 3) that the experts and undergraduates would differ on the domain-general skills. The contrast between the High- and Mid-Ability undergraduates provided a test of whether any observed differences between experts and undergraduates could be attributed to differences in general reasoning ability.

Because there were many measures being examined, we will not describe the details of the statistical tests in the prose whenever figures are presented. Instead, the figures will indicate which of the pair-wise comparisons reached statistical significance. A mark indicates that the group's performance on that measure differed significantly from the group immediately adjacent. Three levels of significance were used: *** for $p < .01$, ** for $p < .05$, and * for $p < .2$. The nontraditional $p < .2$ is included for two reasons. First, the expert groups were small, and so comparisons involving those groups have relatively low power, and thus traditional cutoff values would lead to unacceptably high probabilities of type-II errors. Such a high rate of type-II errors would be particularly grievous here because the key predictions are for no differences between Domain- and Task-Experts on domain-general skills. Second, of particular interest was the trend analyses across skills in which the existence or non-existence of many marginal results could prove telling. Fisher Exact tests were used for binary variables and *t*-tests were used for continuous variables.

Before examining the results from the skill-specific analyses, we begin with three general results regarding the overall performance of the groups. First, there is the performance on the overall goal of the discovery task: to determine which theory of the

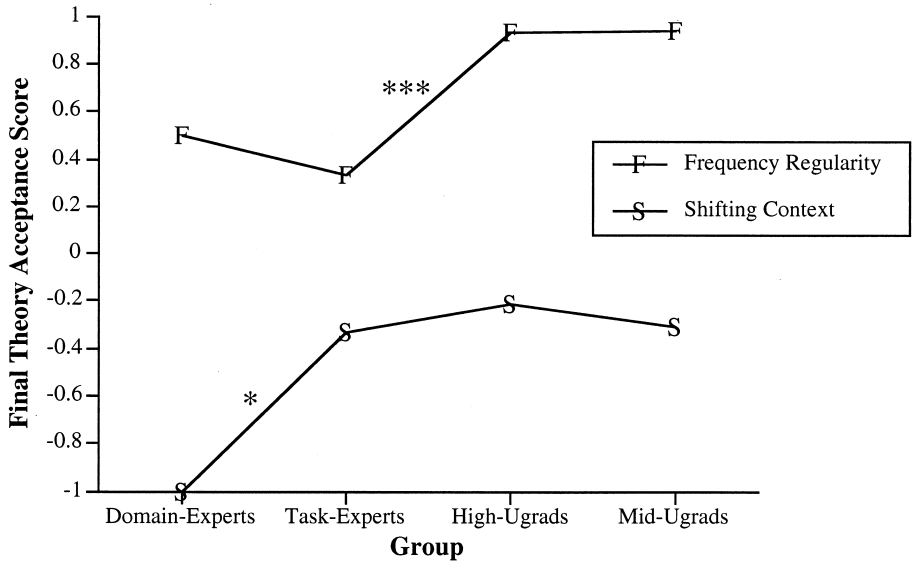

Figure 5. Mean acceptance score for each group for each of the two theories. ***, $p < .01$; *, for $p < .2$.

two (Frequency Regularity and Shifting Context) provides a good explanation for the spacing effect. The memory model built into the interface was strongly inconsistent with the Shifting Context theory and generally consistent with the Frequency Regularity theory. We coded the participants' final conclusions to examine whether the participants were able to discover and correctly interpret these results. One point was given for accepting a theory, zero points for no conclusion, and -1 for rejecting the theory.[8] Figure 5 presents the means for each group on each theory. Somewhat surprisingly, the undergraduates were more likely than the experts to accept the Frequency Regularity theory. This occurred because there were several results that were inconsistent with the Frequency Regularity theory, and the experts were more likely to notice these inconsistencies: 1) the Frequency Regularity theory implies an added advantage of exact matches of delay to spacing, which did not occur; and 2) the Frequency Regularity theory could not explain why there was no spacing effect for the stem completion test task. For the Shifting Context theory, the Domain-Experts all discovered that this theory was incorrect, whereas far fewer of the other participants were able to come to this conclusion.

Turning to time-on-task, the three groups spent approximately an equal amount of time to complete the task (36.0, 38.0, 36.3, and 34.0 min for the Domain-Experts, Task-Experts, High-Ability, and Mid-Ability undergraduates, respectively, ps $> .5$). However, Domain-Experts conducted marginally fewer experiments (2.8) than did the Task-Experts (4.8; $p < .2$), who in turn conducted about as many experiments as the undergraduates (5.6 and 5.7, respectively; $p > .6$). As we shall see below, this occurred because the Domain-Experts conducted a small number of complex experiments, whereas the other groups conducted a larger number of simple experiments.

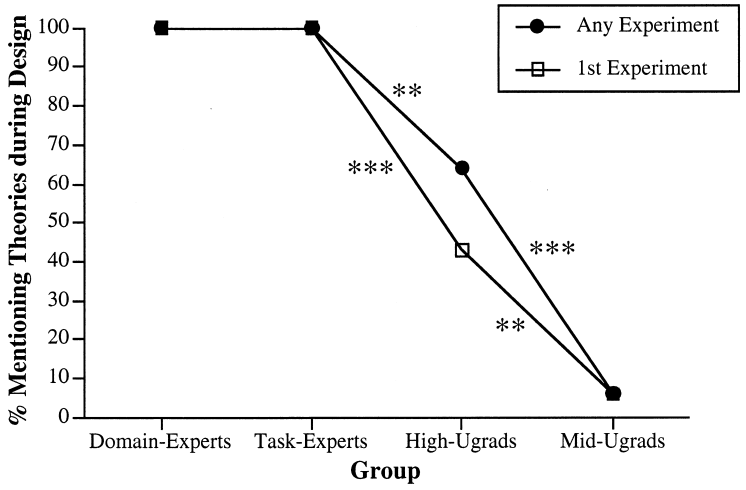

Figure 6 The percentage of participants in each group who mentioned the theories during experiment design in the first experiment or in any experiment. ***, $p < .01$; **, for $p < .05$.

Design—Domain-General Skills

In designing an experiment, the participants had two basic tasks: deciding which variables to vary and what values to use for each variable. There were also two kinds of skills: skills directly relevant to testing the given theories and practical rules-of-thumb.

Design Experiments to Test the Given Theories. Using the verbal protocols, we classified the participants according to whether or not they mentioned either of the two theories (Frequency Regularity and Shifting Context) during the course of design experiments, either during the first experiment or during any experiment. Note that this is a very lax criterion for measuring the use of theories in experiment design—the theory need only be mentioned in passing. All of the Domain-Experts and Task-Experts mentioned the theories, starting with the very first experiment (see Figure 6). However, only 64% of the High-Ability undergraduates and 6% of the Mid-Ability undergraduates mentioned the theories during *any* of the experiments, significantly fewer than the Task- and Domain-Experts.

Were these references to the theories correlated with the use of the theories to guide experiment design? It is possible that the Experts only mentioned the theories but did not use them, and the undergraduates used the theories but did not name them directly. To examine this issue, the variables manipulated in the undergraduates' first experiment was analyzed as a function of whether or not they mentioned the theories (see Figure 7A). These two groups of undergraduates ran rather different first experiments—the two profiles correlated only $r = .13$. The undergraduates who mentioned the theories focused on the Source Context, Spacing, and Delay variables; the variables that are most obviously relevant to the theories. By contrast, the undergraduates who did not mention the theories

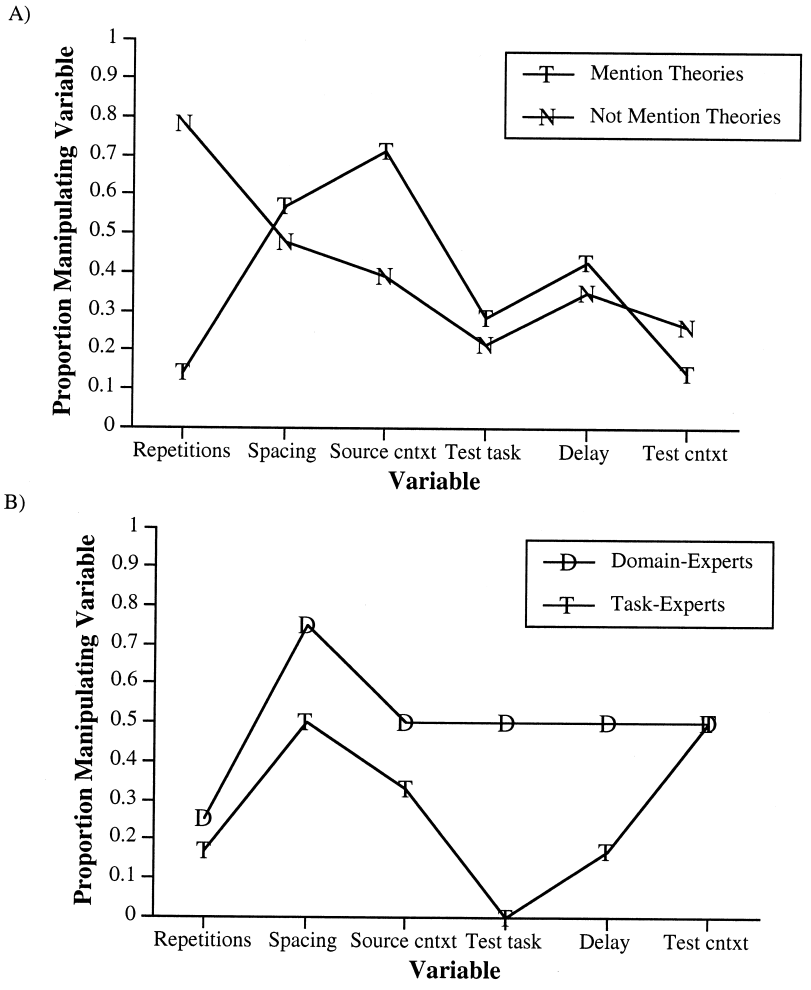

Figure 7. Proportion of participants varying each of the variables in the first experiment. A) Undergraduates only as a function of whether or not they mentioned the theories in their first experiment. B)Domain- and Task-Experts.

primarily varied Repetitions, the upper-leftmost variable in the interface. Moreover, relative ordering of variable use in this group is highly suggestive of a left-to-right, top-to-bottom strategy, which is much more consistent with simply varying variables without regard to their relevance to the theories.

Figure 7B presents the corresponding variable use in the first experiment for the Domain- and Task-Experts, who all mentioned the theories in their first experiment. While the Experts focused on different variables than did the undergraduates, perhaps reflecting different views of what variables were relevant to the theories, the Experts did prefer Spacing and Source Context (the variables of obvious relevance) and avoided Repetitions,

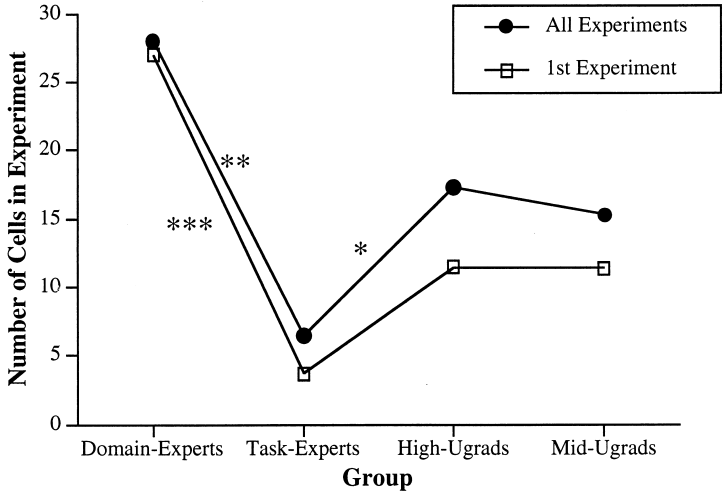

Figure 8. Mean number of factorial design cells per experiment in the first experiment and across all experiments. ***, $p < .01$; **, $p < .05$; *, for $p < .2$.

the variable of least apparent relevance of the theories. The two Expert profiles correlated $r = .52$ with one another and $r = .58$ (Domain-Expert) and $r = .23$ (Task-Expert) with the theory-oriented undergraduate profile. The Expert profiles correlated poorly with the non-theory-oriented undergraduate profile ($r = -.47$ and $r = .00$ for the Domain-Expert and Task-Expert profiles, respectively). Differences between the Experts on variable selections will be discussed further in a later section. Overall however, it appears that all the Experts knew that theories should be used to guide experiment design, whereas many of the High-Ability and almost all of the Mid-Ability undergraduates did not.

Keep Experiments Simple (when necessary). One general principle of experiment design is to keep experiments simple, especially as a first approach. Figure 8 presents the mean experiment complexity, defined as the mean number of cells in the design of each experiment, for participants in the various groups. The Domain-Experts designed more complex experiments than did the Task-Experts, and the High-Ability undergraduates designed more complex experiments than did the Task-Experts. The High- and Mid-Ability undergraduates produced equally complex experiments. These differences are reflected in both the number of variables that the participants manipulated (Task-Experts manipulated two or fewer variables, the other groups manipulated two or more variables) and the number of levels per manipulated dimension (Task-Experts typically included only two levels in their manipulations, the other groups included two or three levels equally often). Moreover, half of both undergraduate groups attempted to design experiments with more than four factors, whereas none of the Task-Experts attempted such complex designs. Thus, it appears that Domain-Experts do not need to keep experiments simple and that undergraduates do not know that they should keep experiments simple.

Use Sensible Variable Values. There are several general heuristics that one can use in

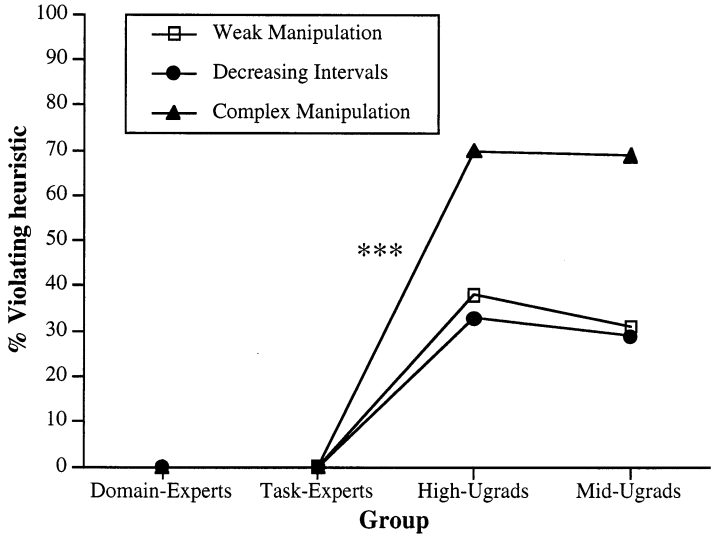

Figure 9. The percentage of participants in each group making value selections that violate three domain-general heuristics. ***, $p < .01$.

selecting values on the various dimensions. First, one should avoid very weak manipulations (insufficient range). In this task, selecting two repetitions values that are adjacent (e.g., 2 and 3, or 4 and 5) violates this heuristic. Second, when selecting three values on an interval dimension, the difference between the first and second interval should not be larger than the difference between the second and third interval (because noise values on psychological measures typically increase as the raw value increases). In this task, this constraint can be violated within the repetitions (by selecting 2, 4, 5), spacing, and delay variables. Third, manipulations should be as simple as possible. In this task, when selecting pairs of values for the spacing and delay variables, it is better to use the same numerical value and vary the unit (e.g., 1 hr versus 1 day) or vary the value and use the same unit (e.g., 5 min versus 15 min) than it is to vary both simultaneously (e.g., 2 min versus 5 days). Both groups of undergraduates were significantly more likely to violate all three of these heuristics (mean violation rate of .38 and .33 for the High- and Mid-Ability groups, respectively) than were the Experts, who never violated these heuristics (see Figure 9).

Keep General Settings Constant across Experiments. Another general heuristic of experimental design is to use the same constant values across experiments (Schauble, 1990; Tschirgi, 1980). It makes comparisons across experiments easier, and it capitalizes on the success of previous experiments. Violations of this heuristic were counted by examining the situations in which a variable was not manipulated in consecutive experiments and then determining whether the same constant value was used in both experiments (e.g., hold repetitions constant at 2 across multiple experiments). Three measures of the tendency to keep values constant were considered: whether the participant ever

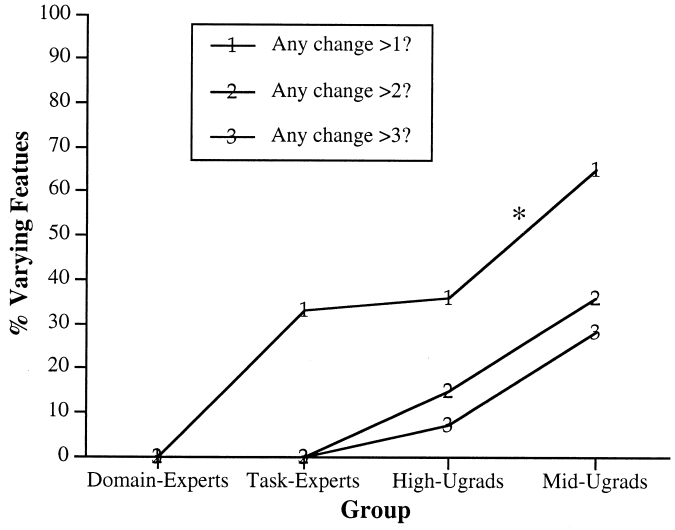

Figure 10. The percentage of participants in each group varying more than one, two, or three values from one experiment to the next. *, $p < .2$.

changed more than one unmanipulated variable value (i.e., minor violations), whether the participant ever changed more than two values, and whether the participant ever changed more than three values (i.e., major violations). Across the different measures of value variation, the Domain-Experts, Task-Experts, and High-Ability undergraduates did not differ significantly (see Figure 10). By contrast, the Mid-Ability undergraduates were significantly higher on all measures of value variation, with almost one third of them varying three or more constant values, suggesting that many of them did not understand this heuristic.

Design—Domain-Specific Skills

Knowledge of Potentially Interacting Variables. Although the principle of manipulating variables that are likely to interact with the variables of interest is likely to be a domain-general one, knowledge of which variables may interact is domain-specific. Within this task, one variable serves as a good test of this knowledge: test task. To non-domain-experts, there is no clear relevance of test task to the testing the two theories. However, domain-experts in this area know that many results do vary with test task (e.g., implicit tasks, such as stem completion, versus explicit tasks, such as free recall), and so it is important to examine the interactions of the other variables with test task. As expected, Domain-Experts were found to manipulate test task more often than did Task-Experts, in both the first experiment or during any experiment (see Figure 11). Moreover, for all variables, Domain-Experts were more likely than Task-Experts to study the interaction with test task.

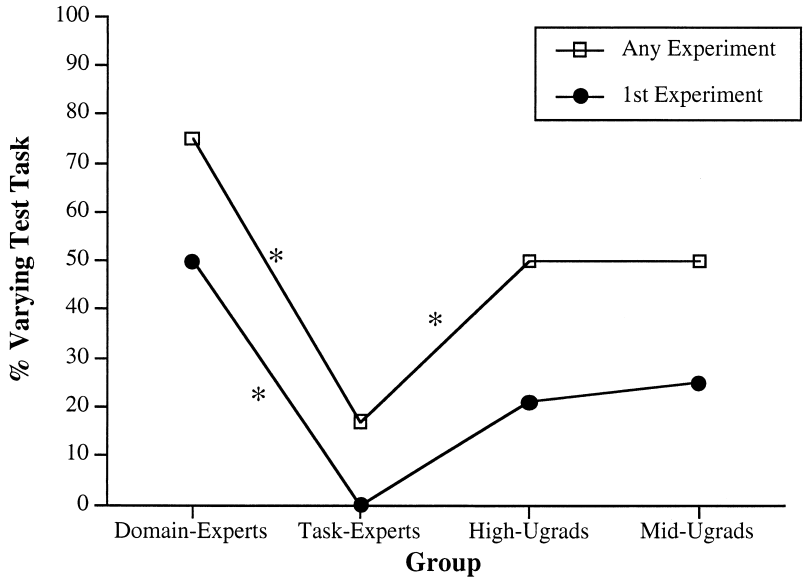

Figure 11. The percentage of participants in each group varying the test task variable at any point or on the first experiment. *, $p < .2$.

Choose Variable Values Useful in the Given Domain. In this particular memory domain, there is a domain-specific heuristic useful for selecting values: the spacing values should be in the same range as the delay values (because the Frequency Regularity theories makes important predictions in the overlap case). The participants' experiments were coded as to whether they ever violated this heuristic. Domain-Experts never violated this heuristic, whereas Task-Experts and both groups of undergraduates frequently violated this heuristic (see Figure 12).

Avoid Floor and Ceiling Effects. In designing experiments, it is a good heuristic to try to avoid floor and ceiling effects because they make the interpretation of null effects and interactions very problematic. Whereas the knowledge that floor and ceiling effects should be avoided may be domain general, the knowledge of how to avoid them is likely to be domain specific. To examine whether the groups were differentially effective at avoiding floor and ceiling effects, we coded which experiments produced outcome values that were all over 90% correct or all less than 10% correct. As expected, none of the Domain-Experts ever ran an experiment that produced such floor or ceiling extreme outcomes (i.e., a priori knowledge helped them avoid extremes). By contrast, 17%, 14%, and 25% of Task-Experts, High-Ability, and Mid-Ability undergraduates, respectively, produced such experiments (the differences between groups were not statistically significant). Overall, most participants were able to avoid such unfortunate outcomes, but only the Experts were able to avoid them entirely.

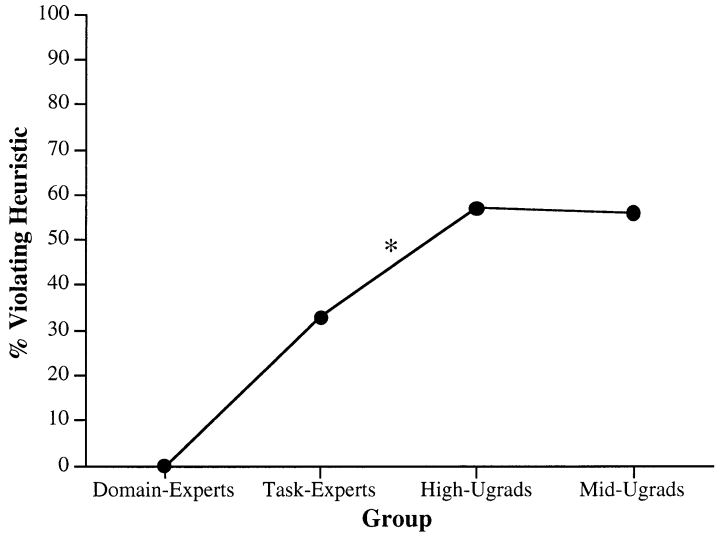

Figure 12. Proportion of participants in each group making value selections that violate a domain-specific heuristic (chose spacing and delay values in the same range). *, $p < .2$.

Making Predictions—Domain-General Skills

After selecting a table structure, the participants were required to make numerical predictions for all of the cells in their current experiment. We examined the use of general knowledge about how to make predictions.

Make Caricature Predictions of Interactions. It is usually unnecessary to make precise quantitative predictions for interactions; instead, caricatures of interactions are usually sufficient. Occasionally, psychologists (and other scientists) do make precise quantitative predictions. However, those are always the predictions of precise mathematical or computational theories rather than simple verbal theories such as those under test in this particular task.

The three basic caricatures of two-way interactions are additive effects, cross-over interactions, and effect/no-effect interactions (i.e., effect at one level, no effect at another level). To see whether the groups differed in terms of whether they predicted qualitative or quantitative interactions, the participants' predictions for each possible two-way interaction in each experiment were classified into one of the three caricatures or a simple quantitative interaction (i.e., effects are always in the same direction, but the magnitude of the effect varies). Domain-Experts and Task-Experts were much less likely to predict simple quantitative interactions than were undergraduates (see Figure 13). Thus, undergraduates were less likely to use caricatures of effects.

One interpretation of this result is that undergraduates were actually just as likely to try to make caricature predictions, but their predictions were simply sloppy, especially for experiments with three or four variables. Arguing against this alternative explanation is

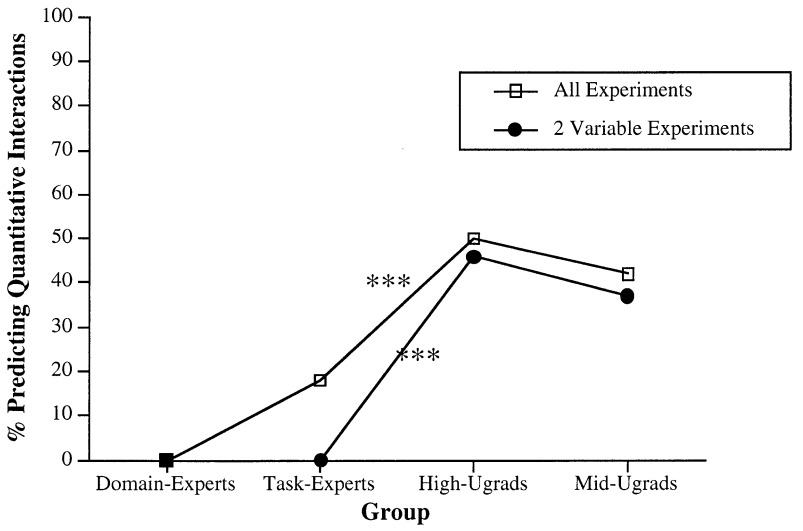

Figure 13. The mean percentage of interactions that were predicted to be quantitative (i.e., not simple qualitative). ***, $p < .01$.

the fact that there were no group differences in the proportion of additive interactions, which would be the most likely source of a quantitative interaction produced through sloppy predictions. However, to further assess this possibility, the analysis was redone using only the predictions for experiments in which only two variables were manipulated. Even in the case of the simple experiments, undergraduates made significantly less use of caricature interactions (see Figure 13).

Interpreting Outcomes—Domain-General Skills

The final task for the participants in this study was to interpret the outcomes of the experiments they designed. Two types of skills were examined with respect to outcome interpretation: encoding the basic results of the experiment (i.e., main effects and interactions) and relating the results to the two theories under test.

Encoding Main Effects. A very basic interpretation skill is the ability to correctly encode main effects from a table of data. One measure of participants' ability to encode main effects is the correctness of their final conclusions about each variable. Because not all participants varied all six variables, this measure was defined as making correct conclusions conditional on having manipulated that variable at least once. To be coded as correct, the participant had to specify the correct direction of the effect. Incorrect responses included "I don't know". For all six variables, there were relatively few differences across groups other than that the Domain-Experts had the highest over-all performance levels. The differences between groups were not statistically significant. The means accuracies across all six variables were 94%, 82%, 78%, and 73% for the Domain-Experts, Task-Experts, High-Ability, and Mid-Ability undergraduates, respectively.

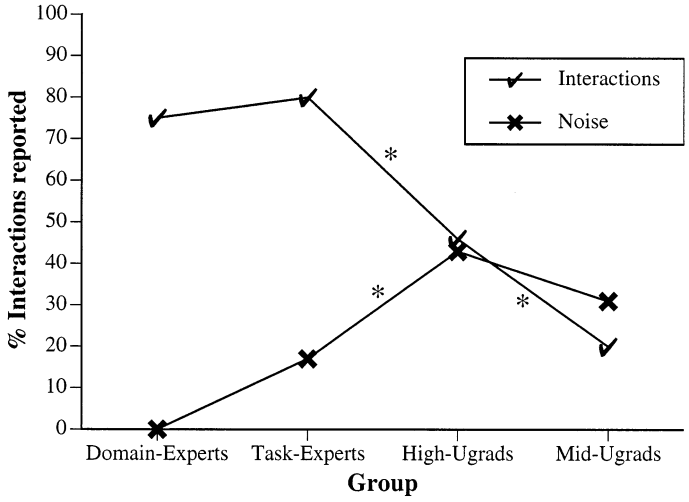

Figure 14. The percentage of participants in each group making correct conclusions about each interaction given opportunity to observe the interaction (Interactions) and percentage of participants making extraneous interaction conclusions (Noise). *, $p < .2$.

Encoding Interaction Outcomes. Another important general outcome interpretation skill is the ability to encode interactions. In this task, there were two two-way interactions. First, there was a quantitative Spacing x Delay interaction, such that the spacing effect was larger at longer delays. Second, there was an effect/no-effect Spacing x Test Task interaction, such that there was no spacing effect with stem completion. As with the main effect analysis, participants' final hypotheses were coded for correctness on these two interactions, and only those participants who had conducted the relevant experiments were included in this analysis. Overall, the Domain-Experts and Task-Experts were equally able to correctly encode these interactions (see Figure 14). By contrast, the High-Ability undergraduates were less able to encode the interactions, and the Mid-Ability undergraduates rarely encoded the interactions.

Ignoring Noise Levels. In addition to being able to encode interactions when they exist, there is also the skill of noting non-interactions (i.e., not being deceived by small levels of noise). To see whether the groups differed in their ability to note non-interactions, the participants' final conclusions were coded for descriptions of non-existent interactions. The Domain-Experts and Task-Experts almost never made such errors, whereas the undergraduates made a significant number of such errors (see Figure 14). If one computes participants' abilities to encode the existing interactions corrected for false alarms (i.e., hits − false alarms), the effect of task expertise is quite large. In fact, the undergraduates show no discriminibility over noise at all.

Make Conclusions in Light of Theories Under Test. After encoding the basic results of each experiment, the participants should have attempted to relate the experimental evidence to the theories under test. To investigate potential differences across groups in

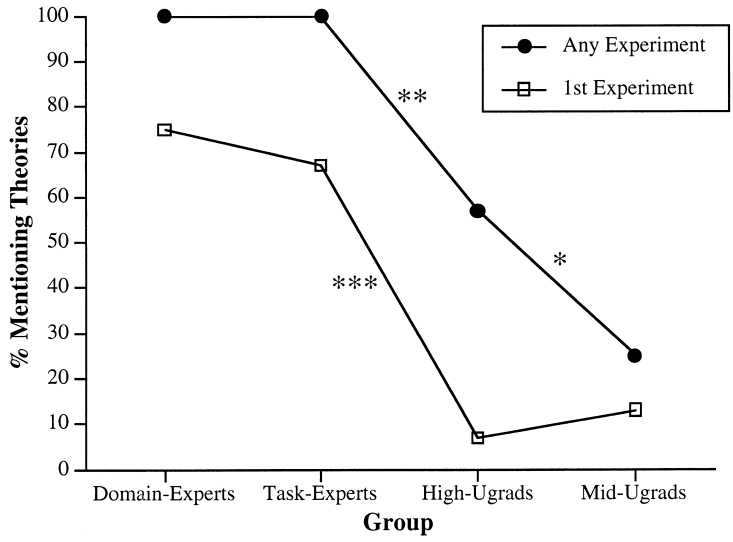

Figure 15. Proportion of participants in each group who mention the theories during outcome interpretation (during the first experiment or during any experiment). ***, $p < .01$; **, $p < .05$; *, for $p < .2$.

this skill, we coded for the presence of conclusions made about the two theories while interpreting outcomes (during the first experiment or during any experiment).

The Domain-Experts and Task-Experts all mentioned the theories at some point, and usually mentioned a theory during the interpretation of the first experiment. By contrast, even the High-Ability undergraduates did not all make any mention of the theories, and they mentioned theories much less often in interpreting the outcome of the first experiment (see Figure 15). Thus, it appears that many of the undergraduates did not use the theories in designing or interpreting the experiments.

Use Evidence in Support of Conclusions about Theories. A general principle in science is to use empirical evidence to support one's conclusions. However, research by Kuhn (1989 and 1991) suggests that many children and adults prefer to rely on theoretical arguments to justify theories. To examine whether individuals in any of the groups made such errors in the context of this obviously scientific task, we coded the participants' final conclusions about the two theories as to whether their conclusions for either theory relied exclusively on opinion rather than empirical evidence. A very lax coding criterion was used: If the participant only referred to their own experience or expectations in making their conclusions about the two theories, this was coded as relying on opinion; if the participant referred at all to the experimental results from the task, this was coded as relying on evidence. Additionally, in making their responses, participants could decline to make a conclusion about a particular theory. Two undergraduates declined to make conclusions about either theory and so were not included in this analysis.

None of the Domain-Experts or the Task-Experts relied exclusively on non-empirical

justifications. By contrast, over a third of both the High- and Mid-Ability undergraduates (.39 and .33, respectively) relied exclusively on non-empirical justifications for at least one of the theories, which was significantly more often than the experts ($p < .05$). These particular results are very important when paired with the failure to mention the theories during experiment design and interpretation. Specifically, they rule out a plausible alternative interpretation of those other results: that the undergraduates did not make use of the theories because they felt they did not understand them. In fact, the undergraduates appeared to have understood the theories sufficiently well that they could relate the theories to their own beliefs and experiences and made strong use of this relation.[9]

In sum, although the undergraduates were just as able to encode the main effects of each experiment, it appears that many of the undergraduates did not understand the purpose of experimentation. They did not refer to the theories in designing or interpreting experiments, and they did not refer to the experiments to justify their conclusions about the theories.

IV. GENERAL DISCUSSION

The primary goals of this study were to identify new kinds of scientific reasoning skills and examine their empirical generality. Toward this goal, the results section presents analyses of group differences on numerous skills. Over-all, consistent patterns across the skill types seemed to emerge: Domain-Experts differed from Task-Experts primarily in terms of domain-specific skills, and Experts differed from High-Ability undergraduates primarily in terms of domain-general skills.

To examine these patterns quantitatively, aggregate, repeated-measures analysis of variance was conducted across measures of each skill. There were three domain-specific measures and 10 domain-general measures (one measure per skill). The measure for each skill that was selected varied from 0 to 1, with 1 reflecting best performance. For some measures, this was the binary outcome of whether they violated the heuristic or not. For other measures, it was a proportion (e.g., the proportion of main effects correctly encoded). In the cases where multiple such measures of the skill existed, either the measure reflecting initial performance was used, if possible (e.g., mentioning theories in the design of the first experiment); otherwise an aggregate measure was used. For the 4% of the data for which a measure was undefined for a given participant (e.g., for encoding interactions if the relevant interaction experiments were never conducted), the grand mean for that measure was used.

As Figure 16 reveals, the over-all group effects were quite clear. For domain-specific skills, the Domain-Experts had very high performance levels, far above the levels of the other three groups [$F(1,36) = 4.74$; $p < .04$], and there were no differences among these other three groups. For domain-general skills, Domain- and Task-Experts had equally very high performance levels [$F(1,36) < 1$; $p > .8$], far above the performance levels of the High-Ability undergraduates [$F(1,36) = 29.5$; $p < .0001$], who in turn performed only very slightly better than the Mid-Ability undergraduates [$F(1,36) = 1.2$; $p > .25$]. That the undergraduates' performance levels were so much lower than the Experts' performance

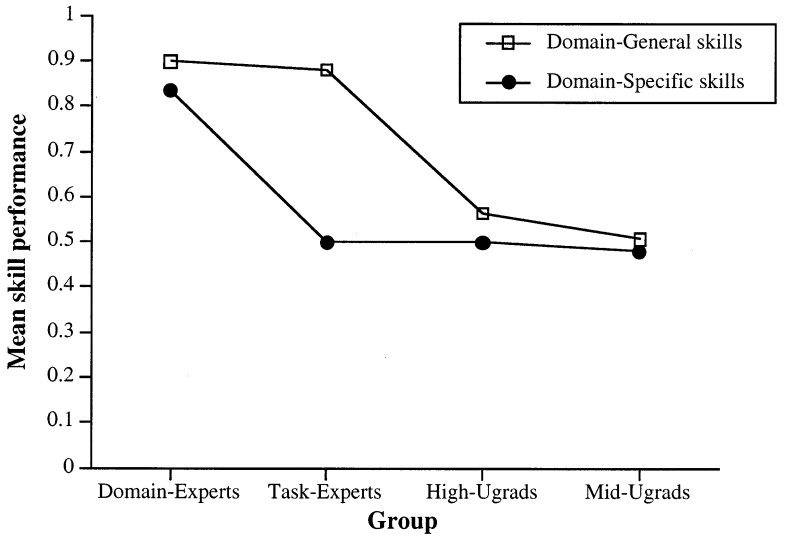

Figure 16. Mean skills performance in each group for domain-general and domain-specific skills in aggregate.

levels demonstrates that these skills were measures of expertise—only the Experts had the skills. Moreover, the lack of differences between High- and Mid-Ability undergraduates suggests that the differences could not be attributed to differences in general reasoning ability.[10]

In sum, the two Expert groups differed primarily on domain-specific skills, and the Task-Experts differed from the undergraduates primarily on domain-general skills. Thus, contrary to a general reasoning ability model or a model in which expertise is highly situated or domain-specific, it appears that expertise in scientific reasoning consists of both domain-specific and domain-general skills. Moreover, the domain-generality of these skills can be predicted by using a simple model that rests on procedural/declarative distinctions.

How dissimilar were the domains of the Domain- and Task-Experts? From the perspective of a physicist or computer scientist, the domains of expertise (cognitive and primarily social psychology) may have appeared very similar if not identical; all are psychology of one form or another. Although there are clear similarities (especially in data analysis methodologies and in shared coursework at the undergraduate level), it is also important to note the differences. For example, it is possible to have relatively little shared coursework. Psychology curricula have very little in the way of particular required classes and there are many subdomains in psychology. Cognitive and social psychologists, although typically in the same departments and sharing the same building, attend completely different conferences and publish in completely different core journals. Most important, they have different research questions, different theories, and many different methodologies, although some methodologies are shared. In sum, the comparison is closer to researchers from different domains than researchers from the same domain who happen to be interested in different particular issues.

The results of this study produced striking similarities to and differences from the findings of the Voss et al. (1983) and Shraagen (1993) studies. We found, as did Voss et al. and Shraagen, that the overall quality of solutions was still best for Domain-Experts. Thus, their results extend to situations in which participants were allowed to iterate through the discovery process (thereby making use of feedback as they would in actual scientific tasks). Also similar to both of those previous studies, our study found that there were many skills that expert scientists share across domains.

In contrast to the previous studies, our study catalogued many additional domain-general experiment design skills and also added prediction and outcome interpretation skills. However, much work remains to be done in building a complete model of expertise in scientific reasoning. For example, our data display task was quite simple, providing little examination of expertise in this area. It is also important to note that there are many other tasks within scientific reasoning that were not tapped by our task (e.g., developing research questions, proposing theories, etc.).

The general pattern of results for the domain-general skills present a picture of expertise that contrasts with the current view of domain-expertise: that domain-expertise consists primarily of a large quantity of domain-specific facts, skills, and schemata acquired only through thousands of hours of practice (e.g., Ericssonet al., 1993; Gobet & Simon, 1996). Instead, our findings suggest that expertise in some domains may also consist of many domain-general skills.

How might our results be reconciled with existing models of expertise? A potentially important factor determining which kinds of experiences and abilities underlie expertise in a given domain may be the relative familiarity of typical problems seen in the domain. Many of the previous studies of expertise involved well-defined problem tasks, such as chess (e.g., Chase & Simon, 1973; Gobet & Simon, 1996), in which good search heuristics provided little help to the computationally limited human mind, whereas being able to recognize good problem states (based on previous experience with those states) provided a strong advantage. Other studies involved giving experts very simple problems that were highly familiar to them (e.g., Chi & Koeske, 1983; Larkin, 1980), problems which could also be solved using recognition processes. By contrast, scientific discovery, by definition, involves tasks that are quite novel to the experts, and thus expertise in such a domain cannot rely heavily on recognition processes. However, it is likely that there is significant domain-specific knowledge that experts do need and that was not tapped by our tasks, such as knowledge of previous theories and experiments in the literature.

This kind of model of expertise for scientific reasoning has several important consequences for artificial intelligence and attempts to automate scientific discoveries (cf. Valdes-Perez, 1995). The large presence of domain-general skills in our results suggests that substantial aspects of scientific reasoning could be automated by using computational programs that are fairly domain-general and, hence, more widely applicable. Towards this goal, we have identified several general heuristics that scientists use. While some of the heuristics may appear obvious to many scientists, they were not so obvious to undergraduates. Moreover some of the skills were ones that scientists may not be explicitly aware

of using (e.g., keep-experiments-simple or keep-general-settings-constant-across-experiments) and, thus, are unlikely to have been included in an artificial intelligence model.

We have presented these domain-general heuristics (and the domain-specific heuristics) in the form of rules implemented in a production system (Table 1), a framework used in several previous models of scientific discovery (e.g., Cheng, 1990; Holland, Holyoak, Nisbett, & Thagard, 1986; Kulkarni & Simon, 1988). This implementation allowed us to make predictions about the domain-generality of the examined skills. However, other implementations or conceptions are possible. A more general, alternative conception in the scientific discovery literature is the search space framework (e.g., Klahr & Dunbar, 1988; Langleyet al., 1987; Schunn & Klahr, 1995 and 1996), in which scientific discovery is viewed as problem solving in a search space, or several search spaces. The search is typically for an experiment to conduct, or for a hypothesis that explains, the experiment's outcomes. In this conception, the domain-specific and domain-general skills that were examined here are search heuristics, methods narrowing down selections of particular experiments or hypotheses. For example, the skill of keeping experiments simple can be conceived of as a search heuristic that narrows the search to only experiments that are relatively simple. The domain-generality of these search heuristics is simply cast in terms of the domains to which they are applicable. The production system framework can be viewed as a more specific instantiation of problem space search with different emphases. For the current discussion, the advantage of the production system framework over the general search space framework is that the former highlights the differences between the search heuristics themselves (the procedures or productions) and the knowledge that the search heuristics use (the declarative facts or chunks). For issues of learning, computational implementation, and transfer of skills to new domains, these differences between procedural and declarative can prove to be very important.

The results of this study also have educational implications. It appeared that there were several design and prediction skills that few of even the High-Ability undergraduates had mastered. At the level of design, the undergraduates made poor values selections for various variables (e.g., selecting a poor range of values). Given their problems in selecting experiment features, it was particularly problematic that the undergraduates also violated the design heuristic of keeping experiments simple. It is possible that the undergraduates underestimated the difficulty of the task: Schunn (1995) has demonstrated that undergraduates do regulate the complexity of their experiments according to their expectations and experiences with task difficulty. At the level of prediction skills, the undergraduates were less likely to make caricature predictions. This difference may have reflected a poor understanding of the purpose of the prediction task: to assess the predictability of verbal theories, which do not make precise quantitative predictions. At the level of interpretation skills, undergraduates were only marginally less able to encode main effects, but were much less able to encode interactions and ignore noise levels.

The most striking of the undergraduate differences was the fundamental lack of appreciation of the purpose of this scientific task: to obtain empirical evidence that could distinguish between two theoretical accounts of an empirical phenomenon. Counter to the purpose of the task, many of the undergraduates did not use the theories in designing the

experiments, nor did they relate the results of the experiments to the theories. Although it may be that some of these undergraduates thought of the theories but merely did not report them in the verbal protocols, the lack of mention of the theories in the verbal protocols was correlated with other differences in the kinds of experiments they designed. Yet, it is also unlikely that the undergraduates did not mention the theories because they felt they did not understand them. They understood the theories sufficiently to use theoretical rather than empirical justifications for their final conclusions for the theories. Moreover, it seems unlikely that performance differences could be attributed to motivational differences because those undergraduates not mentioning the theories worked at the task for just as long as the experts and the other undergraduates.

That the undergraduates frequently relied exclusively on non-empirical justifications (i.e., personal beliefs) to support their conclusions suggests that these undergraduates have basic misconceptions about the role of evidence in scientific activity. Kuhn (1989 and 1991) argued that many children and adults have a confusion between theory and evidence. That they treat the two as the same. Our results are certainly consistent with that interpretation. However, our results also showed that these undergraduates lacked other basic domain-general skills, such as keeping experiments simple, and lacked the ability to interpret main effects and interactions in tables.

In addition to these problems true of all the undergraduates, there were also several difficulties that were unique to the Mid-Ability undergraduates, but were not apparent in the overall trends. For example, the Mid-Ability undergraduates were even less likely to use the theories in designing experiments. The Mid-Ability undergraduates were also much more likely to violate the domain-general heuristic of holding values constant across experiments.

In sum, this research has shown that our expert scientists have acquired through their extensive experiences a broad set of skills that are not specific to their particular research areas. Moreover, these skills are not just a matter of general reasoning ability and do transfer to doing research in other related areas.

Acknowledgments: We thank David Klahr, Marsha Lovett, and Greg Trafton, Nancy Nersessian, Gary Olson, and James Greeno for comments made on earlier drafts of this paper. This research was supported by grant N00014–96-1–0491 from the Office of Naval Research to the second author.

NOTES

1. None of the undergraduates had previously taken a psychology research methods course.
2. To simplify the task, we did not describe the full spacing effect, which is an interaction with delay such that smaller spacing is better at very short delays but longer spacings are better for longer delays.
3. If the participants asked, they were told that there were 20 participants per cell and 20 words per list, which provided an estimate of the noise levels.
4. If the context was different for each study repetition, then same test context meant same as one of the study contexts, whereas different test context meant different than any of the study contexts.
5. However, completing the prediction phase was not enforced by the interface, and occasionally participants conducted an experiment without completing their predictions.

6. This open-ended formulation of domain-specific skills is due to an implementational freedom: domain-specific productions can always be written either as implicitly specific to a domain (e.g., the following production might only be useful in one domain: if Y is true then do Z) or in a more general form that requires the retrieval of a domain-specific fact (e.g., if facts X and Y are true then do Z).
7. One of the skills, keep-experiments-simple, was thought to be less relevant to the Domain-Experts. For this skill alone, the Domain-Experts were not combined with the Task-Experts in comparison to the High-Ability undergraduates.
8. Reliability on this measure and others coded from the verbal protocols was assessed by re-coding 20% of the data. Reliabilities on all measures was 88% or greater.
9. It is possible that the undergraduates had an incorrect understanding of the theories. In fact, the experts often differed in their exact interpretation of the theories. However, the analyses were structured so that they did not depend on a particular interpretation of the theories.
10. Of the eight Domain-General skills for which there were clear effects of Task-Expertise, only three showed any trend of an effect of reasoning ability (i.e., High- versus Mid-Ability).

REFERENCES

Adelson, B. (1990). Modeling software design within a problem-space architecture. In R. Freedle (Ed.), *Artificial intelligence and the future of testing*. Hillsdale, NJ: Erlbaum.

Adelson, B., & Soloway, E. (1985). The role of domain experience in software design. *IEEE Transactions on Software Engineering, 11,* 1351–1360.

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Erlbaum.

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

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

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

Anderson, J. R., & Lebiere, C. (1998). *Atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

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

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Cheng, P. C.-H. (1990). Modeling scientific discovery. Unpublished dissertation, The Open University, Milton Keynes.

Chi, M. T. H., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology, 19,* 29–39.

Clancey, W. J. (1993). Situated action: A neuropsychological interpretation response to Vera and Simon. *Cognitive Science, 17,* 87–116.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern analyzing skills in amnesia: Dissociation of knowing how and knowing that. *Science, 210,* 207–210.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science, 17,* 397–434.

Dunbar, K. (1994). How scientists really reason. Scientific discovery in real-world laboratories. In R. J. Sternberg, & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365–395). Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A., Krampe, R. Th, & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Gobet, F., & Simon, H. A. (1996). Recall of random and distorted chess positions: Implications for the theory of expertise. *Memory & Cognition, 24*(4), 493–503.

Greeno, J. G. (1988). Situations, mental models and generative knowledge. In Klahr, D., & Kotovsky, K. (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 285–318). Hillsdale, NJ: Erlbaum.

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

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12,* 1–48.
Kuhn, D. (1989). Children and adult as intuitive scientists. *Psychological Review, 96*(4), 674–689.
Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge Press.
Kulkarni, D. & Simon, H. A. (1988). The process of scientific discovery: The strategy of experimentation. *Cognitive Science, 12,* 139–176.
Landauer, T. K. (1975). Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology, 7,* 495–531.
Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations in the creative process.* Cambridge, MA: MIT Press.
Larkin, J. H. (1980). Skilled problem solving in physics: A hierarchical planning model. *Journal of Structural Learning, 6,* 271–297.
Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life*. New York: Cambridge University Press.
Mitroff, I. I. (1974). *The subjective side of science.* New York: Elsevier.
Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmational bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29,* 85–95.
Qin, Y., & Simon, H. A. (1990). Laboratory replication of scientific discovery processes. *Cognitive Science, 14,* 281–312.
Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.
Ryle, G. (1949). *Concept of mind.* London: Hutchinson.
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., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*(9), 859–882.
Schunn, C. D. (1995). *A Goals/Effort Tradeoff Theory of experiment space search.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.
Schunn, C. D., & Anderson, J. R. (1998). Scientific discovery. In J. R. Anderson & C. Lebiere (Eds.), *Atomic components of thought* (pp. 385–427). Mahwah, NJ: Erlbaum.
Schunn, C. D., & Klahr, D. (1992). Complexity management in a discovery task. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 177–182). Hillsdale, NJ: Erlbaum.
Schunn, C. D., & Klahr, D. (1993). Self- Versus Other-Generated Hypotheses in Scientific Discovery. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 900–905). Hillsdale, NJ: Erlbaum.
Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society* (pp. 106–111) Hillsdale, NJ: Erlbaum.
Shoenfeld, A. H. (1985). *Mathematical problem solving*. New York: Academic Press.
Shraagen, J. M. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17*(2), 285–309.
Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard Press.
Suchman, L. A. (1987). *Plans and situated action: The problem of human-machine communication.* New York: Cambridge University Press.
Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development, 51,* 1–10.
Tweney, R. D. (1990). Five questions for computationalists. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation* (pp. 471–484). San Mateo, CA: Morgan Kaufmann.
Valdes-Perez, R. E. (1995). *Generic tasks of scientific discovery.* Paper presented at the 1995 AAAI Spring Symposium Series on Systematic Methods of Scientific Discovery, Stanford, CA.
Voss, J. F., Tyler, S. W., & Yengo, L. A. (1983). Individual differences in the solving of social science problems. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition* (Vol. 1, pp. 205–232). New York: Academic.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12,* 129–140.
Wason, P. C., & Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content.* Cambridge, MA: Harvard University Press.