General and Specific Expertise in Scientific Reasoning

Christian D. Schunn
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
schunn+@cmu.edu

John R. Anderson
Departments of Psychology & Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
ja+@cmu.edu

Abstract

Previous research on scientific reasoning has shown that it involves a diverse set of skills. Yet, little is known about generality of those skills, an important issue to theories of expertise and to attempts to automate scientific reasoning skills. We present a study examining what kinds of skills psychologists actually use in designing and interpreting experiments. The results suggest: 1) that psychologists use many domain-general skills in their experimentation; 2) that bright and motivated undergraduates are missing many of these skills; 3) some domain-general skills are not specific to only scientists; and 4) some domain-specific skills can be acquired with minimal domain-experience.

Introduction

What are the reasoning skills required for making scientific discoveries? Previous psychological research on scientific reasoning has produced a rich and varied set of findings regarding the nature of scientific reasoning skills (e.g., Dunbar, 1994; Klahr & Dunbar, 1988; Kulkarni & Simon, 1988). While such research has indicated many dimensions to scientific expertise, it is often difficult to determine whether the features that define scientific expertise are due to differences in general ability, familiarity with the research question, or familiarity with the research methods. One approach that offers some progress on these issues uses a quasi-experimental design which involves a common, representative scientific reasoning task and then systematically manipulates various forms of expertise. This approach has two clear exemplars: Voss, Tyler & Yengo's (1983) study of scientific reasoning in political science, and Shraagen's (1993) study of study of scientific reasoning in experimental psychology. In both of these studies, at least two kinds of experts were used—scientists working within the domain of the research question given to them, and scientists from the same discipline (e.g., political scientists or experimental psychologists) but with a different domain of expertise and therefore unfamiliar with the research question. A third group of participants included undergraduate novices familiar neither with the scientific method nor the particular domain of the research question. Thus, these studies contrasted domain expertise with task expertise. Both studies found that, while there were effects of domain expertise, there were also effects of task expertise, indicating that there are domain-general components to scientific reasoning.

While the Voss et al. and the Shraagen 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, and generating or comparing results to hypotheses (cf. Schunn & Klahr, 1995). Second, the Voss et al. and Shraagen 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 iteratively attack a problem (Tweney, 1990)—scientific questions are rarely answered in one experiment (and especially not in the first one).

The current study was designed to address these issues. As with the Voss et al. and Shraagen studies, our study contrasts domain-experts with task-experts and task novices. However, in contrast to the two previous studies, the current study had two new features. First, it makes use of a new computer interface that simulates the outcomes of experiments, which in turn, allows the participants to see the results of their experiments and design multiple experiments based on the feedback they receive. Second, this study examines the processes by which the participants examined the experimental outcomes and measures what they concluded from the outcomes.

Three general questions were of focus in our study. 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. 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 empirical generality of these reasoning skills should provide important information about the nature and origins of these skills. Third, assuming there are both domain-general skills and domain-specific skills, will these skills transfer to the current task? Recent theories of learning and transfer have suggested that transfer is non-trivial matter (e.g., Clancey, 1993; Greeno, 1988; Lave, 1988; Suchman, 1987). For example, the Task-Experts may not be able to apply their domain-general skills because the current task is not situated in their familiar domains.

The problem given to the participants was to find the cause of the spacing effect, a problem from within cognitive psychology, specifically memory. This problem met three important constraints: 1) the solution must be unknown to the domain-experts, as 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) despite being easy to understand and not yet solved, the solution must be obtainable through experimentation.

Methods

Participants

There were three sources of participants: Cognitive psychology faculty studying memory (Domain-Experts; N=4), Social and Developmental psychology faculty not studying memory (Task-Experts; N=6), and Carnegie Mellon undergraduates (N=30) from a variety of backgrounds. The Domain-Experts and Task-Experts were a mix of senior and junior faculty at strong research universities.

Materials & Procedure

At the beginning of the experiment, all the participants were given a step-by-step introduction into the main task on the computer. The instructions described the spacing effect—that spaced practice produces better memory performance than massed practice—and two theories which have been proposed to explain 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. Thus, 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. The second theory was the frequency regularity theory, which stated that the mind tries to estimate how long memories will be needed based on regularities in the environment and, in particular, adjusts forgetting rates according to the spacing between items. The participants' primary task was to determine which theory provided a better account of the spacing effect—this goal was presented to them at the beginning, in the middle, and at the end of the instructions.

Since we were interested in the process by which people plotted and interpreted data in addition to how they designed experiments, we designed a computer interface, called the Simulated Psychology Lab, that produced experimental outcomes and allowed participants to iterate through the process of design, plot and interpret. The interface was designed to support 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 on six dimensions, any of which could be manipulated, and up to four factors could be simultaneously manipulated 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 source task factors that the participants could experiment with included 1) repetitions—the number of times that the list of words was studied; 2) spacing—the amount of time spent between repetitions; and 3) source context—whether the participants were in the same context for each repetition or whether they changed contexts on each repetition. The test factors 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. For each variable, the participants could either hold the variable constant or vary it. Values had to be selected on all dimensions, including the dimensions that were held constant in the given experiments; no default values were used. There was no restriction on the order of value selection, and participants could go back to change their selections for any of the variables at any point in time up until they selected to run the experiment.

The participants made predictions and were given outcomes in a table format with all cells being shown at once. A table format was used rather than a graphical format because it was thought that the table format was less difficult to understand and manipulate for the undergraduate participants. Before being given the table, participants had to decide on which dimension each manipulated factor would be plotted. After deciding on the table structure, participants

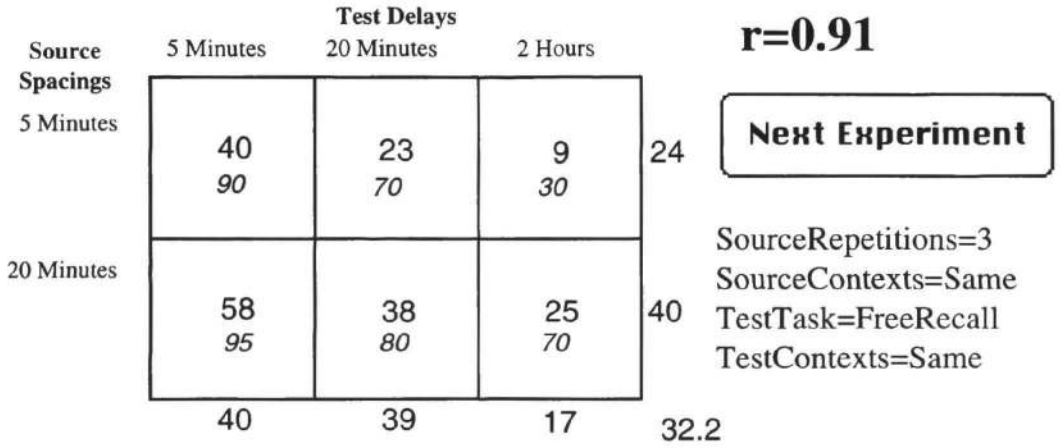

Figure 1: The interface for displaying the experiment outcomes. Main entry is the actual outcome, predictions are in italics.

made numerical predictions for their tasks. The number to be predicted for each cell in their design was the mean percent correct for that cell. 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), this particular form of a prediction task was used 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 was less intrusive; 3) it provided further data about the participants' theories and beliefs about each of the variables; and 4) it provided some cost to large experimental designs to simulate the increasing real-world cost of larger experimental designs.

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 1). In addition, the outcome tables also displayed the participant's predictions for each cell in italics. 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. The actual results displayed were generated by a mathematical model that is roughly consistent with results from research on memory and the spacing effect.

Participants worked at the task until they felt that they had found out what the cause of the spacing effect was or 40 minutes had elapsed—this time-limit was selected because most participants were able to complete the task in this amount of time, and because we were more interested in discovery processes than in final products.

The primary data gathered in this experiment was keystroke data as the participants generated, plotted, and interpreted experiments. However, the participants were also asked to give a think-aloud verbal protocol throughout the task. Moreover, at the end of the task, participants were asked to verbally report their conclusions about the spacing effect—i.e., whether either of the two theories given to them at the beginning of the task explained the spacing effect. The participants were also asked to give conclusions about the effects of each of the variables.

Results

The goal of this paper is to examine the presence and absence of different kinds of domain-specific and domain-general experimentation skills in psychology, specifically in the domain of cognitive processes in memory. Rather than bias the results in favor of one outcome or another, we include all the skills that we examined. The skills are divided into four general classes of skills: designing an experiment, displaying data, making predictions, and interpreting outcomes. Within each class, skills are divided into domain-general (skills expected to be useful in at least psychology generally) and domain-specific (skills expected to useful only in memory experiments). The left half of Table 1 presents a complete list of the skills examined.

The experiment design skills were evaluated using the experiments generated by the participants (as reflected in the keystroke protocol), and, in the case of the first design skill, what the participants said while they designing the experiments. The display and prediction skills were evaluated using the participants display and prediction choices (as reflected in the keystroke protocol). The outcome interpretation skills were evaluated using the participants final conclusions' about the six factors and the two theories for the cause of the spacing effect. The interpret skills were evaluated conditional on opportunities to learn (e.g., correct conclusions regarding a main effect or interaction were only evaluated conditional on having conducted the relevant

Table 1: Complete list of skills examined by skill type and skill level, as well as the direction and statistical significance (.2, .05, and .01) of the pairwise comparisons between Domain Experts and Task Experts, All Experts and High-Ability undergraduates, and High and Mid-Ability undergraduates.

Type	Level	Skill	DE v TE	E v HA	HA v MA
Design	General	Design experiments to test the given theories	0	+++	+++
Design	General	Keep experiments simple	- -	++	0
Design	General	Use sensible factor value	0	+++	0
Design	General	Manipulate factors relevant to the theories under test	0	0	++
Design	General	Keep general settings constant across experiments	0	0	++
Design	General	Avoiding floor and ceiling effects	- -	0	0
Design	Specific	Knowledge of variables likely to interact	+	0	0
Design	Specific	Choose factor values useful in the given domain.	0	++	0
Display	General	Display continuous factors within rather than across tables	+	0	0
Display	General	Display factors consistently across tables	0	0	0
Display	Specific	Display the true spacing effect within a table	0	0	++
Predict	General	Correctly predict the direction of known effects	0	+	0
Predict	General	Make caricature predictions of interactions	0	+++	0
Predict	Specific	Making ball-park estimates of effects	+	0	0
Interpret	General	Make conclusions in light of theories under test	0	+++	+
Interpret	General	Use evidence in support of conclusions about theories	0	++	0
Interpret	General	Encode main effects	0	0	0
Interpret	General	Encode interaction outcomes	0	+++	0

Table 2: The mean number of experiments and mean time on task in minutes (with standard errors).

Group	# Experiments	Time on task
Domain-Experts	2.8 (0.8)	36.0 (5.7)
Task-Experts	4.8 (0.7)	38.0 (3.2)
High-Ability	5.6 (1.0)	36.3 (1.6)
Mid-Ability	5.7 (0.7)	34.0 (1.8)

experiments).

To examine the empirical generality of the skills, we structure each of the analyses by contrasting performance for each of the 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.

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 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 (mean=728, N=14), and those below were called Mid-Ability undergraduates (mean=586, N=16). Since our undergraduates all had Math SATs above the US national median, we used the label Mid rather than Low.

Before examining the results from the skill-specific analyses, we begin with two general results regarding number of experiments generated and time on task for each of the groups. The three groups spent approximately an equal amount of time on the task (see Table 2). However, Domain-Experts conducted fewer experiments than did the Task-Experts and undergraduates. This occurred because the Domain-Experts conducted a small number of complex experiments, whereas the other groups conducted a larger number of simple experiments.

The statistical analyses focus on three pairwise comparisons: Domain-Experts versus Task-Experts, Domain and Task-Experts versus High-Ability undergraduates, and High-Ability versus Mid-Ability undergraduates. Because there are so many skills being examined, we cannot describe the detailed results for each of the skills. Instead, here we will present two characteristic results, and then focus on the patterns across the skills.

The primary goal of the task was to test the two given theories for the cause of the spacing effect (the Shifting Context theory and the Frequency Regularity theory). Yet, many of the undergraduates did not seem to understand how this goal might be achieved. Table 3 presents the proportion of participants mentioning either theories while they were designing experiments. Both Domain and Task experts mentioned the theories from the very beginning of the task,

Table 3: Proportion of participant mentioning either of the theories during experiment design (first experiment or ever).

Group	1st Exp.	Ever
Domain-Experts	1.00	1.00
Task-Experts	1.00	1.00
High-Ability	.43	.64
Mid-Ability	.06	.06

whereas a large proportion of the undergraduates did not mention the theories during the design of any of their experiments. A similar pattern held for outcome interpretation: all Experts related the results back to the two theories under test, whereas the majority of the undergraduates did not. This lack of focus on the theories under test was also reflected in the kinds of variables that the undergraduates included in their experiments—they were more likely to included irrelevant factors such as repetition.

An 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. The 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 Table 4). By contrast, the High-Ability undergraduates were less able to encode the interactions, and the Mid-Ability undergraduates rarely encoded the interactions. 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 participant's 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 Table 4). In fact, the undergraduates were overall just as likely to report non-existent interactions than to report existing interactions.

These two sets of results illustrate a general pattern: Domain and Task-Experts were near ceiling on domain-general skills, and undergraduates often did not possess these skills. Across all the skills, the following pattern emerged: Domain-Experts differed from Task-Experts

Table 4: Proportion of participants (and Ns) making correct conclusions about each interaction given opportunity to observe the interaction and proportion of participants making extraneous interaction conclusions.

Group	Spacing x Delay	Spacing x Test Task	False Alarms
Domain-Experts	1.00 (2)	.50 (2)	.00 (4)
Task-Experts	.75 (4)	1.00 (1)	.17 (6)
High-Ability	.44 (9)	.50 (4)	.43 (14)
Mid-Ability	.11 (9)	.33 (6)	.31 (16)

Table 5: Effect directions by skill level.

Pair	Level	+++	++	+	0	-	- -
Domain-Experts	General			1	**11**		2
vs. Task-Experts	Specific			**2**	2		
Experts vs.	General	**5**	**2**	**1**	6		
High-Ability	Specific		1		**3**		
High-Ability vs.	General	**1**	**2**	**1**	10		
Mid- Ability	Specific		1		**3**		

primarily in terms of domain-specific skills; Experts differed from High-Ability undergraduates primarily in terms of domain-general skills; and High-Ability undergraduates different from Mid-Ability undergraduates in terms of domain-general skills.

To examine these patterns quantitatively, we classified those pairwise comparisons for each skill according to its direction (expected direction vs. unexpected direction) and statistical significance ($p<.01$, $p<.05$, $p<.2$, and no effect), using Fisher Exact tests for discrete variables and t-tests for continuous variables. The expected directions for the three pairwise comparisons of interest were: Domain-Experts>Task-Experts, Experts>High-Ability undergraduates, and High-Ability undergraduates>Mid-ability undergraduates. Combining direction and statistical significance produced seven categories of possible effects: expected very strong (+++), expected strong (++), expected weak (+), no effect (0), unexpected weak (-), unexpected strong (- -), and unexpected very strong (- - -). Table 1 presents the results of these statistical tests for each skill.

The results of this aggregation analysis are presented in Table 5. The shaded areas are the regions predicted by an Expertise-type by Skill-type interaction. As can be seen by comparing the density of effects in the shaded versus unshaded areas, while there are some exceptions, the overall effects of skill generality by group comparisons are as expected. In the comparisons between the two expert groups, more of the domain-specific skills showed positive differences (50%) than did the domain-general skills (7%), $\chi^2(1)=4.1$ $p<.05$. In the comparisons between the Experts and the High-Ability undergraduates, more of the domain-general skills showed positive differences (57%) than did the domain-specific skills (25%), although because of the small number of domain-specific skills, this trend was not statistically significant, $\chi^2(1)=1.3$ $p<.25$. Finally, in the comparisons between the High and Mid-Ability undergraduates, few of skills of either type showed positive differences (29% and 25% for domain-general and domain-specific skill respectively, $\chi^2(1)<1$).

In sum, the two Expert groups differ primarily on domain-specific skills, the Task-Experts differ from the High-Ability undergraduates primarily on domain-general skills, and the two groups of undergraduates differ rarely, and equally often on domain-general and domain-specific skills. Thus, contrary to a general reasoning-ability model or a situated action model, it appears that expertise in scientific reasoning consists of domain-specific and domain-general skills for design, display, prediction, and interpretation skills.

Discussion

The demonstrations of empirical generality of scientific reasoning skills in this paper depend on several assumptions. The first assumption is that the lack of differences between the expert groups were not due to low N problems. In support of this assumption, the Task-Experts' performance was near ceiling on the measures of domain-general skills (e.g., see Table 3) suggesting that they did indeed possess these skills. Moreover, we also used a fairly liberal criteria in assessing statistical trends, and found a weak trend on only one of the 14 domain-general skills. The second assumption is that the task that we used was representative of a real scientific task. In support of this assumption, scientists were shown to have skills useful in our task, and these skills were typically not present in bright, motivated undergraduates.

The results of this study contained striking similarities and differences to the findings of the Voss et al. (1983) and Shraagen (1993) studies. Similar to both of those previous studies, our study found that there were many skills that expert scientists share across domains—our study catalogued additional domain-general experiment design skills and also added prediction, plotting, and outcome interpretation skills.

The general pattern of results across the skills present a picture of expertise that contrasts the current view of domain-expertise—that domain-expertise consists primarily of a large quantity of domain-specific facts and skills acquired only through thousands of hours of practice (e.g., Ericsson, Krampe, & Tesch-Römer, 1993; Gobet & Simon, 1996). Instead, our findings suggest that expertise in some domains may also consist of many domain-general skills, and that domain-specific skills can occasionally be readily acquired by bright individuals without numerous hours of experience.

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 like chess (e.g., Chase & Simon, 1973; Gobet & Simon, 1996) in which there was little role for good search heuristics and a relatively large role for recognizing good problem states (based on previous experience with those states). Other studies involved giving experts very simple problems that were highly familiar to them (e.g., Chi & Koeske, 1983; Larkin, 1980), in which the problems could also be solved using recognition processes. By contrast, scientific discovery, by definition, involves tasks which are quite novel to the experts, and thus expertise in such a domain cannot rely heavily on recognitional processes.

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 many aspects of scientific reasoning could be automated 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.

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 factors (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, since Schunn (1995) found that undergraduates do regulate the complexity of their experiments according to their expectations and experiences with task difficulty. At the level in interpretation skills, undergraduates were only marginally less able to encode main effects, but 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 this scientific task: to obtain empirical evidence which could distinguish between two theoretical accounts of an empirical phenomenon. Counter to the purpose of the task, the majority of the undergraduates did not use the theories in designing the experiments nor did they relate the results of the experiments to the theories. While 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. Moreover, it seems unlikely that performance differences could be attributed to motivation differences as the undergraduates not mentioning the theories worked at the task for just as long as the experts and the other undergraduates.

In sum, this study have provided new information regarding the nature of expertise in science: 1) there are skills that are common across domains and do transfer, contrary to the predictions of theories of situated learning and action; 2) many of these domain-general reasoning skills are a function of task expertise and are not a simple function of general reasoning ability, contrary to a simple general-reasoning ability model of expertise; 3) some of these skills have already been acquired by bright undergraduates, contrary to an intensive domain-specific practice model of expertise; and yet 4) many undergraduates still lack fundamental aspects regarding the function of scientific experimentation.

Acknowledgments

This research was supported by ONR grant N00014-96-1-0491 to the second author. We would like to thank David Klahr, Marsha Lovett, Greg Trafton, and two anonymous reviewers for comments made on earlier drafts of this paper.

References

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

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

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

Dunbar, K. (1994). How scientists really reason. Scientific discovery in real-world laboratories. In R.J. Sternberg, & J. Davidson (Eds.) *Mechanisms of Insight.* MIT Press.

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

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

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

Kulkarni, D. & Simon, H.A. (1988). The process of Scientific Discovery: The strategy of Experimentation. *Cognitive Science,* 12, 139-176.

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: CambridgePress.

Schunn, C. D. (1995). *A Goals/Effort Tradeoff theory of experiment space search.* Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Schunn, C. D., & Klahr, D. (1993). Self- Vs. Other-Generated Hypotheses in Scientific Discovery. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society.*

Schunn, C. D., & Klahr, D. (1995). A 4-space model of scientific discovery. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society.*

Shraagen, J. M. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17,* 285-309.

Suchman, L. A. (1987). *Plans and situated action: The problem of human-machine communication.* New York: Cambridge University Press.

Tweney, R. D. (1990). Five questions for computationalists. In J. Shrager & P. Langley (Eds.), *Computational models of scientific discovery and theory formation.* San Mateo, CA: Morgan Kaufmann.

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). New York: Academic.

Valdes-Perez, R.E. (1995) Generic tasks of scientific discovery. Paper presented at the *1995 AAAI Spring Symposium Series on Scientific Discovery.*