4

Acquiring Expertise in Science: Explorations of What, When, and How

Christian D. Schunn
George Mason University

John R. Anderson
Carnegie Mellon University

Relatively little is known about the skills that practicing scientists actually use. This lack of knowledge makes the design of science curricula rather difficult. How can we train students to be scientists if we do not know what it means to be a scientist? In this chapter, we ask two central questions about the nature of expertise in science: 1) Are there general skills that scientists from different domains share?, and 2) If there are any general skills, are these skills just ones that any intelligent adult would have, or are they the result of training and practice in scientific activities? To answer these questions, we present a study of expert research psychologists working on a scientific discovery problem taken from psychology. Then we turn to a more practical question: 3) If there are general skills not possessed by average intelligent adults, are they being covered in undergraduate education? As a preliminary answer to this question, we present an evaluation of several research methods courses in psychology at one university.

Are There General Skills That Scientist From Different Domains Share?

Within cognitive psychology, there is some debate about the generality of expertise. On the one hand, there is discussion of general problem solving procedures and general characteristics of experts in a domain. For example, it has been argued that experts (at least in some domains) use forward reasoning whereas novices use backward reasoning (Larkin, 1980)—i.e.,

reasoning from goals to givens versus from givens to goals. On the other hand, the most prevalent view of expertise in cognitive psychology is one of domain-specific pattern recognition skills (Chase & Simon, 1973; Chi & Koeske, 1983; Ericsson & Charness, 1994; Gobet & Simon, 1996; Hayes, 1985; Johnson & Mervis, 1997; Larkin, 1980). For example, chess experts are thought to have learned tens or hundreds of thousands of patterns of chess positions. It is this pattern recognition expertise that is thought to underlie their superior performance in chess, rather than differences in general intelligence or sophisticated strategies. Consequently, their expertise is extremely domain specific. For example, while chess experts can play several games of chess simultaneously while blindfolded, their impressive memory for chess positions disappears when random positions are used that could not normally occur in a game (Chase & Simon, 1973).

Moreover, cognitive psychology is filled with examples of failure to transfer knowledge from one domain to another domain. When two domains have superficial differences, people tend to have a very difficult time spontaneously noticing underlying similarities between the domains (Duncker, 1945; Gentner & Toupin, 1986; Holyoak, 1985; Holyoak & Thagard, 1995; Ross, 1989).

Does this perspective of expertise apply to science as well? In particular, are expert scientists experts in only their narrow area of specialization, or is there a set of general skills shared by different kinds of scientists? Expertise in science shares many of the characteristics of expertise in other domains. For example, focused practice of an extended period (usually 10 years) is required before a scientist attains world-class expertise (Hayes, 1985). This 10-year-rule has been found in all other domains of expertise (Ericsson, Krampe, & Tesch-Römer, 1993). One might expect there to be a large amount of domain-specificity to this expertise. Moreover, one might expect that the higher the level of expertise, the more separation between different sciences. For example, with higher levels of training, a chemist might have increasingly less in common with a psychologist.

Yet, science has some properties that seem to make it different. While pattern recognition is also likely to be important in science, more conceptual and procedural components are also likely to be important. Thus, the research on less complex tasks like chess may not generalize to science. Moreover, some research on transfer has found that experts can be more able to transfer knowledge from their domain of expertise than are novices, at least in simple problem solving contexts (Novick, 1988).

Models of science education also rest on certain assumptions about

the nature of expertise in science. On the one hand, there is discussion of the scientific method, as if there is some central aspect that all or most sciences share. On the other hand, fairly early on in a student's high school education, science is taught separately by discipline. For example, there are physics, biology, and chemistry classes, and physics, biology, and chemistry labs, rather than general science classes or general science labs. Underlying this design is an assumption that relatively little can be taught in a generic fashion about science. Thus, we have these opposing views of science as very general and science as very specific.

In sum, it is possible that scientists do not share skills across the different scientific domains. It is also possible that, even if they share skills, they will not be able to apply those skills in a domain outside their narrow specialization. This chapter reports a study that examines this issue and finds that scientists do share skills in common and can transfer their skills to other scientific domains.

If There Are Any General Skills, Are These Skills Just Ones That Any Intelligent Adult Would Have?

In the cognitive and developmental psychology literature, there is a long tradition of viewing individuals (children or adults) as intuitive scientists (Klahr & Dunbar, 1988; Kuhn, 1989; Piaget, 1952). Under this view, people are thought to naturally explore their world, developing and testing hypotheses by conducting simple experiments. For example, the infant learns about gravity and cause and effect by systematically dropping objects from a high chair. Or a chef learns about what factors produce a good cake by trying different ingredients or different cooking methods. Thus, it is quite likely that scientists from different domains do share some general scientific reasoning skills because they share problem solving weak methods (e.g., hill-climbing and means-ends analysis) and some basic scientific reasoning skills with all (or most) humans.

However, there is also a long tradition in the cognitive psychology literature of describing in intricate details the logical reasoning errors that humans tend to make. The average university undergraduate has been found to make basic reasoning errors in syllogistic reasoning (Johnson-Laird, 1972), conditional reasoning (Wason, 1968), probabilistic reasoning (Cheng & Holyoak, 1985), and scientific reasoning (Kuhn, 1989; Wason, 1960). It is typically assumed (although not always found (Mahoney, 1979)) that well-trained scientists would not make these kinds

of reasoning errors.

Assuming scientists are better reasoners than the average university undergraduate, this difference in reasoning ability is not necessarily a result of scientific training and practice. In addition to having had much training and experience in scientific reasoning, scientists also tend to have higher general intelligence levels, even before their training began. This difference is not to say that science requires high levels of intelligence. In fact, there is some controversy about whether intelligence measures can predict whether a scientist will be successful (Sternberg & Williams, 1997). However, there is a simple observation that, in selecting individuals for science training, IQ surrogates like the SATs and the GREs are used quite heavily. This reason alone is sufficient to produce higher levels of general intelligence in scientists compared to the general population.

In sum, even if scientists share abilities in common with one another, these commonalities may not be a consequence of training and experience in science, nor are they necessarily specific to science. This chapter reports a study that examines exactly this issue and finds that scientists do share abilities that are specific to science and are not attributable to general reasoning ability differences.

If There Are General Skills Not Possessed By Average, Intelligent Adults, Are They Being Covered in Undergraduate Education?

Given evidence for general skills that scientists share amongst one another, the question arises: Where did they get those skills? One obvious alternative is that these skills are acquired as a result of thousands of hours of practice conducting and interpreting experiments. Another alternative is that many of these skills are acquired in formal education at either the undergraduate or graduate level. Since the issue is domain-general skills, one might expect that they should be included in existing courses on research methodology. It would certainly be efficient to focus on domain-general skills which students are likely to use independent of which particular scientific domain they end up pursuing.

There are several reasons, however, why existing research methods courses might not cover these domain-general skills. First, there may be an overemphasis on domain-specific information. For example, many psychology departments have separate courses called cognitive research methods, social research methods, tests and measurements, developmental research methods, etc. Similarly, chemistry departments have labs in

physical chemistry, organic chemistry, etc. Even when a department has a more general research methods class, it is still tied to a particular research domain (e.g., psychology vs. physics. vs. chemistry vs. biology). Thus, it is possible that these domain-specific research methods courses would neglect the domain-general research skills.

Another possibility is that existing research methods courses might cover domain-general skills, but not the ones actually used by scientists. Research methods curricula are typically developed from armchair, self-reflective analyses of the skills used in scientific settings rather than detailed, systematic observation of practicing scientists. Research on expertise has shown that experts are often unaware of the components of their expertise. Many skills begin as conscious, declarative knowledge, and then, with enough practice, they become effortless, unconscious, procedural skills (Anderson, 1983, 1993; Anderson, Fincham, & Douglass, 1997).

This chapter reports an evaluation of several research methods courses at one research-oriented psychology department. The study examined 1) whether the domain-general skills are taught in such courses, and 2) whether the skills are acquired by the students in such courses. The study found that many of the skills were not covered explicitly in the courses. Moreover, while there was some improvement on some of the skills, even the most basic and central skills showed only modest improvements.

The Studies

Overview

The studies reported here are really a case study of expertise in and teaching of research psychology. However, the list of skills examined is not logically tied to psychology, and we suspect they will generalize to many other sciences. For example, the ability to read tables of data is a skill that one would expect to generalize to many other sciences. The first study examines whether psychologists from different subdomains of psychology share skills in common, as evidenced by their performance on an experimental design and outcome interpretation task. This first study was reported in great detail in Schunn and Anderson (1999). The second study examines whether undergraduate psychology majors show improvement

in those very skills identified in the first study using the same task as was given to the experts. Because these two studies examine performance on the same skills in the same tasks at four points along an expertise continuum, we will present the studies as one large study with four groups.

The Subjects

Most expert/novice studies confound two different types of expertise: Expertise in the tasks being studied (skills) and expertise in content domain (knowledge). In this particular study, the general task is designing and interpreting experiments, and the specific content domain is the cognitive psychology of memory. To isolate domain-general skills, we have two kinds of experts. The first group consists of cognitive psychologists who study memory. Thus, they are experts in both the content area and the general task. For convenience, we will call these the Domain Experts. The second group consists of social and developmental psychologists who do not study memory. Thus, they are experts only in the task but not in the content area. We will call these the Task Experts. A non-psychologist might expect that there is little different between cognitive and developmental or social psychologists. However, they are quite different domains in terms of the journals in which they publish, the primary conferences they attend, the theories they use, and the methodologies they use.

The third group is undergraduate psychology majors in their second year of study. None of them have yet taken a research methods course, and thus they are experts in neither the task nor the domain. We will call these the Pre-RM group (pre-research methods). The fourth group is undergraduate psychology majors in their third and fourth year of study immediately after having taken a research methods class in psychology. We will call these the Post-RM group.

There were four Domain Experts. They were highly productive research faculty at a well-established, top-tier, research university. All had conducted many studies and written many journal articles in the area of the cognitive psychology of memory. There were six Task Experts. They were also highly productive research faculty and were taken from the same department as the Domain Experts. None of the Task Experts had worked in the domain of memory (cognitive or otherwise). The two groups of experts were equivalent in terms of the number of years since Ph.D. and in the number of publications overall (approximately 65) and in the last year.

There were twenty-two students in the Pre-RM group. All were psy-

chology sophomores at the same university as the experts. None had previously taken a research methods course. The students were recruited from five different lower level psychology classes to ensure a breadth of student interests. There were forty-one students in the Post-RM group. They were recruited from five different psychology research methods courses: two developmental research methods, two social research methods, and one cognitive research methods. All students were paid $8 for their participation in the study. There was no relationship between volunteering for the study and grade received in the course.

In cross-sectional designs, one must always worry about selection artifacts. A somewhat atypical feature of the undergraduate psychology curriculum at this particular university made this particular cross-sections design cleaner than most. At this university, psychology majors are required to take two research methods courses (from the three available) in order to graduate. However, the number of offered classes and the maximum enrollment size in these classes is severely limited and thus only seniors and a few juniors are able to enroll. Consequently, none of the sophomores could have taken a research methods class, but all of these sophomores must eventually take several research methods classes before graduating. Therefore, we need not worry about the Pre-RM being systematically different from the Post-RM groups in orientation towards research (vs. clinical psychology) or in area preferences (cognitive vs. social vs. abnormal) or in general intelligence.

The Task

The task given to the faculty and students was designed to be representative of experimentation and analysis in psychology, but to also have certain extra properties. In particular, it was important to have the task be sufficiently realistic and complex that experts in the area would not know the answer and yet would feel that the question was answerable. If the experts could simply retrieve the answer from memory, then it is unlikely that we would see evidence of the strategies they use in their own scientific research, which, by definition, involves a solution that cannot be simply retrieved from memory. However, it was equally important to have the terminology and issues be understandable to individuals outside of that area of expertise. If the other groups misunderstood the issues and terms, their problem solving behavior would be different from that of the Domain Experts for very uninteresting reasons.

The task that met these criteria was as follows. People were given a description of a simple and pervasive phenomenon from the cognitive psychology memory: the spacing effect. The spacing effect is simply the advantage of studying that is spaced out in time over studying that is lumped all together. For example, on a later test, students who study for one hour straight typically remember less than students that study three different times for 20 minutes. Even though the total amount of study time is the same, the group with more spaced out study episodes remembers more at a later test. This effect is very simple to describe to people who are not cognitive psychologists. However, even cognitive psychologists do not yet know why the spacing effect occurs. The task given to all the subjects was to design an experiment to determine the cause of the spacing effect.

To make the task more similar to the one faced by the experts in the area (who already had many theories for the cause of the spacing effect), everyone was given a description of two different theories for the spacing effect. These two theories are the most common ones proposed by experts in the area. The first theory, called the shifting context theory, assumes that memories are associated with the context under study and that context gradually shifts with time. Under this theory, the spacing effect occurs because spaced practice produces associations to more divergent contexts, which in turn are more likely to overlap with the test context. The second theory, called the frequency regularity theory, states 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 decay more rapidly (because they are not expected to be needed again after long delays) whereas items learned with long intervening spaces decay more slowly (because they are expected to be needed again at long delays).

Everyone was given much longer, more concrete descriptions of the two theories, and the study did not continue until they felt that they understood the spacing effect and the two theories. The specific goal given to the subjects was to determine the cause of the spacing effect. They were told that the answer could be that one, both, or neither theory was correct (as is the typical case in science).

The Simulated Psychology Lab

The most straightforward way to proceed would be to give everyone a pencil and paper and have them design an experiment or series of experiments that would test between the theories for the spacing effect. This technique would make the problem realistically open-ended. Any kind of experiment could be proposed. However, it would not be possible for the subjects to see the results of their proposed experiments. In this way, the task is very much unlike real science. Scientists very rarely answer any questions with only one experiment, and certainly never the first experiment that gets designed. Instead, scientists design an experiment, run it, and construct a new experiment based on the problems revealed by the outcomes of the first experiment. Much of their expertise lies in being able to interpret the outcome of one experiment and use the information to design a better experiment. Additionally, scientific discoveries involve designing good experiments *and* correctly interpreting their outcomes. Using a pencil and paper task, we could not examine the skills that scientists possess for interpreting the results of their experiments.

To achieve these goals, a computer environment, called the Simulated Psychology Lab (SPL), was developed. The SPL environment simplifies the experimental design process by presenting the individual with a large but limited number of experiments that can be designed. The advantage of SPL is that it allows people to see the results of the experiments that they designed. Thus, we can observe how scientists iterate through the cycle of experiment design and outcome interpretation.

The hypothetical experiments that one could create within SPL were simple list learning experiments. The hypothetical subjects would get a list of items to study for a later test. The list could be studied several times under a variety of contexts. The later test could occur in a variety of contexts at various different times. There were six variables that could be manipulated within this basic scenario. Two variables were highly relevant to the theories under test: *spacing* between study repetitions (from one minute to 20 days), and *source context* (whether it was the same context for each study repetition). Three variables were moderately relevant: *test context* (whether it was the same context as during study), *delay* (the time between the last study episode and the test, from one minute to 20 days), and *test task* (whether the test was free recall, recognition, or stem completion). Finally, there was one irrelevant variable: *repetitions* (the number of times each word is studied; two, three, four, or five times).

Participants selected variable settings using a simple mouse-controlled interface. For each of the six variables, they could select whether to manipulate that variable (or hold it constant) and what particular values to pick (for each condition or for the constant value). When a variable was manipulated, two or three different levels could be used. Participants could only vary up to four variables simultaneously in any given experiment. With the six variables, there were almost 400,000 unique experiment settings that could be generated.

To provide a concrete example, a participant might have selected to conduct the following experiment: Vary study spacing (five minutes versus 20 minutes), test delay (five minutes, 20 minutes, or two hours), and source context (same versus different rooms), and hold constant repetitions (three), test task (free recall), and test context (different room). In this sample experiment, there are 12 different conditions (2 x 3 x 2).

The participants were given outcomes in a table format with all cells being shown at once. Tables rather than graphs were used because tables were thought to be 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).

The table of results included a display of the variables held constant and their values (see Figure 1).

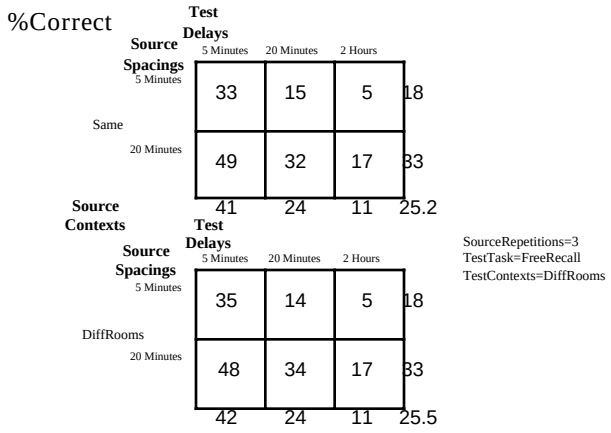

Figure 1. The interface used for displaying the outcomes of experiments.

To facilitate comparison across rows, columns, and tables, the row, column, and table means were also provided. Figure 1 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.

In order to lead the participants to treat the outcomes of their experiments as real data (with noise levels and other properties of real data), the participants were told that the computer had access to a large database of experiments and would simply present the results of those experiments on the screen. However, in reality, a mathematical model was used to produce the results of each experiment. The model was selected to be roughly consistent with existing results from research on memory and the spacing effect. The model also incorporated random noise at a typical level for these kinds of experiments (i.e., between -2% and +2% added to each cell).

Participants worked at the task—iterating through the process of experiment design, choosing a table structure, and viewing outcomes—until they felt that they had found out what the cause of the spacing effect was or 40 minutes had elapsed. The primary data gathered in this experiment were keystroke data as the participants generated experiments, choose 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, the 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.

Terminology

To avoid confusion between this overall study and the experiments that the participants designed and analyzed, the following conventions will be used. The participants designed *experiments*; they took part in this *study*. The participants viewed their results in the form of tables; we will present analyses of their aggregate behavior in this study in the form of graphs.

Skills Examined

Finding the general skills used by scientists is a very open-ended research goal. What particular skills should be investigated? To generate a list of general skills, we constructed a computational model of scientific discovery behavior in the SPL domain (Schunn & Anderson, 1998). The model uses the ACT-R production system framework (Anderson, 1993; Anderson & Lebiere, 1998), which captures skills as if-then rules (in contrast to Minstrell's Facets, this volume). The model is capable of designing appropriate experiments to test the two theories for the spacing effect and analyzing the data to examine whether the data is consistent with each theory. From this model, we extracted twelve skills that 1) did not appear to be specific to the particular domain, and 2) could be examined with the behavioral data provided by our participants. To keep the current story brief, we will focus on six representative skills in this chapter (see Table 1; see Schunn and Anderson (1999) for the full list of 12 skills). There are three skills associated with designing experiments and three skills associated with interpreting outcomes.

Table 1. List of skills examined by skill type, along with English form of the skills in the computational model that implements them.

Skill	Detailed Description of the Skill
Experiment design	
Design experiments to test theories	If given theories to test, then set goal to test some aspect of theory
Keep experiments simple	If variable is not relevant to hypotheses under test, then hold variable constant
Keep settings constant across experiments	If not varying a variable, then pick the value used in the previous experiment
Interpret outcomes	
Encode interactions	If effect of variable X is significantly different at different , levels of Y then conclude there is an interaction
Ignore small noise levels in data	If an effect or interaction is very small , then ignore it
Relate data to theories under test	If finished encoding the results of an experiment, then relate results to theories under test

Table 1 also lists a description, in English, of exactly what the skill entailed in the computational model. The skills included in this list are basic skills that are applicable in a broad range of scientific settings. Thus, one could argue that they are especially important targets for science edu-

cation. While some of the skills may seem quite obvious and simple to the reader, we shall see that they were not so obvious and simple to the undergraduates in the study. The following sections describe the results for each of these skills, grouped by type of skill (experiment design versus interpret outcomes).

Group Comparisons

To answer the main questions raised in this chapter, there are three key comparisons between the four groups (that will be made repeatedly on the range of skills just described). If the domain and Task Experts perform equally well on a given skill, then this is evidence that 1) this skill is domain general (at least across different areas of psychology); and 2) that expert scientists can apply their expertise to problems in other content domains. If the Task Experts do not perform equally well on a given skill, then this would suggest that either the given skill is not domain-general (i.e., not used by experts in multiple domains) or that expert scientists are limited in their ability to apply their skill expertise outside of their domain of expertise.

Comparing the two expert groups with the Pre-RM group establishes whether skills shared among the scientists are also shared with non-scientists. If the Pre-RM group also performs well on a given skill, then this argues that the skill is common to most adults (or at least those found in university settings). If the Pre-RM group performs much more poorly on a given skill, then this suggests that the skill is not common to most adults and is the result of formal training and/or extensive practice in science.

Of course, there are always other possible explanations for differences between the Pre-RM group and the experts. For example, the groups also differ in age, personality types, overall intelligence levels, etc. The difference that seems most plausibly related to performance differences in this domain is an overall intelligence difference. To address this issue, the undergraduates were also asked about their SAT scores, which ranged in this sample from levels close to the average population to levels very close to those of the faculty's. We then analyzed whether SAT scores predicted performance differences within the undergraduate groups. If SAT scores do not predict performance differences on any of the skills, then overall intelligence differences is not likely to be the cause of performance differences between the undergraduates and the experts.

Finally, comparing the Pre-RM group with the Post-RM group examines whether the existing research methods courses teach the students any of the general skills that they were missing. The instructors were shown a list of the skills examined and were asked whether these skills were covered in their course, and (whether or not it was explicitly covered) how likely it was that the students would possess those skills at the end of the course. Thus, we can examine whether these skills were explicitly covered in the research methods courses and how that related to whether the students had acquired the skills.

Stylistic Notes

There are two things to note about the format of the results section. First, we will not present inferential statistics in the text. All the appropriate inferential statistics were computed and only the statistically significant ($p<0.05$) results are discussed as differences. Second, since there is a natural set of three pairwise comparisons between the four groups (Domain Experts versus Task Experts, experts versus Pre-RM undergraduates, and Pre-RM versus Post-RM undergraduates), the groups are always presented in the order Domain Experts, Task Experts, Pre-RM, and Post-RM to facilitate these comparisons.

Results

Overall Results

Before examining the performance of the different groups on each of the skills, we will mention the overall performance levels of each group. First, there is the performance on the overall goal of the discovery task: To determine which theory of the two theories (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.[1] 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. Figure 2 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. 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.

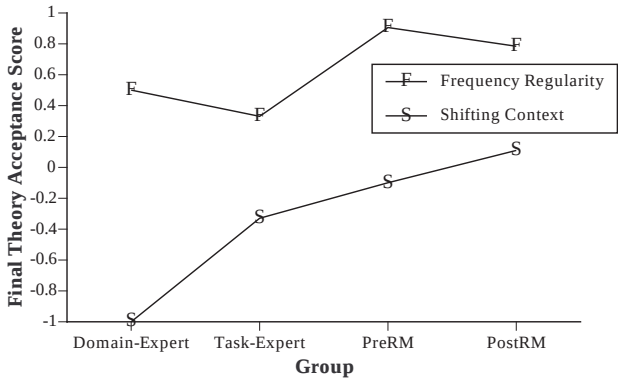

Figure 2. Mean acceptance score for each group for each of the two theories.

Turning to time-on-task, the three groups spent approximately an equal amount of time to complete the task (36.0, 38.0, 31.2, and 34.2 minutes for the Domain Experts, Task Experts, Pre-RM and Post-RM undergraduates respectively). However, Domain Experts conducted fewer experiments (2.8) than did the Task Experts (4.8) who in turn conducted about as many experiments as the undergraduates (5.5 and 5.8 respectively). 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.

---

1. There was no effect of source context, which is strongly inconsistent with the shifting context theory, and generally consistent with the frequency regularity theory. However, a strong form of the frequency regularity theory implies that there should be a matching effect between delay and spacing such that performance is best when delays exactly match study spacing—this was not to be found in the data.

Experiment Design Skill 1: Design Experiments to Test Theories

Not all experiments done by scientists test hypotheses (see Okada and Shimokido, this volume). However, when there are theories to test, the details of the theories should be taken into account when designing the experiment. Although this would seem obvious to the reader, as we shall see, this was not so obvious to the students. 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 designing experiments, either during the first experiment or during any experiment. Note that this is a very lax criterion for measuring use of theories in experiment design—the theory need only be mentioned in passing. Below is an example of what one participant said before designing her first experiment:

> Alright. The first thing that I think about these two theories, as I understand this, what is it, frequency regularity theory, it doesn't say anything about context at all. It says the thing that matters is whether you have, um, close or far intervals. So. One way to attack the problem is to show that there is a context effect.

This (Domain Expert) participant not only mentioned one of the theories but also mentioned how it influenced her design. In this coding scheme, only the mention of the theory was necessary. Here is an example of what a participant not mentioning the theories would say (also before the design of the first experiment):

> Ok. Click on repetitions... I'm going to set it. Number of different repetitions. I'll have them do all the same number of repetitions at one. And I'll set the repetitions to... four... Ok... Now, there's spacings. Number of different spacings ... uh. I'll set that at…two…and... Do the first one ... in minutes for ... fifteen. And the second one, in hours.

As we can see, this participant is simply selecting options within the interface, apparently without thinking about the theories under test.

As one would expect, all of the Domain Experts and Task Experts mentioned the theories, starting with the very first experiment (see Figure 3). However, fewer than half of the Pre-RM undergraduates mentioned the theories during *any* of the experiments. Thus, they appeared not to understand the experimentation should be guided by theories at hand. Taking the research methods courses did appear to help: The proportion of undergraduates mentioning the theories in the design of the first experiment almost doubled from Pre-RM to Post-RM. However, there

remained a fairly substantial proportion of Post-RM undergraduates who did not ever mention either of the two theories during the design of their experiments.

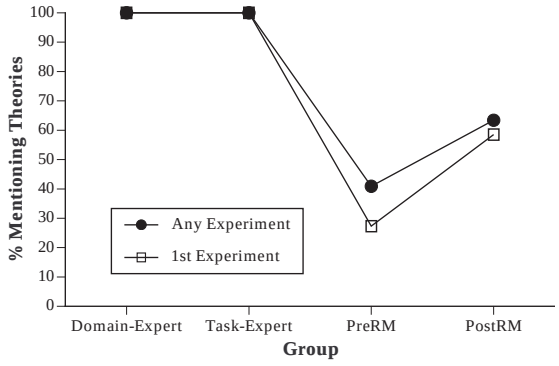

Figure 3. The percentage of participants in each group who mention the theories during experiment design in the first experiment or in any experiment.

Were these mentionings of the theories correlated with using the theories to guide experiment design? It is possible that the undergraduates used the theories but did not name them directly. To examine this issue, the variables manipulated in the undergraduates' first experiment were analyzed as a function of whether or not they mentioned the theories (see Figure 4). These two groups of undergraduates ran rather different first experiments.

The undergraduates that 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 not mentioning the theories 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.

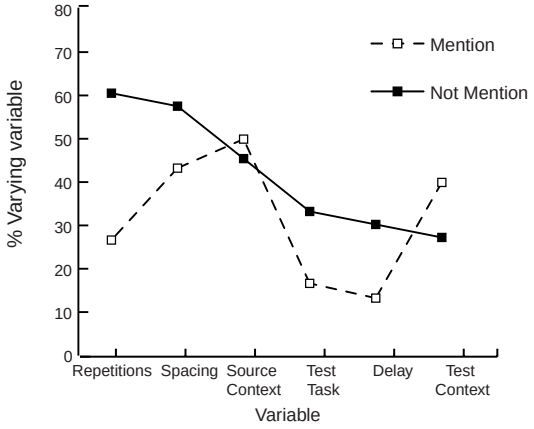

Figure 4. Proportion of undergraduates varying each of the variables in the first experiment as a function of whether or not they mentioned the theories in their first experiment.

Figure 5 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 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, the variable of least apparent relevance of the theories.

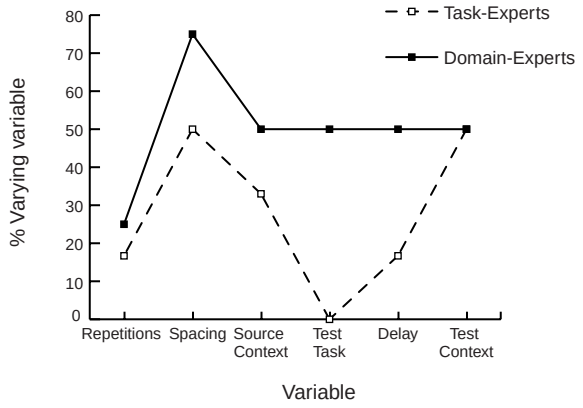

Figure 5. Proportion of domain and Task Experts varying each of the variables in the first experiment.

Experiment Design Skill 2: Keep Experiments Simple (When Necessary)

One general principle of experiment design is to keep experiments simple, especially as a first approach. One rough measure of the complexity of an experiment in this context is the number of cells in an experiment. For example, the experiment in Figure 1 has 12 cells (2 x 3 x 2). Figure 6 presents the mean experiment complexity for participants in the various groups, defined as the mean number of cells in the design of each experiment.

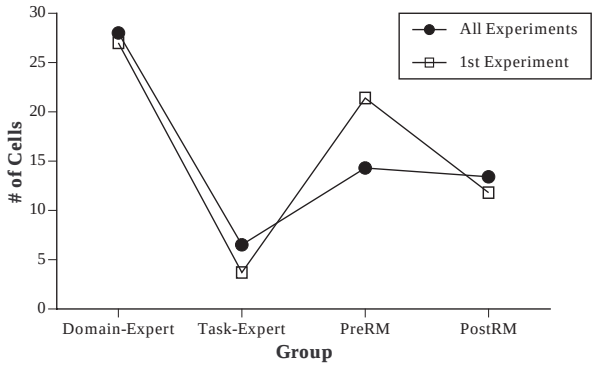

Figure 6. Mean number of factorial design cells per experiment in the first experiment and across all experiments.

The Domain Experts designed more complex experiments than did the Task Experts, and the Pre-RM undergraduates designed more complex experiments than did the Task Experts.[2] These differences are reflected, both in the number of variables that the participants manipulated (Task Experts manipulated two or fewer variables, the other groups manipulated two or more variables), and in the number of levels per manipulated dimension (Task Experts typically included only two levels in their manipulations, the other groups two or three levels equally often). Moreover, 40% of the Pre-RM undergraduates 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

2. Readers outside of cognitive psychology may be shocked by the size of the experiments generated by the Domain Experts. However, in this area of cognitive psychology, with many short-duration trials, it is not uncommon to have designs with 20+ cells.

do not need to keep experiments simple, and that the Pre-RM undergraduates do not know that they should keep experiments simple. There was a small influence of the research methods courses on the undergraduates: They appeared less likely to start with very complex experiments and they were less likely to try to design an experiment with more than four factors (27%). However, there was no impact on the mean complexity across later experiments.

Experiment Design Skill 3:Keep General Settings Constant Across Experiments

In the current context, it is not possible to examine the traditional experiment design issue of avoiding confounds (see Klahr, Chen, and Erdosne-Toth, this volume) because the SPL interface forces participants to use full factorial designs—it is impossible to design a confounded experiment. However, it is possible to study a related general heuristic of experimental design: use the same constant values across experiments (Schauble, 1990; Tschirgi, 1980). By continuing to use the same constant values, it makes comparisons across experiments easier and it capitalizes on the success of previous experiments. To illustrate this issue, consider the following example sequence of experiments. Suppose that a participant decides in the first experiment to manipulate only the repetitions variable, holding the other variables constant. This manipulation and the constant values chosen for the other five variables are listed as Experiment 1 in Table 2. Suppose that the participant finds little effect of repetitions but wants to see whether a stronger manipulation of repetitions would have a more noticeable effect (e.g., two versus five repetitions). What values should be selected for the other variables? In particular, should the participant use the same constant values again (as in Experiment 2 of Table 2), or should the participant select new constant values (as in Experiment 2' of Table 2)? The general wisdom, as we shall see, is to keep most if not all the values the same.

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 spacing constant at 10 minutes across multiple experiments).

Table 2. Example experiments illustrating the coding of feature changes from experiment to experiment. Experiment 2 changes zero features from Experiment 1. Experiment 2' changes five features from Experiment 1.

Variable	Experiment 1	Experiment 2	Experiment 2'
Repetitions	2 vs. 3	2 vs. 5	2 vs. 5
Spacing	10 minutes	10 minutes	2 days
Source context	Same room	Same room	Different room
Test task	Recall	Recall	Recognition
Delay	1 day	1 day	20 minutes
Test context	Same room	Same room	Different mood

Since there are occasionally good reasons for changing one or two constant values (e.g., to examine whether the result generalizes to a different task or population, or to address floor or ceiling effect problems), we focused on the frequency of extreme changes: Changing more than two constant values from one experiment to the next. Figure 7 presents the percentage of participants in each group ever changing more than two constant values.

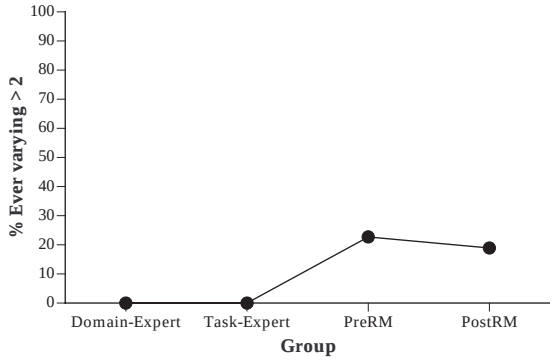

Figure 7. The percentage of participants in each group varying more two values (for variables held constant) from one experiment to the next.

Experts appear to be sensitive to this heuristic and never changed such a large number of values. By contrast, a sizeable minority of undergraduates did undergo such extreme changes. Moreover, this estimate of the number of heuristic violators is likely to be a large underestimate: Many of the undergraduates conducted complex experiments that by definition reduces the number of constant variables that could be changed.

Thus far, we have seen clear evidence of domain-general skills of experimental design: Skills that the experts share among one another and are not possessed by untrained undergraduates. There is also some evidence for some learning of these general skills in an undergraduate research methods course. Now we shall turn to the case of outcome interpretation skills.

Outcome Interpretation Skill 1: Encode Interaction Outcomes

A very basic interpretation skill is the ability to correctly encode main effects and interactions from a table of data. However, most of the variables in this task had main effects that one would have expected. For example, more repetitions produced better recall, longer delays produced worse recall, etc. Thus, examining performance on main effects is not likely to produce insight into the participants' abilities. By contrast, the interactions in this task were less obvious. 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. 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 the Task Experts were equally able to correctly encode these interactions (see the upper curve in Figure 8). By contrast, the undergraduates were half as likely to encode the interactions, and this ability did not improve with a research methods course.

Outcome Interpretation Skill 2: Ignore Small Noise Levels in Data

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 (see the lower curve of Figure 8). The Pre-RM also rarely made such errors.

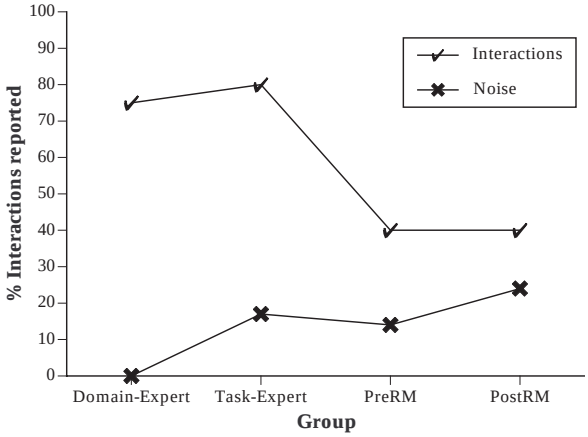

Figure 8. The percentage of participants in each group making correct conclusions about each interaction given opportunity to observe the interaction (Interactions) and percentage of participants making extraneous interaction conclusions (Noise).

However, the presence of any errors shows that undergraduates were willing to pay attention to and report interactions. Thus, the difference in the ability to report interactions is not likely to be due to an unwillingness to discuss interactions. Interestingly, there was a slight increase in the number of false interactions reported following the research methods course.

Outcome Interpretation Skill 3: Relate Results to Theories

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 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, only half of the Pre-RM undergraduates ever made any mention of the theories, and they mentioned theories much less often in interpreting the outcome of the first experiment (see Figure 9). Thus, it appears that many of the undergradu-

ates did not use the theories in designing or interpreting the experiments. As with mentioning theories in the design of experiments, there was a slight improvement with a research methods course in the mentioning of theories during outcome interpretation.

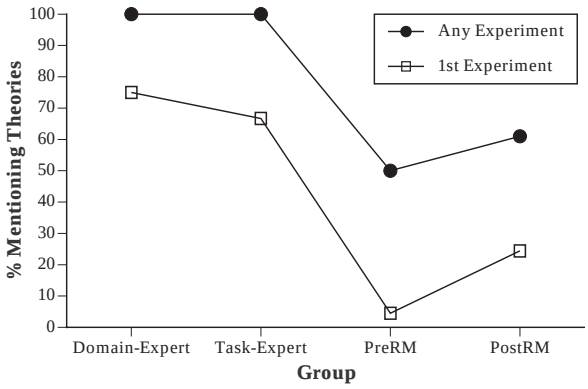

Figure 9. Percentage of participants in each group who mention the theories during outcome interpretation (during the first experiment or during any experiment).

One interpretation of the lack of mention of the theories is that the undergraduates did not understand the theories and thus did not mention them. There is data from the students' final conclusions that addresses these issues. After conducting all their experiments, the participants were asked about their conclusions for each theory. If the students did not understand the theories, then they would be unlikely to provide any conclusions regarding the theories. However, over 90% of the undergraduates made conclusions about the correctness of each theory. Of these, most provided some empirical justification to support their claims (e.g., the shifting context theory is correct because there is a context effect). Yet many of these comments about the relationship between the theories and the evidence appeared to be constructed only when asked. That is, when first asked about conclusions, the students tended not to comment on the theories but instead comment on the effects of the various variables (e.g., there is an effect of spacing and delay, but not of source context). When subsequently asked about the theories, the students made comments like: "Oh, the theories. What were they again? Let me think about that."

Providing further support that the undergraduates understood the theories, approximately 10% of the undergraduates who made conclusions about the theories provided no empirical justifications for their

claims—instead they simply referred to their own personal experiences and beliefs (e.g., the shifting context theory is correct because that is what works for me). Thus, not only did the undergraduates understand the theories well enough to make conclusions regarding them, but they also understood them so well to apply them to their own lives. Of course, in this context, it was not particularly appropriate to rely exclusively on personal experience and belief as a justification. This is further evidence that some of the undergraduates did not understand the basic role of experimentation in theory testing.

General Discussion

The chapter began with three questions. The results from the studies provide partial answers to these questions. First, it appears that there are general skills shared by different kinds of scientists, or at least different kinds of psychologists. On all of the dimensions, the Domain Experts and Task Experts performed quite well and equally well. Thus, this study has identified a core set of experiment design and outcome interpretation skills that psychologists share, independent of their research styles, training background, and research domain.

Second, these core skills are not ones already possessed by all intelligent adults. The experts performed significantly better than the undergraduates. Moreover, when the undergraduates were split by their SATs,[3] there were no performance differences on these dimensions. Thus, at least within these ranges of intelligence and for these skills, overall intelligence appears not to play a role. What is surprising from these results is that so many of the undergraduates were missing such fundamental skills. These were bright students at a strong private university and presumably had already been exposed to many science content courses. Yet, they appeared to be unclear on the important and basic relationship between theory and experimental data.

The implication for science education from these results are clear. We have identified important skills that can be applied across a wide range of domains and are not yet possessed by untrained adults. Therefore, we

3. The undergraduates were divided using a median split of 1240 combined math plus verbal. The mean combined scores of the two groups were 1152 and 1340. Thus, there was a fairly large difference in scores between the two groups, and the higher group had scores more than adequate for entry into graduate school in psychology.

have a clear target for instruction. What is not so clear is how to teach these skills.

Third, and relevant to this last point, it appears that current undergraduate research methods classes in psychology address only some of these skills. On only half of the six skills were there signs of improvement as a result of taking a research methods course. Moreover, on all six dimensions, there was still room for improvement—there were still significant differences between the Post-RM and expert groups for all six skills. This lack of improvement cannot be attributed to one bad instructor or one bad curriculum. The Post-RM group included students from five different classes, taught by five different instructors, teaching, in some cases, very different curricula.

There are several interpretations of the mediocre improvement of the Post-RM group. For example, it is possible that the students had been taught the relevant skills but did not see the relationship between their class material and the SPL task. In the psychological literature on analogical reasoning, this might be called a failure to spontaneously notice the deep similarity between the two domains. At the end of the SPL task, the students were asked whether their psychology course had helped them do the task, and if so, what aspects. Three-quarters of the Post-RM students thought their course had helped, whereas only half of the Pre-RM students thought their course had helped. When the Post-RM students thought the courses had helped, the two most common aspects that students mentioned were various aspects of designing the experiment (being systematic, avoiding confounds) and various aspects of interpreting the outcomes (organizing the tables, reading the tables for main effects and interactions). The Pre-RM group mentioned design aspects 25% of the time and never mentioned interpretation aspects. Many of them felt their course had helped but could not name a particular way in which it had helped. The Post-RM group mentioned design aspects 50% of the time and interpretation aspects 33% of the time. Thus, many of the Post-RM group were more likely to see a connection to their research methods class and could be more articulate about that connection.

Of particular interest are the Post-RM students that did not feel that their research methods course had helped them. Why did they feel that the course that should be so directly relevant to the current task had not helped? A few said that they already knew how to do this kind of task before taking the research methods. However, most said that the current task was too different from what was covered in the course.

This pattern of responses suggests that noticing the relationship between their research methods course and the SPL task may be part of the problem, but it is not likely to be the primary reason for their poor performance on the SPL task. This raises the question: What material was covered in the courses? Perhaps the skills examined in this study were not the ones covered in those courses. To address this issue, the instructors were shown a list of the skills examined, and were asked to rate the skills on two dimensions: 1) Were these skills covered in their course, and 2) whether or not a skill was explicitly covered, how likely it was that the students would possess those skills at the end of the course?

Looking at the correlations among instructors' responses, there were great similarities in what was covered, but essentially no agreement in what the students should be able to do. This lack of agreement about what students could do was true for both the items that the instructors felt they covered and for the ones they felt they did not cover. The instructors used a scale of zero (never) to four (often) for the taught question and zero (none of the students) to four (all of the students) for the "should possess" question. The mean ratings for each skill are indicated in Table 3. As can be seen in the table, the instructors felt that five of the six skills were taught in their class, and that at least some of their students should possess each of the six skills.

Table 3 also indicates that improvement on each skill as a result of taking these research methods classes. An effect size measure was used—it divides the difference in group means by the standard deviation in group performance (i.e., an effect size of 1 is a one standard deviation improvement).[4] As noted earlier, only three of the six skills showed significant improvements. The one skill that the instructors unanimously agreed was not covered (keep settings constant across experiments) was among the skills that showed no improvement. Thus, we have an explanation for the lack of improvement on one of the skills (and we have evidence that studies of the current type can provide new insights into what skills should be included in research methods courses). However, the other two skills that showed no improvement were rated as covered—in fact, one of the skills (encode interactions) had the highest ratings on the taught dimension. Thus, some of the skills showed no improvements despite (apparently) being covered in the courses.

---

4. For the skills in which two measures had been gathered (first experiment/all experiments), the first experiment measure was used because it seemed to most cleanly represent transfer from the course rather than learning that occurred during the study.

Table 3. For each of the six skills examined, the improvement from Pre-RM and Post-RM groups (difference in group means divided by group standard deviation), and the mean instructor ratings on whether the skills were taught and whether the students should possess those skills.

Skill	Improve (Effect Size)	Taught (0-4)	Should possess (0-4)
Design experiments to test theories	0.62	2.5	2.5
Keep experiments simple	0.51	2.5	2.3
Keep settings constant across experiments	0.09	0.0	2.0
Encode interactions	0	3.8	3.3
Ignore small noise levels in data	0	2.5	2.5
Relate data to theories under test	0.52	3.5	3.3

In sum, the courses produced at best small improvement on these core skills, and the variability in improvements can only be partially explained by what was explicitly not covered in the courses. What can be done to improve the situation? It may be that these skills, while basic and simple to describe, are not so simple to learn. This could be because they have many, many component skills. For example, encoding interactions in tables may have many component skills relating to the many types of interactions one may find. In support of this interpretation, the computational model that we developed required a surprisingly large set of If-Then rules to search a table and encode interactions. We are investigating this interpretation further using eye-tracking studies of how experts and undergraduates scan tables of data.

Another reason for why the skills could be difficult to learn is that they involve deep misconceptions rather than simple lack of knowledge. For example, understanding the basic relationship between theory and evidence may involve a deep misconception. Kuhn (1989; 1991) has argued that many children and adults have confusion between theory and evidence—that they treat the two as the same. The work by Lehrer, Schauble, and Petrosino (this volume) suggests that many students do not understand the larger context in which experiments fit. The current findings are consistent with those views.

There is another line of research suggesting that many teenagers have an epistemological stance in which all beliefs are viewed as equally valid—every belief is just someone's opinion. It is often only with extended

undergraduate and graduate school experience that many individuals appear to acquire the more sophisticated view that while nothing can be known with 100% certainty, some views are more credible that others given the current evidence (Kitchener & King, 1981; Kitchener, King, Wood, & Davison, 1989). One could see how an individual with the perspective that all beliefs are equally valid could find the task of designing experiments (i.e., collecting data) to select among theories as a fundamentally confusing, if not wrong-minded, activity (similar to a claim currently made by deconstructivists).

Designing for Science: The Simulated Psychology Lab

A more optimistic interpretation of our results is that, while current instruction has produced little improvement on these skills, alternative forms of instruction might produce more consistent and strong improvements. We propose that the Simulated Psychology Lab might contain the seeds of such an alternative form. In this chapter, we have presented thus far two uses of the SPL task. First, we presented it as a research tool for understanding what skills experts use in designing and interpreting experiments. This has advantages for instruction in that it can help identify which skills need to be covered (e.g., keep settings constant across experiments).

Second, SPL can be used as an assessment tool for understanding what undergraduates learn or do not learn in research methods and other psychology classes. The essay and multiple choice exams that courses typically use are not likely to be good tests of the complex procedural skills required in experimental design and outcome interpretation. The project-based assessment that research methods courses also use has its own problems, too. The students are typically scaffolded through the design and interpretation process to such a heavy extent that it is often unclear what was the student's skill and what was the teacher's skill. By contrast, SPL offers a way to test the complex skills involved in experimental design and outcome interpretation.

The third and new use of SPL that we propose is one of a teaching tool. With the use of computer projection screens that are now readily available in university settings, the instructor can bring the SPL task into the classroom and use it as a teaching tool. As a group activity, experiments can be designed and outcomes can be interpreted. The students can quickly see the experiment design cycle at an appropriately detailed level.

Misconceptions can be addressed and correct behavior can be modeled by the instructor. On the experiment design end, the advantage of SPL is that the consequences of various design decisions can be quickly and concretely explored. For example, the rapid growth of the number of cells in a factorial design becomes quickly clear. On the outcome interpretation end, the SPL task can be used to show how the results of one experiment can be used to inform the design of the next experiment. For example, floor or ceiling effects can be used to calibrate the difficulty of test items in an experiment (and they make clear the importance of pilot experiments!).

SPL is written in an educationally-targeted programming environment, called cT (Sherwood & Sherwood, 1988), that is cross-platform (i.e., Mac, PC, unix). The SPL program is freely available for distribution from the first author. The interface was designed such that it could be generalized to other scientific domains. For example, by changing the names of the variables to be manipulated and by specifying the equations determining outcomes, the SPL task can be used in other domains such as social psychology, physics, sociology, etc. The only requirement is that factorial experimental designs be appropriate for the domain. Looking to the future, we are currently developing an HTML variant that allows more flexibility in both the range of experiments that can be designed and the kinds of outcome analyses that can be conducted (e.g., graphs, inferential statistics, etc.).

Acknowledgements

This work was supported by grant N00014-96-1-0491 to the second author from the Office of Naval Research. The authors would like to thank Kevin Crowley and Takeshi Okada for comments provided on earlier versions of the manuscript. Correspondence regarding this chapter may be addressed to the first author at Department of Psychology, MSN 3F5, George Mason University, Fairfax, Virginia 22030-4444. Electronic mail may be sent to schunn@gmu.edu.

References

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

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

Anderson, J. R., Fincham, J., & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23(4), 932-945.*

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

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.

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391-416.

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

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58*.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*(8), *725-747*.

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

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

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277-300.

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.

Hayes, J. R. (1985). Three problems in teaching general skills. In S. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills, Vol. 2* (pp. 391-406). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation (Vol. 19)*. New York: Academic Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: MIT Press.

Johnson, K. E., & Mervis, C. B. (1997). Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General, 126*(3), *248-277*.

Johnson-Laird, P. N. (1972). *Psychology of reasoning: Structure and content*. Cambridge, MA: Harvard University Press.

Kitchener, K. S., & King, P. M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. *Journal of Applied Developmental Psychology, 2*(2), 89-116.

Kitchener, K. S., King, P. M., Wood, P. K., & Davison, M. L. (1989). Sequentiality and consistency in the development of reflective judgment: A six-year

longitudinal study. *Journal of Applied Developmental Psychology, 10*(1), 73-95.

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

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96(4)*, 674-689.

Kuhn, D. (1991). *The skills of argument*. Cambridge, MA: Cambridge Press.

Larkin, J. H. (1980). Skilled problem solving in physics: A hierarchical planning model. *Journal of Structural Learning, 6*, 271-297.

Mahoney, M. J. (1979). Psychology of the scientist: An evaluative review. *Social Studies of Science, 9*, 349-375.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 14*(3), 510-520.

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

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 15*, 456-468.

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.

Schunn, C. D., & Anderson, J. R. (1998). Scientific discovery. In J. R. Anderson & C. Lebiere (Eds.), *Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schunn, C. D., & Anderson, J. R. (1999). The generality/specificity of expertise in scientific reasoning. *Cognitive Science, 23*(3), *337-370*.

Sherwood, B. A., & Sherwood, J. N. (1988). *The cT language*. Champaign, IL: Stipes Publishing.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychology? A case study. *American Psychologist, 52*(6), *630-641*.

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

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. (1968). Reason about a rule. *Quarterly Journal of Experimental Psychology, 20*, 273-281.