CHAPTER 10

Laboratory Methods for Assessing Experts' and Novices' Knowledge

Michelene T. H. Chi

Introduction

Expertise, by definition, refers to the manifestation of skills and understanding resulting from the accumulation of a large body of knowledge. This implies that in order to understand how experts perform and why they are more capable than non-experts, we must understand the representation of their knowledge, that is, how their knowledge is organized or structured, and how their representations might differ from those of novices. For example, if a child who is fascinated with dinosaurs and has learned a lot about them correctly infers attributes about some dinosaurs that was new to them by reasoning analogically to some known dinosaurs (e.g., the shape of teeth for carnivores versus vegetarians), we would not conclude that the "expert" child has a more powerful analogical reasoning strategy. Instead, we would conclude that such a global or domain-general reasoning strategy is available to all children, but that novice children might reason analogically to some other familiar domain, such as animals (rather than dinosaurs), as our data have shown (Chi, Hutchinson, & Robin, 1989). Thus, the analogies of domain-novice are less powerful not necessarily because they lack adequate analogical reasoning strategies, although they may, but because they lack the appropriate domain knowledge from which analogies can be drawn. Thus, in this framework, a critical locus of proficiency lies in the representation of their domain knowledge.

This chapter reviews several methods that have been used to study experts in the laboratory, with the goal of understanding how each method reveals the structure of experts' knowledge, in contrast to that of novices. The theoretical assumption is that the structure or representation of experts' knowledge is a primary determiner of how experts learn, reason, remember, and solve problems.

This chapter has three sections. It starts by briefly reviewing the historical background to studies of the experts' representations. The second section describes four general types of methods that have been commonly used to study expert knowledge. Finally, I briefly summarize what these methods can uncover about differences in the knowledge representations of experts and novices.

A Brief History on Representation in the Study of Expertise

The studies of representation in expertise have historically been intimately related to the type of problems being used. In early research on problem solving, the study of representation was carried out in the context of insight-type problems, such as Duncker's (1945) candle problem. The goal of this problem is to mount three candles at eye level on a door. Available to use for this problem are some tacks and three boxes. Participants were presented with the tacks either contained in the three boxes or outside of the boxes so that the boxes were empty. The solution requires that one re-represents the function of the boxes not as a container but as a platform that can be mounted on a wall to hold a candle. All the participants presented with the empty boxes could solve the problem, whereas less than half of the participants given the full boxes could solve it.

The key to all of these kinds of insight problems is to re-represent the problem in a way to either release a constraint that is commonly assumed, or to think of some new operator, that is again not the conventional one. So in the case of the candle problem, one could say that the conventional functional attribution that one applies to boxes is use as a container. Solving the problem requires thinking of a new function or affordance for boxes, in this case, as objects that can hold things up rather than hold certain kinds of things inside.

Although insight problems investigated the role of representation in the understanding phase of problem solving (i.e., how the elements, constraints, and operators of a problem are encoded and interpreted), insight problems did not lend themselves well to the study of expertise. That is, since expertise is defined as the accumulation of a large storehouse of domain knowledge, it is not clear how and/or what domain knowledge influences the solution of insight problems.

A next generation of problem-solving research explored both knowledge-lean (puzzle-like) problems (such as the Tower of Hanoi) as well as knowledge-rich problems (such as in chess). Even though chess is arguably more knowledge-rich than the Tower of Hanoi problem, it shares similarities with puzzles and other "toy" domains in that the understanding phase of the representation had been assumed to be straightforward (But see Ericsson, Chapter 13, and Gobet and Charness, Chapter 30). That is, for a domain such as chess, the understanding phase of the representation needs to include the chess pieces, the permissible operators (or moves) for each kind of chess piece, and the goal state of checking and winning. In short, the understanding phase of the representation had been assumed to not clearly discriminate experts from novices.

If understanding is not the phase that affects the choice of efficient moves, then what is? One obvious answer is how effectively a solver can search for a solution. The classical contribution by Newell and Simon (1972) put forth the idea that what differentiates experts from novices is the way they search through "problem spaces." A problem space includes not only the elements, the operators, but also all the possible or permissible "states" created by the application of operators to the elements, which are entailed by the permissible strategies for guiding the search through this problem space. In this perspective, a representation is a model of the search performance of a solver on a specific problem (Newell & Simon, 1972). Thus, a "problem representation" consists of:

1. An understanding phase – the phase in which information about the initial state, the goal state, the permissible operators, and the constraints is represented (so for chess, that would be the pieces and their positions on the chess board, the moves allowed and disallowed for each kind of chess piece, etc.), and

2. A search phase – the phase in which a step-by-step search path through the problem space is represented.

Because the understanding phase had been assumed to be straightforward, differences between experts and novices are assessed via comparing differences in the search phase. A variety of different search heuristics have been identified, such as depth-first versus breadth-first searches, backward versus forward searches, exhaustive versus reduced problem-space searches, and so forth.

This view – that differences in search strategies or heuristics accounted for differences in expertise – was also applied to knowledge-rich domains for which the understanding phase may not be so straightforward. A perfect example is the work of Simon and Simon in the domain of physical mechanics. In this research, Simon and Simon (1978) compared the problem-solving skills of an expert and a novice by representing their solution paths in terms of a sequence of equations (a set of productions or condition-action rules) that they used to solve a physics problem. Based on this sequencing, the expert's representation was characterized as a forward-working search (working from initial state toward the desired end state in a series of steps), whereas the novice's representation was characterized as a backward-working search (working from the desired end state back to the initial state). Thus, the postulated representational difference between the expert and the novice was restricted to the search phase, even though the understanding phase may be a more crucial component for this knowledge-rich domain.

The revelation that search may not be the entire story came from the work of de Groot (1966). He found that world-class chess players did not access the best chess moves from an extensive search; rather, they often latched on to the best moves immediately after the initial perception of the chess positions. For example, de Groot could not find any differences in the number of moves considered, the search heuristics, or the depth of search between masters and less-experienced (but proficient) players. What he did find was that the masters were able to reconstruct a chess position almost perfectly after viewing it for only 5 seconds. This ability could not be attributed to any superior general memory ability, for when the chess positions were "randomized," the masters performed just about as poorly as the less-experienced players. This finding suggests that the masters' superior performance with meaningful positions must have arisen from their ability to perceive structure in such positions and encode them in chunks.

The findings that chess experts can perceive coherent structures in chess positions and rapidlly come up with an excellent choice of moves suggest that the understanding phase must be more than merely the straightforward encoding of the elements and permissible operators to apply to the elements. Moreover, the application of different search heuristics cannot be the characterization that differentiates the experts from the novices in the search phase. Thus, what differentiated the experts and the novices' problem representation is determined by the representation of their domain knowledge, of chess in this case. This recognition led Chase and Simon (1973a, b) to the identification and characterization of the structures or chunks of meaningful chess patterns in memory. Thus, the work of de Groot (1966) and Chase and Simon (1973a, b) represented a first attempt at representing not just a *problem* solution, but knowledge of the *domain*. Subsequent work on expertise attempted to focus on how domain knowledge is represented in a way that leads to better solutions.

For example, we have shown that expert physicists' representation of their domain is more principle based, whereas novices' representations are more situation or formula based (Chi, Feltovich, & Glaser, 1981). Thus, the expertise work in the 80's reemphasized the understanding phase of representation, but it differed from the earlier work on insight and other knowledge-lean problems in that the focus was on the structure and organization of domain knowledge, and not merely the structure of the problem.

The next challenge for researchers is to combine the understanding phase and the search phase of a representation in order to understand how it differentiates experts from novices. In addition, new challenges are also presented when expertise is being investigated in real-world domains. Many complexities are involved when one studies expertise in real-world domains, where problems are complex and dynamic, so that the "space" is constantly changing with contextual dependencies and contingencies. In this kind of real-world scenarios, the space-search model of problem solving does not always apply as an explanatory mechanism. It is also essentially mute about problem finding, which is a main phenomenon in real-world problem-solving (see Klein, Pliske, Crandall, & Woods, 2005).

Empirical Methods to Uncover Representational Differences

The nature of expertise can be ascertained in two general ways. One way is to see how they perform in tasks that are familiar or *intrinsic* to their domain of expertise. For example, selecting the best chess move, generating the optimal blueprint, or detecting a cancerous mass on X-rays are tasks that are intrinsic to the domains of chess playing, on being an expert architect, and on being an experienced radiologist. This has been referred to as the study of performance at "familiar tasks" (Hoffman, 1987; Hoffman, Shadbolt, Burton, & Klein, 1995). Although these tasks might be abridged or in many ways adapted for empirical investigation under conditions of experimental control and the manipulation of variables, they are nevertheless more-or-less representative of what the domain experts do when they are doing their jobs.

Alternatively, one can use *contrived* tasks (Hoffman, 1987; Vicente & Wang, 1998) that are likely to be either unfamiliar to the practitioner, or that depart more radically from their familiar intrinsic tasks. Contrived tasks serve different purposes so that there is a continuum of contrived tasks, based on the degree of modifications to the familiar task in order to "bring the world into the laboratory," as it were (Hoffman et al., 1995). However, there is a set of standard tasks that are commonly undertaken in psychological laboratories, such as recall. Recall of chess positions, for example, can be considered a contrived task since chess experts' primary skill is in the selection of the best moves, not in recalling chess patterns. Although experts do recall games for a number of reasons (e.g., knowledge sharing), asking them to recall chess patterns can be thought of as a contrived task.

It is often the case that asking experts to perform in their familiar intrinsic tasks will show only that they are faster, more error free, and in general better in all ways than the novices. Their efficiency and speed can often mask how their skills are performed. Asking experts to perform contrived tasks, on the other hand, can have several advantages. First, a contrived task is often one that can be undertaken just as competently by a novice as an expert. Thus, it is not merely the completion, efficiency, or correctness of performance at a contrived task that is being evaluated, but rather, what the performance reveals about the knowledge structure of the individual, whether an expert or a novice. More importantly, a contrived task can shed light on experts' shortcomings (see Chi, Chapter 2), whereas an intrinsic task will not, by definition of expertise. A key limitation of contrived tasks, however, is that if the contrived task departs too much from the familiar task (e.g., lacks ecological validity and/or representativeness), then the model of performance that comes out may be a model of how the person adapts to the task, not a model of their expertise.

In this section, I describe four contrived tasks that have been used most extensively in laboratory studies of expertise with the goal of uncovering representational differences. The four methods are: recalling, perceiving, categorizing, and verbal reporting. Studies using these four methods are grouped on the basis of the tasks that were presented to the participants, and not the responses that they gave. For example, one could present a perceptual task and ask for verbal reports as responses. However, such a task would be classified here as a perceptual task and not a verbal reporting task. Clearly there are many combinations of methods and many optional ways to classify a task used in a specific study. The choice here reflects only the organization of the presentation in this chapter. Moreover, many studies use a combination of several methods.

RECALL

One of the most robust findings in expertise studies comes from using the method of free recall. Experts excel in recalling materials from their domain of expertise, such as better, faster, and more accurate recall, in domains ranging from static chess positions (Chase & Simon, 1973a) to dynamic computer-simulated thermal-hydraulic process plant (Vicente, 1992). The classic study by de Groot (1966) in the domain of chess involved presenting chess players with meaningful chess boards for a brief interval, such as 5 seconds, to see how many pieces they could recall by reproducing the arrangements of the pieces on a blank board. Chess masters were able to recall the positions almost perfectly (consisting of around 25 pieces). Less experienced players, on the other hand, typically recall only about 5 to 7 pieces (Chase & Simon, 1973a). However, when de Groot (1966) asked the players to find the best move, the masters and the less experienced players did not differ significantly in the number of moves they searched nor the depth of their search, even though the masters were always able to find and select the best move. Likewise, Klein, Wolf, Militello, and Zsambok (1995) found that the first move that expert chess players consider is significantly better than chance. Furthermore, chess experts do not differ from class-C players in the percentage of blunders and poor moves during regulation games, but do differ during blitz games. In fact, the experts showed very little increase in rate of blunders/poor moves from regulation to blitz, but the class-C players showed a big difference (Calderwood, Klein, & Crandall, 1988).

These findings suggest that it is not the experts' superior search strategies that helped them find the best move. Neither can the master players' superior recall be attributed to any differences in the memory capacities of the master and less experienced players, since masters can only recall a couple more pieces when the pieces are randomly placed on the chess board (Chase & Simon, 1973a).

This same pattern of results was also obtained when Go (or Gomoku) players were asked to recall briefly presented Gomoku (or Go) board patterns. Both Go and Gomoku utilize the same lattice-like board with two different colored stones, but the object of the two games is very different: In Go the goal is to surround the opponent's stone and in Gomoku it is to place five stones in a row (Eisenstadt & Kareev, 1975). The success of players in recalling board configurations suggests that it is the meaningfulness of the configurations that enables the strong players' better recall.

In order to understand how experts and novices might organize their knowledge to result in differential recall, Chase and Simon (1973a,b) incorporated two additional procedures in conjunction with their recall procedure, both aimed at segmenting the sequence in which players place the chess pieces during recall. The first procedure tape-recorded players as they reproduced chess pieces from memory and used the pauses in their placement of pieces to segment the sequence of placements. The second procedure was to modify the task from a recall to a visual memory task. In this modified visual task, players were simply asked to copy chess positions. The head turns they made to view the positions in order to reproduce the chess positions were used to segment the sequence of placements, that is, to reveal how the game arrays were "chunked." The results showed that players recalled positions in rapid bursts followed by relatively longer pauses (i.e., > 2 seconds), and they reproduced a meaningful cluster of pieces after a head turn. Because the master players recalled and reproduced a greater number of pieces before a long pause and a head turn, respectively, these two results, together, suggest that chess experts had many more recognizable configurations of chess patterns in their knowledge base, and these configurations (based on power in controlling regions of the board) were comprised of a greater number of pieces. The representational differences between the masters and less proficient players were that the masters had a greater number of recognizable patterns (or chunks) in memory, and each pattern on average contained a greater number of pieces.

More important, when memory performance was reanalyzed in terms of experts and non-expert chunks, the number of chunks recalled by experts and non-experts were now about the same, implying that their basic memory capacity is not that different after all, validating the finding of the depressed expert-recall performance for randomized board arrangements. The findings of equivalent recall for randomized positions and equivalent recall in terms of number of patterns, together, confirm that both expert and non-expert players are subject to the same short-term memory capacity limitations, but the limitation is not the point. The point is how people come to create meaningful chunks.

The recalled chess patterns (as determined by segregated pauses and head turns), when analyzed in detail, showed that they tended to consist of commonly occurring patterns that are seen in regular routine playing of chess, such as clusters in attack and defense positions. It seems obvious that such "local" patterns may be used to form representations at a higher level of familiar "global" patterns. Direct evidence of such a hierarchical representation can be seen also in the domain of architecture. Using the same recall procedure, looking at pauses, Akin (1980) uncovered a hierarchical representation of blueprints, with such things as doors and walls at the lowest level and rooms at a higher level, and clusters of room at the highest level.

The chunking of patterns into a hierarchical representation applies not only to games and architecture, but to other domains, such as circuit fault diagnosis. Egan and Schwartz (1979) found that expert circuit technicians chunk circuit elements together according to the function, such as chunking resistors and capacitors because together they perform the function of an amplifier. Here too, chunking leads to superior recall for experts as compared to non-experts. Moreover, the skilled electronic technicians' pattern recall was faster and more accurate, again suggesting that the local patterns formed higher-order patterns.

The recall superiority of experts can be captured not only in visual tasks, but also in verbal tasks. Looking at a practical domain, Morrow, Mernard, Stine-Morrow, Teller, and Bryant (2001) asked expert pilots and some non-pilots to listen to Air Traffic Control messages that described a route through an air space. Participants were then asked to read back each message and answer a probe question about the route. Expert pilots were more accurate in recalling messages and in answering the question than non-experts.

In sum, several different types of recall-related contrived tasks provide some insight into the experts' and non-experts' representation of their domain, such as patterns of familiar chunks, clusters of circuit elements with related function, and hierarchical organization of chunks.

PERCEIVING

Perception tasks address the issue of what experts versus non-experts perceive in a given amount of time (Chase & Chi, 1981). A good example of a perceptual task is examining X-ray films. Although the goal of examining X-ray films is usually to diagnose disease, one can also determine what experts and novices see (literal stimulus features) and perceive (meanings of the features or patterns of features). Lesgold et al. (1988) asked four expert radiologists with 10 or more years of experience after residency, and eight first-to-fourth year residents to examine X-ray films for as long as they wished, commenting on what they saw as well as verbally expressing their diagnoses. Although diagnosis is the familiar intrinsic task, the participants were also asked to undertake a more contrived task, which was to draw contours on the films showing what they believed to be the problematic areas, as a way of identifying the relevant features they saw. (The films showed diseases such as multiple tumors or collapsed lung.) Two of the four experts, but only one of the eight residents, diagnosed the collapsed lung film accurately. Did they see the features in the films differently? Both experts and residents saw the main feature, which was the collapse of the middle lobe, producing a dense shadow. However, this feature can lead only to a tumor diagnosis; the correct diagnosis of collapsed lung must require seeing the displaced lobe boundaries or hyperinflation of the adjacent lobes. Residents did not see the more subtle cues and the relations among the cues.

In addition to the accuracy of the diagnoses, the researchers looked at two kinds of coding of the protocols. The first coding was the diagnostic findings, which referred to the attribution of specific diagnostic properties in the film. For example, one finding might be "spots in the lungs." The second coding was the meaningful clusters. A cluster is a set of findings that had a meaningful path or reasoning chain from each finding to every other finding within the set. That is, the participants would relate the features logically to entail a diagnostic explanation. For example, if the participants commented that such spots might be produced by blood pooling, which in turn could have been produced by heart failure, then such a reasoning chain would relate the findings into a cluster. The results showed that the experts identified around three more findings per film, and had about one more cluster than the residents. This suggests that the experts not only saw more critical features on a film than the residents, but perceived more interrelations among the features.

Moreover, experts had finer discriminations. For example, the tumor film showed a patient with multiple tumors. For this tumor film, residents tended to merge local features (the tumors) as "general lung haziness." That is, they interpreted the hazy spots in the lungs as indicating fluid in the lungs, suggesting congestive heart failure, whereas experts saw multiple tumors. Residents also saw the heart as enlarged, while the experts did not. Residents also interpreted the cues or features they saw rather literally. For example, a large size heart shadow implied an enlarged heart, whereas experts might adjust their evaluation of the heart to other possibilities, such as a curvature in the spine.

The results of this study show basically that experts perceive things differently from non-experts. There are many other studies that show the same kind of results (see Klein & Hoffman, 1992). This includes the perception tasks of reproducing chess board patterns as discussed earlier. Reitman (1976) also replicated the Chase and Simon (1973a) study for the game of Go. In addition to asking participants to reproduce patterns of Go stones as quickly and accurately as possible while the stimulus board pattern remained exposed throughout the trial, she also asked the Go experts to draw circles (on paper transcriptions of the real game positions) showing stones that were related, and if appropriate, to indicate which groups of stones were related on yet a higher strategic level. The results showed that the experts partitioned the patterns not into a strictly nested hierarchy, but rather into overlapping subpatterns, as one might expect given the nature of Go – a given stone can participate in, or play a strategic role in, more than one cluster of stones. Although there were no novice data on penciled partitioning, the expert's partitioning into overlapping structures suggests this more interrelated lattice-like (versus strictly hierarchical) representation.

The perceptual superiority of experts applies to dynamic situations as well, such as perception of satellite infrared image loops in weather forecasting (Hoffman, Trafton, & Roebber, 2005), or watching a videotape of classroom lesson (Sabers, Cushing, & Berliner, 1991). For example, when expert and novice teachers were asked to talk out loud while watching a videotaped classroom lesson that showed simultaneous events occurring throughout the classroom, the experts saw more patterns by inferring what must be going on (such as "the students' note taking indicates that they have seen sheets like this . . . "), whereas the non-expert teachers saw less, saying that "I can't tell what they are doing. They are getting ready for class." In short, the explanations experts and non-experts can give reveal the features and meaningful patterns they saw and perceived.

A related task is detection of the presence of features or events accompanied by measurement of reaction times. For example, Alberdi et al. (2001) asked some more- and some less-experienced physicians to view traces on a computer screen showing five physiological measurements, such as heart rate, transcutaneous oxygen, etc. The traces represented both key events, such as developing pneumothorax, as well as more secondary but still clinically noteworthy events. Although the less-experienced physicians were almost as good in detecting and identifying the key events, they were significantly worse than the more-experienced physicians in detecting the secondary events. The more-experienced physicians were also significantly better at detecting artifacts. This suggests that they were not only better at detecting secondary events, but that they also made finer discriminations between meaningful events versus literal stimulus features.

It should perhaps be pointed out that such results do not arise from experts having better visual acuity. Nor do the results mean that the experts' perceptual superiority is necessarily visual (vs. analytical). That is, expertise involves perceiving more, not just seeing more. To deny the first interpretation, one can show that novices' visual acuity is just as good as experts in some other domain for which they have no expertise. However, expertise can enhance sensitivity to critical cues, features, and dimensions. Snowden, Davies, and Roling (2000) found expert radiologists to be more sensitive to low contrast dots and other features in X-rays. This increased sensitivity can be driven "top down" by more developed schemas (rather than a better developed acuity) since greater experience with films means they have more familiarity with both under- and overexposed films. To disprove the second interpretation – that perceptual superiority is necessarily visual – one can show that experts can excel in perception even if the materials are not presented visually, as in the case of chess masters playing blindfolded chess (Campitelli & Gobet, 2005) and expert counselors forming an accurate model of a client from listening to a transcript of a counseling session (Mayfield, Kardash, & Kivlighan, 1999).

In sum, this section summarized perception tasks and related contrived tasks such as asking experts and novices to circle Go patterns or draw contours of X-ray films. The point of these studies is not merely to show *whether* experts are superior in performing these kinds of tasks, but to uncover their underlying representations and skills that derive from practice and perceptual learning, such as more interrelated clustering of findings on X-ray films and their representation of secondary events.

CATEGORIZING

Sorting instances according to categories is a simple and straightforward task that can be readily undertaken by experts and non-experts. One procedure is to ask participants to sort problem statements (each problem typed on a 3 × 5 card) into categories on the basis of similarities in the solution or some other functional categories. Chi et al. (1981) solicited the participation of physics graduate students (who technically would be apprentices or perhaps journeymen on the proficiency scale, but probably not fully expert) and undergraduate students (who had completed a semester of mechanics with an A grade, making them "initiates" and not really novices). They were asked to sort 24 physics problems twice (for consistency), and also to explain the reasons for their sorting. One would not necessarily expect quantitative differences in the sortings produced by the two skill groups, such as the number of groups, or the number of problems in the groups – since anyone could sort problems on any of a nearly boundless number of dimensions or criteria. The real interest lies in the nature of the sortings. Based on analyses of both the problems that the participants categorized into the same groups as well as their explanations for the sortings, it became apparent that the undergraduates grouped problems very differently from the graduate students. The undergraduates were more likely to base their sorting on literal surface features, such as the presence of inclined planes or concepts such as friction, whereas the graduate students were much more likely to base their sorting on domain principles that would be critical to the solutions (e.g., such as problems that involve Newton's Second Law or the laws of thermodynamics such as conservation of energy). This finding was further replicated by a specially designed set of problems that had either the same surface features but different deep principles, or different surface features but the same deep principles. The same results emerged, namely, that undergraduates sorted according to the surface features and graduates tended to sort according to the deep principles.

One interpretation of such results is that the undergraduates' schemas of problems are based on physical entities and literal formulas, whereas experts' schemas are more developed and organized around the principles of mechanics. This means that the explicit words or terminologies and diagrams used in the problem statements are connected (in experts' reasoning) to the basic principles. However, that connection is not necessarily direct. For instance, an inclined plane per se does not by itself indicate a Newton's-Second-Law problem for an expert physicist. An additional study asking participants to cite the most important features in a problem statement showed that the words in the problem statements are mediated by some intermediate concepts, such as a "before and after situation." Thus, the words in a problem interact to entail concepts, and experts' solutions may be based on these higher-level concepts (Chi et al, 1981; Chi & Ohlsson, 2005).

Much research followed that replicated the basic finding of shallow versus deep representations for novices versus experts. For example, when expert and novice programmers were asked to sort programming problems, the experts sorted them according to the solution algorithms, whereas the novices sorted them according to the areas of applications, such as creating a list of certain data types (Weiser & Shertz, 1983). Similarly, when expert and novice counselors were asked to categorize client statements from a counseling script as well as to map the relationships among the categories, novices tended to categorize and map on the basis of superficial details, such as the temporal order of the client statements (Mayfield et al., 1999), whereas the expert counselors tended to categorize and map on the basis of more abstract, therapeutically relevant information. Similarly, Shafto and Coley (2003) found that commercial fishermen sorted marine creatures according to commercial, ecological, or behavioral factors, whereas undergraduates sorted them according to the creatures' appearance.

Many variations of the sorting task have also been used. One variation is to ask participants to subdivide their groups further, to collapse groups, or to form multiple and differing sortings in order to shed light on the hierarchical structure of their knowledge representations (Chi, Glaser, & Rees, 1982). For example, by asking a young dinosaur "expert" to collapse his initial categories formed about different types of dinosaurs, the child would collapse them into two major superordinate categories– meat-eaters and plant-eaters (Chi & Koeske, 1983)– suggesting that the superordinate categories are somewhat well defined.

Another variation is a speeded category-verification task. In such a task, a category name appears first, followed by a picture. Participants press "true" if the picture matched the word, such as a picture of a dog with the term "animal," and "false" if it does not match, and reaction latencies can be measured. Moreover, the words can refer to a superordinate category such as "animals," a basic-object-level category such as "dog," or a subordinate category such as "dachshund." The basic-object level is normally the most accessible level for categorizing objects, naming objects, and so forth (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). It has a privileged status in that it reflects the general characteristics of the human perceiver and the inherent structure of objects in the world (i.e., frequency of experience and word use). The basic-object level is also the first level of categorization for object recognition and name retrieval.

Dog experts showed the typical pattern of responses for their non-expert domain, such as birds, in that their reaction times were faster at the basic level than at the superordinate or the subordinate levels (Tanaka & Taylor, 1991; Tanaka, 2001). However, in their domain of expertise, the experts were just as fast at categorizing at the subordinate level as they are at categorizing at the basic-object level. For example, dog experts can categorize a specific dog as a dachshund as fast as they can categorize a dachshund as a dog. This downward shift in the creation of a second, more specific basic level in a hierarchy means that the experts' hierarchies are more differentiated even at the subordinate level (see also Hoffman, 1987). Moreover, this finer subordinate-level discrimination is evident even in child "experts" (Johnson & Eilers, 1998).

In sum, the categorization tasks described here, consisting of sorting and category verification, can reveal the structure of experts' knowledge, showing how it is more fully developed and differentiated at both the subordinate levels and the superordinate levels.

VERBAL REPORTING

One of the most common methods in the study of expertise is to elicit verbal reports. (It should be kept in mind that verbal reporting and introspection are different in important ways. Verbal reporting is task reflection as participants attend to problems. It is problem centered and outward looking. Introspection is to give judgments concerning one's own thoughts and perceptions.) Verbal reporting, as a category of task, can be done either as an ongoing think-aloud protocol (Ericsson & Simon, 1984; see Ericsson, Chapter 13), as answers to interview questions (Cooke, 1994), or as explanations (Chi, 1997).

These three techniques are quite different. For concurrent think-aloud protocols, the participants are restricted to verbalize the problem information to which they are attending. In interviews, especially structured interviews, the questions are usually carefully crafted (i.e., to focus on a specific topic or scenario) and are often sequenced in a meaningful order (see Hoffman & Lintern, Chapter 12). Explanations, on the other hand, are given sometimes to questions generated by a peer, by oneself, or by an experimenter. Explanations can be retrospective and reflective. (Differences between think-aloud protocols and explanations are elaborated in Chi, 1997.) Not only are there different ways to collect verbal reports, but there are other important issues that are often debated. One issue, for example, concerns whether giving verbal reports actually changes one's processing of the task (Nisbett & Wilson, 1977), and another issue is whether different knowledge elicitation methods elicit different "kinds" of knowledge from the participants – the "differential access hypothesis" (Hoffman et al., 1995).

Not only can verbal reports be collected in several different ways, but they can be collected within the context of any number of other tasks, such as a perception task, a memory task, or a sorting task, as some of our earlier examples have shown. Thus, providing verbal reports can be a task in its own right – as in the case of a free-flowing, unstructured interview (Cullen & Bryman, 1988), or simply asking the participant to say what he or she knows about a concept (Chi & Koeske, 1983). But a verbal protocol can also be solicited in the context of some other task (such as solving problems or analyzing documents). However, to be consistent with the heuristic of this chapter, the studies below are grouped in this section according to the main task presented to the participants. In this regard it is worth noting that in some domains, giving a concurrent or retrospective verbal report is part of the familiar intrinsic task (e.g., coroner's audio record during autopsies; and during weather forecasting briefings, forecasters think aloud as they examine weather data).

The most difficult aspect of verbal report methods is data analysis. That is, how does one code and analyze verbal outputs? Again there are many methods; they can only be alluded to here (see Chi, 1997; Ericsson & Simon, 1984, for explicit techniques, and Ericsson, Chapter 13). Typically, think-aloud protocols are analyzed in the context of the cognitive task, which requires a cognitive task analysis in order to know the functional problem states that are to be used to categorize individual statements. The goal of protocol analysis then is to identify which sequence of states a particular participant progresses through, and perhaps a computational model is built to simulate those steps and the solution procedures. For explanations, coding methods involve segmenting and judging the content of the segments in terms of issues such as whether it is substantive or non-substantive (Chi, Siler, Jeong, Yamauchi, & Hausmann, 2001), principle oriented (deep) or entity oriented (shallow) (Chi et al., 1981). Note that an analysis of verbal data means that the content of the data is not always taken literally or word-for-word. That is, we are not asking experts and novices their subjective assessment of how they performed, or how they have performed. This is because much of expert knowledge is not explicit nor subject to introspection.

How people perform can be captured by the coding scheme. A study by Simon and Simon (1978) provides a good example. They collected concurrent protocols from an expert and a novice as they were solving physics problems. The researchers coded only the equation-related parts of the protocols. By examining what equations were articulated, and when, the researchers were able to model (using a production-system framework) each participant's problem-solving procedure and strategy. The researchers showed that the expert solved the problems in a forward-working strategy, whereas the novice worked backward from the goal (as one would predict on the basis of studies described earlier in this chapter). The same forward-backward search patterns were obtained also in the domain of genetics with experts and novices (Smith & Good, 1984).

In a different kind of domain and task, Wineburg (1991) asked historians and history students to give think-aloud protocols while they constructed understanding of historical events from eight written and three pictorial documents. The participants' task was to decide which of the three pictures best depicted what happened during the Battle of Lexington at the start of the Revolutionary War, the event presented in the documents. Statements in the participants' picture-evaluation protocols were coded into four categories: description, reference, analysis, and qualification. Both experts and students provided descriptive statements, but the experts made more statements that fell into the other three categories. This is not surprising since the experts obviously had more to say, being more knowledgeable. What is more interesting is to identify the first category for which both the experts and novices described the picture using the same number of statements. The quality of those descriptions was different. Historians noted 25 of the 56 possible key features in the paintings that had a bearing on the historical accuracy of the paintings, whereas the students noted only four features on average. Moreover, in selecting the most accurate painting, historians did so on the basis of the correspondence between the visual representations and the written documents, whereas the students often chose on the basis of the quality of the artwork, such as its realism and detail. This suggests that the experts' representations were much more meaningfully integrated.

Interviewing techniques can include both open-ended questions and more direct questions. For example, Hmelo-Silver and Pfeffer (2004) asked experts and students both direct questions about aquaria, such as "*What do fish do in an aquarium?*" and open-ended questions, such as thinking out loud while attempting to *"Draw a picture* *of anything you can think is in an aquarium.*" Since biological systems and devices often can be characterized by their structure, behavior, or function (Gellert, 1962; Chi, 2000, p. 183; Goel et al., 1996), the protocols were coded according to statements relating to those three categories. There were no differences between the experts and the novices in the number of statements referring to the structures, but there were predictable and significant differences in the number of statements referring to behaviors and functions. The novices often did not offer additional behavioral or functional information even when probed. This suggests that the experts represent the deeper features (i.e., behavior and function), whereas novices think in terms of literal features (i.e., the structure).

In sum, the goal of these verbal reporting methods is to capture the underlying representations of the experts and novices, such as whether their searches are forward versus backward, whether their understanding of pictures and text are integrated versus literal, or whether their understanding manifest deep (behavioral and functional) versus shallow (structural) features.

Representational Differences

If the difference in representation (reflecting the organization of knowledge and not just the extent of knowledge) is one key to understanding the nature of expertise, then in what ways do the representations of experts and novices differ? In this section, I briefly address dimensions of representational differences, as captured by the empirical tasks of recalling, perceiving, categorizing, and verbal reporting described above. Each of these tasks has revealed ways in which representations of experts and novices differ.

KNOWLEDGE EXTENT

An obvious dimension of difference is that experts have more knowledge of their domain of expertise. More knowledge must be measured in terms of some units. Without being precise, a "bit" of knowledge can be a factual statement, a chunk/familiar pattern, a strategy, a procedure, or a schema. Chase and Simon (1973a, b) estimated an expert chess (master-level) player to know between 10,000 and 100,000 chunks or patterns, whereas a good (Class-A) player has around 1000 chunks; and Miller (1996, pp. 136–138) estimated college-educated adults to know between 40,000 to 60,000 words. Hoffman et al., (in press; Hoffman, Trafton, & Roebber, 2006) estimate that it would take thousands of propositions to capture the expert weather forecaster's knowledge just about severe weather in one particular climate. Regardless of how one wishes to quantify it, clearly, one can expect experts to know more than non-experts (including journeymen and especially compared to apprentices, initiates, and novices). Indeed, this is one definition of expertise. The recall task summarized earlier also revealed how the number of chunks and the chunk sizes differ for experts versus non-experts.

Aside from the sheer number of "bits" (however these are defined) in their knowledge base, a related concept to the dimension of size is *completeness*. Completeness has a different connotation than the idea of merely greater amount or extent of knowledge. In real-world domains knowledge is always expanding. Any notion of "completeness" becomes very slippery.

In terms of frame theory, one can conceive of completeness in terms of the availability or number of slots, or necessary slots. For example, a tree expert might have slots for "susceptibility to different diseases" with knowledge about potential diseases (values) for each kind of trees, whereas a novice might not have such slots at all. The earlier-described finding from a perception task showed that the more- (but not the less-) experienced physicians were able to recognize secondary events on traces of physiological measurements (Alberdi et al., 2001), can be interpreted to indicate that the more-experienced physicians had more complete frames or schemas. Greater amount of knowledge might also refer to more details in the experts' representation than in novices', for a particular domain.

Another way to discuss knowledge extent is in terms of the *content*. Experts might not have just more production systems than nonexperts for solving problems, but they might have different production systems, as shown by Simon and Simon's (1978) study of physicists using a verbal-reporting task. For example, experts might have rules relevant to the principles, whereas novices might have rules relevant to the concrete entities in the problem statement (Chi et al., 1981). This can mean that the experts' production systems are deeper and more generalizable.

In sum, differences in the size or extent of the knowledge as a function of proficiency level can be uncovered in a number of contrived tasks that have been discussed in this chapter.

THE ORGANIZATION OF KNOWLEDGE

The hierarchical representation of knowledge can be inferred from the way experts cluster in their recall, as in the case of recalling architectural plans (Akin, 1980) and circuit diagrams (Egan & Schwartz, 1979). If we therefore assume that representations are sometimes hierarchical (depending on the domain), then in what further ways are the experts' representations different from novices?

One view is that non-experts might have missing intermediate levels. For example, using a recall task, Chiesi, Spilich, and Voss (1979) found that individuals with high or low prior knowledge of baseball were equally capable at recalling individual sentences that they had read in a baseball passage. However, the experts were better at recalling sequences of baseball events because they were able to relate each sequence to the high-level goals such as winning and scoring runs. This suggests that the basic actions described in the individual sentences were not connected to the high-level goals in the novices' understanding. Perhaps such connections have to be mediated by intermediate goals, which may be missing in novices' hierarchical structure. The same pattern of results was found in children's representation of knowledge about "Star Wars." The "Star Wars" game can be represented in a hierarchical structure, containing high-level goals such as military dominance, subgoals such as attack/destroy key leaders, and basic actions, such as going to Yoda (Means & Voss, 1985).

Similar findings have been obtained also in studies of medical domains, in which physician's diagnostic knowledge has been represented in terms of hierarchical levels (Patel & Arocha, 2001). In such a representation, studies using a perception task show that physical observations are interpreted in terms of *findings*, which are observations that have medical significance and must be clinically accounted for. At the next level are *facts*, which are clusters of findings that suggest prediagnostic interpretation. At the highest level are *diagnoses*. Novices' and experts' representation can differ in that novices can be missing some intermediate-level knowledge, so that decisions are then made on the basis of the *findings* level, rather than the *facts* level.

A third way to conceive of differences in hierarchical representations of experts and novices is a in the level of the hierarchy that is most familiar or preferred for domains in which the hierarchical relationships is one of class-inclusion. Expert versus nonexpert differences arise from the preferred level within the hierarchy at which experts and novices operate or act on. According to Rosch et al. (1976), to identify objects, people in general prefer to use basic-object-level names (bird, table) to superordinate-level names (e.g., animals, furniture). People are also generally faster at categorizing objects at the basic-object level than at the superordinate or subordinate levels (e.g., robin, office chair). Experts, however, are just as facile at naming and verifying the subordinate-level objects as the basic-level, suggesting that the overall preferential treatment of the basic level reflects how knowledge about the levels are structured, and not that the basic level imposes a certain structure that is more naturally perceived. Using a sorting task, this differentiated preference for experts and novices has been replicated in several domains, such as birds (Tanaka & Taylor, 1991), faces (Tanaka, 2001), dinosaurs (Chi et al., 1989), and geological and archaeological classification (Burton et al., 1987, 1988, 1990).

Just as the notion of knowledge extent can be slippery (because knowledge is never static), so too the notion of hierarchical memory organization can be slippery. For example, instead of conceiving of non-experts' memory representation as missing the intermediate levels, another view is that their representations are more like lattices than hierarchies (Chi & Ohlsson, 2005). (Technically, a lattice would involve cross connections that would be "category violations" in a strict hierarchy or "is-a" tree.) It is valuable to look at an extreme, that is, domains where everything can be causally related to everything else, and neither hierarchies, lattices, nor chains suffice to represent either the world or knowledge of the world, such as the weather forecaster's understanding of atmospheric dynamics (e.g., thunderstorms cause outflow, which in turn can trigger more thunderstorms). We do not yet have a clear understanding of how dynamic systems are represented (Chi, 2005). On the other hand, for a domain such as terrain analysis in civil engineering, much of the expert's knowledge is very much like a hierarchy, highly differentiated by rock types, subtypes, combinations of layers of subtypes, types of soils, soil-climate interactions, etc. (Hoffman, 1987).

In sum, although any inferences about knowledge representation need to be anchored in the context of a specific domain, contrived tasks such as recalling, perceiving, and categorizing can allow us to differentiate the ways experts' and novices' knowledge is organized.

"DEPTH" OF KNOWLEDGE

Representational differences can be characterized not only by extent and organization, but also by dimensions such as deep versus shallow, abstract versus concrete, function versus structure, or goal-directed versus taxonomic. Such differences have been revealed using a sorting task, to show, for example, that physicists represent problems at the level of principles, whereas novices represent them at the concrete level of entities or superficial features (Chi et al., 1981), or that landscaping experts sort trees into goal-derived categories (e.g., shade trees, fast-growing trees, etc.), whereas taxonomists sort trees according to biological taxa (Medin, Lynch, Coley, & Atran, 1997).

Such differences can be revealed also in perception tasks. For example, a patient putting his hands on his chest and leaning forward as he walks slowly is interpreted by novices merely as someone having back pain (a literal interpretation), whereas a more expert physician might interpret the same observation as perhaps suggesting that the patient has some unspecified heart problem (Patel & Arocha, 2001). Differences can also be revealed in a verbal reporting task, such as explaining the behavior/function of fish in an aquarium versus explaining the structure of fish (Hmelo-Silver & Pfeffer, 2004). Differences can be revealed in a task that involves explaining causal relationships – a novice's explanations might focus on the time and place of an historical event, whereas an expert's explanations might focus on using the time to reconstruct other events (Wineburg, 1991).

In short, all four of the task types reviewed here can reveal differences between experts' and novices' representations in terms of depth.

CONSOLIDATION AND INTEGRATION

A fourth dimension of representational differences between experts and non-experts is that the experts' representation may be more *consolidated*, involving more efficient and faster retrieval and processing. A related way to characterize it might be the integratedness or coherence of a representation, that is, the degree to which concepts and principles are related to one another in many meaningful ways (e.g., Falkenhainer, Forbus, & Gentner, 1990; Schvaneveldt et al., 1985). One interpretation of integratedness is the interaction of features. Evidence for this interpretation can be seen in physics experts' and non-experts' representations (Chi et al., 1981), in which they identify features that are combined or integrated to form higher-level concepts in a sorting task, as well as in physicians' ability to form clusters of observations for their prediagnostic interpretation in a perception task (Patel & Arocha, 2001).

For example, given a physics problem statement and asked to identify the features that determine their basic approach to the solution, novices will solve a problem on the basis of the explicit concrete entities mentioned in the statement, whereas experts will solve a problem on the basis of derivative features (such as a "before and after" situation), in which the interactions of the concrete entities in the problem statement are integrated to describe the problem situation as "before and after" (see Chi et al., 1981, Table 11). Tabulating the frequencies with which the two experts and novices cited concrete entities (such as spring, friction) versus higher-level dynamic features (such as a "before and after" situation, or a physical state change), there were 74 instances in which the experts cited dynamic features versus 21 references to concrete entities, whereas the reverse was true for novices, who cited 39 instances of concrete entities versus only two instances of dynamic features. The more integrated nature of the experts' knowledge base was also reflected in the reasoning chains that expert radiologists manifested in their diagnoses, cited earlier (Lesgold et al., 1988).

In short, recall, perception, and categorization tasks can all reveal differences in the consolidation and integration of representations.

Conclusion

The goal of this chapter was to describe and illustrate the kind of laboratory methods that can be used to study the nature of expertise. The four general types reviewed – recall, perception, categorization, and verbal reports – are domain independent, or contrived tasks. These are tasks that are not necessarily expressive of the skills of the experts because they do not precisely mimic the tasks the experts usually perform. But these tasks, used often in the laboratories or under controlled conditions (although they can be used also in cognitive field research), are suggestive of the ways that the mental representations of experts and novices can differ. The recall paradigm has revealed the differences in experts' and novices' representations in terms of chunks (coherent patterns) and organized structure; perception tasks have revealed phenomena of perceptual learning and differences in the salience of relevant features and the interrelatedness or integration of cues into meaningful patterns; and both the sorting and verbal reporting tasks have revealed differences in the depth and structure of knowledge representations.

There are of course important deeper and lingering issues that this chapter has not covered. A key issue is how exactly do the experts' knowledge representations facilitate or inhibit their performance for a specific skill. Some treatment of this issue just for the task of memory recall can be gleaned from papers by Ericsson, Delaney, Weaver, and Mahadevan (2004) and Vicente and Wang (1998). Moreover, although our interest focuses on understanding "relative expertise" (see Chi, Chapter 2), with the assumption that novices can become experts through learning and practice, in this chapter I have said little about another important issue of *how* one can translate differences in the representations of novices and experts into instruction and training (i.e., how we can train novices to become experts).

Acknowledgement

The author is grateful for the support provided by the Pittsburgh Science of Learning Center.

References

Akin, O. (1980). *Models of architectural knowledge*. London: Pion.

Alberdi, E., Becher, J. C., Gilhooly, K., Hunter, J., Logie, R., Lyon, A., McIntosh, N., & Reiss, J. (2001). Expertise and the interpretation of computerized physiological data: Implications for the design of computerized monitoring in neonatal intensive care. *International Journal of Human-Computer Studies*, 55, 191–216.

Burton, A. M., Shadbolt, N. R., Hedgecock, A. P., & Rugg, G. (1987). A formal evaluation of knowledge elicitation techniques for expert systems: Domain 1. In D. S. Moralee (Ed.), *Research and development in expert systems, Vol. 4*. (pp. 35–46). Cambridge: Cambridge University Press.

Burton, A. M., Shadbolt, N. R., Rugg, G., & Hedgecock, A. P. (1988). Knowledge elicitation techniques in classification domains. In Y. Kodratoff (Ed.), *ECAI-88: Proceedings of the 8th European Conference on Artificial Intelligence* (pp. 85–93). London: Pittman.

Burton, A. M., Shadbolt, N. R., Rugg, G., & Hedgecock, A. P. (1990). The efficacy of knowledge elicitation techniques: A comparison across domains and levels of expertise. *Journal of Knowledge Acquisition*, 2, 167–178.

Calderwood, R., Klein, G. A., & Crandall, B. W. (1988). Time pressure, skill, and move quality in chess. *American Journal of Psychology, 101*, 481–493.

Campitelli, G., & Gobet, F. (2005). The mind's eye in blindfold chess. *European Journal of Cognitive Psychology*, 17, 23–45.

Chase, W. G., & Chi, M. T. H. (1981). Cognitive skill: Implications for spatial skill in large-scale environments. In J. Harvey (Ed.), *Cognition, social behaviors, and the environment* (pp. 111–136). Hillsdale, NJ: Erlbaum.

Chase, W. G., & Simon, H. A. (1973a). Perception in chess. *Cognitive Psychology*, 4, 55–81.

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

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *Journal of the Learning Sciences*, 6, 271–315.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Hillsdale, NJ: Erlbaum.

Chi, M. T. H. (2005). Common sense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences, 14*, 161–199.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 1* (pp. 7–75). Hillsdale, NJ: Erlbaum.

Chi, M. T. H., Hutchinson, J., & Robin, A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly*, 35, 27–62.

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

Chi, M. T. H., & Ohlsson, S. (2005). Complex declarative learning. In K. J. Holyoak, & R. G. Morrison, (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 371–399). Cambridge: Cambridge University Press.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25, 471–533.

Chiesi, H. L, Spilich, G. J., & Voss, J. F. (1979). Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18*, 257–273.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies, 41*, 801–849.

Cullen, J., & Bryman, A. (1988). The knowledge acquisition bottleneck: A time for reassessment? *Expert Systems*, 5, 216–225.

De Groot, A. (1966). Perception and memory versus thought: Some old ideas and recent findings. In B. Kleinmuntz (Ed.), *Problem solving: Research, method, and theory* (pp. 19–50). New York: Wiley.

Duncker, K. (1945). On problem-solving. (L. S. Lees, Trans.). *Psychological monographs*, 58 (Whole No. 270). (Original work published, 1935.)

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory & Cognition*, 7, 149–158.

Eisenstadt, M., & Kareev, Y. (1975). Aspects of human problem solving: The use of internal representations. In D. A. Norman & D. E. Rumelhart (Eds.), *Exploration in cognition* (pp. 308–346). San Francisco: Freeman.

Ericsson, K. A., Delaney, P. F., Weaver, G., & Mahadevan, R. (2004). Uncovering the structure of a memorist's superior "basic" memory capacity. *Cognitive Psychology, 49*, 191–237.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis*. Cambridge, MA: MIT Press.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1990). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1–63.

Gellert, E. (1962). Children's conception of the structure and function of the human body. *Genetic Psychology Monographs, 65*, 193–405.

Goel, A. K., Gomez de Silva Garza, A., Grue, N., Murdock, J. W., Recker, M. M., & Govinderaj, T. (1996). Towards designing learning environments. In C. Frasson, G. Gauthier, & A. Lesgold (Ed.), *Intelligent tutoring systems: Lecture notes in computer science* (pp. 493–501). Berlin: Springer-Verlag.

Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and function. *Cognitive Science, 28*, 127–138.

Hoffman, R. R. (1987, Summer). The problem of extracting the knowledge of experts from the perspective of experimental psychology. *The AI Magazine, 8*, 53–67.

Hoffman, R. R., Coffey, J. W., Ford, K. M., & Novak, J. D. (in press). A method for eliciting, preserving, and sharing the knowledge of expert forecasters. *Weather & Forecasting*.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62*, 129–158.

Hoffman, R. R., Trafton, G., & Roebber, P. (2006). *Minding the weather: How expert forecasters think*. Cambridge, MA: MIT Press.

Johnson, K., & Eilers, A. T. (1998). Effects of knowledge and development on subordinate level categorization. *Cognitive Development, 13*, 515–545.

Klein, G., Pliske, R. M., Crandall, B., & Woods, D. (2005). Problem detection. *Cognition, Technology, and Work, 7*, 14–28.

Klein, G. A., & Hoffman, R. R. (1992). Seeing the invisible: Perceptual-cognitive aspects of expertise. In M. Rabinowitz (Ed.), *Cognitive science foundations of instruction* (pp. 203–226). Mahwah, NJ: Erlbaum.

Klein, G., Wolf, S., Militello, L., & Zsambok, C. (1995). Characteristics of skilled option generation in chess. *Organizational Behavior and Human Decision Processes, 62*, 63–69.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). Expertise in a complex skill: Diagnosing X-ray pictures. In M. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise* (pp. 311–342). Hillsdale, NJ: Erlbaum.

Mayfield, W. A., Kardash, C. M., & Kivlighan, D. M. (1999). Differences in experienced and novice counselors' knowledge structures about clients: Implications for case conceptualization. *Journal of Counseling Psychology, 46*, 504–514.

Means, M. L., & Voss, J. F. (1985). Star Wars: A developmental study of expert and novice knowledge structures. *Journal of Memory and Language, 24*, 746–757.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49–96.

Miller, G. A. (1996). *The science of words*. New York: McGraw-Hill.

Morrow, D. G., Menard, W. E., Stine-Morrow, E. A. L., Teller, T., & Bryant, D. (2001). The influence of expertise and task factors on age differences in pilot communication. *Psychology & Aging, 16*, 31–46.

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Patel, V. L., & Arocha, J. F. (2001). The nature of constraints on collaborative decision-making in health care settings. In E. Salas, & G. A. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 383–405). Mahwah, NJ: Erlbaum.

Reitman, J. S. (1976). Skilled perception in Go: Deducing memory structures from interresponse times. *Cognitive Psychology, 8*, 336–356.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Sabers, D. S., Cushing, K. S., & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal*, 28, 63–88.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & DeMaio, J. C. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies*, 23, 699–728.

Shafto, P., & Coley, J. D. (2003). Development of categorization and reasoning in natural world: Novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 641–649.

Simon, D. P., & Simon, H. A. (1978). Individual differences in solving physics problems. In R. Siegler (Ed.), *Children's thinking: What develops?* (pp. 325–348). Hillsdale, NJ: Erlbaum.

Smith, M. U., & Good, R. (1984). Problem solving and classical genetics: Successful versus unsuccessful performance. *Journal of Research in Science Teaching*, 21, 895–912.

Snowden, P. T., Davies, I. R. L., & Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: A functional role for improvements in adults' visual sensitivity? *Journal of Experimental Psychology: Human Perception and Performance*, 26, 379–390.

Tanaka, J. W. (2001). The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology: General*, 130, 534–543.

Tanaka, J. W., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23, 457–482.

Vicente, K. J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. *Memory and Cognition*, 20, 356–373.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review*, 105, 33–57.

Weiser, M., & Shertz, J. (1983). Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies*, 14, 391–396.

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, 83, 73–87.