6. Theoretical Implications from Protocol Analysis on Testing and Measurement

K. Anders Ericsson
University of Colorado at Boulder

Ericsson, K. Anders, "6. Theoretical Implications from Protocol Analysis on Testing and Measurement" (1987). *The Influence of Cognitive Psychology on Testing*. 10.
https://digitalcommons.unl.edu/buroscogpsych/10

6 Theoretical Implications from Protocol Analysis on Testing and Measurement

K. Anders Ericsson
University of Colorado at Boulder

One of the goals of psychology has always been to describe, understand, and measure individual differences. The diversity of human behavior makes it particularly challenging to seek to identify general and stable underlying elements that correspond to systematic individual differences. A major problem in the efforts to identify such elements is that the elements cannot be observed directly. The primary method has been to use the current psychological theory to develop procedures to measure such hypothetical elements. In this chapter I present a new theoretic framework, based on verbal reports from subjects, for identifying and measuring individual differences. I argue that this framework is superior to the previous ones; hence, I briefly review some of the earlier approaches to measurement of individual differences.

When scientific psychology was first established over 100 years ago, the predominant method of investigation consisted of eliciting introspective verbal reports from trained observers. During the introspective era, the research was directed toward uncovering the basic sensations and cognitive processes that provided the building blocks of the varied and complex human experiences. Within this theoretical perspective, it was assumed that observable individual differences in normal cognitive functioning were a consequence of differences in basic cognitive processes. It was furthermore assumed that individual differences in performance on simple tasks, like simple reaction time, letter cancellation, and sensory discrimination, would directly reflect individual differences in the corresponding basic processes. However, the first studies of individual differences on simple tasks showed disappointingly low correlations among tasks as well as to grades in school and other indices of ability (Guilford, 1967).

Particularly damaging for this view of simple tasks reflecting basic processes was the finding that substantial improvement in performance was observed with practice (Binet cited in Varon, 1935). Although subsequent successful attempts to measure intelligence reliably relied almost exclusively on *complex* tasks involving comprehension, the view that individual differences are due to differences in basic processes was never completely discarded. Exceptional ability (exceptional memory) was consistently interpreted as a result of differences in such basic processes.

The behavioristic era had interesting implications for measurement, in that a theory of cognitive structures was explicitly rejected. Among extreme behaviorists, all individual differences were attributed to differences in learning, or exposure to relevant experiences. Hence, measurement of basic cognitive functions would be meaningless. The measurement of individual differences in complex tasks had to be conducted in an inductive mode, where stable patterns of individual differences were discovered empirically rather than deduced theoretically. Lacking a cognitive theory, a general theory of measurement was developed and refined through the years. This theory of measurement was incorporated as an integral part of the methodology of experimental psychology. A central problem with the behavioristic approach was to understand what the observed performance on a test actually measures.

Using the computer as a metaphor, theories of human information processing were proposed in which the focus was placed on the intermediate processing stages necessary to produce observable behavior. Many of the old concepts of attention and different types of memory stores were reintroduced in these theories with more explicit definitions and characteristics. The emphasis of these models on process rather than final responses led to a concern for observations providing information about the process, like latencies, eye-fixations, and verbal reports. It became important to use converging evidence from many different types of observations to identify the ongoing cognitive processes.

For the purpose of this chapter one could divide contemporary cognitive research into the mainstream of cognitive psychology, which only uses traditional performance measures, like accuracy and latency, and other research emphasizing supplementary data on the cognitive processes, like eye-fixations and verbal reports. The aim of the first category of research has been to provide a finer grain analysis of what the psychometric test measures. Some of this research has measured individual differences on tasks assumed to provide pure measures of critical capacities according to current cognitive theory. These pure measures were then related to compound abilities like verbal IQ (Hunt, 1978). Other researchers, notably Sternberg (1977), have analyzed the latencies and errors for performing tasks similar to test-items on psychometric tests, to identify measurements of critical information processes. In both the above approaches the composite performance (reaction time and accuracy) is factored into components using theoretical assumptions, which cannot be directly tested and evaluated

within this framework. At least one of the reasons for the remarkable impact of these theoretical efforts on research on individual differences and testing is the methodological compatibility between test theory and these theories of cognition.

Another research approach within cognitive psychology has been directed toward understanding the detailed structure of cognitive processes. The aim has been to develop models of cognitive processes at a level where one can simulate the observable behavior of subjects by a computer program. The pioneering work of Newell and Simon (1972) showed that building such models required very detailed information about subjects' cognitive processes. The method used by Newell and Simon (1972) to elicit such detailed knowledge about subjects' cognitive processes was to instruct subjects to "think aloud," i.e., verbalize their thoughts, as they solved the presented problems. In a recent review of research using verbal reports, Ericsson and Simon (1984) showed that this methodology has been successfully applied to research problems in all major areas of psychology—memory, decision-making, educational psychology, instruction, and clinical psychology. Although much of that research has implications for measurement of individual differences, I know of only a limited number of studies using verbal reports of cognitive processes to directly address issues of measurement and assessment of individual differences.

AN OUTLINE OF THE CHAPTER

The goal of this chapter is to argue for the importance of verbal report data in understanding what current psychometric tests actually measure and for the usefulness of verbal report data in the design of future test instruments. The argument has three parts. First, I need to present a convincing case that particular kinds of verbal reports provide valid data and that a rigorous methodology for the analysis of verbal reports is available. Then, I present a theoretic framework that relates verbal report data to other, more traditional kinds of data, like correctness of response and latency. Finally, I show that studies using verbal reports have significantly altered our understanding of the processes measured by prevailing tests.

The chapter has three major sections that roughly correspond to the different parts of the argument. The first section provides an introduction to how verbal reports on cognitive processes can be used as valid data. This section summarizes my work with Herbert Simon (Ericsson & Simon, 1980, 1984) and describes a model of how some types of verbal reports yield reliable data on the sequence of thought in tasks. I briefly show how these forms of verbal reports differ from other disreputable forms of verbal reports, like introspection and rationalization.

The second section presents a theoretical framework for identifying and encoding sequences of cognitive processes from verbal reports. Hence, protocol

analysis provides a tool for gaining empirical data on the sequence of cognitive processes elicited in a given task for a certain subject. Such data is shown to give us an empirical method for determining what process or sequence of processes are mediating performance in a test. Theoretical assumptions of mediating processes can therefore be empirically evaluated in a more direct manner. This section also describes inductive approaches, where important cognitive processes are abstracted from the verbal protocols to give generalizable accounts of cognitive mechanisms in different domains.

In the final section, I illustrate how verbal reports have extended our understanding of individual differences. For example, within the context of tests measuring spatial ability, I demonstrate differences in strategies used by subjects of high- and low-spatial ability and how verbal reports can improve our understanding of what available psychometric tests actually measure. In another example, I show how verbal reports can give insights into the structure of practice-effects and the structure of exceptional memory.

Let me first turn to an introduction to the analysis of verbal reports on cognitive processes.

PROTOCOL ANALYSIS AND VERBAL REPORTS

The use of verbal reports on cognitive processes has a long history filled with many methodological controversies. The early pioneers of psychology used introspective reports in an attempt to describe the sensory images underlying perception and thinking. Following several contradictory findings by different research laboratories, the introspective method was seriously criticized. Many moderate psychologists (for example, Woodworth, 1938) suggested that introspective analysis (which directed attention toward underlying sensations) was misguided, and said this method should be replaced by verbal reports that expressed thoughts. A careful historic review shows that the founder of behaviorism, Watson, rejected introspection, but accepted verbalization of thinking. In fact, Watson (1920) was the first investigator to publish an analysis of the verbalized thoughts of a subject while he was "thinking aloud." Even so, the rejection of introspection by behaviorists was so complete that it generalized to any use of verbal reports.

With the emergence of information-processing models of cognition, several researchers started to consider verbal reports as a means to get more direct and detailed access to the cognitive processes of subjects. In contrast with most early introspective studies, these investigators collected extensive performance data and hence were able to evaluate the veridicality and converging validity of verbal reports. With his newly developed blank-trial technique, Levine and his associates (Frankel, Levine, & Karpf, 1970; Karpf, & Levine, 1971) showed essentially perfect correspondence between verbally reported concepts and specific

judgments about instances. In studies of memory, subjects' verbal reports on mediating associations were found to have remarkable effects on memory performance (for a review see Montague, 1972). Newell and Simon's (1972) analyses of verbal reports during problem solving was the most extensive and intensive use of such data. On the basis of verbal reports they were able to construct computer programs powerful enough to both solve problems and regenerate essential aspects of the reported thought processes. Newell and Simon (1972) instructed their subjects to verbalize their thoughts concurrently, i.e., "think aloud," whereas subjects in the memory studies often recalled their thoughts retrospectively. Other investigators using other kinds of instructions found that subjects giving verbal reports performed differently from subjects who were not required to give verbal reports—thus throwing some doubt on the validity and representativeness of verbalized thought.

The basic concern of Ericsson and Simon (1980) was to propose a model in which the cognitive processes responsible for verbalization of thoughts in attention(heeded thoughts) could be explicated. In its most general and abstract form, information processing theory (Anderson & Bower, 1973; Newell & Simon, 1972; Simon, 1979) postulates that a cognitive process can be seen as a sequence of internal states successively transformed by a series of information processes. Moreover, each of these successive states can be described in large part in terms of the small number of information structures, or chunks, attended to, or available in the limited-capacity short-term memory store (STM). Information in the vast long-term memory (LTM) and in the sensory memories (of brief duration) can be accessed, but the results of these access processes will be attended to (heeded) and available in STM. In Fig. 6.1 I have illustrated a sequence of successive states, showing how new thoughts are expressed verbally as they enter attention, and hence become observable as verbalization segments.

The general relation between heeded thoughts, i.e., thoughts in attention, and the observable verbalizations is much easier to understand in the context of specific examples. In Table 6.1, the thinking-aloud protocol of a subject mentally multiplying 36 times 24 is given. Most of the verbalized information consists of generated intermediate steps, like "4," "carry the 2," "144." There is no difference in principle between these intermediate steps and the final result, "864." Even when one asks students to answer questions like, "What is the number of windows in your parents house?," their thinking-aloud protocols are remarkably similar. A representative example of such a thinking-aloud protocol

FIG. 6.1.   A thought process represented as a sequence of states of heeded information. Each state is associated with verbalization of new information entering attention.
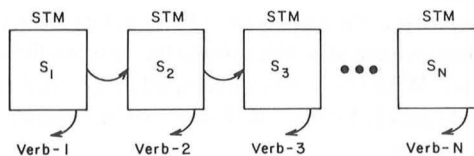
TABLE 6.1
A Transcript of a Thinking-Aloud Protocol From a Subject
Mentally Multiplying 36 Times 24

OK	36*
36 times 24	24
um	144
4 times 6 is 24	720
4	864
carry the 2	
4 times 3 is 12	
14	
144	
0	
2 times 6 is 12	
2	
carry the 1	
2 times 3 is 6	
7	
720	
720	
144 plus 72	
so it would be 6	
6	
864	

* On the right side, the same multiplication is performed
using the traditional paper and pencil method.

is given in Table 6.2. Notice that the subject verbally expresses intermediate steps (heeded thoughts) rather than explaining or describing her thought processes.

From this model of concurrent verbalization it is clear that the subject has to have time to complete the verbalization of the heeded information before new thoughts enter attention. For tasks where subjects have extensive experience, the

TABLE 6.2
A Transcript of a Thinking-Aloud Protocol From a Subject
Recalling the Number of Windows in Her Parent's House

Let's see, there's 3 windows in the living room,
3 windows in my room, 1 in the bathroom, 2 in the
sewing room, that's 5 and 3, 6, 7, 8, 13--4 in the
kitchen which would make 17; 3, 4 in the TV room
which would make 21, 2 in my brother's room, 23; and
1 in the upstairs bathroom, 24; and 3 in my parent's
room, 27; and then 1 in the attic, 28

Adapted from Ericsson and Simon (1984).

sequence of thoughts is so closely connected that a concurrent sequential verbalization of the spontaneously occurring thoughts is not possible. In such situations, the subjects can report their thoughts in retrospect by recalling the sequence of thoughts just after the completion of the task. When the time taken to complete the task is relatively short (about 5–30 seconds) our model predicts a rather complete retrospective report of all heeded thoughts. For tasks with longer duration, concurrent reports (thinking-aloud) will be more detailed than the corresponding retrospective reports.

One would not expect either retrospective reports or thinking-aloud protocols to change the cognitive processes under study. If the essence of the cognitive process is the sequence of heeded information, then thinking-aloud doesn't change that sequence. A large number of studies have compared subjects thinking aloud with silent subjects doing the same task (for a review see Ericsson & Simon, 1984). None of these studies has shown evidence for changes in structure of the process due to thinking aloud, as measured by ability to solve problems, type of solution, eye-movement pattern, etc. Several studies have shown that subjects thinking aloud take more time than silent control subjects. This follows from our model, because verbally expressing a thought takes additional time.

A recent analysis (Deffner & Ericsson, 1985) of the temporal structure of subjects thinking aloud showed that they verbalize their thoughts rapidly in speech bursts (at 100–150 words per minute), while most time is spent in silence. If the time spent actively verbalizing is measured and then subtracted, the mean solution time is no different for silent and think-aloud subjects. Hence, it appears that the effect on solution time can be accurately predicted by assuming a local slowing-down of cognitive processes during verbalization.

Ericsson and Simon's (1980, 1984) analysis of studies that do show effects of concurrent verbalizing demonstrates that these studies used quite different instructions to subjects. Typically, subjects are required to verbalize motives or reasons for their actions and thoughts. From subject's thinking-aloud protocols on the same or similar tasks we know that only a subset of the generated thoughts are based on deductions or retrievals with explicit premises verbalized. Forcing a subject to provide reasons for all reported thoughts would therefore clearly change the subject's thought processes. This means, of course, that the sequence of heeded thoughts is changed, which in turn influences performance and the structure of the solution process. For example, many students are accustomed to the situation of solving a mathematics problem at the blackboard in front of class. Some subjects confuse the instruction to think aloud with such a systematic generation of explanations, and investigators of mathematical problem solving explicitly tell subjects: "Do not try to explain anything to anyone else. Pretend there is no one here but yourself. Do not tell about the solution but solve it" (Krutetskii, 1976, p. 93). It is useful to give subjects "warm-up" tasks, where thinking aloud is particularly easy. Examples of such tasks are mental multiplication and anagram problem solving.

In a well-known paper criticizing the validity of verbal reports, Nisbett and Wilson (1977) showed that in many studies of social psychology, subjects report incorrect reasons in response to why-questions. For example, a subject selecting among a set of displayed stockings will argue, if asked, that the selected stocking is better in terms of some of its physical attributes. Such reasons are given by subjects even when the displayed stockings are identical, although they are not informed of that. Nisbett and Wilson (1977) argued that in responding to the why-question, subjects do not try to remember their thoughts while the associated behavior was generated, but theorize and try to infer reasons for their behavior. Our model of verbal report is consistent with Nisbett and Wilson's argument as long as the subjects generate the incorrect reasons without recalling their corresponding thoughts during the task.

In some situations, the why-question is asked after such a delay following the corresponding behavior that subjects cannot recall their thoughts or are not willing to spend the effort required for successful retrieval. In other situations, the behavior is elicited without mediating thoughts and hence there are no thoughts to be retrieved and used in answering the why-question. For example, when normal subjects generate a word starting with "a," a high proportion simply report that "apple" emerged. When you ask such subjects why they generated "apple" rather than any of the other words starting with "a," they may not be unwilling to speculate. Often they suggest that perhaps they learned the association between "a" and "apple" while learning the alphabet. Regardless of the truth of these subjects' hypothesis, I can agree with Nisbett and Wilson (1977) that the validity of these subjects' speculations about their own behavior would not be any greater than that of subjects speculating about the reasons for other people's selection behavior.

Our model of verbal report also provides considerable guidance for how verbal reports should be encoded and what inferences can legitimately be made. During the era of introspection, experienced and respected observers made observations on their own thought processes. These observations were assumed to represent facts—a subject reporting X would imply that X was true. Within our framework, we would argue that the fact is that *the subject reports X*. The rather uncontroversial inference we want to make is that the subject attends to X.

Let us clarify this by returning to the protocol on mental multiplication. A traditional psychologist might only accept the validity of the verbalization of the final answer. From the verbalization of the final answer we infer that the final answer was generated and heeded. In an analogous way we infer from the sequence of verbalized intermediate products a corresponding sequence of heeded information. The verbally reported thoughts are data, and a model is needed to account for how relevant thoughts are generated—hence a full model would regenerate the heeded information. In many cases, one will find that a simulation model able to regenerate the verbally reported intermediate steps will be powerful enough to generate the final solutions to the presented problems.

It may appear that I am unduly cautious in accepting inferences drawn from verbal reports. However, much of the poor reputation of verbal reports comes from the debatable validity of psychodynamic analysis of dreams and fantasy. Furthermore, all too often general statements like, "I always do X," are interpreted to be unconditionally true, and when inconsistent performance data are obtained, the inference is made that all verbally reported information is questionable. Herbert Simon and I interpret such verbalizations to simply indicate that the subject at that time believes (correctly or incorrectly) that he always does X.

Traditionally, subjects have been interviewed at the end of experiments and test-taking sessions about their strategies and thought processes *during* the experiment. At the end of the experiment subjects have poor memory for their actual sequences of thoughts leading to specific solutions. Furthermore, investigators often encourage subjects to describe a general strategy that encourages them to make inferences and speculate rather than attempt to recall specific memories for actual solutions. It is not surprising that strategies assessed through such postexperimental probing provide a poor fit to the subjects' performance during all phases of the experiment.

For most tasks it is easy to determine what constitutes a thought. In Table 6.3 I have reproduced a protocol from a subject solving an anagram problem, where

TABLE 6.3
A Transcript of a Thinking-Aloud Protocol From a
Subject Solving the Anagram 'NPEHPA' Recorded by Sargent (1940)

N-P, neph, neph	
Probably PH goes together	C:PH*
Phan	A:phan
Phanny	A:phanny
I get phen-ep	A:phan-ep
no. Nap-	A:nap
Phep-an, no	A:phep-an
E is at the end	C:E(end)
Phap-en	A:phap-en
People, I think of	A:people
Try PH after the other letters	C:PH(end)
Naph, no	A:naph
I thought of paper again	A:paper
E and A sound alike	
couldn't go together without a consonant	
Try double P	C:PP
happy	A:happy
Happen	A:happen

*On the right side encodings of the verbalized thoughts are given. Adapted by Ericsson and Simon (1985).

the object is to rearrange the letters to form a single English word. On the right-hand side of Table 6.3 I have given corresponding encodings of the verbalized thoughts. There are two types of task-relevant thoughts. First, the subject selects likely letter combinations and decides where in the solution word they are likely to occur. I denote these constraints or cues as C:### (position). Second, the subject generates alternative possible solution words (denoted by A:###). These encodings can then be used as data for further model-building and hypothesis-testing.

By necessity, this description of the model for verbal report generation and protocol analysis, developed by Simon and myself, is brief. The interested reader should consult the more extensive discussion of these issues in our recent book (Ericsson & Simon, 1984). In spite of its brevity, I hope I have conveyed to you that protocol analysis stands on sound methodological ground and that findings from analyses of verbal protocols can be accepted as facts in our attempts to understand the human mind.

IMPLICATIONS OF VERBAL REPORTS FOR MEASUREMENT AND THEORETICAL ABSTRACTIONS

Verbal reports on cognitive processes provide a much more detailed description of the cognitive processes in a task than the traditional forms of data, i.e., response accuracy and latency. The stuation is structurally similar to the differences between observations made by the naked eye and the same observations made with a microscope or a telescope. Objects appearing to be similar or even identical to the naked eye are demonstrated to either remain identical or to appear very different with the availability of more information about their detailed structure and components. There are two rather different approaches to systematizing the newly acquired, detailed information. The first method is to focus on objects assumed to be similar or identical, and examine their detailed properties to validate or refute the assumption of similarity. This method examines theoretical assumptions in essentially a hypothesis-testing mode. The other method is primarily inductive and considers the detailed information directly. From the detailed information, critical entities are identified and attempts to form meaningful abstractions are made. In this section, I examine how these two methods can be and have been used to relate verbal reports on cognitive processes to compound measures, such as reaction time and response accuracy. I start by examining some theoretical assumptions about the similarity of cognitive processes elicited by a given task or collection of test items.

Traditionally, investigators select test items such that some or all elicit the same process or sequence of processes for all tested subjects. This selection is based on intuition or some considerations based on informal or formal theories. In Fig. 6.2 I have illustrated the data recorded for three individuals on two test
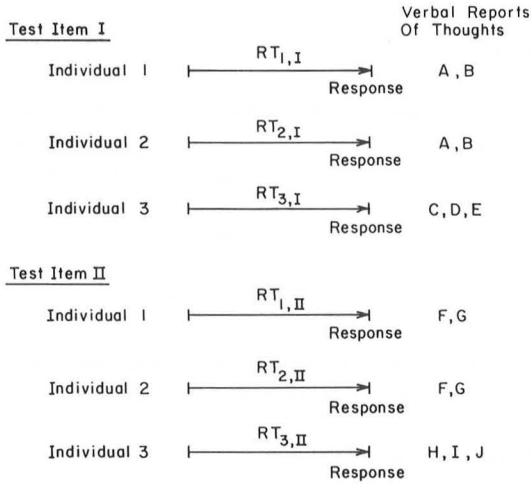
FIG. 6.2.   Traditional data (latency and response) and verbal report data from three subjects' solutions to two test-items.

items. For each subject and test item, both the correctness of the response and the reaction time to respond are recorded.

If the theory used for item selection is correct, then we would be entitled to aggregate the data over test items to attain a more accurate measure of accuracy and latency for the measured process. There are only limited techniques for testing the assumption that all test-items evoke the same process or sequence of processes. Only the lack of positive correlation between different subjects' performance on two items would provide evidence against the assumption. Even small positive correlations would be consistent with the theoretical claim.

The situation is quite different when verbal reports on the cognitive processes are available. According to our earlier-presented model, I assume that for every heeded thought there is a process responsible for its generation. Hence, when I talk about a sequence of heeded thoughts, the corresponding sequence of processes is implicit. In the lefthand panel of Fig. 6.2 I have abstractly represented sequences of verbally reported thoughts for each test item. For solutions to the same test item one can compare the sequence of heeded thoughts directly. Such a comparison for the two test items indicates that two of the subjects relied on the same sequence of thoughts, whereas the third subject relied on a different sequence. The fact that all the subjects' thought sequences differ across test items is to be expected as the content of the two test items are different.

By introducing the theoretical idea of processes one can argue that a different sequence of thoughts are the reflection of the same sequence of processes. It is necessary that the processes are explicitly defined prior to the empirical analysis. In Fig. 6.3 I have illustrated a number of processes, which would characterize
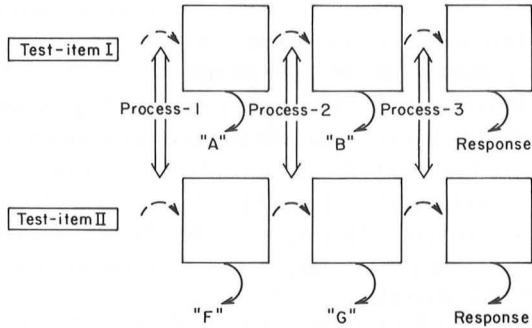
FIG. 6.3. Illustration of how two different sequences of verbally reported thoughts can be described as the realization of the same sequence of processes.

the relation between generated thoughts and different content in test items (input). Through the assumption of processes, I can now argue that the aggregation of test items provides a legitimate estimate of the speed of the component processes. However, for only two of the subjects the same sequence of processes are measured. It is important to note that processes are theoretical entities, which may or may not correspond to some unitary psychological process.

Out of the large number of possible relations between sequences of reported thoughts to two test items, we can identify two cases where one can legitimately argue that the protocol information is consistent with the claim that the same process or sequence of processes is measured by the two items.

The first case is the extreme case, where no mediating thoughts are verbalized for either of the two test items. Such a lack of mediating thoughts would be expected for highly automatic reactions, like naming familiar objects, reading, etc. It is commonly assumed that rapid reactions (faster than 2 seconds) assure no mediating states. However, I later present evidence showing that such a view is not correct.

The second case is the most interesting, where the sequences of reported thoughts for the two test items can be seen as the realization of the same sequence of processes. For example, a subject performing a mental addition of two 2-digit numbers can follow the same process sequence even though the specific numbers are different for the two test items. By assuming the existence of a general addition process for any two digits, one can see those two different mental additions as two realizations of the same general process sequence. Even in this uncontroversial example one can question the theoretical status of the general addition process. There is evidence showing that the simple addition of two digits takes different amounts of time depending on the digits involved (Miller, Perlmutter, & Keating, 1984). For adult subjects the differences are small enough to make the abstraction of general adding processes completely acceptable.

It is unlikely to find classes of test items for which the second case is absolutely true. It is reasonably likely that situations will be found where equivalence of the cognitive processes on the different test items is a good approximation. Through the collection of verbal reports on the cognitive processes on test items, it si possible to identify blatant violations of the assumptions of measurements of the same general process sequence, by identifying systematically different strategies among the tested subjects. Before I turn to a discussion of how different investigators have analyzed verbal protocols to abstract general processes, we briefly consider an example of analysis of verbal reports for a task with fast latencies (less than 2 seconds).

Ericsson and Simon (1984) reviewed the relatively extensive evidence showing that subjects' retrospective verbal reports provide reliable information to predict the latencies for a variety of task domains. The validity of retrospective verbal reports extended to tasks with average latencies of less than 2 seconds. Systematic attempts to derive a processing model to predict the observed reaction times on the basis of retrospective reports are much more rare. Two English investigators, Hamilton and Sanford (1978), studied subjects who made simple judgments of whether two presented letters, like "RP" or "MO," were in alphabetical order or not. In accord with previous investigators, they found that the reaction times were longer when the two presented letters occurred close together in the alphabet as opposed to when they were far apart. From the reaction-time data alone, one would infer a uniform retrieval process, where factors internal to the retrieval process required more time for order decisions for letters occurring close together in the alphabet. Retrospective verbal reports for subjects doing individual decisions indicated two types of cognitive processes. For some of the trials, subjects reported no mediation or direct access of their order judgment. For the other trials, subjects reported they ran through brief segments of the alphabet before making a decision of order. For example, when the letter-pair "MO" was presented, a subject reported retrieving "LMNO" before the subject reached the decision that the letters were in alphabetical order. In another case a subject reported retrieving "RSTUV" before rejecting the letter-pair "RP" as not being in alphabetical order. In a subsequent analysis of the reaction times, Hamilton and Sanford (1978) found very different relations with the separation of the two letters for trials with direct access, versus trials with retrieval of segments of the alphabet. For trials with retrieval, the observed reaction time was a linear function of the number of retrieved letters. The estimated rate of retrieval corresponded closely to rates obtained in studying simple recital of the alphabet. For trials with reports of direct access, no relation of reaction time to the amount of separation of the two letters was found. Hamilton and Sanford (1978) concluded that the original effect was due to a mixture of two quite different processes, and that closeness of the letters influenced the probability that recall of letters would be necessary before an order decision could be made.

Hence, even in simple tasks with rapid responses, one can see variability in cognitive processes or reported thought sequences leading to differences in observed reaction times. As the complexity of the task increases, the range of possible thought sequences giving the correct response increases dramatically. With practice on a task, the availability of short-cuts and emergence of different and more efficient representations and corresponding strategies makes the space of possible thought sequences mediating correct solutions intimidatingly large. In a later section I more directly address the issues of assessing the availability of strategies and representations for subjects. The conclusion I draw at this point is that protocol analysis provides a sensitive measure to help us define equivalent classes of processes for which proper measurement of the speed of component processes is valid. Consistent individual differences in mean reaction time cannot and should not be interpreted as evidence for stable characteristics of basic processes. For many types of test items, considerable diversity in frequency of use of short-cuts and strategies is possible.

On the detailed level of description provided by verbal reports, the variability between individuals appears so large that any attempt to search for general theories of cognitive activities might appear futile.

Before turning to the final section with applications of protocol analysis to tests and measurements, I briefly review research from three areas of general psychology where protocol analysis has been related to such general models. The three areas are problem solving, decision making, and memory. In each of these areas, I show how detailed descriptions of processes can be reconciled with abstract and general, sometimes mathematical, descriptions of processes.

It is appropriate to start with a discussion of problem solving, because it was the analyses of problem solving by Newell and Simon (1972) that led them to produce the first computer simulations of cognitive processes. In their pioneering work of subjects proving theorems in propositional logic they collected thinking-aloud protocols from subjects solving such tasks. The verbalized thoughts were identified as being results of induced general information processes, which could be explicated as routines in a computer program. Newell and Simon (1972) also induced a general organization of problem solving, which they called means-ends analysis. They found that a simulation model of human problem solving was sufficiently powerful to produce the solution, and at the same time the mediating steps of the program corresponded closely to the verbally reported thoughts of subjects. The correspondence of subjects' verbal reports and the theory or simulation model was on the level of types of intermediate steps rather than *exact* order of intermediate steps leading to the solution.

Subsequent evidence for means-ends analysis has been demonstrated for a wide range of problems (Ericsson & Simon, 1984). For example, I can illustrate the similarity of verbalized thought across different subjects for the 8-puzzle. In the 8-puzzle, subjects are presented with a 3x3 matrix of numbered tiles as shown in Fig. 6.4. By sliding one of the directly adjacent tiles into the empty
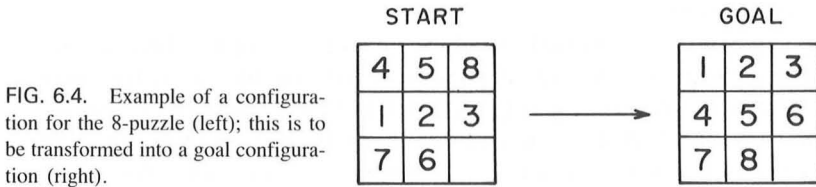
START

GOAL

FIG. 6.4. Example of a configuration for the 8-puzzle (left); this is to be transformed into a goal configuration (right).

4	5	8
1	2	3
7	6	

⟶

1	2	3
4	5	6
7	8	

space, the arrangement of tiles can be changed. Subjects are instructed to move tiles until they attain the goal configuration given on the right in Fig. 6.4.

According to means-ends analysis, subjects should solve this problem by finding differences between the goal configuration and the current arrangement of tiles. From an analysis of the task it is possible to a priori predict the space of possible thoughts (problem space) that subjects will generate in response to a problem like the 8-puzzle. The first difference they encounter is that the tile with number 1 is not in its correct location. In Table 6.4 I have illustrated a small sample of the times subjects verbalized their intention to put Tile 1 in its correct place.

The verbalizations in Table 6.4 differ in exact wording but the thought is the same. Once they placed Tile 1 they would proceed to place Tile 2, etc. A more complete account of subjects' problem solving in the 8-puzzle is given by Ericsson (1975).

Means-ends analysis appears to provide a general account of subjects' behavior on problems with which they are naive or unfamiliar. With expertise and considerable experience, the structure of the problem solving is quite different and becomes a function of the subject's extensive knowledge of the task domain (Chi, Feltovich, & Glaser, 1981; Larkin, McDermott, Simon, & Simon, 1980).

TABLE 6.4
Examples of Verbalizations to Attain Correct Placement of Tile 1

```
I'm going to try          to get 1
I must                        get 1 up there
I shall try               to get 1    here
                  first       get 1    there instead of 4
      I shall             have 1 up
      I want to      have the 1 up there
                         move 1 up where it should be
thinking of moving            1 up at once
that I shall              get 1    here
                  to get 1 up (and get 2 there, first 1 up)
in any case               get 1    in place first and foremost
now I want to        have 1 up right from the beginning
try to get them in order
to start with                 1 upmost to the left and get it in
```

The demonstration of general problem-solving methods has received considerable attention from educators, who have explored the possibility of teaching students such methods. In the final section I discuss this attempt to describe individual differences among subjects in terms of the availability of such general methods to subjects. Training subjects to use means-ends analysis appears to be somewhat misguided, as virtually all subjects exhibit such a method spontaneously in unfamiliar tasks.

Another domain with consistent patterns of cognitive processes is decision making. In the paradigmatic decision-making situation, a subject is presented with a set of alternatives. Each alternative is characterized by different attributes on several common dimensions. The prevailing model of how decisions are made is that all attributes are combined using a mathematical weighting function to form a single evaluation score. Deciding which is the "best" alternative in the set would then correspond to selecting the one with the highest evaluation score. Few, if any, investigators have argued that such a mathematical formula mirrors the cognitive processes of human subjects making decisions.

Verbal protocols of subjects making decisions have shown cognitive processes quite different from a sequential full evaluation of each alternative (Payne, 1976; Svenson, 1979). Instead, subjects begin by rejecting alternatives because they have unacceptable values on important dimensions. When only a couple of viable alternatives remain, subjects switch to a more intensive analysis, where differences on some dimensions are traded off or compared to differences on other dimensions. Other data, recording what information subjects attend to, have provided converging support for the existence of these different processes. Analogous to the previously discussed work on problem solving, general processes sufficiently powerful to account for the observed behavior have been identified.

Research on how subjects evaluate alternatives (judgment) has found that verbally reported categorical decisions can describe a series of judgments equally as well as an empirically fitted linear regression model (Einhorn, Kleinmuntz, & Kleinmuntz, 1979). In one of their experiments, Einhorn et al. (1979) observed a subject thinking aloud while judging many cereals on a five-category scale. From the thinking-aloud protocols they identified a number of rules used by the subject. These rules predicted the subject's categorizations of a new set of cereals remarkably well, in fact as well as a regression model identified for the first set of judgments. Einhorn et al. (1979) established the correspondence between these different types of models by showing how a linear regression model can closely approximate categorical rules as reflected in a verbal report. This last result is particularly important as it demonstrates that prevailing mathematical models can be reconciled with the more detailed evidence from verbal protocols.

The research in both problem solving and decision making has shown the types of cognitive processes revealed through protocol analysis provide a sufficient and general account of subjects' performance. The consistency across

subjects is intriguing, and many investigators have argued that general information processing constraints lead subjects toward adopting such processes and strategies. At least, these analyses show that the adopted processes are compatible with the well-known limits of attention and short-term memory.

Studies of memory and retention have always been one of the central areas in general psychology. Ever since Ebbinghaus (1964, 1885) invented the nonsense syllable, there has been explicit concern to study pure memory, that is, memory and retention uncontaminated by previous knowledge. During the behaviorist era, few investigators challenged the assumption that no intermediate processes were involved during memorization. In the 50s, it was demonstrated that nonsense syllables were differentially difficult to memorize and that this difficulty could be independently predicted from the meaningfulness of the nonsense syllable (Noble, 1952). In the 60s and early 70s, investigators asked subjects to verbally report their thoughts during study. These investigators found a remarkable diversity of different mediating thoughts reported by different subjects. I have extracted some examples of mediating thoughts in Table 6.5 from studies of Martin, Boersma, and Cox (1965), and Prytulak (1971).

The central issue concerned whether different reported mediators during study of items were related to subsequent recall performance on the corresponding items. Several different encoding schemes were developed to use explicit criteria for the goodness of the generated mediating responses, like those in Table 6.5. Although the biggest difference appeared between some mediating response versus no mediating response (rote rehearsal) these encoding schemes were also able to capture differences between types of mediating responses. This extensive research is fully reviewed by Montague (1972). Subsequent research in which subjects formed meaningful associations via visual images or constructing sentences have demonstrated very large effects compared to uninstructed subjects'

TABLE 6.5
Examples of Mediating Thoughts in Memorizing Individual
Nonsense Syllables and Paired Associates

Individual CVC[a]	Reported Mediator[a]
CAZ	case
CIB	sibling
BUH	bunch
JEK	jerk

Paired Associate[b]	Verbal Report[b]
Sagrole - Polef	Each word contains an OLE. Sagrole begins with S and Polef with F, thought of State Police
Rennet - Quipson	Changed Rennet to Bennet and saw Quips in Quipson--thought; Bennet Cerf Quips on TV

[a]From Prytulak, 1971.
[b]From Martin, Boersma, and Cox, 1965.

performance (Bellezza, 1981; Bower, 1972). From an individual difference perspective, it is interesting that some subjects report using such effective means for memorization without instruction (Bower, 1972).

Detailed descriptions of the associations making up the memory trace are by no means inconsistent with current mathematical theories of memory. These theories represent memory traces as associations of different strengths. Verbal reports allow us to assess the micro-structure of these associative bonds.

Summary

Verbal reports on cognitive processes in a task provide a series of intermediate steps (heeded thoughts), which are generated by corresponding cognitive processes. Hence, verbal report data can be used to confirm that subjects' responses to test items are generated by the same sequence of component processes. In the case that verbal reports show different sequences of processes for a set of test items as evidenced by short-cuts or different strategies, average reaction times and accuracies for items in a test will *not* measure differences in stable characteristics of assumed underlying processes, and thus these average results of the test reflect a composite of factors and cannot be interpreted as a pure measure of anything.

For the domains of problem solving, decision making, and memory, systematic analysis of verbal reports allows for the abstraction of postulated cognitive processes. These cognitive processes, like forming meaningful associations or interactive visual images (memory), or means-ends analysis (problem solving), were generally found for all individuals in the corresponding task domain and appeared to account for previous findings based on traditional performance data. These and other demonstrations (Ericsson & Simon, 1984) that generalizable aspects of cognitive processes can be induced from analyses of verbal reports give considerable confidence that similar analyses of cognitive processes elicited by tests will be successful.

Protocol Analysis in Assessment and Measurement

The purpose of this final section is to select a small number of important measurement issues and illustrate how protocol analysis has been applied to further our understanding. The first issue concerns how one can identify actual and valid cognitive processes. The fact that it is possible to verbally describe a hypothetical cognitive process does not assure its empirical validity. After a brief historical review of earlier attempts to identify processes and representations of general applicability, I concentrate on more recent efforts to specify such general processes in the analysis of mathematical ability.

The second issue regards the importance of differences in strategies for performance on psychometric tests. I focus on some recent research on tests measur-

ing spatial ability. In the introduction I mentioned that one of the few individual differences consistently explained by differences in basic processes concerns exceptional abilities, especially exceptional memory (Wechsler, 1952). Drawing on my collaborative research with Bill Chase, I will examine whether such memory processes are basic and direct, as evidenced by a lack of mediating states in the verbal reports.

Use of Verbal Reports to Assess Individual Differences

Some of the earliest studies using verbal reports identified general differences between subjects' reported cognitive processes and representations. The importance of differences in cognitive processes was shown by Heidbreder (1924), who found consistent differences in concept formation between subjects actively generating and testing hypotheses and subjects more passively waiting until hypotheses occurred to them. The importance of differences in representation were demonstrated by several independent studies of human maze learning, which found striking differences in learning rate as a function of the mode of encoding (motor, spatial, or verbal) reported by subjects when they had to memorize solution paths. (For a review see Ericsson & Simon, 1984.)

Since the publication of Bloom and Broder's influential study of problem solving in 1950, research on individual differences using the verbal report methodology primarily has focused on identifying general and task-independent processes and strategies. Although the results of this research on general processes have been rather disappointing, it is worthwhile to describe some of these studies and discuss reasons for the lack of success of such approaches. Later I discuss other research focusing on more task-specific processes and knowledge.

Bloom and Broder (1950) were interested in processes of thought and reasoning rather than simple fact retrieval, as emphasized in many educational tests. By selecting test items requiring reasoning, they found intriguing differences between think-aloud protocols of subjects with high and low aptitude scores. Low aptitude subjects tended not to be able to represent the problem in such a way that their relevant knowledge could be retrieved or used for inferences in generating solutions. The weakness of low aptitude subjects was taken as a focus for a remedial program for training low aptitude subjects. The training program was successful, and Bloom and Broder (1950) attribute its success to training in general cognitive processes. However, the lack of methodological controls in their study makes their results only suggestive.

In the domain of mathematics, similar ideas have been explored with explicit concern for methodological issues. Many have the belief that mathematical ability is something more general than a composite of specific abilities to solve types of mathematical problems. Polya (1957) is one of the few theorists who has explicitly proposed general methods (heuristic questions) in mathematical problem solving. Examples of such heuristic questions are ''What is the unknown?'',

"Will a figure help?", "Have I solved a related problem before?", "Can I see that it is correct?" (Polya, 1957).

In his pioneering dissertation, Kilpatrick (1968) took these questions and attempted to describe cognitive acitivity that would provide evidence for the existence of such general problem-solving heuristics. After considerable exploratory work, he identified a revised set of heuristics relevant to the mathematical problems solved by 8th-grade subjects. From the thinking-aloud protocol of each subject, Kilpatrick (1968) would determine if evidence for the application of any one of the heuristics was available.

Kilpatrick's attempt to predict mathematical problem-solving performance (time, percent age correct) from the frequency with which heuristics were used failed. Ericsson and Simon (1984) have summarized similar negative results of several other studies using encoding schemes based on Polya's work (1957).

In examining the failure to identify heuristics, it is important to realize that the hypothesized processes were not induced or abstracted from the protocols, but derived theoretically. Even more important is the fact that these heuristics were not (and possibly could never be) explicated in such detail that one would know when and exactly how to apply them. It is implicitly assumed in Kilpatrick's (1968) aggregation procedure that application of any one of the heuristics will always be helpful in solving any problem. A subsequent study Gimmestad (1977) showed that application of various heuristics was differentially useful for
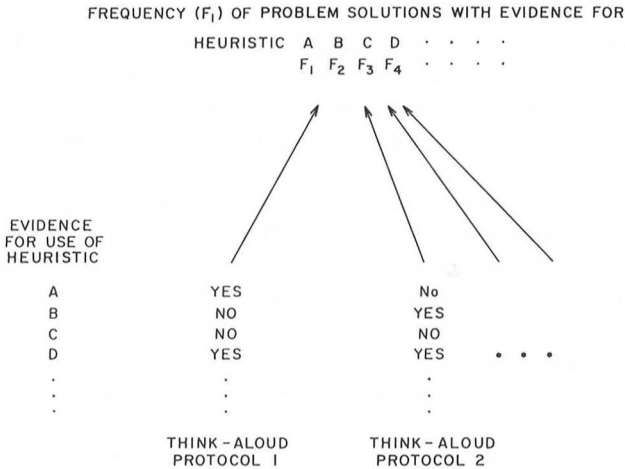

FIG. 6.5. The aggregation of information about judged use of specified heuristics for a given subject. Each thinking-aloud protocol is first scored with a dichotomous decision regarding use of a given heuristic. An aggregate measure is obtained for each subject by counting the number of problems where a given heuristic was used.

solving different problems. In fact, application of some heuristics was found to be *negatively* related with success on some problems.

The best evidence against the implicit generality of these heuristics comes from studies of training subjects in applying these heuristics. There appears to be little or no transfer of heuristics to problems different from practice problems (Lucas, 1972). However, some transfer in use of heuristics has been observed for problems similar (but not identical) to the problems used in training (Schoenfeld, 1979). It appears safe to conclude that application of general heuristics requires knowledge of when and how to apply them. This knowledge is necessarily relatively specific to types of problems.

Studies assessing the use of heuristics have provided important additional data on factors determining performance on mathematics tests. Webb (1975) found that basic tests of mathematical achievement accounted for 40% of the variance on mathematics tests, which was considerably more than any predictor related to the use of heuristics.

In their classic work on problem solving, Newell and Simon (1972) argued for the importance of knowledge on effective problem solving, and for the specificity of problem-solving methods. Lesgold (1984) reviewed evidence from a wide range of domains and demonstrates the importance of specific knowledge in the acquisition of skill for each domain.

In parallel with the studies relying on Polya's heuristics, other researchers have studied mathematical problem solving with much more emphasis on knowledge and domain-specific methods. Hinsley, Hayes, and Simon (1977), for example, showed that subjects would reliably sort algebra word problems in categories or problem types (e.g., mixture problems, distance/rate/time problems). From an analysis of thinking-aloud protocols they found that subjects appeared to categorize a given problem early during the solution of that problem and use knowledge about that type of problem to aid in the solution process. Subjects' ability to sort mathematical problems into types with the same mathematical structure was shown by Silver (1979) to be predictive of subjects' performance on a related mathematics test, even after IQ scores and scores on tests of mathematical concepts and computation were controlled for. Similarly Kennedy, Eliot, and Krulee (1970) analyzed students' thinking-aloud protocols while solving algebra problems in content-defined steps, which were determined separately for each problem. Their major result was that students of lower ability were less able to generate the necessary physical inferences from the information in the problem statement, rather than having any basic deficits in knowledge about algebra and mathematics.

The most successful attempts to identify individual differences come from rather complete analyses of very simple and specific tasks. Children in school are taught explicit procedures to solve different types of problems in arithmetic. By matching the target procedure against the observed sequence of processing steps

it has been possible to identify school childrens' systematic errors and misconceptions. In some early work, Buswell and John (1928) identified around 150 types of errors from students solving arithmetic problems aloud.

The importance of verbal reports for assessing many types of errors becomes clear from the three types of errors in division shown in Table 6.6.

In more recent work, several investigators (Brown & Burton, 1978; Brown & VanLehn, 1980; Young & O'Shea, 1981) have developed simulation models that can account for and describe errors in the subdomain of subtraction problems, with reference to general rules for carrying out the subtraction procedure. These attempts do not rely on verbal reports, but induce the type of error from consistent patterns of incorrect results on several problems. This means that diagnosis of errors can be conducted automatically through a computer program, which also can serve as a tutor by explaining to the student the nature of his or her specific types of errors.

This brief review of studies on individual differences in mathematical ability shows essentially no evidence for the mediation of very generalizable cognitive processes. The protocol data suggest the importance of cognitive processes related to problem types as well as specific procedures and knowledge. However, protocol analysis can only provide a partial answer to the question of how general or specific the cognitive processes are that generated the thoughts given in the verbal reports. It can provide a lower bound for the generality, in that when subjects verbalize recognition of specific types of problems, like ''distance-time-rate'' problems in mathematics or ''conservation of energy'' problems in mechanics, the inferred processes need to be equally general. The inferences about the generality of processes generating intermediate steps/thoughts is an empirical issue that can only be clarified by observing subjects' solutions to a specified

TABLE 6.6
Three Examples of Verbal Reports From Students Thinking Aloud While
Dividing two Numbers (Shown to the Left)

Used Remainder Without New Dividend Figure

	306	Another pupil said, "16 into 57 goes 3 times [multiplied
16	576	and subtracted]; 16 into 9 won't go [wrote 0 in the quotient];
	48	16 into 96 goes 6 times."
	$\overline{96}$	
	$\underline{96}$	

Added Remainder to Quotient

	442	The pupil said, "2 into 9 is 4 times and 1 over; 2 into 6,
2	964	3 times and 1 is 4; 2 into 4, 2 times."

Began Dividing at Units' Digit of Dividend

	26	One boy said, "7 into 42, 6; 7 into 15, 2 and 1 over."
7	31542	He was puzzled because 7 would not go into 3 and 26
		did not look right but could think of no other method.

Each verbal report illustrates a common type of error.
From Buswell and John, 1928, pp. 184, 186.

class of problems and tasks. It is clear that verbal reports will be indispensable data in these empirical tests of generality.

Assessing Strategies in Tests of Spatial Ability

In a recent article in *Psychological Review,* Just and Carpenter (1985) present a very interesting analysis of cognitive processes involved in the performance measured by a psychometric test of spatial ability. Examples of a couple of test items from a cube comparison test are illustrated in Fig. 6.6.

The task is to decide if the two drawings *could* or *could not* be views of the same cube. The general psychological process generally assumed to account for subjects' ability to make correct judgments is called mental rotation. Just and Carpenter (1985) went further, defining several types of possible strategies for solving this task and developing complete simulation models in the form of computer programs. For my intended discussion of the verbal reports on cognitive processes in this task, a brief description of three of these strategies is sufficient.

The first strategy corresponds to mental rotation of the cube along the standard axis of the cube. In order to rotate the cube at the left to overlap with the corresponding cube on the right, one might first rotate the E towards the top and then turn the cube so the E will match in orientation (see Fig. 6.7-I). A second, and in many cases more efficient, strategy would be to select a nonstandard rotation axis as illustrated in Fig. 6.7-II. With such a selection of a rotation axis a *single* rotation is sufficient.

With the third strategy, orientation-free descriptions, subjects encode the information for the presented cube on the left as two symbolic descriptions where
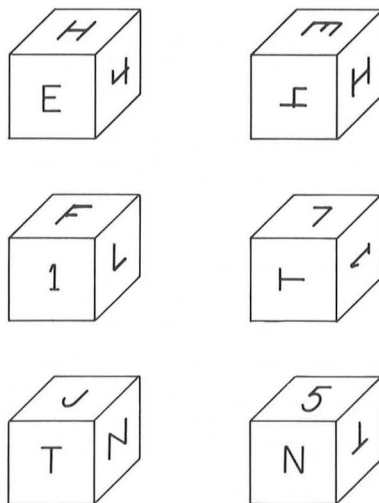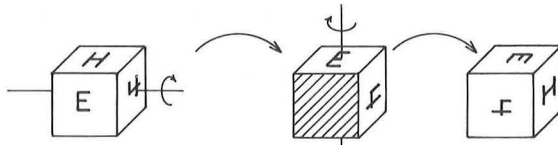

FIG. 6.6.   Examples of three possible test-items from the cube comparison test.

I A sequence of rotation along standard axis

Example of verbal report:
"If you first rotate the E on the front to the top and
then turn the cube so that the E will match (in orientation)."

II A single rotation along non-standard axis

Example of verbal report:
"I spun it around the corner of the three sides until
the letters matched up."

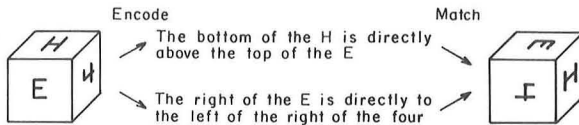III Orientation-free descriptions

FIG. 6.7. An illustration of three different strategies for solving items in the cube comparison test.

one of them could be "the bottom of the H is above the top of the E." This encoded information of one of the cubes can be validated or invalidated by comparing it to information provided in the second cube. In comparing the retrospective reports of subjects with high scores on spatial tests to subjects with low scores, Just and Carpenter (1985) found reliable differences in reported cognitive processes. Three of the high-ability subjects used predominantly non-standard rotation axes when applicable, whereas low-ability subjects used standard axes. One of the high-ability subjects relied on orientation-free descriptions.

From analyses of the temporal sequence of eye-fixations, Just and Carpenter (1985) could validate the verbally reported cognitive processes as responsible for the different pattern of latencies for high- and low-ability subjects. In addition, the high-ability subjects using the orientation-free description displayed a different pattern of latencies from subjects using the other two strategies. Just and Carpenter (1985) argued for the importance of determining and describing strategies to better understand spatial ability as measured by psychometric tests. They also noted that "trivial" changes in aspects of cube comparison tests can change the strategies subjects use. Just and Carpenter (1985) collected verbal reports

from subjects taking the original Thurstone version of the cube comparison test, which differs only in that arrows, circles, and pluses are used instead of letters. For that version, subjects predominantly used the strategy of orientation-free descriptions rather than the strategies using rotation.

The role of verbal reports in identifying strategies is even more clear in earlier studies of spatial ability. In two earlier studies (Barratt, 1953; French, 1965) subjects were asked to think aloud and verbalize their solution processes to sample items from many psychometric tests, which they had previously taken under standardized conditions. The methods for extracting strategies for solving items from specific tests were only briefly described, but given that high inter-rater agreement of encoding was obtained, the findings should be considered seriously. Barratt (1953) showed that assessed solution methods or strategies were reliably related to performance on several psychometric tests measuring spatial ability. In his original dissertation Barratt (1952) provides more detail about his methods of assessing subjects' strategies. For example, Barratt (1952) identified about half of the subjects as mentally rotating *whole* figures in the Figures Test on the basis of verbal reports like these:

> Subject #18: ". . . . I would look at all these various choices here, and I would take the problem and try to switch it around, turn it around in the same form as these here; after I turn it around, I see that they can be made to coincide. . . ."
> Subject #44: "I'm trying to turn the figure around in a way that it is in the same position that the key problem would be. . . ." (pp. 58–59)

Most of the other subjects appeared to rotate only *parts* of the figures as indicated by the following verbal reports:

> Subject #4: ". . . . The semicircle is pointed in one direction, and the V is to the bottom of it, and if the figure were the same way, well, the semicircle would be pointed in the same direction, or if it were laying down or opposite, the semicircle, uh, the V would always be to the left. . . ."
> Subject #79: ". . . . I'd look at this V here; I would look for ones that would be this way if turned this way . . . . . I would look at this bar on the bottom; that would be my distinguishing mark here; the bar is turned around in B, etc. . . . ."
> (Barratt, 1952, p. 59)

French (1965) divided his subjects into two groups on the basis of their strategy for solving items in a given test. For each group the intercorrelations on all psychometric tests were recomputed separately. Subsequent factor-analysis of each group showed remarkably different factor structures for several of the strategies. French (1965) summarized his findings by saying "Systematizing is a tendency which leads a person to use specialized or symbolic thought processes; this changes what the tests measure, and consequently affects the correlations between the tests" (p. 28).

The research on performance of tasks measuring spatial ability is particularly interesting as it illustrates how quite different sources of data (reaction time, eye-movement data, verbal reports, training studies and experiments) provide converging support for the importance of strategies in accounting for individual differences (Lohman & Kyllonen, 1983; Snow & Lohman, 1984). It also nicely demonstrates the need for information-rich data, like eye-movements and verbal reports, to fully describe complex cognitive entities such as strategies.

EXCEPTIONAL ABILITY vs. ACQUIRED SKILL

Given the reports on successful elicitation of verbal reports on cognitive processes described in the two preceding sections, one might rightfully ask which abilities are basic and yield no or uninformative verbal reports. In the introduction I mentioned that exceptional abilities, like exceptional memory, have consistently been attributed to innate differences in the structure of memory. Implicit in the definition of exceptional basic abilities is the claim that normal subjects cannot attain such abilities even after extensive practice. Furthermore, it is claimed that demonstration of such abilities in, for example, a memory task, will not allow the subject to report any mediating cognitive processes. In the first part of this section I describe some research I conducted with the late Bill Chase examining practice on a specific task. I then discuss analyses of people with alleged exceptional memory.

Effects of Practice on Performance on Memory Tests

Bill Chase and I intentionally selected digit span, because several investigators had proposed that digit span provided the best measure of the *fixed* capacity of short-term memory (STM). The fast rate of presentation of digits was assumed to force subjects to exclusively rely on STM in this memory task.

Our research approach consisted of providing subjects with extensive practice on the digit-span task and monitoring any improvements by requesting retrospective verbal reports from a selected portion of the trials. All significant changes in the reported thoughts were validated by a specially designed experiment (Chase & Ericsson, 1981, 1982; Ericsson, Chase, & Faloon, 1980).

The focus of this account is on our first subject (SF), who discovered the means to improve his memory performance. SF was selected to be a representative and average college student with respect to intelligence and memory ability. His original digit span was about seven.

During each session SF was read random digits at the rate of one digit per second; he then recalled the sequence. If the sequence was reported correctly, the next sequence was increased by one digit; otherwise it was decreased by one digit. The performance on the last sequence in the preceding session determined

the length of the digit sequence presented on the first trial on the following session. Figure 6.8 shows SF's average digit span as a function of practice for over 200 practice sessions distributed over 2 years.

Figure 6.8 shows that SF increased his digit span from 7 to over 80 digits. A naive interpretation of this dramatic increase in memory performance is that SF simply extended his short-term memory by a factor of 10. In comparison, subjects with alleged exceptional memory have digit spans of *less* than 20 digits. The relation to exceptional memory is discussed later.

However, after most of the digit-span tests, SF gave a retrospective verbal report on his cognitive processes during the trial. From an analysis of these verbal reports, we find that SF's memory performance can be accounted for in terms of an acquired skill rather than expansion of some basic capacity. The main findings were confirmed by experimental tests.

During the first session with the digit-span task, the verbal reports show that SF relied almost exclusively on rehearsal of all presented digits to remember them. In the second session he started trying to commit the first three digits of a series to memory and to rehearse the remaining digits of the presented series. Once the rehearsed digits had been committed to memory, he would retrieve the first three and initiate recall. The primary mode of encoding was repetition of digits and different numerical relations.

During Session 5, SF reported that he suddenly realized that a 3-digit sequence could be interpreted as a running time for a mile. For example, 418 could be a 4-minute, 18-second mile-time. His average digit span for this session jumped four standard deviations from the session before. SF was a long-distance runner with extensive knowledge of both specific and general categories of
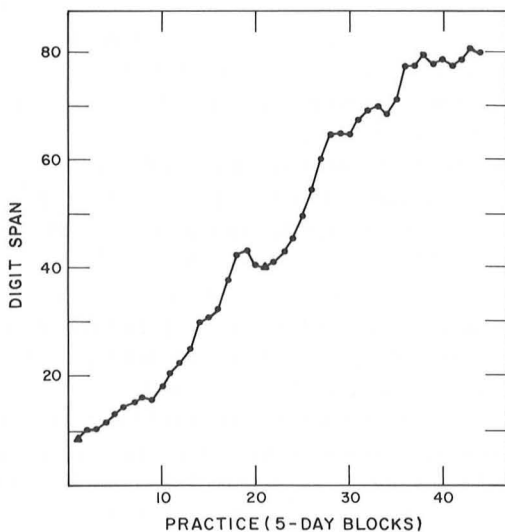

FIG. 6.8. Average digit span for SF as a function of practice.

running times for a large number of different races. During the following sessions, SF retrieved a set of races (¼-mile, ½-mile, ¾-mile, mile, 2-mile) that would cover the range of most 3-digit numbers from 100 to 959. However, no 3-digit numbers with a middle digit of 6, 7, 8, or 9, (e.g., 483, 873) can be interpreted as meaningful running times. In one experiment we presented digit sequences made up of only such uncodable 3-digit sequences to SF and his memory-span was reduced almost to the level prior to practice. Later SF started to encode 4-digit groups as running times. The different types of encodings are illustrated in a typical retrospective report given by SF shown in Table 6.7.

Finally, SF used an encoding as ages of people for digit groups that could not be meaningfully encoded as running times or dates.

In parallel with the emergence of new and more effective encodings of 3- and 4-digit groups, SF started to store up to four different groups in memory in addition to the four to five digits in the rehearsal buffer. In order to recall these digit groups in their correct order, SF encoded the order of presentation of each digit group as first, middle, or last. At the time of recall, SF could use this as the main cue to retrieve the encoded digit groups in the presented order. The encoding of these additional cues, integrated with memory traces for the purpose of subsequent retrieval, we call *retrieval-structure*. In order to be able to store more groups in memory SF introduced a new level of organization, and used two super-groups to organize encoded digits as either 4-digit groups or 3-digit groups. This hierarchical organization is illustrated in Fig. 6.9, and was evidenced in SF's retrospective verbal reports on how he encoded the digit sequence, as well as in the pauses and intonation patterns of his recall of the digit sequence. Before our experimental study of SF ended, he had extended his retrieval structure to successfully hold 84 digits.

TABLE 6.7
An Example of SF's Retrospective Reports From a Digit-Span Trial

Presented sequence: 4 1 3 1 7 7 8 4 0 6 0 3 4 9 4 8 7 0 9 4 6 2
Segmented digit groups: 4131 - 7784 - 0603 - 494 - 870 - 9462
Retrospective report:
Starting from the beginning.
I made the four thirteen point one a mile time.
I just remembered the seventy-seven eighty-four.
Ok? Ok? Right. Seventy-seven eighty-four.
Then...then...then I...
(Any pattern?)
What?
(Any pattern?)
No. No. Nothing. Just like seventy-seven eighty-four.
Ok. Then I made the oh six oh three, I made that a mile time.
Then I remembered the four nine four and the eight seven oh.
I just had to remember those.
Then I remembered the nine forty-six point...two!
It's definitely point two, two-mile.
I said, so I said to myself "What did you run it in?"
I ran it in nine forty-six point two. Nine forty-six point two. Right.

The digits, presented orally at 1 second/digit, are shown at the top of the table along with SF's segmentation into digit-groups for this trial. Adapted from Chase and Ericsson (1981).
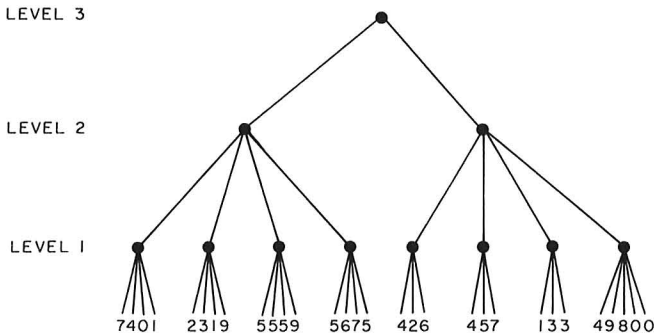
FIG. 6.9.   Proposed hierarchical organization of SF's memory encoding of thirty presented digits. The first level contains mnemonic encodings of digit groups and the second level consists of super-groups, where the relative location of several digit groups are encoded.

SF did not rely on short-term memory for his recall of the digits. His digit span was essentially unaffected by performing other tasks in the interval between the presentation and the recall of a digit sequence, even when these interpolated tasks required the full capacity of short-term memory. More conclusive evidence for storage in long-term memory is obtained from SF's ability to recall about 90% of 200–300 presented digits after the session.

Finally, SF's memory skill did not lead only to an ability to remember a larger number of digits. In a self-paced situation, SF showed that shorter digit lists (10–50 digits) could, after practice, be memorized at more than twice the original presentation rate.

In sum, SF's final performance is based on radically different cognitive processes and capacities than his initial performance prior to practice. In our study of three additional subjects practicing the digit-span task, we found evidence for the same components of skill. Two subjects given fewer practice sessions surpassed the magical limit of 20 digits. The third subject attained a digit span of more than 100 digits and is still improving with further practice. The fact that our subjects could attain digit spans surpassing subjects with alleged exceptional memory after only 50–100 hours of practice raises the possibility that the exceptional subjects were simply misdiagnosed.

Alleged Exceptional Memory Ability

When people attribute to exceptional subjects an innate ability, there is little or no evidence to substantiate such an inference. In fact, such attribution is based on the lack of alternative explanations (Ericsson & Faivre, in press). Some of the affirmative evidence comes from the subject's own verbal descriptions. The famous subject S of Luria (1968) reported storing visual images of matrices without any mediational activity involving meaning. The exceptional memory of

Inaudi was alleged to be based on sound (Binet, 1894). More commonly subjects report a complete lack of mediation, which is often interpreted as evidence for innate basic ability. These general verbal descriptions are quite different from the thinking-aloud protocols and retrospective reports I advocated earlier. Further, there appears to be a conflict of interest that might bias and contaminate the verbal reports from exceptional subjects. In our culture a mysterious ability is deemed more interesting than one that is understood (cf. an act by a magician before and after the detailed steps of the act are explained). If one's livelihood depends on the income from public performances of one's ability, which is the case with several people of alleged exceptional memory, one's willingness to describe any available details of the cognitive processes might be reduced.

First, I report on some comparisons between the memory performance of our trained subjects, whose memory structure is known, and the performance of subjects with alleged exceptional ability. I then describe some analyses of other memory experts using protocol analysis.

Binet (1894) analyzed the digit memory of two mental calculators and a mnemonist. The emphasis on memory for digits was fortunate for Chase and me because it provided an interesting test for our trained subjects (SF and DD). One of the tasks Binet used was memorization of a 25-digit matrix. Luria (1968) reported on memorization of a 50-digit matrix by his subject, S. Ericsson and Chase (1982) compared the trained and the exceptional subjects on time taken to memorize each of these two matrices, and found that the trained subjects could memorize the digits as fast or faster than the exceptional subjects. After the digit matrices were committed to memory, the subjects were asked to recall the digits from the matrix in a wide range of different orders (backward and forward recall of rows, recall of columns of digits starting at the bottom, etc.). It had been argued by Binet (1894) that the observed recall times could differentiate between auditory and visual memory representations. A reanalysis of these recall times showed a remarkable similarity between all exceptional subjects and our two trained subjects. In fact, relying on the retrospective verbal reports of one of our trained subjects, Chase and I constructed a mathematical model of the retrieval, which described the retrieval times of *all* subjects (exceptional or trained) with remarkable accuracy (Ericsson & Chase, 1982).

When Luria (1968) argued that his subject, S, had an exceptional memory, it was based on a combination of performance data and verbal descriptions from S on how he memorized information. A review of a surprisingly large number of case studies of memory experts shows that the subjects showing the most exceptional memory performance do *not* claim to have structurally different memories (Ericsson, 1985). Extensive laboratory studies of Professor Rueckle (Mueller, 1911, 1913, 1917) and of a professional mnemonist, Isahara, (Susukita, 1933, 1934) provide detailed accounts of their methods for memorization directly consistent with the three attributes of acquired memory skill discussed earlier (Chase & Ericsson, 1982). For example, a contemporary analysis of a waiter with exceptional memory for dinner orders showed that thinking-aloud protocols and

designed experiments could uncover the mnemonic associations and retrieval structure used to store the information in long-term memory (Ericsson & Polson, in press). The empirical evidence indicates that extraordinary memory performance is due to acquired memory skill regardless of claims for exceptional ability (Ericsson, 1985).

When exceptional memory performance is demonstrated by mentally retarded subjects, such performance is often assumed to reflect "pure" memorization without mediation. From verbal reports of some mentally retarded subjects with exceptionally good memories, however, we find evidence that these retarded subjects are able to use mnemonics in a manner similar to that of trained memory experts. Jones (1926), for instance, analyzed a subjects's (IQ = 75) memorization of digits under laboratory control. The following is a verbal protocol taken from the subject as he memorized the number 30249385274. It bears a striking resemblance to those of our trained digit-span experts.

> 30 is the number of days in a month. 249—if that were 149 it would be the distance from Chicago to Peoria, Illinois. 385—I once paid $3.85 railroad fare going from Cheyenne, Wyoming to Wheatland, Wyoming. 274—I can remember that by putting a 6 in front of it for the time being. 6274 is the seating capacity of the Hippodrome. (Jones, 1926, p. 372.)

On a more general level it appears that most people with remarkable skills are surprisingly unable to describe them and the corresponding cognitive processes. However, the same subjects are able to give detailed concurrent or retrospective verbal reports while performing specific tasks in their domain of expertise. In the beginning of this section I raised the question of what performance or ability is basic, or at least unmediated by reportable cognitive states. At this time I don't know where the boundary will fall, although the documented existence of unmediated retrieval and recognition processes provides a lower bound (Ericsson & Simon, 1984). The clear importance of mediating cognitive processes in perceptual skills and many exceptional abilities in mentally retarded subjects (Ericsson & Faivre, in press) shows that many investigators' intuitions about the location of such a boundary have been incorrect and require a serious reevaluation.

CONCLUDING REMARKS

In this chapter I have shown how data from verbal reports can be represented in the same theoretical framework as traditional performance measures, such as reaction time and correctness of response. The intermediate states of cognitive processes (revealed by encodings of verbal reports) provide detailed descriptions of the processes. The claims that certain tests measure specific cognitive processes can be empirically evaluated by examining verbally reported thought sequences. Drawing on three different areas of research, I have argued for the

richness and validity of verbal reports and how the verbal report data have been used to change commonly held views of underlying processes.

The issues of measurement are much broader, and in this final section I describe the relevance of protocol analysis to measurement issues. The arguments in this chapter can easily be extended to apply to the central issue of understanding the correlation between scores on different tests. Understanding what individual tests measure is a prerequisite for understanding the observed correlation between scores on two different tests. Protocol-analysis would allow us to evaluate the importance of two different sources of correlation. The first possibility is that superior test performance is due to the application of the same process or knowledge for both tests. The second and complementary possibility is that superior performance on one of the tests is determined by quite different processes and knowledge from those of the other test. Accounting for correlations due to the second possibility would require an account much different from the first.

In identifying broad issues of future measurement research, I was very influenced by Gene Glass' (1985) recent critique of current measurement research. One of his central arguments was that the information provided by tests was too general and measured abilities on such an abstract level that test scores did not provide any useful or diagnostic information to educators and the people concerned with remedial training. To describe a subjects' cognitive processes for some task requires a lot of information if this description should provide an educator with possible incorrect processes and strategies, lack of relevant knowledge, etc. Such an assessment goal is quite different from the traditional measurement of stable capacities or processing characteristics. In the body of my chapter I have tried to illustrate how protocol data can supply such information. However, the relation between verbally reported knowledge and teachability of the corresponding cognitive processes is more complex than it might appear at first glance.

It is clear that uncovering mediational elements in cognitive processes responsible for some superior or inferior performance on a task raises the possibility of improving the inferior subjects' performance through instruction. This does not, however, imply that subjects following the instruction instantly attain the superior performance of the subjects spontaneously exhibiting that strategy. Furthermore, we know that mediational cognitive processes are involved in many forms of expert performance, which can be attained only after years of practice by highly motivated students. Hence, stable individual differences in tasks are by no means irreconcilable with the existence of mediating cognitive processes. In our earlier discussion of memory skill, we showed that normal, motivated subjects could obtain exceptional memory performance after 50–100 hours of practice. The major obstacle subjects had was the fast presentation rate. To deal with the limited time available to develop retrieval structures, they needed to speed up their encoding processes. This is particularly well-illustrated by our second subject, who was instructed in the cognitive processes used by our first subject.

Given that the second subject was the running partner of the first subject, we can assume that his knowledge about running times was comparable. Although the second subject improved faster during the initial training, the advantage disappeared at a digit span of around 20 digits. This suggests that instruction can effectively guide the subject toward the correct sequence of cognitive processes, but that acquiring the necessary speed and integration requires practice. In this and other respects, mental skills resemble sports and other motor skills.

In the discussion of individual differences in mathematical ability, we noted that global strategies and general heuristics identified by experts did not provide a good conceptual system, either for characterization of individual differences or for instruction. Much better success was obtained with descriptions using domain-specific methods and different types of organization of knowledge.

The realization that any accurate characterization of individual differences in some ability requires a rather detailed description of knowledge and solution methods is important, yet somewhat disappointing. It is important because it should stimulate a closer collaboration between educators and people involved in measurement and assessment. Furthermore, it could lead to the emergence of standardized, individualized testing, with thinking aloud for the purpose of specific assessment of deficiencies as well as computer-based assessment. It is disappointing in that the task of measuring generalizable stable individual differences appears difficult or even impossible. Differences in available specific knowledge and strategies will always confound and cover any basic differences. By extracting information about strategies through verbal reports, we will explicitly address such influence and hence understand better what tests actually measure.

There is, of course, a rather different view, which argues that general individual differences are made up of differences in acquired methods and organized knowledge. The dramatic improvements after practice on all types of tasks (especially simple tasks used to measure basic capacities and processes) appear to provide strong support for this emphasis on skill. The research exploring effects of extensive practice has shown that practice does not simply make the performance quantitatively better but also leads to qualitative changes in performance. This means that many abilities assumed to require such structurally different characteristics might still be a function of practice—extensive practice. Within this skill-based view of individual differences, verbal reports and other descriptions of processes, like eye-fixations, will be absolutely essential in allowing us to characterize the components and organization of performance.

REFERENCES

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. Washington, D.C.: V. H. Winston.

Barratt, E. S. (1952). *An analysis of verbal reports of solving spatial problems*. Unpublished doctoral dissertation. University of Texas, Austin.

Barratt, E. S. (1953). An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *The Journal of Psychology, 36,* 17–25.

Bellezza, F. S. (1981). Mnemonic devices: Classification, characteristics and criteria. *Review of Educational Research, 51,* 247–275.

Binet, A. (1894). *Psychologie des grands calculateurs et joueurs d'echecs.* Paris: Librairie Hachette.

Bloom, B. S., & Broder, L. J. (1950). *Problem-solving processes of college students.* Chicago, IL: The University of Chicago Press.

Bower, G. H. (1972). Mental imagery and associative learning. In L. W. Gregg (Ed.), *Cognition in learning and memory.* New York: Wiley.

Brown, J. S., & Burton, R. R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2,* 155–192.

Brown, J. S., & VanLehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4,* 379–426.

Buswell, G. T., & John, L. (1928). Diagnostic studies in arithmetic. *Supplementary Educational Monographs, 30.*

Chase, W. G., & Ericsson, K. A. (1981). Skilled memory. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16). New York: Academic Press.

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

Deffner, G., & Ericsson, K. A. (1985). *Sprechtempo und Pausen bei lautem Denken.* Paper presented at Tagung experimentell arbeitender Psychologen. Wupperthal, West Germany.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* H. A. Ruger and C. E. Bussenius trans. New York: Dover Publications. (Originally published, 1885.)

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86,* 465–485.

Ericsson, K. A. (1975). Instruction to verbalize as a means to study problem solving processes with the Eight Puzzle: A preliminary study (No. 458). *Reports from the Department of Psychology.* Stockholm: University of Stockholm.

Ericsson, K. A. (1985). Memory skill. *Canadian Journal of Psychology, 39,* 188–231.

Ericsson, K. A., & Chase, W. G. (1982). Exceptional memory. *American Scientist, 70*(6), 507–515.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science, 208,* 1181–1182.

Ericsson, K. A., & Faivre, I. A. (in press). What's exceptional about exceptional abilities? In L. K. Obler & D. Fein (Eds.), *Neuropsychology of talent.* New York: Guilford.

Ericsson, K. A., & Polson, P. G. (in press). Memory for restaurant orders. In M. T. T. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

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

Ericsson, K. A., & Simon, H. A. (1985). Protocol analysis. In T. A. van Dijk (Ed.), *Handbook of discourse analysis* (Vol. 2). New York: Academic Press.

Frankel, F., Levine, M., & Karpf, D. (1970). Human discrimination learning: A test of the blank-trials assumption. *Journal of Experimental Psychology, 85,* 342–398.

French, J. W. (1965). The relationship of problem-solving styles to the factor composition of tests. *Educational and Psychological Measurement, 25,* 9–28.

Gimmestad, B. J. (1977). An exploratory study of the processes used by community college students in mathematics problem solving (Doctoral dissertation, University of Colorado at Boulder, 1976). *Dissertation Abstracts International, 37,* 7590A. (University Microfilms No. 77–11300)

Glass, G. V. (1985). Testing old, testing new: Schoolboy psychology and the allocation of intellectual resources. In B. Plake & J. C. Witt (Eds.), *The future of testing* (pp. 9–27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Hamilton, J. M. E., & Sanford, A. J. (1978). The symbolic distance effect for alphabetic order judgements: A subjective report and reaction time analysis. *Quarterly Journal of Experimental Psychology, 30,* 33–43.

Heidbreder, E. (1924). An experimental study of thinking. *Archives of Psychology, 11* (Whole No 73).

Hinsley, D. A., Hayes, J. R., & Simon, H. A. (1977). From words to equations: Meaning and representation in algebra word problems. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive Processes in Comprehension*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunt, E. (1978). Mechanics of verbal ability. *Psychological Review, 85,* 109–130.

Jones, H. E. (1926). Phenomenal memorizing as a special ability. *Journal of Applied Psychology, 10.* 367–377.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92,* 137–171.

Karpf, D., & Levine, M. (1971). Blank-trial probes and introtacts in human discrimination learning. *Journal of Experimental Psychology, 90,* 51–55.

Kennedy, G., Eliot, J., & Krulee, G. (1970). Error patterns in problem solving formulations. *Psychology in the Schools, 7,* 93–99.

Kilpatrick, J. (1968). Analyzing the solution of word problems in mathematics: An exploratory study. Doctoral dissertation, Stanford University, 1967. *Dissertation Abstracts International, 28,* 4380-A (University Microfilms No. 68-6442)

Krutetskii, V. A. (1976). *The psychology of mathematical problem solving*. IL: University of Chicago Press.

Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Lesgold, A. M. (1984). Acquiring expertise. In J. R. Anderson & S. M. Kosslyn (Eds.), *Tutorials in learning and memory: Essays in honor of Gordon Bower*. San Francisco: W. H. Freeman.

Lohman, D. F., & Kyllonen, P. C. (1983). Individual differences in solution strategy on spatial tasks. In R. F. Dillon & R. R. Schmeck (Eds.), *Individual differences in cognition, Vol. 1*. New York: Academic Press.

Lucas, J. F. (1972). An exploratory study on the diagnostic teaching of heuristic problem-solving strategies in calculus. Doctoral dissertation, University of Wisconsin, 1972. *Dissertation Abstracts International, 32,* 6825A. (University Microfilms No. 72-15368)

Luria, A. R. (1968). *The mind of a mnemonist*. New York: Avon.

Martin, C. J., Boersma, F. J., & Cox, D. L. (1965). A classification of associative strategies in paired-associate learning. *Psychonomic Science, 3,* 455–456.

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparisons of operations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10,* 46–60.

Montague, W. E. (1972). Elaborative strategies in verbal learning and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 6). New York: Academic Press.

Mueller, G. E. (1911). Zur Analyse der Gedachtnistatigkeit und des Vorstellungsverlaufes: Teil I. *Zeitschrift fur Psychologie, Erganzungsband 5.*

Mueller, G. E. (1913). Neue Versuche mit Rueckle. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane, 67,* 193–213.

Mueller, G. E. (1917). Zur Analyse der Gedachtnistatigkeit und des Vorstellungsverlaufes: Teil II. *Zeitschrift fur Psycholgie, Erganzungsband 9.*

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.

Noble, C. E. (1952). An analysis of meaning. *Psychological Review, 59*, 421–430.

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance, 16*, 366–387.

Polya, G. (1957). *How to solve it*. Garden City, NY: Doubleday-Anchor.

Prytulak, L. S. (1971). Natural language mediation. *Cognitive Psychology, 2*, 1–56.

Schoenfeld, A. H. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education, 10*, 173–187.

Silver, E. A. (1979). Student perceptions of relatedness among mathematical verbal problems. *Journal for Research in Mathematics Education, 10*, 195–210.

Simon, H. A. (1979). *Models of thought*. New Haven, CT: Yale University Press.

Snow, R. E., & Lohman, D. F. (1984). Toward a theory of cognitive aptitude for learning from instruction. *Journal of Educational Psychology, 76*, 347–376.

Sternberg, R. J. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Susukita, T. (1933). Untersuchung eines ausserordentlichen Gedaechtnisses in Japan (I). *Tohoku Psychologica Folia, 1*, 111–154.

Susukita, T. (1934). Untersuchung eines ausserordentlichen Gedaechtnisses in Japan (II), *Tohoku Psychologica Folia, 2*, 14–43.

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human Performance, 23*, 86–112.

Varon, E. J. (1935). The development of Alfred Binet's psychology. *Psychological Monographs, 46*(Whole No. 207).

Watson, J. B. (1920). Is thinking merely the action of language mechanisms? *British Journal of Psychology, 11*, 87–104.

Wechsler, D. (1952). *The range of human capacities*. Baltimore: The Williams & Wilkins Co.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

Young, R. M., & O'Shea, T. (1981). Errors in children's subtraction. *Cognitive Science, 5*, 153–177.