(Egan, 1983; Green & Gilhooly, 1989; Thorndyke & Stasz, 1980). However, it is critical that some form of verbal report be collected to assess the actual sequences of thoughts and that subjects' behaviors be analyzed and modelled individually (cf. Mathews et al., 1988, 1989; Stanley et al., 1989). Some fields of psychology, such as psychophysics and judgment and policy making, have been committed for a very long time to analyzing the performance of individual subjects (Hammond, McClelland, & Mumpower, 1980).

The recent discoveries of "implicit" rules mediated by memory for previously experienced exemplars and episodes would have been virtually impossible without the use of verbal reports. With increased reliance on task analysis to specify the logically possible methods for achieving the observed performance, there will be more effective encoding and use of verbal reports to assess the sequence of thoughts of individual subjects.

Over twenty years ago Newell and H. A. Simon (1972) concluded, on the basis of task analysis and analysis of verbal reports, that available knowledge and acquired skills impose the most important constraints on human performance. Recent research has clearly supported that claim. Knowledge of a domain, such as baseball or soccer, predicts ability to comprehend and remember a text about that domain far better than do standard measures of reading ability and verbal IQ (Recht & Leslie, 1988; Schneider, Körkel, & Weinert, 1989).

Acquired knowledge and skill in a specific domain can dramatically change the normal limits of cognitive processing, as we showed in the discussion of expert performance. Working memory can be extended beyond short-term memory with acquired skilled memory, anticipatory processing can circumvent limits set by simple reaction time, and recognition processes can identify relevant information and patterns during brief exposures. Future studies face a challenging problem of assessing the knowledge and skill that mediate this performance. Methods like those discussed in the section on expert systems will need to be used in conjunction with each other. Theory will need to account for all the different methods and their mediating processes in a common framework. As suggested by our discussion of Type 3 verbalization in an earlier section, such types of verbal reports as post-session recall and interviews, using Type 1 and Type 2 verbalization, would perhaps provide the best data for attempts at unification. It is also necessary to develop additional methods for validating and analysing the knowledge extracted, and characterizing it beyond task analysis.

Expert performance and acquired skill is mediated by highly domain-specific mechanisms, and the extent of transfer beyond tasks in the domain appears to be severely limited (Singley & J. R. Anderson, 1989). Studies of far and near transfer are a promising method to examine experimentally the structure

and accessibility of acquired knowledge and skill. It is nearly impossible to assess in what form prior knowledge and skill mediate transfer performance using more traditional types of data. Verbal reporting, on the other hand, allows monitoring of mediating thoughts and has been used successfully to describe and identify the mechanisms of transfer (Ericsson & Polson, 1988b; Robertson, 1990).

If this last decade is any indication of the prospects for verbal reports and protocol analysis, we are looking forward to a new decade of major advances in our understanding of the human mind.

PROTOCOL ANALYSIS

1

INTRODUCTION AND SUMMARY

After a long period of time during which stimulus-response relations were at the focus of attention, research in psychology is now seeking to understand in detail the mechanisms and internal structure of cognitive processes that produce these relations. In the limiting case, we would like to have process models so explicit that they could actually produce the predicted behavior from the information in the stimulus.

This concern for the course of the cognitive processes has revived interest in finding ways to increase the temporal density of observations so as to reveal intermediate stages of the processes. Increasingly, investigators record the directions of the subject's gaze (eye movements), and the intermediate behaviors (movements or physical manipulations of stimulus material) that precede the solution or criterion performance. Since data on intermediate processing are costly to gather and analyze, it is important to consider carefully how such data can be interpreted validly, and what contribution they can make to our understanding of the phenomena under study.

One means frequently used to gain information about the course of the cognitive processes is to probe the subjects' internal states by verbal methods. These methods are the topic of this monograph.

USING VERBAL REPORTS: SOME ISSUES

There are several issues that we must deal with if we are to use subjects' reports as fundamental data in psychological experiments. First, we must respond to the strong doubts that have been expressed by many psychologists in the past about the suitability of subjects' verbalizations as scientific data. Second, we must consider the processing that must take

place in order to transform subjects' behaviors (whether verbal or not) into data. Third, we must examine how the encoding of behavior into data can be made objective and univocal, so that the resulting data will be "hard" and not "soft." Fourth, we must be explicit about the theoretical presuppositions that are necessarily embedded in the encoding process. Finally, we must specify the processes that allow us to go backward from the data to the behavior and thence to inferences about the subjects' thought processes.

We offer a few comments on each of these five issues. They will reappear frequently as recurrent themes throughout the monograph.

Doubts About Verbal Data

Since the triumph of behaviorism over "introspectively" oriented competing viewpoints, verbal reports have been suspect as data. More precisely, behaviorism and allied schools of thought have been schizophrenic about the status of verbalizations as data. On the one hand, verbal responses (or key punches that are psychologically indistinguishable from verbal responses, except that they are made with the finger instead of the mouth) provide the basic data in standard experimental paradigms. In a concept attainment experiment, the subjects say (or signal) "yes" or "no" when a possible instance is presented to them. In a problem solving experiment, they report the answer when they find it. In a rote verbal learning experiment, they say "DAX" when the stimulus syllable "CEF" is presented. The actual performance measures commonly used–latencies and numbers of items correct– are derived from these responses, and the former depend for their validity on the veridicality of the latter.

On the other hand, modern psychology has been dubious about verbalizations produced by subjects along the route to their solutions or final responses. Even more dubious has been the status of responses to experimenter probes or retrospective answers to questions about prior behavior. All of these sorts of verbal behavior are frequently dismissed as variants of the discredited process of introspection (Nisbett & Wilson, 1977). Introspection, it has generally been argued, may be useful for the discovery of psychological processes; it is worthless for verification. As Lashley (1923, p. 352) said, in a vigorous and widely cited attack on the method, "introspection may make the preliminary survey, but it must be followed by the chain and transit of objective measurement."

Extracting Data from Behavior

The notion that verbal reports provide possibly interesting but only informal information, to be verified by other data, has affected the ways in which verbalizations are collected and analyzed. If the purpose of obtaining verbal reports is mainly to generate hypotheses and ideas, investigators need not concern themselves (and generally have not concerned themselves) with methodological questions about data collection. As a result, there is little published literature on such issues, the data-gathering and data-analysis methods actually used vary tremendously, and the details of these methods are reported sketchily in research publications that make use of such data.

If we are to make rapid and continuing progress in understanding human cognitive processes, this state of affairs is wholly unsatisfactory. In the first place, no clear guidelines are provided to distinguish illegitimate "introspection" from the many forms of verbal output that are routinely treated as data–as passing the chain and transit test (see the examples above). On what theoretical or practical grounds do we distinguish between the subject's "yes" or "no" in a concept attainment experiment and his assertion that the hypothesis he is entertaining is "small yellow circle"? In the second place, no distinctions are made among such diverse forms of verbalization as thinking-aloud (TA) protocols, retrospective responses to specific probes, and the classical introspective reports of trained observers. All are jointly and loosely condemned as "introspection."

Soft versus Hard Data

Some investigators call verbal reports and verbal descriptions "soft data" in contrast to simple behavioral measures like latency or correctness of response, which are referred to as "hard." What does this distinction mean? In science one would like to maintain as clear a separation as possible between data and theory. Data are supposed to derive directly from observation; theories are supposed to account for, explain, and predict these observation-based data. Data are "hard" when there is intersubjective agreement that they correspond to the facts of the observed behavior.

Even psychoanalytically or existentially oriented psychologists will

accept response latencies as data–even though being possibly irrelevant data for explaining behavior. When, however, an analyst codes a five-second description of a dream as "oral fixation," many psychologists would argue that this encoding is not a datum but a subjective interpretation of the data (i.e., of the verbal description of the dream). Surely, theory-laden inferences were required to derive the encoding from the verbal protocol. Data are regarded as "soft" to the degree that they incorporate such inferences, especially when the theoretical premises and rules of inference are themselves not completely explicit and objective. The problem with "soft" data is that different interpreters making different inferences will not agree in their encodings, and each interpreter is likely, wittingly or not, to arrive at an interpretation that is favorable to his theoretical orientation.

The hard-soft distinction is orthogonal to the distinction between verbal and non-verbal. The same problems of inference can emerge in observers' attempts to understand non-verbal events (e.g., sequences of physical movements, pieces of music). Such events may require as much interpretation as is required to understand verbal sequences.

Technological advances have enhanced our ability to treat verbal protocols as hard data. Until tape recorders were generally available, it was common practice for experimenters to take selective notes of verbalizations, paraphrasing and omitting whatever was "unimportant." In analyzing such notes further, it was impossible to distinguish the inferences from the original verbalizations. Using encodings of verbal protocols as data has often been made even more difficult because the theories employed, explicitly or implicitly, in the encoding were formulated in very general terms. The search for general mechanisms also led to overall interpretations of entire protocols with little concern for encoding and explicating individual protocol statements.

More recent research based on explicit information processing models of the cognitive process has caused thinking-aloud verbalizations to be viewed in a new light. It is now standard procedure to make careful verbatim transcripts of the recorded tapes, thus preserving the raw data in as "hard" a form as could be wished. At the same time, information processing models of the cognitive processes provide a basis for making the encoding process explicit and objective, so that the theoretical presuppositions entering into that process can be examined objectively.

Theoretical Presupposition in Encoding

Clyde Coombs, in his book *A Theory of Data*, shows that raw data go through a typical sequence of steps on the route from initial observation to the edited and encoded form in which they are used to test theories or make predictions. These steps, which are not neutral with respect to theory, can be seen in the processing of protocol data as they can with other kinds of data. At the first step, theory delimits a small portion of the universe of potentially observable behavior as being relevant. This judgment of relevance determines what behaviors should be recorded. At the next step, these behaviors are encoded in a manner that is again determined on theoretical grounds.

In the case of verbal behavior, the process begins with tape-recording, containing essentially all the auditory events that occurred during the experimental session. In producing from the tape a written transcript, some selection is required. After the temporal information, repetitions, and stress have been used to segment and parse the verbal stream, most of this information is usually eliminated from the transcript, except as it is captured by punctuation. We will refer to this transcription step as *preprocessing.*

At the next step, the preprocessed segments are encoded into the terminology of the theoretical model. This is often achieved by first determining coding categories, a priori, and then having human judges make the coding assessments. If each of the segments is to be treated as an independent datum, then the encoding of that segment must be made on the basis of the information contained in it, independently of the surrounding segments. In Chapter 6 of this book, we will discuss at some length methods for carrying out this kind of local encoding, and the conditions that must be met to make it possible.

Verbal protocols have been analyzed in two rather different ways. One method claims *not* to require the analysis of meanings, while the other does require it. In the first kind of analysis, subject and experimenter have agreed, by prior instruction, upon specific signals, which may be speech signals or button presses, for their communication. These signals are mostly arbitrary–a subject could say "cef" instead of "yes"; communication is possible only because of the agreement established between subject and experimenter. To analyze the recorded verbalizations under these conditions, the experimenter has only to categorize each speech signal into one of the agreed-upon categories. In theory, if not in practice, a coder should not even need to know the subject's

language–assuring that no meaningful analysis of inferencing is involved. A large number of paradigms in psychology use this kind of analysis. For example, studies using scales and multiple-choice alternatives can all be seen as instances of this method.

In the second kind of analysis, the observed verbalizations are analyzed in terms of their meanings. Even in this case, the theory buiding the analysis limits the encoding to selected aspects and features rather than the full meaning of the verbalization. For example, in a typical concept attainment task, each instance or stimulus can be represented as a unique combination of features. Each distinct concept can be represented by some particular configuration of features. Then encoding simply requires the mapping of the verbalizations onto these concepts and features usually a rather unequivocal matter. Although the space of logically possible different concepts may be very large, it is severely limited compared with the variability of natural language. Thus a verbalization like "red circles are cef's" can normally be encoded as identical with "blood-colored round ones are cef's."

The context of a particular theory and experiment greatly constrains the range of possible interpretation and allows the meaningful analysis of verbalizations to be selective and incomplete. If a theory of concept attainment is limited to the language of hypotheses, many verbalizations will not be encoded at all-statements like, "I wonder what I should do. I'll just guess on this one." Many examples can be cited of this kind of meaningful analysis, where verbalizations are mapped onto a priori formal alternatives. The analysis of memory for meaningful text has been studied by Kintsch (1974) and many others. Newell and Simon (1972) analyzed tasks, identifying formally defined knowledge states in terms of which subjects' thinking-aloud protocols could be encoded.

Many analyses of verbalizations do *not* fit the above scheme, including most analyses that seek to arrive at an understanding of the verbalizations. In less formal kinds of analysis, the encoding scheme is not defined formally and a priori, but the search for interpretations proceeds in parallel with the search for an appropriate model or theory. We recognize clearly the need for and value of such interactive processes in the search for theories in new domains, but in our own account here we will be concerned primarily with situations where the theoretical terms are fixed before the actual encoding begins.

Inferring Thought Processes From Behavior

It is sometimes believed that using verbal data implies accepting the subjects' interpretation of them or of the events that are reported. This issue of trust has its origins in our everyday experience and use of language. In order to communicate effectively with other people, we accept their word for many facts. If someone says that he has bought a new car, we generally accept his statement as true instead of asking him to produce the sales contract or a receipt. In a similar vein we trust people–at least our friends–to answer questions correctly and to give us the best advice they can. However, if the issue is important to us or we suspect ulterior motives in the responses, we may demand more details and may review all the available evidence ourselves. The same thing holds in scientific research; few scientists will accept another scientist's claim of finding conclusive evidence for ESP without wanting an independent review of the evidence.

Subjects' reports of their own mental states and mental processes raise slightly different issues of trust. According to a naive theory of consciousness, subjects have the sole *direct* access to their own mental states and processes. The subjective feeling of one's ability to report one's own mental experiences veridically is strong. For a great many reasons, this confidence is not shared by experimental psychologists, who have shown that under numerous circumstances such self-reports are unreliable.

However, the issue of the reliability of self-reports can (and, we think, should) be avoided entirely. The report "X" need not be used to infer that X is true, but only that the subject was able to say "X"–(i.e., had the information that enabled him to say "X.") By following this path, we can even show that there is an inverse relation between how much subjects need to be trusted and how much information they verbalize. For the more information conveyed in their responses, the more difficult it becomes to construct a model that will produce precisely those responses adventitiously–hence the more confidence we can place in a model that does predict them.

Consider, for example, the following possible interchanges between experimenter and subject:

1. Do you know the name of the capital of Sweden? *Yes.*
2. Which of these three, Oslo, Stockholm, or Copenhagen, is the capital of Sweden? *Stockholm.*
3. Name the capital of Sweden. *Stockholm.*

4. (A retrospective report as to how the subject arrived at an answer to Question 1): *First I tried to picture where Sweden is located on a map of Europe, then Oslo came to mind, but I remembered that it is the capital of Norway. Then Stockholm popped up and I remembered that is where the Nobel prizes are awarded; then I felt sure I could answer "yes."*

In the first case we have to trust the subject if we want to infer that he actually knows the capital, whereas in the third case it is unlikely that he could generate the correct name unless it were accessible from memory. The primary difference between second and third cases is that, for the second, one could conceive of a number of processes other than memory retrieval (e.g., guessing) that would account for the response. The fourth response, the retrospective report, also verifies that the subject has the name in memory together with some redundant information about it that gives him confidence in his answer. Of course we do not have to believe that he has given a veridical report of the process whereby he generated the name, although there is nothing implausible about the sequence of associations he reports.

Consider next a more controversial example, which has played a role in the psychological literature on learning without awareness. After a learning experiment, the experimenter asks the subjects whether they were aware of any relation between the stimuli and responses, on the one hand, and the reward contingencies on the other. Yes/no responses to this question are informative only if we trust the subjects. If a subject, however, describes the stimulus-response contingency for reward, we can be reasonably certain that he had access to this information while he was learning. On the other hand, if a subject is unable to report anything about the contingency, we *cannot* conclude that he wasn't aware of it during the learning process–we have solid evidence neither for nor against awareness during the experiment. Later, we will discuss the problem of making inferences from reports of lack of information.

These examples illustrate that the information externalized in verbal responses often provides the experimenter with data that eliminate the need for trust in the subject. The examples also show that verbal reports may be generated in many ways. To understand the reports, we must understand the processes by which they were generated. In none of these respects do data from verbal reports differ from data based on other types of observations.

Some Basic Assumptions

We can now summarize the basic assumptions that set the stage for our further explorations. Most fundamentally, we see verbal behavior as one type of recordable behavior, which should be observed and analyzed like any other behavior. The cognitive processes that generate verbalizations are a subset of the cognitive processes that generate any kind of recordable response or behavior. Hence, we would look for the same kind of "mechanical" and complete process description of verbal behavior as of other kinds of behavior, and we would not accept magical or privileged processes as explanations for verbalizations.

Whether one can and should trust subjects' verbal reports is not a matter of faith but an empirical issue on a par with the issue of validating other types of behavior, like eye fixations or motor behavior. A single invalid verbal report should not force us to discard analysis of verbal reports generally. Indeed, this monograph will undertake to build a theory of verbalization, so that we can then specify when, where, and under what kinds of instructions informative verbal reports can be obtained from subjects.

Postulating that the cognitive processes underlying verbalization are a subset of all cognitive processes implies that verbalization must comply with the constraints that have been identified, experimentally, to govern all cognitive processes. These information processing constraints will provide powerful guidelines for our attempts to specify how observed verbalizations could have been generated. We wish to account for verbally reported information by proposing a processing model sufficiently powerful to regenerate that information.

Plan of Attack

Our first task is to describe a general theory of cognitive processes and structure, which, we argue, accounts for verbalizations and verbal reports. For reasons that have already been stated, the analysis must be carried out within a framework of theory. This framework must be sufficiently general to permit us to relate, within a unified perspective, all the kinds of data that are commonly used in psychological experiments.

Usually, in choosing between theories, we want to pick the strongest one–the one that will make the strongest predictions. In the

present case, where the theory we choose will influence the way in which we encode and analyze our data, we want to pick the weakest and most "neutral" one that can do the job. The fewer controversial assumptions we incorporate in the theory, the less we will be involved in the circularity of using theory-laden data to test our theories. Nevertheless, there appears to be no way of processing data that does not incorporate *some* theoretical assumptions about the system and processes that generated the data. Our particular strategy will be to set forth the theory in its most general, hence least controversial, form first, then add more specific hypotheses where they are required.

After presenting the theory as an information processing model of cognitive processes, we will survey the literature on verbal reporting and derive from it a taxonomy of reporting procedures. We will follow this survey with an historical review of earlier approaches to verbal reports. We will then take up the major issues surrounding the use and validity of verbal reports, discussing the empirical studies within the framework of a more detailed information processing model.

THE PROCESSING MODEL

Our purpose in presenting a specific processing model is to aid us in interpreting verbal data obtained from subjects and the relation of their verbal to their other behavior. Since the data (including the verbal data) are gathered in order to test theories about the human information processing system, we are engaged in something of a bootstrap operation. We need a model in order to interpret data that are to be used, in turn, to test the model. Under these circumstances, our data-interpretation model should be as simple as possible, and it must not incorporate components that are themselves bones of theoretical contention. The model should be robust (i.e., compatible with a wide range of alternative assumptions about human information processing).

The specifications we are about to present are simple and robust in this sense, and, indeed, summarize the core that is common to most current information processing theories of cognition. Of course they are not entirely neutral, for they would be hard to reconcile with an extreme form of behaviorism that denied the relevance of central processes to the explanation of behavior. But they are not specific to the view of any particular "sect" within the general information-processing tradition. (For fuller discussion of the model, see Newell and Simon (1972, Chapter 14), and Simon (1979, Chapters 2, 3).

General Specification

The most general and weakest hypothesis we require is that human cognition is information processing: that a cognitive process can be seen as a sequence of internal states successively transformed by a series of information processes. An important, and more specific, assumption is that information is stored in several memories having different capacities and accessing characteristics: several sensory stores of very short duration, a short-term memory (STM) with limited capacity and/or intermediate duration, and a long-term memory (LTM) with very large capacity and relatively permanent storage, but with slow fixation and access times compared with the other memories.

Within the framework of this information processing model, it is assumed that information recently acquired (attended to or heeded)* by the central processor is kept in STM, and is directly accessible for further processing (e.g., for producing verbal reports), whereas information from LTM must first be retrieved (transfered to STM) before it can be reported.

This general picture is compatible with all sorts of specific hypotheses that have been put forth with respect to the details of the mechanisms. For example, some theorists propose that what we call "short-term memory" is not a separate, specialized store but simply a portion of LTM that is currently and temporarily activated (Anderson, 1976). Some theorists believe that information in STM extinguishes with passage of time, unless rehearsed; others that it is lost only when replaced. In general, these differences of detail do not affect the model at the level of specificity required for our purposes. The important hypothesis for us is that, due to the limited capacity of STM, only the most recently heeded information is accessible directly. However, a portion of the contents of STM are fixated in LTM before being lost from STM, and this portion can, at later points in time, sometimes be retrieved from LTM.

Our specification of the system is general, but it is not vague. Specific information processing models that incorporate these features have been constructed in the form of computer programs, and these have

*Because the phrase "attended to" is often stylistically awkward, we will sometimes use "heeded" instead. So we will say, more or less synonymously, that information was "attended to," was "heeded," or was "stored in STM."

been shown to produce a variety of behaviors previously observed in psychological laboratories. Verbal predictions of how such a system behaves can, thereby, be tested by using a computer program as a simulator. The principal model of this kind that guides our own thinking about these processes is the EPAM program, due to Feigenbaum (1963) and Simon, and discussed in some detail in Section 3 of Simon (1979).

We assume that any verbalization or verbal report of the cognitive processes would have to be based on a subset of the information held in STM and LTM. From this and the above hypotheses, the taxonomy of verbalization procedures shown in Table 1-1 follows in a straightforward fashion (Ericsson & Simon, 1980).

Table 1-1

A Classification of Different Types of Verbalization Procedures as a Function of Time of Verbalization (Rows) and the Mapping From Heeded to Verbalized Information (Columns)

	Relation between heeded and verbalized information			
		Intermediate processing		
Time of verbalization	Direct one to one	Many to one	Unclear	No relation
While information is attended	Talk aloud Think aloud			
While information is still in short-term memory	Concurrent probing	Intermediate inference and generative processes		
After the completion of the task-directed processes	Retrospective probing	Requests for general reports	Probing hypothetical states	Probing general states

The two dimensions of Table 1-1 represent two major distinctions. First, the time of verbalization is important in determining from what memory the information is likely to be drawn. Second, we make a distinction between procedures where the verbalization is a direct articulation or explication of the stored information, and procedures where the stored information is input to intermediate processes, like abstraction and inference, so that the verbalization is a product of this intermediate processing.

Detailed Specification

We now specify more fully the components of the information processing system that we have just sketched. The model draws upon a variety of sources that are summarized in Newell and Simon (1972, Ch. 14) and Simon (1979, Ch. 2.3). Few of the model's specifications are controversial. It makes no real difference, for example, whether we assume a single homogeneous memory with different modes of activation (e.g., Anderson, 1976; Shiffrin & Schneider, 1977) or several discrete memory stores (sensory stores, STM, and LTM). The important matters, which can be described in either terms, relate to the amounts and kinds of information that can be retained, and the conditions for accessing them and reporting them verbally. We will use the conventional model of multiple memories in our description.

Recognition. Information received from the sensory organs resides for a short time in memories (iconic and echoic memories) associated with the different senses. During this time, portions of the sensory information are directly *recognized* and encoded with the aid of information already stored in LTM. Recognition associates the stimulus, or some part of it, with existing patterns in LTM, and stores in STM "pointers" to those familiar patterns. (The EPAM discrimination net is a model of this recognition mechanism.) Intermediate stages of the direct recognition process (the successive steps of discrimination), which may take only 10 to 100 msec, do not use STM to store their products.

Long-Term Memory. The LTM may be pictured as an enormous collection of interrelated nodes. Nodes can be accessed either by recognition (through the discrimination net), as just explained, or by way of links that associate these nodes to others that have already been accessed. Information accessed in either way is then represented by pointers in STM. Thus, information can be brought into STM from sensory stimuli via the recognition process, or from LTM via the association process. Association processes are much slower than direct recognition processes, requiring at least several hundred msec for each associative step. Associative processes may use STM to store intermediate steps. So, for example, in recalling a name that is not immediately accessible, a person may use a sequence of cues to find an associative path, step by step, to the sought-for name. Such processes may last tens of seconds, or even minutes, and may leave numerous intermediate symbols in STM, where they are temporarily available for verbal reports.

Short-term Memory. The central processor (CP), which controls and regulates the non-automatic cognitive processes, determines what small part of the information in sensory stimuli and LTM finds its way into STM. This is the information that is *heeded* or *attended to.* The amount of information that can reside in STM at one time is limited to a small number (four?) of familiar patterns (*chunks*). Each chunk is represented by one symbol or pointer to information in LTM (Simon 1979, Ch. 2.2). As new information is heeded, information previously stored in STM may be lost.

When a cognitive task (e.g., mental addition of a column of figures) is being carried out, the typical chunks in STM are pointers to the operands, operators, and outputs of the operations that are being performed. Thus, in adding 3 to 4, pointers corresponding to the symbols "3," "4," "PLUS," and "7" might at some time be present in STM. Since, in our culture, adding two digits involves a direct reference to LTM ("table lookup"), no further detail of the process would be heeded in STM or available for verbal reports. On the other hand, if the task were to multiply 17 by 45, STM might hold, at various points in the process "45," "17," "7," "TIMES," "3" (the carry in multiplying 45 by 7), "315" (the first intermediate product), "45," "1," "TIMES," "PLUS," "765."

We hold no brief for the details of the above description, which is intended merely as an example of the *kinds* of information we would expect to be heeded in STM, and to be available, potentially, for concurrent or retrospective reports. The specific details would depend on the particular strategies subjects used and the nature of the chunks they had stored in LTM (Simon, 1979, Ch. 2.4). STM would symbolize the process only down to some modest level of detail (corresponding to elementary processes of a second or two in duration), and we would not expect to find information there about simple, automated processes (e.g., the processes of retrieval from LTM or recognition processes), much less about neuronal events. Thus, the architecture of the control apparatus (CP) determines the fineness of grain of the representation of processes in STM.

Control of Attention. The flow of attention is diverted, from time to time, by interruptions through the higher control mechanism. Intermediate stages in these interruptions, not being symbolized in STM, are not reportable. Sudden movements in peripheral vision, loud noises,

emotions operating through the reticular system are important causes of interruption and shift in attention (Simon, 1979, Ch. 1.3). While information heeded immediately before or after a shift in attention may sometimes allow subjects to give a relatively clear account of the interruption, we would expect such information to be less complete than reports of an orderly process that is induced by the successive content of STM itself (e.g., a thought sequence during which goals in STM are guiding the thought processes).

Fixation. New information is retained in STM during the time the CP is attending to it. In order to create an LTM representation of new information that can later be recalled, associations must be built up by coding and imaging, as well as new tests and branches in the recognition network. These learning processes, including the storage of new information in LTM and the addition of new pathways in the discrimination net for accessing it, are modeled in some detail by EPAM (Simon, 1979, Section 3). Processing of the order of 8 to 10 seconds is required to assemble each new chunk from its familiar components in STM, and to store it in LTM as a new chunk (Simon, 1979, Chs. 2.2, 2.3).

Automation. As particular processes become highly practiced, they become more and more fully automated. (Shiffrin & Schneider, 1977). Automation means that intermediate steps are carried out without being interpreted, and without their inputs and outputs using STM. The automation of performance is therefore quite analogous to executing a computer algorithm in compiled instead of interpretive mode. Automation (and compiling) have two important consequences. They greatly speed up the process (typically, by an order of magnitude) and they make the intermediate products unavailable to STM, hence unavailable also for verbal reports.

TYPES OF VERBALIZING PROCEDURES

The only feature common to the whole range of techniques used to obtain verbal data is that the subject responds orally to an instruction or probe. Because of the flexibility of language, there are virtually no limits to the probes we can insert and the questions we can ask subjects that will elicit some kind of verbal response.

Within our theoretical framework, we can represent verbal reporting

as bringing information into attention, then, when necessary, converting it into verbalizable code, and finally, vocalizing it. The crucial issue for verbal reporting procedures is what information is heeded. There have been studies showing that the response modality does not affect the frequency of different responses. Newhall and Roderick (1936) found no differences in frequencies between verbal reports, button presses with fingers, or pedal presses with the feet. This result indicates that the response is heeded symbolically, and then translated into the appropriate overt form. (See Chapter 5 for further discussion.)

Two forms of verbal reports can claim to being the closest reflection of the cognitive processes. Foremost are *concurrent verbal reports*–"talk aloud" and "think aloud" reports–where the cognitive processes, described as successive states of heeded information, are verbalized directly (see Figure 1-1).

We claim that cognitive processes are not modified by these verbal reports, and that task-directed cognitive processes determine what information is heeded and verbalized. We will evaluate this claim empirically in Chapter 2.

A second type of verbal report is the *retrospective report.* A durable (if partial) memory trace is laid down of the information heeded successively while completing a task. Just after the task is finished, this trace can be accessed from STM, at least in part, or retrieved from LTM and verbalized. Retrospective reports based on information in LTM will require an additional process of retrieval that will display some of the same kinds of error and incompleteness that are familiar from experimental research on memory. Both of these kinds of reports, we claim, are direct verbalizations of specific cognitive processes.

Recoding Before Verbalization

Various processes, and especially recoding processes, may intervene between the time information was heeded by the central processor (CP) and the time a verbalization is generated. When information is reproduced in the form in which it was heeded, we will speak of *direct* or *Level 1* verbalization. When one or more mediating processes occurs between attention to the information and its delivery, we will speak of *encoded* or *Level 2* or *Level 3* verbalization. A number of different kinds of intermediate processes between access and verbalization may modify the information. Among the important kinds are the following:

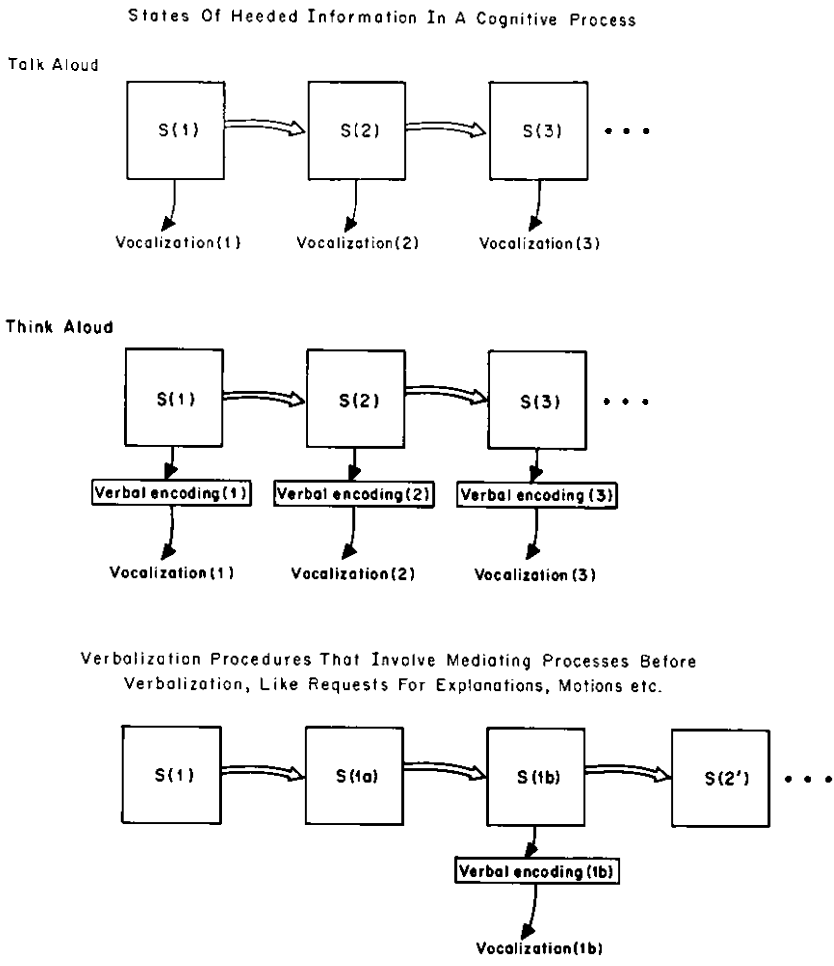

Figure 1-1

The Relation Between the Heeded States of a Cognitive Process and Verbal Reports for Various Types of Verbal Report Procedures

1. Recoding into verbal code (Level 2 verbalization). When the internal representation in which the information is originally encoded is not a verbal code, it has to be translated into that form. Werner and Kaplan (1963) have shown that when subjects generate verbal descriptions of nonverbal stimuli for their own future use, the format is compact and incorporates many idiosyncratic referents. When verbalizations are generated to communicate information to another person, additional processing is required to find referents (Werner & Kaplan, 1963).

2. Intermediate scanning or filtering processes (Level 3 verbalization). When the task instructions ask for verbalization of only selected information, it is necessary to postulate additional processes that test if the heeded information is of the desired type. Such instructions are used, for example, in commentary driving experiments, in which the subjects are asked to report all perceived traffic hazards while they are driving a car (Soliday & Allen, 1972).

3. Intermediate inference or generative processes (Level 3 verbalization) The situation is even more complicated if the experimenter is interested in particular aspects of the situation that a subject would not ordinarily attend to. The issue of whether the instruction to verbalize calls for information not normally heeded by the subjects is central and directly related to the occurrence of intermediate inference and generative processes. Since we will return to this issue in more depth, only a brief summary will be given here of the types of information that are likely to require additional mediating processing for their generation.

In addition to verbalizing their ongoing thinking, subjects are sometimes asked for verbal descriptions of their motor activities, for example, what objects are moved where, or where they are looking. When this information is not heeded directly, as is often the case, the subject is required to observe his or her own internal processes or overt behavior to generate the information.

Experimenters are often interested in subjects' reasons for their overt behavior and consequently ask the subjects to verbalize their motives and reasons, which may not be available directly or even at all. In an excellent review of research on the effects of persuasive messages, Wright (1980) discusses a wide range of biases due to different verbal report procedures.

In sum, with Level 1 and Level 2 verbalization the sequence of heeded information remains intact and no additional information is heeded. On the other hand Level 3 verbalization requires attention to

additional information and hence changes the sequence of heeded information.

Retrospective Reports

In the ideal case the retrospective report is given by the subject immediately after the task is completed while much information is still in STM and can be directly reported or used as retrieval cues. It is clear that some additional cognitive processing is required to ascertain that the particular memory structures of interest are heeded. Our model predicts that retrospective reports on the immediately preceding cognitive activity can be accessed and specified without the experimenter having to provide the subject with specific information about what to retrieve. In this particular case, the subject will still retain the necessary retrieval cues in STM when a general instruction is given "to report everything you can remember about your thoughts during the last problem." This form of retrospective verbal report should give us the closest approximation to the actual memory structures.

Even in this favorable case, some problems arise that are common to all kinds of verbal reports from LTM. First, the retrieval operation is fallible, in that other similar memory structures may be accessed instead of those created by the just-finished cognitive process. The probability of this occurring increases markedly if the subjects have just solved a series of similar problems. However, since most accessed memory structures contain redundant information beyond the cues used for retrieval, subjects may use this additional information to validate the retrieval as well as to increase their confidence in the veridicality of the retrieved information. In a subsequent section we will discuss this type of evaluation further and examine the relevant theoretical and empirical literature.

A second general problem when retrieving cognitive structures is to separate information that was heeded at the time of a specific episode from information acquired previously or subsequently that is associated with it (Mueller, 1911). For example, if a picture reminds one of an old friend, it may be tempting to use the stored information about that friend to *infer* what the person in the picture looked like. (In Chapter 3 we will discuss this issue in more detail.) It may be possible to eliminate this artifact by instructing subjects only to report details that they can remember heeding at the time of the original episode (Mueller, 1911). By imposing a requirement of determinable memory as a basis for reporting, we can

avoid many subjects' tendency to fill in information that they can't remember but "must" have thought.

Inferential or Generative Processes

The most marked difference between concurrent and retrospective reporting is that retrospective reports refer to a cognitive process that is completed and cannot be altered and influenced. Hence, if subjects are requested to report information that was never heeded, they cannot possibly base their responses on direct memory. The subjects can answer that they don't know, but often they will infer and generate an answer on the basis of information provided in the question and other information accessible from LTM. Since retrieval from LTM may be an onerous task, even in situations where the information is potentially retrievable subjects may prefer to generate the information instead.

The most common probe that creates this problem is the why-question: for example, "Why did you do this?" or "Why did you prefer that product?" In an interesting discussion, Lazarsfeld (1935) points to many issues and problems in interpreting responses to why-questions, where subjects select one alternative out of several possibilities. Some of the alternatives may never have been heeded. If we wish to find out: "Why did you buy this book?" we may receive, out of the same concrete experience of the respondent, quite different answers, according to whether we stress "buy," "this," or "book." "If the respondent understood: 'Why did you BUY this book?' he might answer, 'Because the waiting list in the library was so long that I shouldn't have got it for two months.' If he understood: 'Why did you buy THIS book?' he might tell what interested him especially in the author. And if he understood: 'Why did you buy this BOOK?' he might report that he at first thought of buying a concert ticket with the money, but later realized that a book is a much more durable thing than a concert, and such reasoning caused him to spend his money upon the book" (Lazarsfeld, 1935, p. 29).

The example is instructive in showing that a person who did not actually buy the book, and hence had no specific memory of the associated cognitive processes, could give the same or similar answers as *plausible* reasons for someone else's buying a book. Hence, the answers can be generated (inferred) without access to a specific memory trace of the episode.

Directed or Specialized Probing

Verbal probes differ in the comprehensiveness of the topics to be reported and the generality or particularity of the events to be reported. Let us first consider topic specificity. In many studies, the investigator is interested only in particular aspects of subjects' behaviors. Then the verbal probe may be constructed to induce the subjects to generate information specifically relevant to the hypotheses under consideration. In order to help subjects retrieve the desired information from memory and to induce greater completeness of the verbal reports, the question or verbal probe often contains contextual information. To guard against subjectivity in analyzing verbal reports, the investigator often supplies subjects with a fixed set of alternative responses. In contrast, a general instruction to give verbal reports typically asks subjects to tell everything they can remember or are thinking of while performing the task.

In most cases, verbosity and absence of selectivity in subjects' reports is not an important problem. What the subject reports is likely to be less, rather than more, than we should like to hear. In no study known to us using general instructions has the investigator complained that subjects have reported too much information from actual memory.

One common difficulty in probing for specific information, especially when the subjects are offered a fixed set of alternative answers, is to know that the questions conform to the internal representations the subjects are employing in their thought. Probes for types of information that subjects don't have directly accessible, or probes that provide inadequate sets of alternatives may force subjects to intermediate and inferential processing, and hence produce verbal reports that are not closely related to the actual thought process. Moreover, when specific, fixed-alternative probes are used, there is no way to detect from subjects' responses that this has occurred.

Since providing contextual information and prompts to subjects may aid recall from LTM, in studies of LTM the use of prompts and context is frequent and relatively well-motivated. When subjects are asked to report on immediately preceding cognitive processes of relatively short duration, specific probes are more questionable and less useful. In a logical sense, the experimenter gets just as much information from the subject in the third as in the first two of the following three cases.

(1) Directed probe 1
Question: Did you use X as a subgoal?
Answer: Yes.

(2) Directed probe 2
Question: Did you use any subgoals? If so, which?
Answer: Yes, I used X.

(3) Undirected probe
Verbal report: ...I was first trying to get X and I...
when I attained X...

The replies in all three cases provide evidence that the subject used X as a subgoal, yet the evidence is stronger in the third case than in the second, and in the second than in the first. The verbalization in the first case could easily be generated by processes independent of any memory for the actual thought processes. Comparing the second and third cases, the former communicates to subjects what information the experimenter expects them to report. It may encourage subjects to try to infer or guess what particular information the experimenter will accept, and to generate information accordingly.

In many cases, other criteria are available for estimating the validity of the reports. An analysis of the task (Newell & Simon, 1972) will often provide strong indications of the adequacy of verbalized information, especially in cases with many logical possibilities for response.

Finally, different kinds of probes may have different effects upon the behavior of subjects. Requesting a certain kind of information may suggest to subjects what aspects of the task are important. Subjects may also alter their normal ways of processing so as to be able to give the requested information to the experimenter on subsequent trials.

In studies that use retrospective verbalization, subjects are seldom asked what they can remember about specific instances of their cognitive processes. Rather, they are usually asked to retrospect about their thought processes in experiments with many trials or to answer general questions, and thus must try to synthesize all the available information after selective recall. In making judgments, subjects have access to an extremely large base of relevant knowledge. Tversky and Kahneman (1973) have demonstrated that subjects only retrieve a few events or pieces of knowledge and use this sample to infer frequencies and probabilities of events. Although the retrieved sample may often be representative and the inferred probability judgment fairly accurate, there are many factors influencing retrievability that do not reflect frequency. Hence, in many situations such cognitive processes will yield incorrect judgments about frequency. Even though all the specific information retrieved is accurate, the inferred probability may be seriously in error. Nisbett and Ross (1980) have given a recent comprehensive discussion of such biasing factors in human judgment.

Particular and General Reports

If the purpose of retrospective probing were to recover memory traces of subjects' processes, the appropriate instruction would be to ask them to recall their specific thought processes during particular trials of the experiment. For at least two different reasons, such a procedure is rarely used. First, after a series of trials, a subject's memory for individual cognitive processes will be poor and lacking in detail. Moreover, there is a tendency for recurrent cognitive processes gradually to become automatic, so that fewer or none of the intermediate states of the processes for the later trials of the experiment are accessible for recall.

Second, many experimenters, because they are interested in general characteristics of the thought processes and not in the episodic details of the individual trials, probe their subjects with questions of the type, "How did you do these tasks?" Such questions implicitly or explicitly request a general rather than specific interpretation of how the subjects were performing the tasks.

There are several different ways in which subjects might arrive at descriptions of their general procedures, as distinct from reports on specific behaviors during individual trials. One possibility is that subjects are aware of the general procedures, or "programs," they are using, use essentially the same programs on all trials, and can recall and report these directly without reference to the specific behavior they produced. Another possibility is that subjects remember some parts of their processes during particular trials, and generalize this information into a general procedure, which they then report. A different possibility is that subjects remember some specific tasks, regenerate–by redoing them–the processes used for these tasks, and use this information to infer the general procedures they may have used. Finally, subjects may draw upon various kinds of prior information, such as general knowledge on how one ought to do these tasks, to generate a verbal report describing a general procedure or strategy. In this case, the verbal reports may not bear any close relation to the actual cognitive processes (Nisbett & Wilson, 1977).

In areas of applied psychological research where verbal questioning has a long tradition, subjects are usually asked about specific events rather than for general information or conclusions. In the critical incident technique proposed by Flanagan (1954), the subjects were always asked to report their memory for specific events. For example:

> ... pilots returning from combat were asked "to think of some occasion during combat flying in which you personally experienced feelings of acute disorientation or strong vertigo." They were then asked to describe what they "saw, heard, or felt that brought on the experience." (Flanagan, 1954, p. 329)

Interpretive probing, unlike the critical incident technique, cannot be relied upon to produce data stemming directly from the subjects' actual sequences of thought processes. The probing procedures encourage or even require subjects to speculate and theorize about their processes, rather than leaving the theory-building part of the enterprise to the experimenter. There is no reason to suppose that the subjects themselves will or can be aware of the limitations of the data they are providing. Moreover, the variety of inference and memory processes that might be involved in producing the reports make them extremely difficult to interpret or to use as behavioral data.

TWO CHALLENGES TO VERBAL REPORTS

It will be useful, in order to get a perspective on the issues, to use the above analysis to examine two published papers that have sometimes been interpreted as providing strong evidence against trusting verbal reports as data from which cognitive processes can be inferred: the first, a paper reporting a study by Verplanck and Oskamp; the second, the review paper on retrospective verbal reports by Nisbett and Wilson. A discussion of these papers will show how the information processing model we have outlined can help us interpret the findings of experiments on verbalization.

Apparent Inadequacies of Concurrent Verbalization

In an often cited study (Verplanck, 1962), Verplanck and Oskamp claimed to have shown that verbalized rules are dissociated from the behavior they were supposed to control. By having subjects verbalize the rules they were following in sorting illustrated cards, the experimenters could reinforce either the verbal rule or the placement of cards (i.e., behavior). To make the contingencies less noticeable, the criterion trials were followed by additional trials with partial reinforcement. When correct placements were reinforced, the subjects were found to place cards

correctly in 71.8% of the trials; but they stated a correct or correlated rule in only 48.4% of the trials. When correct statement of the rule was reinforced, the subjects stated a correct or correlated rule on 92.8% of the trials, but placed the cards correctly on only 76.8% of the trials.

In a replication and analysis of this experiment, Dulany and O'Connell (1963) were able to show that the above results could be attributed to two artifacts of the original experiment. First, in the case where correct placement was reinforced, by making a correction for guessing (the subjects had a 50-50 chance of placing the card in the correct pile when they didn't know the rule), we can estimate that subjects *knew* the correct answer in 43.6% of the trials–a percentage very close to the 48.4% in which they stated the correct rule.

Second, with respect to the reinforcement of rules, Dulany and O'Connell found that the rules defined by Verplanck and Oskamp were ambiguous for the card illustrations they employed. In fact, naive subjects who were told these rules explicitly misplaced the cards as frequently as did the subjects in the original experiment.

In a detailed analysis of the rules the subjects verbalized on each trial, Dulany and O'Connell found that on all but 11 of 34,408 trials the subjects put the card where they said they were going to. Hence, Dulany and O'Connell impeached rather thoroughly the evidence put forth by Verplanck and Oskamp for believing that the rules subjects verbalized were inconsistent with their behaviors.

Numerous studies provide positive support for consistency between verbalized rules, concepts, and hypotheses and immediately preceding and succeeding behavior, before subjects receive feedback. When Schwartz (1966) asked subjects their reasons for placing a card as they did, the reasons given were consistent with the placements on all but 2 of 1,962 trials. Even more impressive, Frankel, Levine, and Karpf (1970) obtained retrospective reports from subjects about the basis for their responses to four earlier discrimination-learning problems with 30 non-feedback trials each, and found that subjects could provide such reports in more than 90% of the sequences of trials.

Apparent Inadequacies of Retrospective Reports

In a recent extensive review of studies permitting evaluation of retrospective verbal reports, Nisbett and Wilson (1977) have reported evidence that appears at first sight to be very damaging to the utility of verbal

reports for inferring information processes. Since their paper has received widespread attention, it is important that we review their findings carefully. The authors summarize their main empirical findings thus (1977, p. 233):

> People often cannot report accurately on the effects of particular stimuli on higher order, inference-based responses. Indeed, sometimes they cannot report on the existence of critical stimuli, sometimes cannot report on the existence of their responses, and sometimes cannot even report that an inferential process of any kind has occurred.

First, we call attention to the frequent use, in their summary, of the qualifiers "often" and "sometimes." Nisbett and Wilson cite a large number of experiments that support their conclusions, but do not investigate in detail the *conditions* under which these conclusions do and do not hold. Moreover, they do not propose a definite model of the cognitive processes as a framework for interpreting the findings they survey. Their theoretical interpretations are entirely informal, resting heavily on an undefined distinction between introspective access to "content" and to "process," or, as they alternatively state it, (1977, p. 255), between access to "private facts" and to "mental processes." Their summary of the kinds of information to which subjects *do* have access is this (1977, p. 255):

> ... we do indeed have direct access to a great storehouse of private knowledge ... The individual knows a host of personal historical facts; he knows the focus of his attention at any given point of time; he knows what his current sensations are and has what almost all psychologists and philosophers would assert to be "knowledge" at least quantitatively superior to that of observers concerning his emotions, evaluations, and plans. Given that the individual does possess a great deal of accurate knowledge ... it becomes less surprising that people would persist in believing that they have, in addition, direct access to their own cognitive processes. The only mystery is why people are so poor at telling the difference between private facts that can be known with near certainty and mental processes to which there may be no access at all.

Nisbett and Wilson also observe that subjects "are often capable of describing intermediate results of a series of mental operations (1977, p. 255)" (i.e., that they hold in STM and can access the symbols that are inputs and outputs to such operations).

We may compare this list of "private facts" and intermediate results that, according to Nisbett and Wilson, *are* accessible to subjects with the kinds of information that our processing model would imply that subjects could report. The individual knows, they say, his focus of attention, his current sensations, his emotions, his evaluations, and his plans. He knows the intermediate results of his mental operations. But these are exactly the kinds of information that, according to our model, would be held in STM and be available for verbal reports.

Unfortunately, the studies reviewed by Nisbett and Wilson provide little data on what information is heeded during the thought processes, and what information is accessible from STM and LTM at the time of the verbal report. Nisbett and Wilson find that the subjects, when *asked questions about their cognitive processes*, frequently do not base their answers on memory for specific events at all, but "theorize" about their processes (1977, p. 233).

> When reporting on the effects of stimuli, people may not interrogate a memory of the cognitive processes that operated on the stimuli; instead, they may base their reports on implicit, a priori theories about the causal connection between stimulus and response.

In reviewing the studies cited by Nisbett and Wilson, we can profitably raise the question of *why* and *when* subjects do not consult their memories of cognitive processes in answering questions about those processes. It is easy to draw the erroneous conclusion that this independence of verbal answers to questions about cognitive processes from the actual course and results of those processes implies a *general* lack of accessible memory for such processes, or even an unawareness of the information while the process was actually going on. But this sweeping conclusion appears not to be justified.

The accuracy of verbal reports depends on the procedures used to elicit them and the relation between the requested information and the actual sequence of heeded information. Invalid reports, like those discussed and obtained by Nisbett and Wilson, may be due to lack of access to thoughts (their claim), inadequate procedures for eliciting verbal reports, or requesting information that could not be provided even if thoughts were accessible. In a subsequent chapter (Chapter 3) we will describe in some detail what information will be heeded and hence reportable. Although some studies cited by Nisbett and Wilson did probe for such information, we will focus here on the deviations between the verbal report procedures used in many of the studies cited by Nisbett and

Wilson and the procedures that, according to our model, would elicit valid retrospective reports of cognitive processes.

First, many of the verbal reports they discuss could be generated without accessing memory of the corresponding cognitive processes. In some of these studies, the questions presented to subjects contain considerable background information from which answers could be generated without consulting their memories. With questions like, "I noticed that you took more shock than average. Why do you suppose you did?" (Nisbett & Wilson, 1977, p. 237) It is not even clear to us, nor probably to the subjects, that memory for the cognitive process *should* be the information source for the answer. If subjects can generate their answers without consulting their memories (Nisbett and Wilson showed that control subjects could do exactly that), they might often prefer this method to retrieving information from memory.

Second, several aspects of the verbal report procedures reviewed by Nisbett and Wilson made the relevant thoughts less *accessible.* In most of the studies reviewed, the time lag between task and probe was sufficiently great to make it unlikely that the relevant information remained in STM. In Chapter 3 we will review the rather extensive literature from general experimental psychology showing that time and intervening thought activity between the cognitive process and its verbal report, as well as incentive to recall memories of the cognitive process, lead to dramatic declines in the accuracy of the verbally reported information. A recent chapter by Genest and Turk (1981) and a paper by Wright and Kriewall (1980) give references showing that such considerations of accessibility are powerful determiners of the accuracy of verbal reports for cognitive processes in tasks like those discussed by Nisbett and Wilson (1977).

A tendency to generate verbal reports without access to memories will be stronger, the less readily available the memory is. When the probe is not a good retrieval cue for the relevant aspects of the memory, the subject must attempt, through conscious processing, to recall sufficient information to give an appropriate answer. Since retrieval from LTM, even if possible, requires considerable time and effort, subjects, unless explicitly instructed to provide a relatively complete recall, may be disinclined to do so, especially if other ways of producing a response are open to them. A recent study by Wright and Rip (1980) provides strong evidence for an increase in accurate self-report when subjects were explicitly motivated to retrieve memory for thoughts in a judgment task.

Finally, in some studies reviewed by Nisbett and Wilson, subjects were asked to report information that cannot be given even with complete access to the thought processes (cf. why-questions regarding causes), and information that is far from a direct recall of memory of the cognitive processes. Our model predicts that information can be recovered by probes only if the same information would be accessed by undirected requests for concurrent or retrospective reports. For many of the studies in the Nisbett-Wilson review, our model would predict failure to obtain from the probes verbal information about particular instances of processes. For example, in between-subject designs, subjects obviously cannot answer from memory of their processes why they behaved differently from subjects in another experimental condition–the processes did not include such a comparison. Hence, this information can be derived, if at all, only by comparing the descriptions of the processes provided by different sets of subjects in the two conditions. In other studies the subjects were asked how they would have reacted if the experimental conditions had been different in a specified respect. Such probing for hypothetical states can never tap subjects' memories for their cognitive processes, since the information was never in memory. In still other studies, subjects were asked, explicitly or implicitly, to summarize or generalize the processes they used, rather than to report concretely the processes used on each trial.

Several articles have been published making similar criticisms of the Nisbett and Wilson (1977) paper, and raising other objections as well. Of particular interest are the papers discussing the problems with verbal reports in between-group designs. (Smith & Miller, 1978). Some recent studies have shown that in corresponding within-group studies, subjects are able to provide veridical verbal reports (White, 1980, Weitz & Wright, 1979; Wright & Rip, 1980).

In sum, we disagree with Nisbett and Wilson's interpretation that subjects simply were not aware of relevant information during the critical experiments. Instead, we claim that better methods for probing for that awareness (concurrent or immediate retrospective reports) would yield considerable insight into the cognitive processes occurring in *most* of the studies discussed by Nisbett and Wilson. On the other hand, we agree with Nisbett and Wilson's analysis of subject's reports in situations where the subjects do not have access to or for other reasons don't rely on memory for the cognitive processes in question. In such situations, Nisbett and Wilson propose that an experimental subject infers the causes of his own behavior by relying on common-sense theories and observable

events–the same process that an observer would use to infer causes of behavior in an observed subject. By using experimental situations, where common-sense theory would lead to the incorrect assessment of causes, Nisbett and Wilson provide convincing evidence for their interpretation by showing that both experimental subjects and observers agree on the incorrect cause of the experimental subjects' behaviors. (For a nice presentation and extension of these arguments see Nisbett and Ross (1980).)

We think that Nisbett and Wilson's paper has been useful in forcing investigators like ourselves to think carefully about the relation of verbal reports to cognitive processes. Many verbal report procedures are justly faulted by their review. However, their results are consistent with our model of concurrent and immediate retrospective reports.

Concluding Remarks

Our examination of two of the most vigorous challenges to the usefulness of verbal reporting leaves intact our belief that such reports– especially concurrent reports, and retrospective reports *of specific cognitive processes*–provide powerful means for gaining information about such processes. The concurrent report reveals the sequence of information heeded by the subject without altering the cognitive process, while other kinds of verbal reports may change these processes. In retrospective reports of specific processes, subjects generally will actually retrieve the trace of the processes. In other forms of retrospective reporting, subjects, instead of recalling this information, may report information that they have inferred or otherwise generated. Hence, in the chapters that follow, we will pay particular attention to the two special forms of reporting–the one concurrent, the other retrospective–that are most likely to yield direct evidence of cognitive processes.

VERBAL REPORTS OF COGNITIVE STATES AND STRUCTURES

Although this book focuses upon cognitive processes, the model and concepts it employs can be extended to the non-cognitive aspects of verbal behaviors. There are several reasons for undertaking such an extension. It will permit us to identify common problems and issues in areas of psychology, like psychophysics, survey design, and measurement of per-

sonality traits, that traditionally have had little or no interaction with each other. In these areas, too, as in those we have been discussing, behaviorism has muted explicit examination of the status of verbal responses and reports.

First, we will propose a taxonomy of these other kinds of verbal reports, and will discuss briefly some examples of relevant research. Then we will consider two limited topics for more systematic discussion. The first of these is attitude assessment, the second is the historical development of verbal reporting, with particular emphasis on introspection. All of the verbal reports with which we will be concerned in this section are elicited by probes specifying what information is to be reported. Often, also, a set of alternatives is supplied from which the subject has to select a response.

Predictions from our model about the effects of verbal reporting on thought processes will depend on the circumstances under which the verbalizations are induced. We can classify verbalizations according to the memories that are tapped and according to the verbalization instructions the experimenter gives to the subjects. With respect to the memory source of the reported information, we can distinguish among (a) reports of stimuli that remain constant and available to the subject's senses while the report is being made, (b) reports of information retained in STM, and (c) reports of information from LTM. The next three subsections of this section will be devoted to the special problems that arise for each of these three kinds of reports.

Reporting of Sensory Stimuli

At any given moment, a large amount of external stimulation impinges on any human through the sensory receptors (visual, auditory, etc.), as well as from internal visceral sources. Normally this information is not heeded directly, but recognition processes access existing relevant LTM patterns, which provide higher-level descriptions and are in turn heeded. (In Chapter 3 we will discuss these recognition processes and their relation to attention in some detail.) In many circumstances attention can be directed toward the information in the sensory stores (cf. Kahneman, 1973). We can focus on marks on the page we are reading or listen for unusual faint sounds and so on. Many kinds of verbal reporting procedures rely directly on our ability to process sensory information selectively.

In most psychophysical studies, subjects are instructed as to the stimuli as well as the types of responses they will use. They are asked to rate how much pain the experience causes, how loud a certain stimulus is, how far away a certain stimulus is, and so on. This research has had a strong empirical emphasis and has been virtually unaffected by the drastic changes in theoretical views of mainstream experimental psychology. In our historical discussion, we will point to differences between the psychophysical methods and the analytic introspective methods, which also attempted to describe experiences in terms of the sensory units. Now we only want to sketch the relation of the psychophysical methods to our model of verbal reporting.

Since the primary goal of psychophysical research has been to describe the structural relation between physical stimulus and response, little attention has been paid to the mediating processes. However, selective attention is under attentional control and as reportable as is the final reponse. (The study cited earlier showing that subjects can substitute key pressing for verbal reports is a case in point.) The research methodology of psychophysics uses long sessions of trials to seek stable structural relations and highly automatized processes.

In a classic paper, Eriksen (1960) showed that the verbal report is the most sensitive index for basic perceptual processes, like discrimination. Hence, the results from psychophysical methods of report are quite consistent with our model of verbal reporting. We would, however, like to go a step further and argue that detailing the cognitive processes involved in generating psychophysical reports may prove quite useful. First, there is evidence that cognitive structures are involved even in simple judgments, like discrimination. A dramatic example is given by Binet (1969), who showed that the threshold for discriminating touch of two separate points of contact (compared to a single point of contact) could be reduced *10 times* by showing the subject the compass used. Second, different verbal instructions in judgments of size give different results (Carlson, 1977). Subjects give reliably different responses when asked to judge the objective size, the apparent size, and the size of the vertical projection of an object.

Converging support for the use of different cognitive processes in judgment of apparent and objective size was obtained by Epstein and Broota (1975), who found objective size judgments to be slower and a linear function of the distance to the stimulus object, whereas apparent size judgments were faster and unrelated to distance. Brunswik (1956) shows that instructing subjects to analyze the stimulus, as well as asking

them to "be so certain that they could bet on the actual size," clearly influences the judgments. Finally, and probably most important, the observed improvement of psychophysical judgments with practice (see Gibson, 1969) appears to implicate cognitive mechanisms (see Chapter 3). Some recent results by Ericsson and Faivre (1982) show that performance in a perceptual learning experiment can be best described in terms of the acquisition of cognitive structures indentified from retrospective verbal reports.

A related class of learning situations involve control of body functions–like heart rate, audiomotor performance–through biofeedback. In a very interesting review, Roberts and Marlin (1979) discuss the fairly extensive research (with conflicting results) on how reported awareness mediates development of control of these body functions. They define *veridical content* as verbally reported information making reference to "activities or perceptual events that are correlated with target behavior and therefore with feedback presentation" (Roberts and Marlin, 1979, p. 81), and discuss circumstances favorable to the generation of such reports. They point to two main biasing sources. Instructions in these biofeedback tasks often explicitly tell subjects to avoid certain strategies, like regulation of breathing rate. Other instructions give subjects incorrect information (e.g., that rate of heart beats is unrelated to rate of breathing). It is clear that such instructions will bias the subject against reporting such information regardless of their thoughts. These studies indicate that subjects can report the strategies they use for achieving control of visceral functions.

In a subsequent study, Roberts, Marlin, Keleher, and Williams (1982) provide some supportive evidence for the claims of validity of verbal reports made by Roberts and Marlin (1979). In most other studies, subjects are informed what visceral function is to be controlled, but this information may induce inferential processing, and also eliminates the possibility of using statements on what function is involved to validate the verbalized thought. In two studies described by Roberts et al., (1982), subjects were not told which visceral functions were involved and were simply shown an indicator of the function to be controlled. Subjects gave written descriptions of how they achieved control immediately after training. All subjects developing control over the visceral function invariably showed evidence for accurate self-report regarding their processing, as assessed by blind judges of written descriptions.

Our framework for analyzing verbal reports also applies well to psychophysical experiments. Similar methodological and theoretical

issues arise in the two domains, especially with regard to the instructions given subjects. Moreover, there is evidence of subjects' awareness of process even in the reports from these "simple" and "basic" psychophysical tasks.

Reports of Information in STM

Next we will review briefly some types of verbal reporting from STM that are closely related to those already discussed, but which have been used so frequently that they have emerged as separate procedures with separate literatures.

In *thought sampling* an attempt is made to get data on subjects' thoughts while they are performing their daily activities. Subjects are given a portable tone generator, which generates tones at random times. When a tone sounds, the subjects are to stop their normal activity and write down their thoughts, and perhaps additional information.

Genest and Turk (1981) provide a nice review of the emerging research using this method. In most cases the method is non-directive, requesting a report of the heeded thought at the time the tone was heard. Yet, the report is retrospective and often a fair amount of time will intervene before the subject can make his written record. Kendall and Korgeski (1979) propose that subjects should be provided with portable tape recorders so reporting will be more immediate and less disruptive. Genest and Turk (1981) also discuss *event recording*, where subjects are asked to record all instances of a certain type of thought. It is not unlikely that such instructions will lead to conscious monitoring and increase the frequency of thoughts of the observed kind. Unlike thought sampling, event recording is mostly used with maladaptive thoughts, with the aim of identifying their content rather than measuring their frequency.

Another widely used technique is *thought listing*, where the subject is asked to write down all thoughts that occurred during an interval. This technique is in many cases indistinguishable from the retrospective reporting discussed earlier. It is different in emphasizing thoughts as distinguishable elements. Where thoughts are elicited through associations to externally presented information, and are relatively disconnected from each other, one would expect reporting thoughts to be easy and unambiguous. Reports of the lists of thoughts from an interconnected thought activity like mental multiplication will undoubtedly be more difficult. In a nice review, Cacioppo and Petty (1981) note that most of the

studies using thought listing have studied thoughts evoked by persuasive communication.

Reports of Information in LTM

Subjects are often asked to report information that has no relation to their immediately preceding thoughts. The general format is to ask the subject a question and often also to provide a set of alternative answers. According to our model, the subject needs to comprehend the question and retrieve relevant information from memory. Retrieval can in some cases proceed directly from comprehension of the question (i.e., "In what year were you born?"). More often the subject needs to generate retrieval cues to access relevant memory traces (i.e., "How many times have you been to a movie theater in the last two months?"). In Chapter 3 we will consider in more detail the process of retrieval of information and in Chapter 5 we discuss studies of the retrieval process using protocol analysis. Here we wish to show that simply by asking by what processes the subject can make his responses we can arrive at some useful conclusions about these matters.

Our model makes a major distinction between information directly stored in memory and information that is generated and produced. The first class comprises factual information and information about experiences and perceived events and behavior in past situations. The second class comprises information about reactions and behavior in hypothetical situations, including general and abstractly described situations.

Reports of Past Experience. When we ask somebody to report something they should know and the report is not accurate, we may be inclined to distrust the method of asking (i.e., the verbal report). Such evidence is, of course, particularly damaging if we lack methods to validate even occasionally the reported information. In surveys, subjects are often asked many different questions. One question that is fairly easy to validate is "How old are you?". Some studies have shown the reported information to be invalid in as many as 83% of cases (Parry & Crossley, 1950). At first glance, that may be rather surprising, as most people should know their age. Invalid reports might indicate premeditated lying, which of course is always a possibility. But asking for somebody's age is unfortunate, as age changes each year. If we rely on direct retrieval of our age, we may access information stored earlier which is no longer

valid. Bjork (1978) has shown that a similar analysis can account for experiences of children appearing to grow very fast or parents aging very fast. When we see the child or the parent, we access an image of them, which was not the most recently seen image but one stored at an earlier time, hence the too big difference between perception and image. It is, of course, possible to derive one's age from one's birthdate, but the calculation requires mental effort, and can lead to errors and attempts to estimate the answer. This is especially true when the subject does not perceive the need to be completely accurate. Asking for somebody's birthdate would be much better as it remains fixed.

In other cases, the invalidity of reported information can be traced to issues of definition. In answering how many rooms they have in their house or apartment, subjects may differ in their ideas of what constitutes a room. Karlton and Schuman (1980) cite a study of the English census that showed that people were accurate in reporting the number of rooms according to their own definitions, but they simply did not use the census definition.

The problems in obtaining valid reports become more pronounced if the subject doesn't have the relevant information readily accessible in memory. When we ask subjects how often they have been to the doctor or the dentist, experienced various forms of crimes, or made airplane trips during some specified time interval, we would expect them to retrieve all these instances from memory, and attempt to verify that they occurred during the given time interval. However, if only the number of instances is to be reported, we have no way to monitor the subjects' retrieval activity and they may estimate rather than recall the instances.

When subjects are asked to recall instances, investigators have found the retrieved information to be valid. The common error appears to be inability to date instances and hence to determine whether they occurred within the given time interval. For highly salient and retrievable instances, this may lead to overreporting. By asking subjects to recall instances before as well as after the critical time period, such overreporting can be virtually eliminated.

A multitude of issues surround the use of fixed response alternatives to questions (see Schuman and Presser (1981) for an extensive review). In the ideal case, the subject retrieves his response and selects the appropriate response alternative. The results from studies using open-ended questions and fixed responses should then be very similar, but in many cases they are not. In an interesting analysis, Schuman and Presser

(1981) showed that the main source of discrepancy was the unavailability of certain alternatives. By constructing the fixed alternatives from the open-ended responses in a preliminary study they showed that much closer correspondence could be obtained between the two types of responses in a subsequent study. In fact, providing the set of relevant alternatives may reduce retrieval failures and hence enhance the validity of responses. We will talk later about possible effects of bringing to mind certain kinds of information that the subject may not have thought of otherwise. It is interesting that in Schuman and Presser's (1981) study, subjects recalling the most preferred aspect of a job gave most responses with the same frequencies as when they selected the responses from alternatives.

The concern for achieving accurate recall of information is quite explicit in current survey research. Karlton and Schuman (1980) review three methods used by Cannell and his colleagues to achieve more accurate reporting. First, the subject should be given an explicit instruction to recall accurately. We know that people do better when they think carefully about each question, search their memory, and take their time in answering. People also do better if they give exact answers, and give as much information as they can. This includes important things as well as things which may seem small or unimportant (From Cannell et al., 1981, reviewed in Karlton and Schuman (1980, p. 16)). Second, the interviewer should give more sensitive feedback and, in particular, monitor the retrieval process. For example, when subjects gives quick responses, they should be encouraged to think and retrieve more. Last, the interviewer should try to get the subject to make an explicit agreement to respond accurately and completely.

Reports of Hypothetical and General Information. Verbal reports that do not specify a clear relation to retrievable experiences, events, or knowledge are of several kinds. We want to distinguish verbal reports on reactions or behavior in *hypothetical situations* from verbal reports on reactions or behavior towards persons, ideas, and experience *in general* without specification of more specific context or situation.

Occasionally, we find verbal reports about hypothetical situations used in experimental psychology. For example, in a study by Reed and Johnsen (1977), subjects were asked how they would solve a problem if it were presented to them again. Subjects in a study by Nisbett and Wilson (1977) were asked how they would react to a story if some passages had not been presented. However, the most frequent and important use has been opinion-polls, surveys and personality and attitude assessment.

Personality and attitude assessment will be discussed later in a separate section.

We cannot offer a detailed model of the cognitive processes that generate a response to an attitude or opinion question. In fact, given our model, it is quite puzzling how somebody can access and integrate the multitude of relevant aspects and experiences at the time of the question. (The situation is, of course, quite different when the assessment has already been made prior to the question and can be directly accessed.) A possible view is that the question or statement serves as a retrieval cue to access a small subset of selected information, which is evaluated and used as a basis for responding. The consistency of accessed information and response to the same statement at different times will be determined by the organization of LTM–a point we will discuss in more detail in Chapter 3.

A review of the literature shows that such a simple association model has some support, especially for attitudes and opinions that are moderate and refer to non-central issues. Even when people are responding repeatedly to the same items within a relatively short time interval like a year, intercorrelations are relatively low (around 0.40) (Schuman & Presser, 1981). The principal exceptions are strong attitudes to central issues. The most likely locus of the variability between test occasions is in the information accessed.

More direct evidence for the selective cueing of information comes from the extensive body of research showing effects of wording questions in different ways. For example, subjects are much more willing "not to allow" public speeches against democracy than to "forbid" such speeches. Schuman and Presser (1981) shows similar effects for "not allowing" vs. "forbidding" other activities. Even in laboratory studies where subjects are exposed to the same events and information, the wording of the question (i.e., How long was the film? vs. How short was the film?), yields reliable differences, even when the same response alternatives were used. Although we lack evidence about what information was accessed, the direction of the influence is consistent with the hypothesis of selective access of information.

In a situation where subjects' attitudes and opinions are measured, the retrieved information will not simply reflect the current question, for information retrieved on preceding questions will be more accessible and more likely to be retrieved if similar cues are reinstated. The procedure used to study the influence of answering preceding questions is to manipulate the order of presentation of questions and compare the

responses to the same question in the different orders. Although most questions or items do not show such order effects, or at least sufficiently large effects to be statistically reliable, there are several examples where the effects are quite large. Schuman and Presser (1981) discussed two items where the interpretation of the effect appears quite straightforward.

Communist reporter item:
Do you think the United States should let Communist newspaper reporters from other countries come in here and send back to their papers the news as they see it?

American reporter item:
Do you think a Communist country like Russia should let American newspaper reporters come in and send back to America the news as they see it?

The effect of interest is that subjects are more likely to be favorable to letting foreign reporters operate in the USA if they have previously answered the item regarding letting American reporters operate in Communist countries (see Schuman and Presser (1981) for a comprehensive review of similar effects). Interesting effects of previous questions are also shown in a recent study by Bishop, Oldendick, and Tuchfarber (1982). They were able to show that subjects' assessed general interest in politics and the coming election was markedly influenced by preceding questions on facts regarding the election, like names of candidates for President, and voting records of their representatives in Congress.

In two experiments Bishop et al. (1982) showed separately that easy preceding questions (tapping information most people know) led to higher assessed interest levels, and hard preceding questions led to a decrease in assessed interest levels. The effects interacted with the knowledge people had about politics and the election. Highly knowledgeable subjects were unaffected by preceding easy questions, whereas subjects with less information were affected. Hard questions had a rather uniform effect of reducing subjects' assessed interest in the election and politics.

Admittedly, our discussion of cognitive processes in these unstructured verbal report situations is rather speculative. As far as we know, almost no attempts have been made to determine with the aid of verbal reports what information is accessed in such situations and by what processes. However, there is evidence suggesting that this would be feasible. First, open-ended questions where subjects are asked to name aspects and issues have been quite successful. Second, Schuman (1966) showed that asking subjects to elaborate their closed-choice selections for

randomly selected items (random probe technique) was useful in evaluating understanding of the item when it was translated and used in a different culture. Finally, some of the tasks used by Karl Buhler (1908a,b) are rather similar to deciding one's opinion on a statement. He found subjects able to provide informative sequences of thoughts about their opinions. A more complete empirical analysis of the thought processes should yield interesting implications for improvement and redesign of the methods used to assess attitudes, opinions and other general constructs. Below we will consider in more detail psychological research concerned with assessing attitudes and its relation to verbal reporting.

VERBAL REPORTS IN ASSESSMENT STUDIES

In discussing various forms of verbal reporting that claim to elicit currently heeded information or cognitive structures that remain in memory, we have indicated how several different kinds of cognitive processes might generate the reported information. One of our major assertions is that verbal reports can be, and should be, understood in exactly the same way as we understand other kinds of responses.

As a concrete application of our approach to a major area of psychological research, let us see what a first-pass analysis of verbal reports in questionnaire-answering might yield. In particular, let us look at the assessment research aimed at measuring and describing individual differences, especially differences in cognitive structure. This research has adopted many of the ideas advocated by Watson. The approach has been to search for aspects of behavior that remain invariant over some class of situations, yet discriminate one individual from another. Collecting observations and discovering behavioral regularities (especially with the help of correlational techniques) have been emphasized over theoretical analysis. Much of the research has been directed towards useful "real-world" applications: selecting people for education, jobs, and various forms of clinical treatment.

It has been customary to interpret invariant aspects of behavior that are found by assessment in terms of postulated internal states or traits. General abilities, like numerical skill, are associated with traits as are many stable aspects of personality, like aggressiveness. In the approach described above, which we will term indirect assessment, the invariant structures are induced from many specific observations.

Within the same framework, attempts have been made to gain in-

formation about these general invariant structures more directly. Instead of time-consuming observation of subjects' behaviors in concrete situations (e.g., while selecting food dishes), one could ask for their reactions to a verbal description of a general class of situations (e.g., "Do you like to eat fish?") and thus seek to access general preferences directly.

Alternatively, one can ask subjects after they have exhibited some behavior why they did it, hoping to receive a report of a general motive that could explain or predict their behavior over a wide range of situations. This research has taken a basically empirical approach to finding behavioral regularities, and it has not attempted to specify the cognitive mechanisms that both generate behavior and are accessible for verbal questioning. Implicit reference is made to the common-sense notion that people are aware and rational and therefore able to answer questions about the cognitive structures responsible for their overt behavior.

Data

In consistency with the behaviorist viewpoint, the research on direct and indirect assessment has not collected observations and data about the processes that generate the target behavior. Concurrent verbalizations, retrospective reports, and latencies have been collected and analyzed only very sparingly.

Indirect assessment methods have been used primarily and most successfully to assess cognitive abilities. In ability tests a sample of representative tasks for the ability in question is generated, and the subjects' responses are evaluated for correctness. A given subject's ability is then assessed in terms of his or her pattern of success on the items. The cognitive structures are inferred only indirectly.

To create representative situations that will elicit behavior reflecting hypothesized traits–like preference, aggressiveness, and other personality-based characteristics–is much harder. Some research has employed observers to record subjects' behaviors in natural environments or in semi-controlled group interactions. However, most of the research has employed more direct assessment procedures. We have distinguished four methods that produce different kinds of data for assessing traits. Some of these distinctions have been proposed by Olson (1976).

1. In the first type of assessment procedure, observers who study the subjects' behavior in one or several types of situations afterwards make direct estimates of the "levels" of certain traits for the subjects. These ratings by the observers constitute the data.

2. In the second type of procedure, the subjects' memories are probed for their previous behavior or covert reactions in particular classes of situations. The subjects are generally asked to respond with one alternative from a predetermined set, which asks about the frequency (e.g., "never," "occasionally," "often," etc.) of the behavior.

3. A third type of procedure obtains subjects' reactions to verbal stimuli or their predicted actions in general situations verbally described. We give an example of a stimulus and an excerpt from an instruction given to subjects (taken from Mischel, 1968, pp. 61-62):

> I enjoy social gatherings just to be with people. (Item from California Psychological Inventory.)
>
> Your first impression is generally the best so work quickly and don't be concerned about duplications, contradictions or being exact. (From instructions to the Leary Interpersonal Check List.)

The data in this case are the categories (e.g., "very much," "not at all," etc.) that the subject selects, and the responses are not conceptualized as introspective reports about the associated traits.

4. In the fourth type of procedure, subjects are asked for explanations of, or motives for, their observed behavior. When the subjects are asked *how* they were thinking throughout the experiment or *why* they exhibited particular behavior, the experimenter seeks to learn directly from them the underlying cognitive structure that produced the overt behavior.

Effectiveness of Assessment

Indirect assessment of cognitive abilities has been found to be very successful in predicting behavioral differences in real-world situations. This stands in rather stark contrast with the controversies surrounding assessment (direct or indirect) of personality traits and direct assessment of cognitive structures in general. Correlations between different methods and tests for assessing the same personality traits are often unsatisfactorily low (e.g., Campbell & Fiske, 1959). Reported reasons for behavior are unrelated to experimentally induced variations in behavior (e.g., Nisbett & Wilson, 1977). Reported attitudes do not correspond to actual behavior (e.g., Calder & Ross, 1973; Schuman & Johnson, 1976; Wicker, 1969).

In trying to explain the relative success of indirect assessment of cognitive abilities as compared with other types of indirect assessment, we consider the differences in the cognitive processes involved and the cognitive structures accessed. Unfortunately, there has been little research directed toward uncovering processes, most analyses having been made on data representing the final result of the processes. Lacking extensive data on the relevant processes, we will proceed by making assumptions derived from other areas of cognition where more such data have been gathered. First, we will note some general differences between the kinds of cognitive processes evoked by ability tests and the kinds evoked by personality tests and direct assessments. Then, we will turn to a discussion based on a more detailed explication of the cognitive processes underlying probes to assess cognitive structures directly.

Processes Evoked by Assessment

There are at least three marked differences between the cognitive processes evoked by items in an ability test and those evoked by items in a test to assess traits by self-reports. The first difference concerns the likelihood that the relevant cognitive processes and structures are actually evoked or accessed. The "first impression" requested by the instruction, quoted above, for self-reports and the emphasis that responses are neither right nor wrong, stand in stark contrast to the instructions for ability tests, where responses are considered carefully before being produced and are either correct or incorrect. In order to generate the correct answer or response in an ability test, the information has to be processed carefully by the relevant sequence of operations. The probability that a subject can generate the answer by guessing or some short-circuiting procedure like first impression, is small. There is, thus, much more experimental control for ability items than for self-report items over the cognitive processes and structures that are activated to generate a response.

For self-report items, uninteresting response processes (like always agreeing with the statement or simply selecting the socially desirable alternative) often appear to account for a sizeable fraction of the variance. However, by careful selection of questionnaire items, more recent studies have reduced the extent to which subjects can rely on such criteria as social desirability in choosing their answers.

The second difference between ability tests and others derives from the relation between the test item and the "real" situation in which the

actual non-test behavior occurs. Many symbolic tasks occurring in the "real" situations, like arithmetic, are rather accurately represented by a test item. By contrast, verbal description of a social interactive situation is usually a rather poor representation of the situation, which fails to communicate many essential aspects that would influence the actual non-test behavior. We will return to this issue later.

Memories versus Inferences

The frequent low correlations between different assessment tests of the same trait provide grounds for scepticism about verbal reports.

> Inaccuracy has been attributed mainly to a variety of distorting motivational forces, including deliberate faking, lack of insight, and unconscious defensive reactions, all of which presumably produce inaccurate self-descriptions (Mischel, 1968, p. 69).

However, Mischel (1968) points to research supporting other possibilites, which are consistent with the notion that subjects are able to describe and predict specific behavior with minimal interpretation of its meaning. The self-reports described so far require the subjects' global interpretations of their own general behavior patterns, rather than descriptions of specific behavior. Likewise, the attributes assessed by observers are mostly high-level traits that require considerable inference. Correlations could be low between different assessments because they elicited different inferences rather than because of conflict of evidence at the level of description of specific behavior. In support of this view, Mischel (1968) cites research showing that inter-coder reliabilities increase rapidly as the necessity for complex inference decreases. For further discussion of the special problems of assessments by observers see Fiske (1978) and Mischel (1968).

Remembered versus Anticipated Behaviors. In a more detailed analysis of the cognitive processes occuring in assessment procedures, we need to distinguish probing subjects' memories for past processes and occurrences of acts and reactions, on the one hand, from probing for subjects' anticipated responses in verbally described situations or to described classes of objects or people.

First we will address probing for subjects' memories. We will assume that information about specific past behavior and covert reactions is generally stored in episodic form in LTM. This implies that a statement

about occurrences of a certain kind will require access to the memory of all relevant specific occurrences of that kind. For simplicity, we will not discuss the exceptions, where subjects have been asked similar questions before, or have, by their own reflective activity, already generated the corresponding general information, which then can be accessed directly.

From general research on recall, we know that ability to recall specific events–especially with detailed information–deteriorates rapidly with time (see Cannell & Kahneman, 1968). Recall depends very much on the availability of retrieval cues. Since general verbal descriptions of classes of events most often will be insufficient as cues for retrieving specific events, subjects will have to supply additional information to generate more specific cues, like the relevant time period and specific situations in which the activity might have occurred. If the experimenter specifies the relevant time period and particular type of events to be recalled, recall increases considerably (e.g., Biderman, 1967). This type of recall is very time consuming and can hardly take place in the time allotted for filling out a questionnaire, unless the relevant episodes were few and easily retrieved because of recency.

If the subjects were able and motivated to retrieve all relevant episodes, they would face the problem of converting the information into fixed alternatives, like "often," "frequently," etc. Mischel (1968) cites a study by Simpson, that demonstrated that a wide range of percentages were associated with such words, when presented out of context. For example, one fourth of Simpson's subjects associated "frequently" with events occurring over 80% of the time, whereas another fourth associated it with events occurring less than 40% of the time. The processing activity that would be needed for accurate responses to questions about past overt and covert behavior–given the limits of recallability–appears to be incompatible with the relatively fast responses requested.

Causes of Behavior. Let us now turn to the questioning of subjects about the reasons or causes of their behavior. In terms of our model, legitimate probes for reasons and motives for observed behavior in a given process are just one kind of cue for retrieving information selectively from the memory trace of that process. From studies of current verbalization of heeded information, we know that subjects often generate goals in solving problems, hypotheses in concept-formation experiments, and evaluations in decision making. It should be possible to elicit these by probes of *why* a specific overt behavior occurred.

One should not assume that the subjects can assess directly that specific responses were "caused" indirectly by more general goals or

hypotheses. Cognitive processes often involve attention to specific information, which is *not* a specification of heeded general structures like goals. Information is heeded in other cases as a result of direct recognition processes without any intermediate states entering consciousness. In these cases the subject cannot answer a *why* question by direct retrieval from memory.

Much of the research cited by Nisbett and Wilson (1977) and reviewed above concerns experiments where the subjects have been questioned about a long series of experimental trials. When subjects are asked about their average behavior or motives, they obviously cannot answer the questions by retrieving a single motive or episodic memory. The behavior on different trials may correspond to very different cognitive processes, and it may in any event be difficult to retrieve them all from memory. Therefore, it is reasonable to assume that the subject either infers general motives or processes from retrieved selected episodic memories, or tries to rationalize his behavior using other sources of information than the memory of the processes.

Smith and Miller (1978) noted that in many of the experiments cited by Nisbett and Wilson the subjects were asked why their behavior in one condition of the experiment differed from other subjects' behavior in other conditions of the experiment. In such a situation, it is not clear to subjects that their memory is relevant for answering the question, as shown by the following initial step of a typical dialogue:

> Question: I notice that you took more shock than average. Why do you suppose you did?
>
> Typical answer: Gee, I don't really know . . . Well, I used to build radios and stuff when I was 13 or 14, and maybe I got used to electric shock. (Nisbett & Wilson, 1977, p. 237)

The subject appears to understand the assertion to mean that he took more shock than other subjects in the *same* condition, and he therefore probed his memory for explanations that would be independent of the situation, and hence of his processing activity. If the subject, to give a valid report, has to rely on his memory for his earlier processing, it would be necessary for him to have experienced *both* experimental conditions to explain any differences in behavior between them. Inferring what one would do in a new situation should not be confounded with reporting actual memory of completed processes.

Predictive Responses. In the case of asking subjects for their reactions to classes of persons or objects or their expectations of their behavior in verbally described situations, we have little data on what cog-

nitive processes and structures are evoked. It is most plausible to assume that the subject forms some kind of representation or "image" of what is verbally described, and uses this to determine his hypothetical reaction or behavior. LaPiere (1934) questioned the extent to which subjects in many situations are able to represent internally the crucial aspects of the verbally described situations. Any such failure will make their conceived behavior different from actual nontest behavior.

> Thus from a hundred or a thousand reponses to the question "Would you get up to give an Armenian woman your seat in a street car?" the investigator derives the "attitude" of non-Armenian males towards Armenian females. Now the question may be constructed with elaborate skill and hidden with consumate cunning in a maze of supplementary or even irrelevant questions yet all that has been obtained is a symbolic response to a symbolic situation. The words "Armenian woman" do not constitute an Armenian woman of flesh and blood, who might be tall or squat, fat or thin, old or young, well or poorly dressed–who might, in fact, be a goddess or just another old and dirty hag. And the questionnaire response, whether it be "yes" or "no," is but a verbal reaction and this does not involve rising from the seat or stolidly avoiding the hurt eyes of the hypothetical woman and the derogatory stares of other street-car occupants. (LaPiere, 1934, p. 230)

In his classic study, LaPiere (1934) studied attitudes and behavior towards Orientals. Six months after a large number of hotels and restaurants had been visited by an Oriental couple, the same places were sent a questionnaire with the question, "Will you accept members of the Chinese race as guests in your establishment?" The overwhelming majority of the places visited answered "no," with a smaller number saying "under some circumstances." Similar disassociation of verbal responses to symbolic situations from real behavior has been found by, for example, Kutner, Wilkins, and Yarrow (1952).

In information processing terms, LaPiere's hypothesis is that, in the cases where the generated internal representaton contains all relevant aspects appropriately portrayed as in the "real" situation, the behavior and verbally reported behavior will be consistent. When the "real" situation is more or less symbolic, as in the case of voting, accurate predictions can usually be made for actual behavior on an aggregate level from verbal reactions to questions (see Schuman & Johnson, 1976). Similarly, Katona (1975, 1979) has found that sampled subjects' reports of their ex-

pectations of future prices, future income, and so on, give valid information for predicting changes in purchasing behavior for the general population to which they belong. Ajzen and Fishbein (1977) show in a recent review that when the attitude measurement situation corresponds closely to the situation in which the behavior to be predicted occurs, high agreement between attitudes and behavior is found. Fazio and Zanna (1978) have found that extended direct experience with specific entities leads to better defined attitudes (and stable internal representations evoked by the questionnaire items), which can better predict subsequent behavior. When the information in focus of attention is taken into account, attitudes appear to be consistent with each other and with behavior (Taylor & Fiske, 1978).

This brief overview of controversies about direct assessment by verbal probing and questioning shows clearly that a detailed model of cognitive processes and cognitive structures is needed for making decisions on when and how to use this type of assessment procedure.

We know of only two studies that collected concurrent reports (Schneider-Duker & Schneider, 1977) or retrospective reports (Kuncel, 1973) for thinking during responses to personality tests. Although the results from these studies are promising, much more must be done to understand how personality tests should be constructed to measure cognitive structures.

HISTORY OF VERBAL REPORTS AND INTROSPECTION

A good test of the adequacy and usefulness of our analysis of cognitive processes involved in verbal reporting is to see whether such an analysis can shed light on why some forms of verbal report, like introspection, were problematic, while other forms of verbal report, like psychophysics judgments, gave uniform and accepted results. This discussion of the early forms of verbalization will show that many of the difficulties arose from the requirements imposed on subjects in generating the reports.

Early speculations about the human mind and human subjective experiences were closely related to religious and philosophical questions about the nature of man. The human mind was generally viewed as beyond understanding in scientific terms. However, individual philosophers did attempt to inquire about the mechanisms responsible for

acquiring new knowledge and the correspondence between the external world and subjective experience.

The basic source of information for these inquiries was observation by philosophers of their own cognitive processes–that is, introspection. The analyses were directed towards very general issues and questions about the mechanisms and structure of human mind, and were primarily speculative, with little concern for establishing empirical support for the proposed ideas. Speculations and self-observations were inextricably mixed, for they were all the products of the same individual. Although many of the proposals for mechanisms became influential in subsequent theorizing, this type of inquiry gradually became suspect as not conforming to scientific method.

One could observe a similar pattern of speculation for extending our knowledge about the physical environment before a distinctive scientific approach emerged to the analysis of physical phenomena. The scientific approach distinguishes between facts and theories, regarding as facts only "indisputable" observations. Methods of controlled observation and experimental manipulation are essential components of the scientific method. It was several centuries after the emergence of the natural sciences before scientific methods began to be applied to the study of mind and human behavior.

Considerable effort has been devoted in psychology, as in other sciences, to specifying what constitutes "indisputable evidence." Since all observations are made by humans, it was important to secure general agreement on what kinds of observations reflect the external world rather than idiosyncracies of the individual observer. Complex assessments were questioned or discarded as empirical evidence, for they were judged to embody inferences and knowledge not shared by all observers. Complex assessments were also thought to be sensitive to the expectations and subjective biases of observers. By contrast, simple perceptual judgments based on sensory qualities, like colors, were found to be invariant over different observers and, in principle, independent of such biasing factors as differences in knowledge and earlier experience.

Introspection

In the early years of psychology, the direct observation of mind in operation was taken as the primary method for obtaining information about the mind and its contents. William James used introspection (broadly

construed) naturally and unself-consciously as a major tool of investigation.

> *Introspective Observation is what we have to rely on first and foremost and always.* The word introspection need hardly be defined–it means, of course, the looking into our own minds and reporting what we there discover. (James, 1890, p. 185)

Another pioneer, Binet, went so far as to make the definition of psychology contingent in terms of the introspective method.

> Introspection is the basis of psychology; it characterizes psychology in so precise a way that every study which is made by introspection deserves to be called psychological, while every study which is made by another method belongs to some other science. (In Titchener, 1912b, p. 429)

At the turn of the century there was a consensus about the value of naive introspection.

> We need not hesitate to admit, on the other hand, that a roughly phenomenological account, a description of consciousness, as it shows itself to common sense, may be useful or even necessary as a starting-point of a truly psychological description. (Titchener, 1912c, p. 490)

However, as we shall see, naive introspection was soon deemed to be as unscientific as casual observation of natural events would be for the natural sciences. In order to provide facts about the mind, more rigorous and systematic methods of introspection were required.

Structuralism. The main aim of Titchener's research was to gather facts about consciousness (the content of mind), and in the process to uncover its structure. The facts consisted of subjects' direct descriptions of consciousness, whereas inferences and generalizations based on conscious experiences were not accepted.

> But the data of introspection are never themselves explanatory; they tell us nothing of mental causation, or of physiological dependence, or of genetic derivation. The ideal introspective report is an accurate description, made in the interest of psychology, of some conscious process. Causation, dependence, development are then matters of inference. (Titchener, 1912c, p. 486)

Titchener proposed to separate theory from facts by letting the subjects only describe their experienced conscious content, leaving the inferential process to the experimenter.

To the question of how the contents of consciousness should be reported, Titchener proposed a description in terms of the sensory components of thought. There appear to be at least two partly different reasons for this choice. The first is theory-based and should be seen as a hypothesis. Titchener, like Wundt, held the hypothesis that all mental states and experiences could be described in terms of their sensory and imaginal components. Wundt's thesis was that human experience of external stimulation has two phases. First, the invariant sensory attributes of the stimulation are immediately experienced. Then, mediating processes occur, relating the sensory stimulation to existing general knowledge and prior experiences. According to Wundt, it is the result of the second phase that constitutes the cognitive phenomena we call consciousness.

Wundt assumed that we are born with the sensory components of the first stage already fixed, and that they remain unchanged throughout life. Changes in the way we experience the same sensory stimulation are due to changes in the associations evoked by the stimulation (i.e., are the result of the second stage). Assuming that all knowledge is ultimately derived from experience (the assumption of Locke's empiricism), and that all experience corresponds to a conglomerate of sensations, it follows that the structure of mind and consciousness, including thought, could be described in terms of sensory components. In the search for intersubjective invariants and general psychological laws, it was therefore natural to concentrate on the structure of the immediate sensations.

The second reason for Titchener's choice of a vocabulary of consciousness is basically methodological, and derives from the difficulty of transmitting the conscious experience without contaminating it through words with imprecise meanings.

> I quote an illustration from Titchener; a half-trained student reports in an experiment a feeling of "perplexity." Now perplexity is clearly a complex experience. A group of processes is present, some of which we can experience in other contexts, disjoined from each other. True, I have a fair idea of what he has experienced. But only a *fair* idea. The description should be so full and complete that one can imaginatively or sympathetically reconstruct the experience. (English, 1921, p. 406)

Titchener's proposal was that consciousness should be described in terms of its elementary components.

> By the "description" of an object we mean an account so full and so definite that one to whom the object itself is unfamiliar can nevertheless, given skill and materials, reconstruct it from the verbal formula. Every discriminable part or feature of the object is unambiguously named; there is a one-to-one correlation of symbols and the empirical items symbolised; and the logical order of the specifications is the order of easiest reconstruction. This, then, is what we mean by "description" in psychology. (Titchener, 1912a, p. 165)

This procedure is analogous to transmitting a picture as a pattern of dots–as on a TV screen–where no biasing semantic descriptors are required. The analogy may be considered a fair approximation to Titchener's idea, for he says "the record must be photographically accurate" (Titchener, 1909). This view harmonizes well with the conservative criteria for simple perceptual observations used in the natural sciences, and with the notion that introspection is analogous to inspection in physics, but with consciousness as its target of observation.

In their efforts to find the elementary units of thinking, the structuralists searched not only for the elements of thought-content, but also for the elementary processes involved in thinking. Relatively early, Wundt started to pursue research along the lines of Donders, who is seen as the pioneer in the analysis of cognitive processes by means of observed latencies. Donders' central idea was that more complex processes could be viewed as compounded additively from simple reactions and the other cognitive processes. Three different tasks were proposed by Donders to estimate the durations of the most basic cognitive processes (i.e., stimulus discrimination and response selection). The simplest is *simple reaction time*, where the subject responds with a given single response, like a button-press, as soon as a stimulus is presented (*a-reaction*). In the *c-reaction* the subject responds only to a certain type of stimulus with a given single response. The c-reaction was assumed to differ from the a-reaction by requiring an initial discrimination of the stimulus. For the *b-reaction* the subject responds for each stimulus with a different response, and thus is required not only to discriminate but also to select the correct response. Wundt extended this method by proposing an additional reaction that we will discuss in the next section.

Data. Titchener relied primarily on introspective reports given after the completion of the processes, but the latencies of the cognitive processes were also used in his analyses. The introspective reports requested by Titchener were very different from the phenomenal accounts provided by naive introspection. Subjects required extensive practice to

break away from their habits of giving phenomenal accounts. They had an initial tendency to commit the "stimulus error," which was to report information reflecting previous experience and knowledge from the second stage (for example, to report "seeing a book"), instead of reporting the sensory and imaginal components of the thought or presented stimulation. The extent of training required is indicated by Boring (1953), who mentions that Wundt required his subjects to have 10,000 supervised practice trials before they could participate in any real experiments.

In the Structuralist view, the contents of the self-observations or introspections are considered to be facts or data. From an information processing point of view, on the other hand, the fact or datum is that a subject *said or reported "X"*. In the former interpretation we are obliged to trust that the subject is honest and capable and that the words and the sentences are understood in the same way by the subject and the experimenter. In the latter interpretation, it is sufficient to reproduce or account for the report or aspects of it. Taking it literally as an observation is just one of many alternative interpretations.

Another crucial aspect of classical introspection is that in the direct description of the sensory components it wasn't obvious what were to be taken as the elementary units of sensation. Much introspective research activity was devoted, therefore, to determining the characteristics of these units. In this kind of analysis the observers made decisions about which of several proposals for sensory units correctly reported direct judgments and evaluations of hypotheses. This kind of introspective analysis is very different from the direct description advocated by Titchener, and was also particularly plagued by extensive disagreements between different laboratories.

Latencies of cognitive processes were considered interesting as a separate source of data on the structure of thought processes. Donders' proposal, discussed earlier, for three types of reactions was extended by Wundt. He suggested that the c-reaction, where the subject gave a fixed response to only a certain type of stimulus, involved not only a discrimination but also a choice of whether to respond or not. As a consequence of this criticism, Wundt proposed the *d-reaction*, in which the subjects respond as soon as they have made a cognitive discrimination of the stimulus. As the subjects didn't have to make a choice to respond or not (they always responded, as discrimination of a stimulus invariably occurs) this d-reaction would be a pure measure of the time taken to discriminate or to cognize the stimulus.

Issues and Discussion. Titchener's type of introspection was severely criticized on at least two major counts. The Wuerzburgers and the Gestalt psychologists claimed that many aspects of consciousness could not be reduced to sensory and imaginal components, and that, consequently, the method of analytic introspection was inadequate and should be replaced with phenomenal reports. In addition, the researchers at Wuerzburg collected phenomenal evidence rejecting the assumptions underlying the subtraction method for measuring the duration of cognitive processes.

The behaviorists with Watson reacted against the direct observation of consciousness, and claimed that only observable behavior could be used as facts or data. Watson pointed out the lack of reproducibility of analytic introspections from different laboratories (i.e., disagreements on issues like "existence of imageless thought," "whether the primary colors are three or four" and "which are the fundamental attributes of visual sensation"). At the same time he acknowledged the reliable and robust results obtained by introspection in psychophysics. These two lines of critique suggested other methods of study, which we will consider later. First, we will discuss why the difficulties with analytic introspection of thought did not prevent reliable results from being obtained in psychophysical studies. Then we will review briefly the unsuccessful attempts of the structuralists to measure the speed and duration of the basic cognitive processes.

Analytic Introspections. We wish now to describe and reinterpret in information processing terms the cognitive processes involved in making analytic introspections and observations of the sensory and imaginal components of thought. Unfortunately, there is very little explicit discussion of these processes by the introspectionists themselves, and our explication will therefore be partly inferred. The first phase hypothesized by the structuralists, involving the sensory attributes, appears to be very similar to the processes atttributed to the sensory stores in the human information processing model. Classical introspection was aimed at describing the contents of these sensory stores at discrete time intervals, like photographic snapshots (to be interpreted generally to include nonvisual sensations and imagery).

> Observation, as we have said above, implies two things: attention to the phenomena, and record of the phenomena. The attention must be held at the highest possible degree of concentration; the record must be photographically accurate. (Titchener, 1909, p. 24)

From the point of view of attention this means that the subject, if he can, must redirect attention intentionally from the spontaneously emerging thought content of STM to a single sensory store in order to register rapidly the active sensory components. Let us assume for the moment that this is possible. Each recognized pattern in STM would correspond to a very large number of independent sensory components, which would all have to be retained in STM or stored in LTM until they could be reported. However, storage of information in LTM with usable retrieval cues requires considerable time–estimated at 8 seconds for each chunk (Simon, 1979)–which would basically exclude the possibility of storage in LTM in this case.

Span of attention was known by contemporary research to be limited to a small number of elements (see Woodworth, 1938). It was possible to retain much more information if familiar patterns or organizations were recognized, yet encoding in such patterns would violate the notion of a description directly in terms of sensory and imaginal components. This raises the question of how all these sensory components could be registered and then stored awaiting their reporting, as reporting is known to take considerable time.

> Strange to say, a ten-second period of thinking sometimes required as many minutes to recount and make clear to E. (Woodworth, 1938, p. 783)

Evaluating the completeness, objectivity, and veridicality of the "psychological description" of thought contents raises serious methodological problems, since the experimenter lacks external control of, and independent access to, the thought content described. One answer to the problem of the brief availability of thought content is tachistoscopic presentation of visual stimuli. By providing experimental control over the stimuli, this technique allows assessment of the veridicality and accuracy of the "psychological descriptions." In a noted study in 1904, Kuelpe (Chapman, 1932) found that with tachistoscopic presentations of colored letters, an instruction to report certain aspects first (e.g., the colors of the letters) caused a serious decrement in the subsequent reportability of other aspects of the stimulus (e.g., the positions of the letters).

Kuelpe's study doesn't discriminate between incomplete encoding of the stimuli and decay of memory for the information that wasn't reported immediately. In a later study Chapman (1932) demonstrated that informing the subjects about the aspect to be reported prior to the tachistoscopic presentation yielded more accurate reports than informing

the subjects immediately after the stimulus was presented. Still more recently, Sperling (1960) measured the duration of these initial "iconic" recordings of sensory stimuli, and demonstrated that though they endured only a fraction of a second, their content exceeded the capacity of STM. (The units of reporting in Sperling's studies were not elementary sensory components, but letters or digits.)

Clearly, then, reportability depends on what information is heeded, and hence upon the task (Aufgabe). Not only is the capacity for retaining information limited, but ability to report it can be affected by an initial bias to search for particular information. Introspective reports are subject to several sorts of selective bias including the theoretically based training of observers, the uncontrolled use of questions (Humphrey, 1951), and the fact that subjects (often faculty and graduate students) are often not naive to the hypotheses addressed in these studies (Comstock, 1921). Taking these possible biases into account, it becomes difficult to accept the reports as scientifically valid evidence. In fact, it was proposed in the case of imageless thought that the observers simply overlooked the actual images and kinesthetic sensations,

> ...so quick is the process of thought and so completely is the attention of the subject likely to be concentrated on meaning. We have a parallel case in the neglect of after-images and double images ... in everyday experience when other things are in the focus of attention. (Comstock, 1921, p. 211)

Psychophysical Judgments. In contrast with the dubiousness of the method of analytic introspection, high reliability is usually imputed to the results obtained from introspective analysis of psychophysical relations. Yet the standard data in psychophysics are introspections. The explanation for the difference is simple; the experimental situation for making psychophysical judgments of sensory stimuli is very different from the one described above. The observer is instructed in advance when to attend and what to attend to; the stimulus is simple and presented over an extended interval of time. Moreover, the judgments, generally being comparative, are reports of highly encoded stimuli that say nothing about the raw sensory components. Essentially, no additional memory is required for the observation before it can be reported. On the basis of these differences, it is not difficult to accept psychophysical introspections as reliable, but to reject analytic introspections.

Latencies. The Structuralists' research on latencies was criticized on basically the same grounds as was analytic introspection. Some initial research with Wundt's d-reaction, where the subjects responded as soon

as they had discriminated the stimulus, gave very reliable estimates for the times taken to cognize stimuli. Then a series of studies (see Woodworth, 1938) showed the d-reaction to take as much time as the simple reaction. Berger (Woodworth, 1938) explained these results by pointing out that the response in the d-reaction is independent of the discrimination, in that the subject *always* responds, as in simple reactions. Hence, there is no objective criterion to assure that the subject waits until the stimulus is discriminated before responding. In fact, unless the subject is to make his motor response contingent on the result of the discrimination, there seems to be no way to ensure that the motor response is not initiated earlier and in parallel with the perceptual processes. Again it appears that subjects were asked to do an impossible task.

Analyses of latencies were discarded on more general grounds when evidence was found against Donders' crucial assumption that the stages of discrimination and response selection in the b- and c-reaction were simply inserted additively in the a-reaction. Ach and Watt from the Wuerzburg laboratory found from retrospective reports that the processes of preparing for these several reactions were very different in terms of what was attended to prior to the presentation of the stimuli (Woodworth, 1938). These different types of reactions should thus be seen as wholly distinct procedures, and the differences in duration among them could not be used as estimates of the durations of unique component cognitive processes.

Watson's Attack on Introspection

Just at a time when the classical introspectionists were becoming increasingly self-conscious about methodological issues (Titchener, 1912a, 1912b, 1913), Watson (1913), in the influential paper "Psychology as the Behaviorist Views it," launched a total attack on the study of consciousness. He criticized the introspective method and its results, and argued that psychology, as a natural science, could do without introspective data and mental constructs.

It is important to note that Watson's (1913) critique is not directed against all uses of verbal reports as data, but specifically against the analytic methods and results of the classical introspectionists. When he points to the lack of reproducibility of analytic introspections from different laboratories, he refers to the issue of "imageless thought," "whether the primary colors are three or four" and "which the fun-

damental attributes of visual sensations are." Watson is even more disturbed that laboratories try to discredit opposing evidence by attributing it to lack of training of the observers in the competing laboratories.

Although Watson did not mention Comstock's (1921) objection that the observers were not, in general, naive to the hypotheses under study, he did stress the additional problem of communicating meaning. How can we be sure that the introspecting observer uses language in the same way as the interpreting experimenter? Especially when an observer is learning new distinctions of consciousness without any feedback or objective control, there is a problem of ensuring common reference between observer and experimenter. Watson (1920) argues, with evidence, that the introspective verbal report is untrustworthy for scientific purposes.

> After having made as searching analysis as we like upon several players' playing of golf, what will be left out of the individuals' own accounts? Again suppose we take down their overt responses to any questions we may ask and incorporate them into our record. They are of relatively little value. No one since objective studies upon golf have been made trusts the verbal report of a golf player. He will tell you that he never takes his eyes off the ball when making a stroke. The camera shows that he is a prevaricator. (Watson, 1920, pp. 100-101)

It should be noted that the kind of questioning illustrated by this example does not refer to the subject's memory of a specific instance, but to how he thinks he performs activities in general when he is asked about them. Watson made a clear distinction between analytic classical introspection, verbal questioning of a subject, and thinking aloud. His views on the veridicality of the latter kind of verbal report were quite different from his views on the first two. In fact, of course, his view was that thinking consisted primarily of subvocal speech (Watson, 1924), and to give evidence on this point, Watson (1920) demonstrated that thinking can be made overt.

> The present writer has often felt that a good deal more can be learned about the psychology of thinking by making subjects think aloud about definite problems, than by trusting to the unscientific method of introspection. (Watson, 1920, p. 91)

After presenting the first documented analysis of thinking-aloud activity, Watson (1920) summarizes his arguments for the opinion just quoted –that the overt verbalizations in TA correspond to the normally

covert thought activity–by making reference to observations from numerous individuals thinking aloud while working problems. Watson was quite clear about distinctions among modes of verbalization that have since become muddled. These distinctions were also quite apparent to the Gestalt successors of classical introspectionism. In their phenomenological observations, naive subjects were used, and the subjects were allowed to give their own spontaneous descriptions in their own language.

The behaviorists' suspicion of verbal reports was reinforced by their emphasis upon overt performance rather than mediating processes. Even if introspective information was not necessarily incorrect and uninformative, it was unnecessary and could be replaced by appropriate behavioral measures (Watson, 1913). With this point of view, questions of the adequacy and validity of verbal reports, and of methods for obtaining them, were simply irrelevant. It is not surprising, therefore, that this methodology was not studied extensively.

Later Views

When Woodworth (1938), twenty years later, discussed verbal reports, he emphasized the distinction between describing thoughts and expressing them. In response to Titchener's notion of excluding meaning from reports, he presented a case for a more direct and natural reference to complete thoughts (Woodworth, 1938, p. 785):

> Even though reference to the object is a very incomplete description of a particular instant of experience, a series of such statements does describe the *general course* of a thinking process–just as naming the towns through which you have driven maps the route you have taken. If O reports "I thought of A, and B, of C, noticed that I was drifting away from the problem and went back to A," he gives a picture of the course of his thinking (Selz, 1913).

And as a more concrete illustration of the type of verbal report he had in mind, he gave (Woodworth, 1938, p. 786) the following example from Binet (1903, p. 14):

> I thought of the pump in the garden which someone was operating and said to myself that it must be the cook, then I heard a rooster crow and thought of this rooster.
>
> I asked myself whether Polly would be willing to lend me her