Foundations of Cognitive Science

edited by

Michael I. Posner

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

7 Experimental Methods in Cognitive Science

Gordon H. Bower and John P. Clapper

7.1 General Introduction

Cognitive science is a multidisciplinary field with diverse goals and intellectual agenda. Of the fields contributing to cognitive science, it is cognitive psychology that makes primary use of experimental methods to answer its questions. We do not argue that reliable knowledge can only be obtained through experimentation. Science is not identical with experimentation; otherwise astronomy or anthropology could not be sciences, which they manifestly are. Rather each intellectual discipline has its own rules for justifying its claims, hypotheses, and theories. Cognitive science is a somewhat uneasy marriage of the differing methodologies and styles of justification of its contributing disciplines. The relevance of experimental observations to cognitive science largely depends on what *claims* are being made for the "psychological reality" of the principles under discussion—that is, whether the authors propose their claims as descriptions of human mentation.

The goal of sciences is to build theories that enable the explanation, prediction, and control of events in their domain of inquiry. Such theories are like mental models of their empirical domain; they postulate an interlocking network of theoretical concepts to explain and unify known facts and occasionally guide the discovery of new, previously unobserved regularities. Theories are systems of abstract concepts whose properties and rules of operation *correspond* (are analogous) to some empirical system. For us to have confidence in this correspondence, the observations on which the theory is based should be accurate and reliable. Evaluating a theory depends on checking whether its implications are true.

Many facts about the human mind are apparent from introspection, and no special methods seem necessary to uncover them. But introspection alone has had a dismal record of failure as a primary method in psychology, a field that began with only the introspective method a hundred years ago and then abandoned it after years of frustration because of its variability, reactive unreliability, and frequent invalidity. Much of the data obtained by introspection are heavily influenced by

the observers' theoretical preconceptions (Nisbett and Wilson 1977, Nisbett and Ross 1980). Moreover many cognitive processes go on outside of awareness (Lewicki 1986) or occur far too rapidly to be available for conscious report. Thus a science of human cognition can benefit from special techniques for observing, recording, and interpreting mental events. There are a variety of empirical techniques that can yield quite reliable knowledge. We briefly discuss two of them before moving on to experimental methods.

Naturalistic Observation

Naturalistic observation refers to the systematic observation and recording of the behavior of some organism (or social unit) as it occurs in a somewhat "natural setting" without any attempt by the observer to intervene. Developmental psychologists, for example, have conducted many naturalistic studies of infants to describe the normative (average) maturation sequence of physical abilities—at what average age infants can sit up, walk, babble, and so on. Computer-simulation theorists who try to recreate in their program introspections of their own problem-solving (or question-answering) activities are doing a sort of naturalistic observation. So is the linguist who tries to generate a set of counter-example sentences to some proposed linguistic generalization and then tries to formulate a more adequate generalization to cover all the examples.

Although naturalistic observations can provide descriptive generalizations about a class of phenomena (say, the formation of plurals in English nouns), they are weak in supplying evidence for *cause-effect* relations. For instance, we might hypothesize that as the days grow shorter, a species of bird begins to migrate south. But simple observation alone cannot allow us to untangle all the factors other than day-length that might contribute to migration, such as temperature changes, the azimuth of the sun, and so on. Experiments could try to check for the causal power of length of daylight by *controlling* these other *confounding* factors.

Correlational Studies

Correlational studies have a more formal character than naturalistic observations in that they usually try to measure the degree of association (relatedness) of two or more events or attributes. A well-known example is the correlation of cigarette smoking with lung cancer: the amount and duration of smoking correlates with the likelihood of developing lung cancer. The weakness of the correlational approach is that correlation provides no clear evidence for inferring a cause-effect relationship.

A major problem is that one cannot rule out the possibility of an unknown third variable that is causing both of the other two factors to vary together. The tobacco industry has repeatedly argued that the

smoking-cancer correlation provides no evidence for causation, because there may be certain kinds of people who are both cancer prone and also readily addicted to cigarettes and would get cancer at the same rate whether or not they smoked. Even assuming that the correlated variables are causally related, the correlation itself goes both ways and does not itself provide any hint as to the *direction* of the causation. For example, elementary-school students who get better grades are liked more by their teachers, but which is cause and which is effect? Each cause-effect link is plausible on its own merits, or the two might reciprocally cause one another, or both might be caused by a third variable, such as the child's social skills. The point is that correlation does not imply causation, whereas causation almost always implies correlation.

Controlled Experiments

The deficiencies of correlational studies for arriving at conclusions about cause-effect relationships are rectified by the use of experiments. The basic idea of an experiment is very simple: one group of subjects is treated in one fashion, another group in a different fashion, and we measure whether their behavior differs as a consequence. If the two groups were equivalent in all other respects at the beginning of the experiment, then we can justifiably claim that any difference in their behavior at the end of the experiment can be viewed as an effect caused by the different treatments they received. This basic idea is so simple that it bears repetition: in experiments we compare observations under two conditions—an "experimental" condition, which has the crucial procedure, treatment, or factor introduced, versus a "control" condition, identical in all respects to the first group except that the experimental procedure, treatment, or factor is omitted. The comparison of the two conditions enables us to infer whether the experimental treatment causes a difference in behavior in the experimental group.

Experiments are usually conducted to test a specific *hypothesis* about the relation between two variables. The factor that the experimenter manipulates is referred to as the *independent variable*, whereas the behavior that is measured to detect any effects of this manipulation is called the *dependent variable*. Many experiments use several independent and dependent variables in complex arrangements, but the experiments can all be reduced to the same simple reasoning. The hypothesis is a *generalization* or universal statement about the causal relations between variables in the world and the conditions under which these relations can be expected to hold. In a cognitive-science experiment the hypothesis is usually a prediction that a particular change in the conditions under which subjects are observed will cause a specific change in their behavior. If the prediction fails, then the hypothesis should be rejected. On the other hand if the results accord with the hypothesis, then our confidence in that hypothesis will increase (although it is still open to disconfirmation from further experiments).

Measuring Experimental Effects As indicated, experiments are set up to measure changes in people's behavior (cognitive performances) caused by manipulating a particular independent variable. To measure this influence on behavior, we should be able to describe the behavior quantitatively in countable units such as the number of milliseconds required to make a decision or the proportion of answers of a given type. Although qualitative observations (for example, introspective protocols) are useful preliminaries, they should be replaced whenever possible by quantitative measures that can be statistically summarized and compared across experimental conditions. Anyone who has ever confronted the Herculean task of content coding and comparing a large number of unstructured "think-aloud" protocols will appreciate the utility of having behaviors classified into countable categories.

Isolating Causal Effects It is a blunt fact of life that even with constant conditions human behavior is quite variable, both within a given subject as well as between subjects. Consequently it is usually not very informative to compare single observations, either from different subjects or even from the same subject in different conditions. In the face of statistical variability between subjects and conditions, a single subject's behavior provides little assurance of the causal impact of the treatment. Investigators are therefore often forced to examine the behavior of groups of subjects (say, ten to twenty or so subjects). In a *between-subjects* experimental design different subjects are tested in each condition and the average scores obtained from the different groups are compared. In *within-subjects* designs each subject is run in all conditions, then the differences in subjects' performance across conditions are examined to see whether the subjects perform better under some conditions than they do under other conditions. (Which type of experimental design is best for a particular situation depends on many practical factors; see Winer 1971 for details.) By testing groups of subjects in either of these arrangements, extraneous sources of variability such as individual differences in abilities or strategies may be expected to cancel out overall, allowing valid comparisons to be made.

But even when groups of subjects are tested, it is not enough to simply observe that the average performance in one condition exceeds that in another, because any such outcome might also have arisen simply by chance. For instance, in a highly variable performance such as reading comprehension or text memory, the fact that subjects score slightly higher in one condition than another could easily be a random, chance outcome. To cope with the difficulties introduced by such variability, social scientists rely on statistical procedures that evaluate observed differences with respect to that behavioral measure's baseline variability. Statistical procedures allow a precise estimate to be made of how likely an apparent effect is to have occurred by chance alone. Only if the probability of a difference that large occurring by chance is very

small (by convention less than 5 percent) will investigators reject the null hypothesis of "no effect" and report a "positive" finding (for details see statistics texts such as Winer 1971).

The goal of an experiment is to so arrange circumstances that when interpreting the findings one can exclude all plausible alternative hypotheses (or causes of the effect). Thus to conclude that an observed difference between groups is caused by the experimental manipulation rather than some other, extraneous factor, it is important that the conditions differ *only* in the level of the independent variable. (In the world outside the laboratory of course many factors will vary together, which is why it is difficult to justify causal arguments from informal observations alone.) Experimenters go to great lengths to arrange circumstances so that they can be sure that only the independent variable changes across conditions. Standard precautions include randomly assigning subjects to conditions (or testing each subject in all conditions, when this is possible), not informing subjects fully about the experimenter's hypotheses, using equivalent tests in each condition, and so on. Equating groups can often be difficult simply because we may not know all the factors that might vary and cause experimental subjects' behavior to differ from the control condition. As we learn more about important causative factors for a given performance, the complexity and kinds of controls experimenters must arrange to study a *new* variable grow.

The important factors to control differ somewhat across content areas (language, memory, perception, and the like), and accordingly each has its own standard task configurations and experimental designs. An experimental design is a procedure for assigning treatments (procedures or potential causes) to subjects in such a way that we can reach valid inferences about causal relationships. Designs differ along many dimensions, such as the number of causal factors being studied at once, how many "levels" (or values) of each factor are being studied, and whether subjects are tested in all, some, or only one condition. In turn each standard experimental design calls for a particular type of statistical analysis with which to evaluate its results (for details see Winer 1971).

Coordinating Theory with Observables

Experiments examine the relationships between categories of *observable* events, such as the effects of consuming a stimulant on reading comprehension or the influence of just prior colors on perception of a current color. Many lawful generalizations simply relate categories of observable events; examples of such descriptive laws abound in the physical sciences and are occasionally found in the biological and social sciences. Obviously such empirical regularities are useful for controlling and predicting some behavioral phenomena.

A more cherished goal of most scientists, however, is to be able to explain why the empirical law holds. They usually do this by postulating theoretical constructs that do *not* correspond directly to categories of

observable events. For instance, mentalistic constructs such as goals, beliefs, and intentions are not directly observable, but they have strong intuitive appeal and are used frequently in cognitive theorizing. Other common theoretical constructs are memory stores such as short-term and long-term memory, various data structures such as propositions and images, mental processes such as encoding and retrieval, or syntactic and semantic analysis during parsing. None of these are simply "categories of observable events."

Such theoretical constructs provide simple and coherent explanations for a diverse range of empirical phenomena. The goal is to achieve a simpler and more elegant theory than would be possible were scientists to restrict themselves to discussing only categories of observable events. To take just one example, a broad range of cognitive performances are powerfully influenced by the extent to which subjects *attend* to the task at hand. As subjects "pay more attention" to a given task, they usually perform it better. We cannot directly observe attention. However, the construct of attention simplifies a large number of observable relationships. It captures common effects resulting from many dissimilar categories of observable operations, such as orienting instructions, payoffs for good performance, time of day, stimulant drugs, and any other factors that might influence attention.

Positing a link between a theoretical construct and a set of outcomes is obviously more parsimonious than enumerating separate links between many observable independent variables and many dependent variables. It also provides a more *coherent* theory, because the single construct of attention captures the relationship between many categories of observable operations that would not be apparent by simply enumerating them separately. Further the theoretical construct invites extensions; once we learn that some new independent variable affects attention in a particular way, we already know how it will affect the whole range of other behaviors that depend on attention.

Most of the variables of interest to cognitive scientists are unobservable, perhaps because human behavior is so complex that simple descriptive laws are not easily achieved. This fact creates a gap when we wish to check the correspondence between our cognitive theory and human cognition in the laboratory. Because experiments deal with observable events, terms in the theory must be coordinated to observable stimuli, responses, and events in the experimental setting. We can indirectly manipulate a theoretical variable by altering observable factors that are presumed to affect it (for example, induce "thirst" by having the subject eat salty crackers); similarly a theoretical variable can be measured indirectly through the observable behaviors that it affects (for example, the amount of water subjects drink indexes their thirst). In this way observable variables can be used as proxies for unobservable variables in experiments, making experimental testing possible.

To illustrate this process of "operationalizing" theoretical variables,

we review a study by Gluck and Bower (1988). They applied a "connectionist" model to the behavior of human subjects learning to examine the medical symptoms displayed by patients and then diagnose them as having one of several diseases. Connectionist models consist of an interconnected collection of neuronlike computing units, typically divided into a sensory input layer, a motor output layer, and zero, one, or more intermediate layers (see figure 7.1). Information in the form of stimulus activation of units passes forward from the input to the output layer via a set of connections. The response of the system to a given input is determined by the weights (amplifier values) of the connections among the various units. The system is trained to respond properly to a set of stimulus patterns by repeatedly presenting the stimuli one at a time and adjusting the weights between units so that the network gives the correct response to that input pattern.

In Gluck and Bower's experiments 20 college-student subjects saw a series of 250 patients, each characterized by one to four medical symptoms (such as runny nose, stomach cramps, or high fever). The subject first classified each patient as having one or the other of two fictitious diseases ("burlosis" or "mydosis"), was next told the correct disease for that patient, and then proceeded to the next patient. Over the course of the experiment the subjects learned the degree to which the four symptoms were more or less diagnostic of the two diseases. At the end of the experiment the subjects directly estimated the conditional probability of each disease given only knowledge that a patient had a specific symptom (without knowing about any other symptoms).

In applying the connectionist theory Gluck and Bower chose the simplest correspondences possible (see figure 7.1). Presentation of each medical symptom in a given patient was coordinated to activation of a corresponding single element in the sensory input layer, so there were four input units in total. Two response output units were postulated,

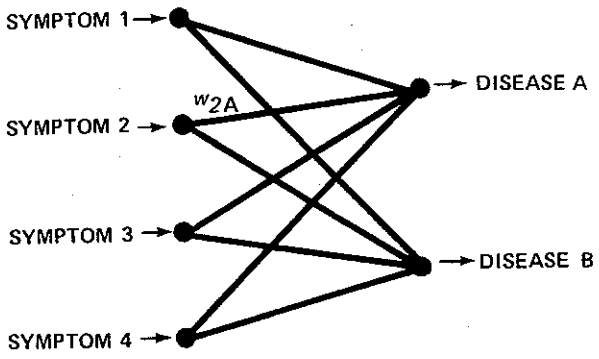

Figure 7.1 A simple connectionist network that learns to diagnose patterns of up to four symptoms as having one of two diseases. The specific w shown denotes the weight of evidence that presence of symptom 2 contributes toward disease A. (From Gluck and Bower 1988. Used with permission.)

corresponding to the two disease categories in the experiment. Each sensory unit was connected directly to both output units, so no intermediate units were assumed. The weights were interpreted as the association strengths from each input unit to each output unit. Each patient corresponded to a pattern of activation (1 or 0) on the four sensory units, which led in turn to a weighted sum of activation on the two response-output ("disease") units. The probability of the subject choosing to classify the patient as having a given disease was assumed to be a logistic function of the difference in activation of the two output units. The more activation on output unit 1 versus 2, the greater the supposed probability that subjects would classify the patient as having disease 1 rather than 2. The theory of Gluck and Bower specified how the connection weights would change trial by trial in adjusting to the corrective feedback. Thus for given symptom-to-disease correlations the theory implied differing strengths for the specific symptom-to-disease connections depending on the training conditions. Gluck and Bower found that the theoretically predicted ordering of association weights was quite accurate across several experiments. That success increased the plausibility of the theory as a whole.

Suppose, however, that the predictions had failed. Then one could question parts of either the theory or the correspondences between theoretical and observable terms set up when applying the theory. Perhaps one sensory unit for each symptom is too simplistic, and we could do better by identifying a symptom with a pattern of activation over a large set of sensory-input units, with different symptoms sharing some input units. Perhaps we need a different output rule or need to postulate one or more layers of hidden units intermediate between the input and output units. Or perhaps the learning rule for adjusting weights trial by trial is wrong and needs to be corrected. Clearly there are many places in the total system where we could assign credit for the failed predictions. Sometimes the misfit can be remedied by altering one coordination. At other times no simple modification in the application seems to rectify the problem, at which point we conclude that the theory is simply not applicable to this situation. Such misfits are often extremely informative, however, in telling us in what way the actual behavior deviates from the idealized system of our model. In any event the applicability of the theory becomes narrowed and its credibility as a whole will be weakened.

The symptom-disease coordination has been described because it illustrates a relatively transparent case. More typical examples of theoretical constructs in cognitive science may seem more complex at first glance (for example, the situational model for a text), but usually reduce upon analysis to similar identifications. All such theoretical concepts must lead to some observational consequences; otherwise they have no empirical cash value. Usually observable indicators relate to a particular theoretical construct; we then try to validate inferences about a theo-

retical construct or a scenario of unobservable events by searching for *converging evidence* from several indicators or consequences of the construct or event. The more that different indicators agree, the more confidence is justified in our inferences about these unobservable events.

Experiments on Human Cognitive Processes

Experiments on cognition generally observe people's behavior as they are performing a specific task, such as perceiving, learning, judging, or remembering something. Many different tasks are used, but a conventional vocabulary has evolved for describing them. Subjects are presented with a *stimulus* in a particular *task context* and must *respond* to that stimulus in some way. For example, an experimenter might present subjects with line drawings of common objects and instruct them to name each object as rapidly as they can. Each picture serves as a "stimulus" and the subject's "response" is to produce a label for the picture; the instructions subjects have received, as well as background factors such as how fast the stimuli are presented, the method of presentation, and so on, can be regarded as part of the task context. The subjects' responses are jointly determined by the stimulus and the task context; if either of these is changed, say, by presenting a picture of a different object or using different instructions, then the subjects' responses change accordingly.

Since the inception of the information-processing approach, performance in cognitive tasks is commonly described by analogy to a computer program that carries out computations on input data and returns some output as a response. The presented stimulus corresponds to the program input, and the subject's response corresponds to the program output; the subject executes a mental "program" to compute the response from the stimulus. The task context is instrumental in determining the particular program the subject runs on the input. For instance, under one set of instructions a subject might respond to a picture of a dog with the corresponding verbal label; under another set of instructions he or she might judge whether it was the same as or different from some stimulus presented earlier.

Assuming cognitive performance can be represented as a kind of program, several types of questions can be asked about it (see, for example, J. R. Anderson 1987, Marr 1982). First, one can simply ask what *function* the program computes (that is, to characterize its input/output relations) or what are the properties of that function. For example, subjects in a typical psychophysics experiment might be presented with tones of various intensities and be asked to give direct magnitude estimates of the loudness of each. The aim of such experiments is to characterize the function that relates physical intensity to perceived loudness (in fact it is a simple power function; see Stevens 1957). A second question asks about the particular algorithm by which

a program function is computed. For example, when subjects multiply two three-digit numbers in their heads, we know that they are computing a multiplication function; what we wish to find out is the mental algorithm by which they compute it and how this algorithm changes with such factors as memory capacity, age, expertise, or intelligence. Because a given function could be computed by any of several algorithms, the cognitive scientist wants to discover which particular algorithm subjects use. For instance, researchers investigating the development of problem-solving skill in such domains as geometry or computer programming usually focus on the particular procedures subjects use to solve problems. Their actual solutions are interesting only to the extent that frequently occurring mistakes provide clues about which problem-solving strategies subjects of a given level of experience tend to use, and what their limitations are. A third question suggested by the computer-program analogy asks about the physical implementation of a given cognitive process in the nervous system. This question has traditionally been regarded as outside the province of cognitive science. As several authors in this book attest, however, cognitive scientists are now beginning to take questions of neural implementation more seriously and to investigate constraints this may place on theorizing at the cognitive level (see, for example, chapter 8 and J. A. Anderson 1983).

Cognitive psychologists believe that, by observing subjects' performance in various laboratory tasks, they can investigate basic properties of human cognition in situations sufficiently simple and transparent that those properties can be revealed. This approach rests on the assumption that humans have a reasonably small set of powerful, general-purpose cognitive operations and capabilities that are applied across a wide variety of situations (see, for example, chapter 2 and Newell and Simon 1972). Probably anyone who thinks that we can have a "science of the mind" must believe something like this. By using relatively simple tasks, it is hoped that these various abilities or operations can be isolated and studied systematically.

It is important, however, to understand that performance in some laboratory tasks often reflects the particular *strategies* subjects use to guide their behavior in that situation, as well as more fundamental cognitive processes. A strategy is like a special-purpose program or sequence of mental operations that the subject constructs to optimize performance in a specific situation. Although the mental components from which such internal programs are built might be basic cognitive structures, the strategies themselves may be quite idiosyncratic and their details may only be relevant to a particular laboratory setting. Because detailed investigation of particular strategies often yields little of general interest, investigators usually attempt to focus on those aspects of task performance that are controlled by basic, or nonstrategic, factors. In such cases the objective is not to fully characterize people's

behavior in a specific task situation but to use the task as a window to more fundamental properties of cognition. This approach is similar to analyzing the performance of a particular computer program to discover properties of the language in which it was written or the machine on which it is running, rather than to characterize that program per se.

Interest in strategies can be illustrated with an example. Much of what people do to solve anagram problems and similar brainteasers is highly strategic; to the extent that these strategies are specific to a particular problem domain, they tell us little about human behavior in general. Nevertheless experimenters may use such tasks to investigate general issues in problem solving. For example, people have severe limitations on how much information they can maintain in active memory. These memory limitations are a major cause of subjects' errors; they also constrain the set of acceptable strategies that subjects can use. Anagrams solved "in the head" can be used to study the role of working memory in problem solving by examining which factors affect the number and types of memory-based errors subjects make, how memory aids improve performance, and so on. We can sometimes identify general heuristics subjects use to solve a variety of problems, and these may generalize beyond the simple situations studied in the laboratory.

In some cases cognitive scientists are interested in studying a specific task itself (for example, text editing) rather than merely using it as a means for studying cognition. Educational research on math, reading, human/computer interaction, or applied memory might be seen as attempting to understand and improve skilled performance in particular task domains. Strategies can be of central interest in applied areas because they play such a large role in determining subjects' performance. A variety of techniques for the analysis of strategies has been developed; a detailed treatment is beyond the scope of this chapter, but see Ericsson and Simon 1980 or Sperling and Dosher 1986.

Characterizing Psychological Processes

Decomposing an empirical system into a set of hypothetical related components and then validating this decomposition through experiments is part of the goal of cognitive science. Most cognitive scientists treat the mind as a system that can be decomposed into a collection of more or less separable subsystems. At a general level this decomposition is reflected in the subdisciplines within the field, for example, the study of memory, language, attention, visual perception, reasoning, and emotion. In turn each of these subdisciplines can be decomposed into subdomains within which more elementary theoretical constructs play prominent roles.

Given some proposed theoretical distinction, say, between short-term and long-term memory or between encoding and comparison stages in a recognition memory task, how is one to evaluate the validity of this distinction? The basic criterion, of course, is that two things are mean-

ingfully different if they have different properties. To take an obviou case, it is useful to distinguish between visual and auditory sensor memories because they have different characteristics and function som what independently. Unfortunately most of the distinctions that co nitive scientists worry about are considerably less transparent than th example.

It is often difficult to decide whether constructs in two similar theorie have the same or different implications. Often the theories are too vagu for this issue to be sharply decided. The basic *empirical* criterion whether or not the proposed components behave sufficiently differentl in experiments to justify distinguishing them. This is ascertained b observing whether they are affected by different independent variable or whether they are affected in different ways by the same independer variable. For instance, psychologists often distinguish between recog nition and recall as involving different retrieval processes. Part of th argument for this distinction is that the two memory indices are affecte in different ways by certain independent variables. For instance, peopl show better recall for common words but better recognition memor for rare words (Crowder 1976).

A number of experimental methods have evolved for evaluating the oretical constructs and claims in cognitive psychology. To familiariz the reader with the basic logic that underlies these methods, we briefl discuss several of them.

Analyzing Representational Types A standard issue in cognitive sci ence is to specify the form in which particular information is represente in the mind. One small aspect of this general issue that has been heavil investigated is the way in which a discrete stimulus (such as a word o letter) is represented in memory immediately following its presentatio and how the form of this representation changes over time. Thes analyses assume that perception of a stimulus can be analyzed into a series (or cascade) of successive stages of encoding, with each stag providing a temporary "internal record" of the stimulus that can b read or utilized by other processes. The internal record (or representa tion) at each stage of the stimulus analysis is called a "code"; the majo goal of this research is to describe these internal codes and thei properties.

A typical theory (see, for example, Posner 1969) is that a discrete stimulus event first gives rise to a specific sensory code in the affected modality (visual, acoustic, and so forth) and that, depending on instructions, this initial encoding can later give rise to associated secondary o tertiary codes. Thus a visually presented form such as an alphabetic letter may be represented (coded) initially as a visual form, then by its name in a phonemic code, and then perhaps by some further classification (for example, as a type of consonant or vowel). Several codes may simultaneously coexist, but some codes typically become available

at an earlier stage of stimulus analysis than do others. Interest usually focuses on discriminating among the various codes that are formed and determining the time at which a particular coding of the stimulus becomes available.

Several techniques have been devised for attacking such questions. Although we will describe their application to simple letter stimuli, the reader should keep in mind that more complex stimuli, such as language samples or pictures of natural scenes, can also be studied using these techniques. A basic method asks subjects to make speeded recognition judgments of identity—that two stimuli are identical (match) at a specified level of description or categorization. Recognition of identity (under a given description) is a basic operation involved in many cognitive tasks. Regardless of the level of description the matching task can always be cast as a question that yields a simple Yes or No answer. By recording how long it takes subjects to answer such questions, we can make inferences about components of the overall recognition process.

One version of the matching task is to decide whether two successive forms are physically identical (for example, AA are, whereas Aa or Ba are not). A second task is to decide whether the two successive forms have the same name (for example, Bb do, whereas AB do not). In the name-identity task the "same" letters can either by physically identical (AA) or not (Aa). In this "same-name" task response time to physically identical pairs (AA's) averages about 70 to 100 milliseconds faster than to pairs having only the same name (Aa's) (Posner and Mitchell 1967). It was thus concluded that subjects were forming a code for the stimulus letters that preserved their physical features and could be used for carrying out a rapid physical match of the second letter to the first.

If the name code becomes available only after the physical code, then it should not affect the speed of processes that use the earlier, physical code. This independence was confirmed in that deciding whether two stimuli are physically different (versus identical) required the same amount of time for same-name letters (Aa, Bb) as for different-name letters (Ab, Ba). Thus the fact that two visual forms had the same name caused no interference whatsoever in deciding that they were not physically identical. Another bit of evidence that the name-code becomes available some time after the physical code comes from examining the kind of similarity that slows down the decisions. Chase and Posner (1965) found that physical-letter-matching speed is much slowed when the nonmatching lures are visually similar (OQ, GC, PR) but not when they have similar-sounding names (BC, DE, VT).

How long is a particular code maintained in active memory? This question can be investigated using the name-matching task. Posner and Keele (1967) imposed a brief delay (from 0 to 2 seconds) between presentation of the first and second letter. The advantage for the physically identical pair, which was about 90 milliseconds when the two letters occurred together (0-second delay), rapidly declines to only a 10-milli-

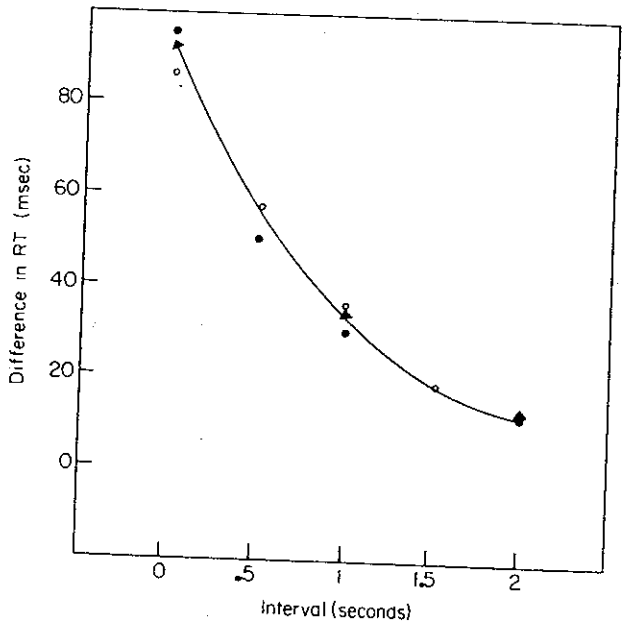

Figure 7.2 Difference in reaction time between name and physical identity "same" responses as a function of the interval between the two letters. The different point symbols represent results obtained under slightly different conditions of presentation of the letter stimuli. (From Posner 1969. Used with permission.)

second advantage by 2 seconds after the presentation of the first stimulus (figure 7.2). The interpretation is that in a name-matching task the subject converts the physical code as quickly as possible into a name code; because it is then no longer being actively maintained, the physical code decays rapidly. Presumably only the name of the first letter remains for comparison with the second letter, so that the advantage for the physical-identity match disappears. Furthermore, if subjects are required to rehearse a short list of letters that are phonemically similar to the first letter of the pair to be tested, then their decision times are slowed substantially. On the other hand this same interfering task produces no impairment when subjects are making physical-identity decisions (Boies, cited in Posner 1969). Thus holding irrelevant phonetic information active in memory interferes with short-term memory for the name of a probe letter but not with its physical code.

These cases illustrate several methods used in analyzing the nature of the memory code (representation) for simple stimulus events. Let us summarize them. We have discussed *judgment times* (for example, that B name-matches b) to infer codes; this method is based on the assumption that retrieval and matching is quicker the more closely the internal representation of the second stimulus matches the memory representation of the first stimulus.

A second set of methods are direct or indirect means of assessing the

similarity of two or more stimuli after they have been coded in a particular form. Internal representations are selective, leaving out some features of the full stimulus pattern (for example, a city map does not tell us the height and color of buildings). Two stimulus patterns that differ in features that are ignored (or squashed) by the coding scheme will have internal representations that appear more similar than the original full patterns. The trademark of similar stimuli is that they may be confused with one another. Thus one method for studying coding uses *confusion errors* in recall or recognition. For example, letters that have been coded phonemically will be confused in recall with other letters that sound like them. Encoding words and statements in terms of their meaning can lead to *false recognitions* of synonyms or paraphrases similar in meaning.

A third set of techniques exploits the fact that material that is encoded similarly to the target material will create greater *interference* in processing, or in accurately remembering, the target material. For example, people's ability to "shadow" (repeat immediately) one message and ignore a second, simultaneous message depends on their similarity along many perceptual and semantic dimensions, such as their voicing, pitch, source, ear of arrival, and topical overlap. Similarly people's ability to retain a set of target items in short-term memory despite attending to other material is better the more dissimilar the modality and encoding of the interfering material is to the target items.

A fourth method occasionally used to infer similarity of encoding is *clustering* of items in free recall. Free recall refers to unconstrained recall of a set of items in any order as they come to mind. When subjects freely recall a list of presented words that belong to taxonomic categories (names of furniture items, animals, cities, and so on), they tend to cluster (recall together) items belonging to different categories. The assumption is that these words were encoded during the learning phase of the experiment in terms of their meaning and category membership and that these relationships are then used to guide later recall. The method is quite general; in practically any unconstrained recall task subjects will tend to recall and cluster together items that they have encoded in similar ways. This method can be used to detect how subjects have subcategorized different domains of their topical knowledge.

Additive Factors Method

In Posner and Keele's (1967) task, subjects see an item, then after a variable delay they see a second item, which they compare with their memory of the first item. Saul Sternberg (1966, 1967) used a similar task in which a small *set* of items (say, H, P, Z) must be held in memory before the probe item is presented. The subject's task is to compare the probe item to the memory set and decide whether it matches some one of the items in a specified way (for example, has the same name). Sternberg found a nearly linear increase in decision time as the memory

set was increased from 1 to 6 elements. (This number of elements can be maintained errorlessly in short-term memory; with more than 6 elements significant errors occur for once-presented lists.) Much is known about this "memory-scanning" task.

Almost more important than the specific task that he introduced, Sternberg suggested a general experimental logic or method for dissecting the different processes involved in such performances. This is the "additive-factors" method, illustrated in figure 7.3. This schematic diagram depicts a series of stages in the processing of a probe item: first, the probe is encoded in a form suitable for comparison to the memory set, then the items in the memory set are retrieved and compared with the probe, with match or mismatch being decided according to a criterion specified by the experimenter (for example, "same name"). Finally, the response (match or no match) is determined on the basis of the accumulated comparisons. If the stages are executed independently and in series, then the total reaction time (RT) should be the additive sum of the times for each of the stages, that is, total RT = encoding time + comparison time + response time.

The additive-factors method provides a way to decide whether two independent variables that influence overall reaction time do so by affecting the same or different components of the model depicted in figure 7.3. The investigator attempts to manipulate independently the duration of different stages by using several independent variables, each of which affects only a single stage. This should be possible if the stages are really independent in terms of their processing times. Variables that affect the same stage may interact, but variables that affect different stages should have additive effects on total reaction time. (A computer analogy would be that the time to search computer memory for a given datum should be independent of how that statement was input to the CPU, whether from a keyboard, magnetic tape, or from memory.)

Sternberg used the additive-factors method to demonstrate that the several stages in his item-recognition task were independent, in the

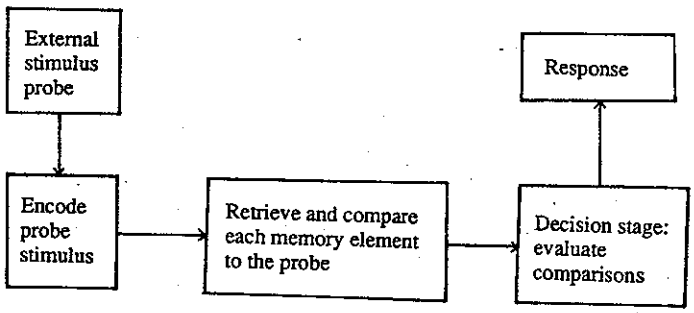

Figure 7.3 A proposed series of mental events that occur during each trial of Sternberg's memory-scanning experiment.

sense that the time required to complete each stage was independent of the time to complete any other stage. As an example, blurring the probe item by presenting it in visual noise increased total reaction times but did so by making the probe harder to encode. The critical point is that this encoding effect was independent of the effect of another variable, namely, memory-set size: degrading the probe had the same slowing effect regardless of how many items were in the memory set; similarly increasing the size of the memory set increased RTs the same amount regardless of whether the probe item was degraded. In other words degrading the probe elevated the intercept of the reaction-time function (by about 100 milliseconds), but did not affect the slope of that function relating reaction time to set size.

On the other hand the presence of interactions among variables signals that they are influencing the same stage in the model. For instance, visual degradation and stimulus probability (of a Yes versus No test probe) interact in their effects on the encoding stage, but neither variable interacts with memory load (Miller and Pachella 1973). This pattern of dissociations supports Sternberg's (1969) claim that probe encoding and memory comparison represent separate, independent stages in the overall recognition process. He argued that the existence of independent stages is empirically established by finding variables that affect each stage: variables from the same set should interact, but those from different sets should have additive effects. This method can be used generally in research on reading or memory retrieval, whenever a theory suggests a serial-stage analysis of some cognitive task.

The additive-factors method is particularly interesting because it provides an elegant and powerful technique for disentangling the component stages of a serial process. But its power carries the price of specialization, which is its somewhat limited applicability; many processes of interest do not fit the simple model of a sequence of clearly independent mental operations. Other methods have been devised for studying concurrent processes; we now discuss one of them.

Dual Tasks One of the fundamental propositions of cognitive science is that the *processing resources* of the mind (or any other cognitive system) are *limited*. The attention (processing resources) required by a task depends on its complexity, its points of decisional uncertainty, and how well practiced it is. It is assumed that the mind has available only a fixed amount of processing resources (analogous to computing cycles per second), and that we allocate more or less to the several tasks in which we are concurrently engaged so as to optimize the utility of our overall performance (see Sperling and Dosher 1986). The resource-allocation models are similar to schedules for prioritizing and assigning CPU cycles to different jobs in a time-sharing computer system.

Dual-task experiments are used to investigate how much processing capacity is required for particular tasks, and how people allocate re-

sources among tasks depending on payoffs. If performance on a given task gets worse the less attention is allotted to it, then we can use that task as an *indirect* measure of how much processing is required by another, concurrent task. Consider an example: as subjects read a text on which their comprehension will be tested, they may be asked simultaneously to listen for a soft tone that comes on at random, unpredictable intervals as they read. When they hear the tone, they are to press a button as quickly as possible. Their detection and average reaction time to the probe tone while reading are compared with those measures obtained when subjects are not reading but only listening for the tone. The increased RT during reading is presumed to be an indirect measure of the processing resources required to read and comprehend the text. If the text is difficult, involving many unfamiliar ideas and complex arguments, people will have to process it more deeply for comprehension, and consequently their signal-detection performance will suffer, as measured by their reaction times to a probe tone (see Kahneman 1973). Such dual-task experiments are common because cognitive scientists are often interested in how much "mental effort" is required by a given task. The logic and pitfalls of the procedure are detailed in Sperling and Dosher 1986.

Signal-Detection Theory In the aforementioned tasks the subject is trying to detect a weak signal, such as a soft tone or faint light. Accordingly this is an appropriate place to mention the standard theory of signal detection, which is widely used in studies of perception, discrimination, and memory. The theory best applies to experiments using discrete trials, on each of which a specified signal is either presented or not; the subject indicates at the end of each trial whether he or she thinks the signal was presented. In the paradigmatic case the signal is a simple tone added to white noise; on no-signal trials the white noise is presented alone. The experimental variables in such tasks are the intensity of the signal relative to the noise, the proportion of signal versus no-signal trials, the payoffs for correct responses, and the penalties for errors.

The theory of signal detection derives from statistical decision theory (Chernoff and Moses 1959, Wald 1950): each trial type, either signal-plus-noise or noise-alone, is assumed to give rise to a collection of internal sensory samples that are distributed according to normal bell-shaped probability curves, with the signal-plus-noise distribution having the higher mean. Subjects are presumed to reach their decision on a given trial by comparing their sensory sample to a criterion and deciding "Yes; signal" if the sample exceeds the criterion and "No; no-signal" otherwise. For given signal and noise distributions the choice of a criterion determines both the probability that a signal will be detected and that no-signal trials will be correctly identified. The criterion is under the subject's control. As the subject shifts the criterion, the

probabilities of correct hits and rejections are constrained to covary; for instance, if the criterion is lowered to detect weaker signals, the person will perforce begin making more "false positives," misidentifying noise-alone trials as "signal" trials. The important point is that perceptual sensitivity (the difference between the means of the two underlying distributions) depends on both performance indices (correct Yes and No responses), and the theory tells us how to carry out this estimation. For details, see Green and Swets (1966).

The framework of signal-detection theory is quite general, and it has been applied to a large number of judgment tasks in which subjects must discriminate among two or more classes of stimuli. Included would be discriminating any property of two auditory or two visual stimuli, the length of two time intervals, the correct pronunciation of two syllables, the genuine versus counterfeit nature of an author's writings, suicide notes, paintings, and so on. The framework has been extended especially to recognition-memory judgments, in which after exposure to a list of learning items, the subject judges a series of other items, deciding whether each is a repeat (or copy) of one of the learning items or is just a distracting lure (Parks 1966). In those cases the extent to which certain lures are "falsely recognized" as repeated old items tells us much about the way the initially learned items were coded and represented in memory.

Having introduced several general methods, we now look in detail at more specific methods in particular content areas. The major content areas of cognitive science comprise studies of perception, attention, learning, memory, skill learning, categorization, language processing, semantic (or world) knowledge, and problem solving. For specific experimental techniques in each one, see Kling and Riggs 1971 and Puff 1982, for example. We limit our discussion to just two of the domains, namely, studies of memory and of language processing because these comprise major foci of interest within cognitive science. Many of the issues and techniques from these two research domains can also be applied to other content areas.

7.2 Experimental Methods I: Learning and Memory

Key Issues in Memory Research

One set of research questions about memory concerns "implementational" issues (following J. R. Anderson 1987)—that is, the fundamental nature of the brain's basic memory mechanisms, regardless of the type of task or materials. Such research focuses on the basic processes for *encoding* information into memory and later *retrieving* it when needed, the nature of the *representational codes* in which this information is stored, whether memory is a unitary system or a collection of separate, more specialized subsystems, and similar issues. Also included here is the general topic of inductive learning, covering such diverse topics as

classical (Pavlovian) and instrumental (operant) conditioning, category learning, and social perception, all of which focus on discovering principles that determine how people abstract general knowledge from individual experiences.

A second major research area aims to characterize people's *knowledge* of the world—the organization of that knowledge in memory; how it is derived from experience; how sensory, motor, and conceptual features of the knowledge are interrelated; and so on. Much recent research on such topics as the structure of natural categories (see, for example, Rosch 1973, Smith and Medin 1981), schemas (see, for example, Bartlett 1932, Rumelhart and Ortony 1977, Brewer and Treyens 1981), and scripts (see, for example, Schank and Abelson, 1977, Bower, Black, and Turner 1979, Graesser, Woll, Kowalski, and Smith 1980) falls under this heading.

A third set of topics relates to more "applied" issues, such as educational tutoring, strategies for improving memory performance, the effects of various drugs on learning and memory, aging and memory, and so on. Of course findings in these areas frequently have strong theoretical implications.

Characterizing Memory Experiments

Memory tasks generally involve three phases: (1) an *acquisition* or encoding phase, during which the person first encounters the material to be remembered, with or without an intention to remember it; (2) a *retention* phase, during which the material must be maintained in memory, often while the subject is engaged in some other activity; and (3) a testing or *retrieval* phase, during which the subject may retrieve the material from memory to perform some task, for example, to judge whether a particular test item had occurred in a set studied during the acquisition phase.

A typical memory task that illustrates these three phases is the "list-learning" task. Here subjects are presented with a series of items—words, pictures, number-letter pairs, and the like—and are asked to commit them to memory. They then rest or engage in some distracting activity for awhile, after which they are tested on their memory of the original list. This test can take a variety of forms, depending on the experimenter's aims and the nature of the stimuli being used. Common examples are *recall* tests, in which subjects are asked to reproduce the items from the list, and *recognition* tests, in which they are asked to discriminate those test items that were present on the original list from some lures (distractor items) that were not.

A surprising number of issues can be studied with different variations on this simple task. Consider just a few examples. One task, called context discrimination, is the experimental analog of remembering where or in what context one has experienced overlapping collections of objects (for example, which friends were met at which parties, which

furniture was seen at which houses). In the laboratory several different lists of items are presented, and subjects are later asked to judge in which list or lists a given test item had occurred (see Anderson and Bower 1974). By varying such factors as the similarity of the list contexts, how closely in time they were presented, and in how many contexts a given item occurred we can learn much about how people reconstruct contexts in which they learned something. A second illustration examines people's learning of words or statements of differing emotional connotations, depending on the learner's current emotional state. Thus when presented with a mixed list of pleasant and unpleasant words or statements, people who are feeling temporarily happy or sad will learn more of those items that are congruent with their emotional state (see Bower 1981). As a third illustration, list learning has been a useful method for comparing the efficacy of various "mnemonic strategies." We can compare memory performance of subjects who have studied the same list of items using different mnemonic techniques (see Bower 1970).

These examples illustrate the versatility of the simple list-learning task. Moreover the basic acquisition-retention-retrieval schema can be varied to produce many useful tasks. Let us consider some of these.

Memory Tasks

Recall Tasks In a recall test subjects are asked to *reproduce* the items they encountered during the acquisition period, in response to some cue. This cue may range in specificity from quite vague to a modest hint, up to one that is highly detailed and specific. In general the more specific the cue, the better subjects' performance will be. Cue specificity can be illustrated by experiments in which subjects learn lists containing words grouped by taxonomic categories (for example, several names of fruit, followed by several names of animals, city names, and so on). When later asked to write all the items they can remember, subjects will perform more poorly than if they are provided with specific category names as cues (so-called cued recall; see Crowder 1976). The effect of cue specificity is similarly illustrated by the beneficial effect of hints on our performance in memory-based brainteasers such as crossword puzzles.

The most straightforward way to analyze recall is simply to compare subjects' overall performance (that is, mean percent of items correctly recalled) across the different experimental conditions and note whether they differ significantly. But a variety of more subtle analytic techniques is often required to interpret the results. One such procedure (mentioned previously) analyzes subjects' recall for their tendency to recall together in sequence ("cluster") items on the basis of their thematic or taxonomic relations. This tendency can be exploited to discover the kinds of interitem relations subjects find most obvious, because mean-

ingful relations often serve as a basis for clustering. Another informative class of errors, as mentioned previously, are recall *confusions*. For example, subjects may mix up the serial order of two list items or mistakenly produce a synonym of a word or a paraphrase of a statement from the acquisition list. The nature of these confusions can reveal much about how subjects represent the stimulus materials in memory.

An often informative by-product of recall tests is a set of items that subjects have mistakenly produced that were not presented during acquisition. Such *intrusion* errors often result from subjects' guessing strategies; having forgotten, they then produce items that seem likely to have been on the original list but were not in fact presented there. When witnesses to crimes (real or staged) are asked to recall the events they saw, their intrusion errors can create serious complications (Loftus 1979). Intrusions in an experiment can similarly complicate comparisons across different conditions, because frequent intrusions make it difficult to estimate subjects' memory merely by counting the number of list items they reproduce at testing. To obtain an accurate estimate of subjects' memory, the intrusions must be subtracted from the overall score to correct for guessing. Intrusions are not all bad news, however, they are often a useful source of information in themselves. Frequently they reflect the background knowledge subjects use to reconstruct the earlier learning episode. For instance, when subjects recall stories about scripted event sequences, such as doing laundry or eating in a restaurant, they often intrude descriptions of events that are central to the goal of the script but were not mentioned in the text (Bower et al. 1979; Graesser et al. 1980). This indicates that people's memory for scripted events is organized around the characters' goals satisfied by the scripted activities.

Recognition Tasks · In recognition-memory tests subjects are presented with a series of test items and for each item must judge whether it was presented during an earlier acquisition period. In other words they must *discriminate* probe items that were originally presented ("targets") from those that were not ("distractors" or "foils"). Whereas in recall tasks subjects are asked to produce the previously learned items in response to some contextual cue, such as "recall all the words from the first list" or "recall the name of the man married to Sally," this situation is reversed in recognition. In recognition subjects are provided with an item and must attempt to remember its association to the acquisition context. This yes-no decision is often supplemented with a confidence rating, say, on a three-point scale. The subjects indicate whether they have high, moderate, or low confidence in the accuracy of their yes-no decisions. Such confidence ratings often provide more differentiating information than the simple yes-no decision alone.

A related format uses multiple-choice recognition tests: presented with a set composed of the old item and N distractors, subjects try to

select the one that they remember as having been presented during the acquisition phase. Even more discriminating information can be obtained by having subjects rank order all the alternatives in the set in terms of their likelihood of being the old item. The experimenter can then calculate the average rank of the old items in subjects' rememberings.

Several factors influence how well subjects perform in recognition tests. One of these is the similarity of the distractor foils to the targets; the more similar the incorrect alternatives are to the correct one, the more difficult the test becomes. It is much easier, for example, to pick out the correct value of pi from the set 48.0, 3.1416, and 6.3842 than from the set 3.4116, 4.3146, and 3.1416.

Performance in yes-no recognition is also influenced by subjects' beliefs about the proportion of old (target) versus new (distractor) items on the test and the relative payoffs for positive versus negative responses (Parks 1966). These factors mainly influence subjects' tendency to guess old or new when they are uncertain about a test item. Obviously if a subject is unsure about an item but knows that most of the items on the test are targets, a positive response is the best bet. Similarly if a larger reward is given for every correct old response, whereas incorrect old responses receive less penalty, a bias toward old responses would be justified. Such factors must be controlled or equated across conditions when designing recognition tests.

As noted before, when intrusion errors are plentiful on recall tests we cannot use the "percent correctly recalled" as a direct index of subjects' memory. A similar problem exists with recognition tests. Subjects' positive responses will be composed of both correct olds ("hits") and erroneous olds ("false alarms"). Similarly the set of negative responses will consist of both correct and incorrect negatives ("correct rejections" and "misses," respectively). None of these response categories, taken alone, tells the whole story about subjects' memory because they are all determined both by subjects' actual memory for the material and their guessing strategies. The mere fact that a subject has a very high hit rate, for example, is not in itself an indication of good memory; it may just reflect the subject's bias to respond old to any item he or she is uncertain about.

This situation requires a method for separating the contributions of subjects' guessing strategies from their actual memory for the material. One such procedure is based on signal-detection theory, which was introduced previously (Parks 1966). Given that some assumptions about underlying familiarity distributions of old and new items are satisfied, signal-detection theory can be used to obtain separate estimates of each subject's response bias and memory sensitivity. For a detailed treatment of signal-detection theory, see Green and Swets 1966.

In addition to analyzing subjects' *accuracy* in recognition tasks, it is useful to measure the *speed* with which they respond to each test item.

This reaction time (RT) measure makes it possible to detect differences among conditions even when subjects are near ceiling levels (100 percent) on accuracy. If subjects respond more rapidly to one type of item than to another, we can assume that these items are more accessible in memory. This property makes the RT measure ideal for studying memory retrieval. If subjects perform with 100 percent recognition accuracy across all experimental conditions, then we know that differential RTs are not caused by subjects' having failed originally to encode more of the items in one condition than another. Rather the difference must be caused by one type of item being easier to retrieve from memory.

Whenever accuracy is less than perfect, however, the experimenter must be alert to the possibility of speed-accuracy trade-offs, in which subjects tolerate more errors in one condition to go faster. This trade-off, familiar to all typists, is suspected whenever subjects are faster in condition 1 than in condition 2, but show greater accuracy in condition 2 than in condition 1. Such trade-offs can occur if subjects adopt a more conservative response strategy for probes in condition 2, causing them to slow down and exercise greater care than they do with probes from condition 1. Speed-accuracy trade-offs can make the results of an experiment considerably more difficult to interpret. Sperling and Dosher (1986) provide an introductory discussion of correcting for speed-accuracy trade-offs.

Judgment Tasks In addition to recognition-memory judgments, several other judgment tasks are commonly used to study memory issues. For instance, subjects can be asked to judge which of two items occurred more *recently* or more *frequently* during an acquisition period, whether the items were printed in uppercase or lowercase letters, were in French or English, what *order* they were presented in, and so on. Such memory judgments can be used to uncover facts about subjects' memory not easily obtainable by recognition or recall measures. For example, the factors that affect subjects' ability to keep track of an item's frequency of occurrence can reveal much about how separate presentations of an item are represented in memory (Hintzman 1976).

Judgment tasks are often used to study the structure of people's world knowledge. For instance, Rosch (1973, 1975, 1977) and Rosch and Mervis (1975) studied the structure of people's knowledge of natural categories (for example, plants, animals) by asking subjects to judge which instances are more representative or *typical* members of these categories. They found that membership in most natural categories appears to be a matter of graded degree rather than an all-or-none relation; this result challenges the Aristotelian view that categories have rigid definitions in terms of necessary and sufficient attributes (see, for example, Rosch 1973, Smith and Medin 1981). Judgments of the *similarities* of pairs of stimuli drawn from a larger set can be subjected to various analyses, such as multidimensional scaling (see, for example, Shepard 1980, She-

pard and Arabie 1979) or cluster analyses (see, for example, Sattath and Tversky 1977), which help investigators identify the major features or dimensions in subjects' internal representations of these stimuli. These topics comprise large research areas of their own, and a detailed treatment of them is beyond the scope of this chapter.

So far we have described several important performance indexes used to assess subjects' knowledge during the retrieval phase of a memory experiment. We now turn to discussion of task factors that operate before testing, that have their influence during the acquisition and retention periods.

Transfer Tasks It is a truism that all learning takes place within a context of knowledge previously acquired, and that this strongly influences how various events are interpreted and remembered. The main purpose of *transfer* tasks is to study the effects of knowledge acquired before (or after) a given learning episode on what is learned during that episode, and how well this learning is retained over time. Because transfer from previous or subsequent learning is an important factor in any learning or retention situation, psychologists have devoted much time to the careful study of this topic.

One situation that illustrates many transfer issues is learning to use a new computer text editor (Anderson and Singley 1987). Imagine recruiting a group of subjects (say, from a secretarial school) with no prior experience with text editors and teaching them first to use one editor (say, Wordstar) to a satisfactory degree of proficiency and then training them to use a second editor (say, E-macs). What performance could we expect on the second transfer task? First, the time subjects require to learn the second editor should be greatly reduced relative to the time they took to learn the first. This is because much of what was learned in mastering the first editor could be directly "transferred" to the second. These common components include not only commands to achieve specific goals but also an entire superordinate goal structure shared by the two text editors—what an editor can do, how to sequence parts of an editing plan, what plans to use in familiar situations such as deleting or adding pieces of text, concepts such as accessing and saving a file, and so on. These are the components common to any text editor because they characterize the general problem of generating and revising text. Not only would learning a first editor prepare learners for the second (a "proactive" effect in the sense that the influence acts forward in time), but practicing the second editor should maintain or improve the retention of those features it shares with the first (a "retroactive" effect).

Not all of the influences are expected to be mutually beneficial, however. If the two editors require different keystrokes or different sequences of basic commands to achieve the same goal, then specific *negative transfer* is likely to arise at just those points of conflict during learning of the second editor. It is as though the specific habits learned

with the first editor come to mind at the moment the learner wants to achieve a certain goal (for example, "delete the right side of this line"), and these habits get in the way and interfere with learning and performing the correct habits for the second editor. This negative transfer at the level of specific keystrokes, however, is usually not sufficient to overcome the large positive transfer at the global level of goal-structures common to the two text editors.

If sufficient time is allowed to pass after learning the second editor, subjects who are brought back to the laboratory and tested will demonstrate forgetting for certain aspects of both editors, especially those different commands or procedures that gave rise to negative transfer during learning of the second editor. The forgetting of some previously learned material caused by learning similar (but not identical) material later is called *retroactive interference;* the forgetting of later material caused by prior learning of similar material is called *proactive interference.* If we examine only the specific parts of the two editors that differ, then the forgetting of these parts would probably be far greater in a group learning both editors than in a control group learning only one. Such interference-caused forgetting has been much studied by psychologists, especially in the so-called verbal learning tradition (see Cofer 1971, Postman 1971).

This example illustrates several points. The first is simply the important role of transfer in any learning situation, especially the effects of prior knowledge on new learning. Even learning of the first editor was doubtless influenced by subjects' general knowledge about how to achieve editing goals, write papers, use computer terminals, and so on. A second point is that transfer can occur at many levels of abstraction, from high-level goal or plan structures to specific microlevel details. Third, transfer can either facilitate learning or impair it, and both effects can operate at different levels simultaneously (for example, global versus detailed). On one hand using prior knowledge as "scaffolding" for new learning underlies our ability to learn efficiently. On the other hand negative transfer (interference) is considered by many to be the primary cause of forgetting in humans. Fourth, transfer tends to operate mainly across corresponding "roles" in different knowledge systems. For instance, changing the "line-delete" command across text editors will affect performance on just that command but not on other commands such as saving a file or advancing the cursor. In fact because the pattern of transfer observed between two (or more) bodies of knowledge depends on how they are represented in memory, transfer studies can be used as diagnostic "microscopes" for studying knowledge representation.

Whereas readers may be personally familiar with phenomena alluded to in the text-editing example, psychologists studying transfer have focused on simpler cases of associative learning to obtain a clearer view of the phenomena. Thus transfer has been studied most intensively at

the level of specific associations (relations) between pairs of explicitly presented, unrelated items. This is usually done using a task in which subjects learn sets of such pairs and are later asked to recall the second item in each pair given the first as a cue. An example is recalling the specific key that must be pressed to execute a particular text-editing command. By presenting several such sets (lists) of pairs and manipulating how the pairs in each list relate to those in the others, one can study many aspects of associative transfer.

To illustrate this, one might ask subjects to learn a list of carefully controlled text-editing commands, such as "line delete → f12" and "cursor right → control-r." Let us denote the pairings learned on the first list generically as A-B pairs, where A is the abstract command goal and B refers to the particular key or keys that must be pressed to execute it. After this list of pairs is learned to a satisfactory level, a second list of commands, supposedly from a different editor, would be learned. The pairs on the second list may bear one of several relations to the pairs on the first list: (1) A-B pairs, consisting of repetitions of items from the first list; (2) A-D pairs, in which a cue term (command goal) from the first list is retained but the response (keys to be pressed) is altered (for example, "line delete → f4"); and (3) C-D pairs, in which neither the cue nor the response terms were presented on the first list; both are unique to the second list. After second-list learning and a specified retention interval have been completed, memory for either list can be tested by asking questions such as, How do you execute a line delete in the first editor?

Most of the forms of transfer that would result from learning two text editors could be studied with this task, although the units of transfer would be simple keystroke associations rather than high-level structural elements. When commands from the first list are repeated on the second (A-B, A-B condition), positive transfer should result. In the A-B, A-D condition, learning the A-D command on the second list should cause retroactive interference in recalling the A-B association from the first list. Prior learning of A-B should also cause negative transfer in how rapidly A-D is learned and proactive interference with subjects' ability to recall A-D later on. Novel commands in the second list (C-D condition) and first-list commands not repeated on the second list would serve as control associations against which transfer would be judged negative or positive by comparison.

The items used in such "paired-associate" experiments can be any material that subjects perceive and encode into memory as units or "chunks." We used text-editing commands in our example, but practically any items can be used—words, letters, pictures, numbers, subjects and predicates of sentences, and so forth (Cofer 1971).

The advantage of having subjects learn associations between explicitly presented items is that we remove uncertainty about *what* subjects are learning and thus can focus attention on how such learning takes place

and how it is affected by transfer. By studying transfer in situations which it is clear what the cue and response items are, psychologi acquire basic information helpful in studying transfer in cases in whi it is not completely obvious how to characterize what subjects a learning. Many realistic learning situations are examples of such case

Transfer tasks can also be used to study the acquisition of generaliz schemas, which are clusters of interrelated properties. As an illustrati suppose that subjects view a series of photographs of unfamiliar i sectlike creatures that vary in their size, color, wing markings, sha and so on. After examining a given instance for several seconds, su jects would try to draw it accurately from memory. Over a series trials we would find two distinct trends in their recalls. First, acro trials subjects will improve because they are learning the "conceptu macrostructure" or "slots" that characterize the population of insect For example, they would learn to include legs, wings, and antenna drawn in particular relationships, as well as constant values of any the attributes from this set (for example, these insects have six leg orange wings, and so on). Figure 7.4 depicts an illustrative memor schema. Second, as subjects examine successive exemplars that diffe

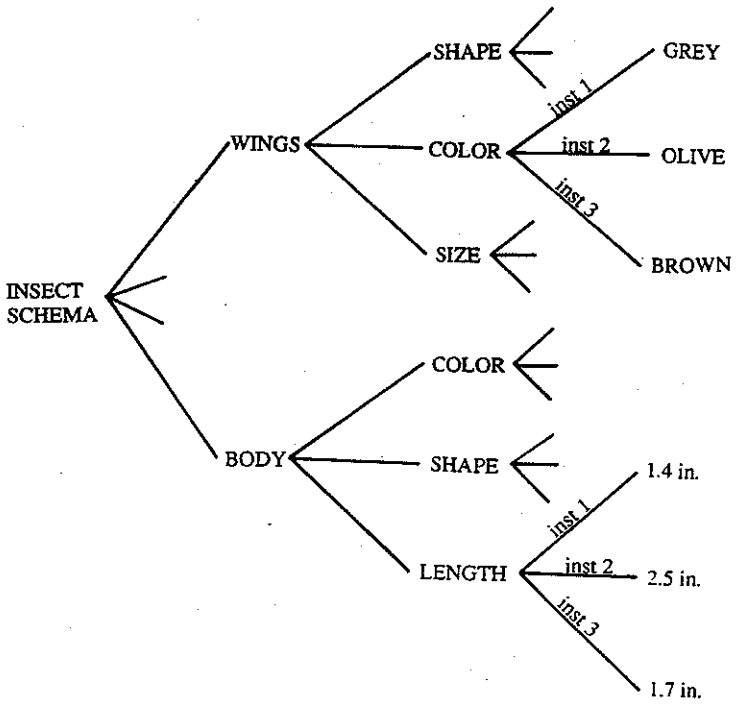

Figure 7.4 Part of a schema representing some of the commonalities and differences among several instances. Exposure to multiple instances (denoted inst 1, inst 2, and inst 3) increases the availability from memory of the superordinate attributes (for example, wings, body) and their interrelations, but causes interference among multiple values of the same attribute (for example, brown or gray wings).

in the values of various attributes, the potential for proactive interference increases so that subjects become more likely to err in recalling the specific details of the most recent insect they have seen (see figure 7.4, Bellezza and Bower 1986, Bower 1974, Thorndyke and Hayes-Roth 1979).

Variations of this schema-abstraction experiment could be conducted with materials from many interesting domains, such as routinized procedures or event sequences, pictures or descriptions of scenes or objects, software packages, series of physics problems that share a common "solution algorithm," and so forth. Many training factors can be varied in such domains to note how the schema acquired becomes attuned to the conditions of learning. Such experiments provide a window to the processes by which people learn and use complex domain knowledge.

Concept-Learning Tasks So far we have mainly discussed rote memory, that is, people's ability to retrieve information from memory in more or less the same form in which it was originally encoded. Although rote memorization is an important topic, it hardly exhausts interesting methods used in studies of learning and memory. We now turn to the topic of *induction* or *generalization*, especially as it has been studied experimentally in *concept-learning* tasks.

A *concept* can be thought of as an internal summary or model that captures some of the commonalities that exist across a particular collection of stimulus patterns or situations. A concept can also be thought of as a "decision rule" for discriminating members from nonmembers. We will use the term *category* to refer to the set of stimuli that are instances of a particular concept. Thus our concept of dogs is our model of what dogs are like; the category "dogs" is the set of objects to which this model applies. Philosophers refer to these as the intension and extension of the concept, respectively.

In concept-learning experiments we examine how subjects learn and apply fairly simple artificial concepts in the laboratory. The goal is to identify general principles that characterize inductive learning across a variety of situations. The prototypical concept-learning experiment has two phases. During the *training* phase subjects are taught to classify a set of stimuli into members and nonmembers of the category. (Actually subjects are usually taught two or more concepts at once, but for simplicity we restrict discussion to the single-concept case). In the *transfer* phase subjects are given new stimuli to classify. The way subjects classify the stimuli in the transfer phase can reveal much about the concept model (or rule) they learned during the training phase. The training and transfer phases need not be explicitly separated. They can be combined by presenting unfamiliar stimuli at several points during the training phase so that subjects' knowledge of the concept is tested as they progress toward learning it.

As noted, training usually involves presenting subjects with a series

of stimuli, some of which are members of the category and some no
On each trial subjects decide whether or not the stimulus is a membe
of the category. The experimenter then gives the subjects "feedback
telling them whether their judgment was correct. Such trials are re
peated until the subject achieves some criterion level of performance
or until the entire stimulus set has been presented for a specified num
ber of cycles. The usual performance indexes for this task are the num
ber of errors before learning and the number of training instance
required before a subject learns the concept; a derivative index is th
proportion of subjects who learn the concept within the allotted trials
These indicators of learning difficulty are influenced by many variable
in the induction situation, allowing us to draw inferences about the
learning process.

Practically any type of concept can be studied using some variation
of this basic procedure. For instance, many studies have investigated
how people infer concepts based on simple rules such as (for geometric
shapes) "all members are blue and square, with other attributes free to
vary." One can also investigate concepts that are probabilistic or
"fuzzy," which specify a set of characteristics that most instances will
possess, but none of which are individually necessary for category
membership. Most "natural" concepts of everyday experience are of
this type (Rosch 1973, Wittgenstein 1953). A common type of stimulus
used to study fuzzy concepts in the laboratory are categories of fictitious
diseases; such diseases are characterized by a typical set of diagnostic
symptoms, but not all of these symptoms are necessarily present in any
particular case of the disease.

Depending on the aspect of inductive learning under investigation,
the features or attributes by which subjects classify the stimuli can be
made more or less obvious. For instance, when teaching subjects to
classify hypothetical patients as instances or noninstances of a particular
disease, the features of each patient are obvious, namely, his or her
symptoms. But in other cases the features can be made quite inconspi-
cuous. Subjects can learn to classify abstract paintings on the basis of
their style or the artist who created them, they can learn to discriminate
the age of different types of wines, and so on. People have remarkable
and little-understood abilities to learn categories of such stimuli, even
though the cues they use to do so may be extremely subtle. This type
of research often focuses on how subjects discover diagnostic cues with
sufficient experience, despite the fact that naive observers are at a loss
to identify them (for a review of perceptual learning, see Gibson 1969).
Often the experts themselves cannot identify or describe the subtle cues
that control their discriminative responses (see Lewicki 1986).

One topic of theoretical interest is to characterize what sort of training
is required to learn different kinds of concepts. Some types of concepts
are learned easily, even in the absence of explicit feedback from a

teacher; these concepts simply "pop out" of one's experience with instances. These concepts have the sort of regularities or correlated features that our mental apparatus is designed to pick up easily. Other concepts are learned only with great difficulty or not at all. For example, people are poor at learning *exclusive disjunctive* concepts, which are defined by such rules as "all members have either feature X or feature Y but not both." Characterizing which factors make a concept easy or hard to learn tells us much about our inductive mechanisms.

During the transfer phase subjects see series of stimuli, some from the training phase and some not previously presented, and are asked to decide whether each stimulus is a member of the category. These are usually designed to test various hypotheses regarding what subjects learned about the category during the previous training phase. For example, by comparing how subjects respond to new versus old instances of the category, one can evaluate the contributions of rote memory for specific instances versus the induction of a general concept. If subjects can classify familiar training instances, but perform at chance when attempting to classify new stimuli, it is reasonable to infer that during training they simply memorized which stimuli were members and which were not. On the other hand if subjects are equally proficient at classifying both new and old stimuli, then it is plausible that they induced a general concept or classification rule during the training phase. Now one can argue in such cases that subjects respond to new examples not by a rule but by their similarity to preceding examples stored in memory, (see, for example, Hintzman 1986, Medin and Schaffer 1978). Nonetheless, the ability to classify new stimuli at least demonstrates subjects' ability to *apply* to novel cases whatever they learned during training.

For some types of concepts measures other than classification may be used. For instance, if subjects learned some procedural concept during the training phase, such as a programming construct like FOR or WHILE loops in Pascal, we might test subjects' ability to apply these concepts to solve actual programming problems rather than test the subjects' ability to classify program-examples of WHILE loops.

Sometimes it is useful to ask subjects to report verbally on any learning strategies they might have used, the attributes they attended to, and so on. The more reliable verbal reports refer to concurrent mental processes; retrospective reports suffer from distortions created by forgetting and hindsight rationalizations (Ericsson and Simon 1980). Verbal reports cannot be taken as a direct readout of what subjects are actually doing to learn the concept; often people believe they are responding to completely different factors than those that are actually affecting their behavior (see Lewicki 1986 and Nisbett and Wilson 1977 for some interesting examples). Often people simply don't know how they have learned something. For instance, when learning a tennis forehand

(which we might consider a "motor concept"), we might not really hav much explicit, verbalizable knowledge of what is being learned; th same is probably the case for expert wine tasters, art critics, medical x ray interpreters, and so on. But it may be just as important to kno that people *do not* have introspective access to their inductive learnin processes in some circumstances as to know that they do have in othe circumstances. This information would allow inferences to be mad about what other factors (for example, conscious learning strategie would affect learning.

Studies of categorization share many features with perceptual studie of pattern recognition. In such studies interest often centers on hov accurately or rapidly the subject can identify or categorize a give stimulus pattern when it is presented under degraded conditions—fo example, for short durations with low signal-noise contrast. For reviev of experimental methods specific to investigations of pattern recogni tion, the reader should consult such texts as Sekular and Blake 1985 o Kling and Riggs 1971.

Knowledge-Based Learning Tasks The previous section describe techniques for studying how people induce simple concepts under con trolled laboratory conditions. We now discuss methods for investigatin the structure of people's knowledge about *real-world* categories of ob jects, events, social groups, and so on. Because many performances o interest to cognitive scientists (for example, language use, reasoning inference, and the like) depend on the application of such knowledge advances here can help advance these other areas as well.

By investigating knowledge that subjects already have rather thar studying artificial laboratory concepts, investigators surrender contro over many important variables relating to the conditions under whic the knowledge was acquired, the nature of that knowledge, and so on. Although this increases the difficulty of designing and interpreting such research, these difficulties can sometimes be offset by the advantage gained by analyzing complex concepts that people have learned in realistic contexts. Despite the difficulties and the newness of this re search area, several useful experimental methods have already evolved.

Research in this area usually involves two or more steps. First, sub jects' intuitive knowledge about the concepts under study are system atically collected. For example, a group of subjects might be asked to list the events that typically occur during routine activities, such as doing their laundry or eating at a restaurant (Bower et al. 1979, Graesser et al. 1980). Alternatively subjects might be asked to rate a number of such events along various dimensions such as their probability of oc currence or their centrality to the activity, their distinctiveness, and so on (see, for example, Galambos and Rips 1982). These ratings can then be used to establish different categories of materials—say, low-centrality

versus high-centrality events in the script—which can be correlated with differences in subjects' performance in later memory tasks using these materials.

A typical experiment of this type would present subjects with several passages describing different stereotyped event sequences, social groups, personality types, physical layouts, and the like and later test their memory for these passages. Subjects might be asked to recall the passages or to write summaries of them. Recognition tests are also frequently used; here the subject reads some statements and judges whether each one appeared in the original passage (one can test either verbatim or gist memory).

In such experiments the errors subjects make on the memory test are often of primary interest. To illustrate this, consider a task in which subjects read narratives describing characters eating at a restaurant (a "script" of Schank and Abelson 1977). When their memories are inexact, people believe that events rated as highly central or "typical" for a given script (such as ordering a meal when eating in a restaurant) were probably presented in the original passage. Thus they will often produce or falsely recognize such highly typical items even when they were not actually mentioned in the passages (see, for example, Graesser et al. 1980). In contrast memory discrimination is much better for "low-typicality" items or (especially) for items that violate subjects' expectations about the topic of the passage. Such findings provide clues about what subjects attend to as they read the passages, how the new information is integrated into previous knowledge, and how this knowledge is applied to reconstruct the episode later during testing.

After finding that some factor such as the "typicality" or "centrality" of events in a script-based passage affects subjects' memory, a next likely project might be to determine the locus of this effect. It could arise during encoding, for example, by subjects attending more to one type of event than another; or the performance difference could arise at retrieval, when subjects use general knowledge to reconstruct gaps in their memory. If the phenomenon is multiply determined, then the factors operating during different stages of memory can be teased apart.

Several techniques can help determine which stage an independent variable influences. One of these is to measure subjects' reading speed for each line of text. If subjects spend more time reading one type of statement than another (say, sentences describing unpredictable versus predictable events), then we may infer that they are paying more attention to, thus encoding more deeply, that type of item. Such an attentional bias could easily influence later memory performance.

A second method for detecting attentional biases at encoding involves *forcing* one group of subjects to attend equally to all items and comparing their memory performance to another group of subjects who are allowed to allocate attention to different parts of the material in any way they

wish. If this manipulation eliminates (or reduces) memory differences observed in the "free-choice" condition, we would infer that differences were wholly or partly due to attentional factors.

Other methods can be used to assess factors at work during retrieval when the subject tries to reconstruct the original stimuli. Retrieval effects can be examined by manipulating the retention interval between study and test. For instance, one group of subjects might be tested immediately after they read each passage, whereas another group might be tested after, say, an hour's delay. If subjects from the immediate-test condition remember both types of items equally well (say, typical versus atypical events in a script-based passage), then we could infer they have at least encoded both types of items. If differences then appear on the delayed test—that is, if subjects forget one type of item more rapidly than another—then we could infer that the effect of materials occurs *after* subjects have initially encoded the passage. For instance, typical items may suffer great interference and forgetting because of previously encountered instances than do atypical items.

In general memory for material based on naturalistic schemas reflects not only the nature of the underlying schema but also the material presented, its conditions of presentation, and other factors. Therefore one is well advised to examine how a schema influences performances across a variety of situations, using different materials, testing procedures, and the like, so that the effects of particular situations can be "cancelled out" and stable theoretical constructs can be identified. The procedures we have described can be combined to help untangle several factors that might be contributing to performance. Brewer and Treyens (1981) provide an especially clear set of methods for doing so.

The goal of all such experimental investigations of course is to obtain empirical findings that one can relate to a theory's predictions. Moreover a good theory helps specify which factors should be investigated and helps interpret experimental results, especially in complex research areas. In turn experimental results force a theory to stay in close correspondence with its empirical domain and provide a check on theorists' intuitions. Every experimenter has a long, expanding list of "blatantly obvious" predictions that turned out to be disconfirmed by his or her results. In trying to explain unexpected results, we are forced to consider important new possibilities that would not otherwise have been entertained. In the end empirical testing may be the only antidote we have to perpetual self-delusion.

We now conclude our discussion of methods for studying the learning of laboratory-presented materials. Our review has omitted several techniques (including autobiographic memory, implicit or unconscious memory tasks, skill acquisition) for brevity. For further information, interested readers may consult a methodology book by Puff (1982) or any of several textbooks on human memory (for example, Baddeley 1976, Crowder 1976, Wickelgren 1977).

7.3 Experimental Methods II: Language Processing

Key Issues in Language Research

The fundamental questions in language research concern the mechanisms by which language is understood and produced. How is a first language acquired and how is it represented in the brain? How is language used to get things done in social communications? Language studies focus not only on speech but also on written language. A dominant trend is the analysis of language understanding on-line, in real time, either for the listener or for the reader. A second trend analyzes speech production, usually for the speaker (not the writer).

Characterizing Studies of Language Processing

Laboratory studies of language processing can be loosely partitioned into three types of tasks, which we discuss in turn:

1. *Judgment* tasks, in which word strings are presented to a supposedly knowledgeable subject who judges the strings according to some criterion specified by the investigator. This is the standard method of linguists who ask their subjects (or themselves) to judge the "linguistic acceptability" of sample sentences. Subjects' judgments in such tasks can reveal much about their linguistic knowledge—for example, their internalized semantic distinctions or syntactic rule systems.

2. Speech *production* experiments, which deal with the processes by which people plan and execute various types of utterances. Here one might examine the pause structure and intonation of phrases as subjects either compose a narrative, recount some episode, or repeat aloud sentences they have just read. A second class of investigations would look for particular types of "slips of the tongue" (for example, spoonerisms such as "bad-cat" spoken as "cad-bat") when the speaker performs under time pressure.

3. *Reception* tasks, in which subjects listen to spoken utterances or read text and then indicate in some manner their understanding of the message. An example might have college students read a short story and then a day later write a summary or abstract of the story from memory. The aim of this research is to characterize the processes and knowledge sources that people use to understand various types of linguistic messages.

Judgment Tasks

Judgment tasks are commonly used to study speakers' linguistic knowledge and their expectations about how the language would be used in various communicative situations. Many types of judgments have been investigated. For example, subjects might be asked to judge what class of words will fit into a sentence frame to make an acceptable sentence. Or they might judge how similar in meaning two sentences are. For

example, if you believe that syntactic transformations should preserve meaning, then you can ask people whether a transform (*Wasn't that John?*) means the same as a base sentence (*Was that John?*). Another common task is to judge the grammaticality of single sentences. A standard procedure for collecting judgments is to ask the subject to rate directly (say, on a seven-point rating scale) the degree to which the stimulus satisfies the specified criterion. A related method is to rank order a set of stimuli from best to worst in terms of the degree to which they satisfy the criterion. Thus subjects might be asked to rank several metaphors on their "aptness."

By observing how subjects judge various stimuli, it is possible to test hypotheses about the subjects' linguistic knowledge or communicative expectations. For instance, a theory of English syntax could be tested by asking subjects to decide whether each of a series of word strings forms a grammatically correct sentence. The stimuli should be chosen such that different syntactic theories will disagree in their predictions about which word strings subjects will find more or less acceptable. This procedure is often applied with only a single subject (usually the investigator), but it is better to use multiple subjects to ensure that the judgments are reliable across speakers. The intuitions of experts are often contaminated by their theoretical preconceptions; the data obtained from naive subjects are sometimes more useful for theorizing about language users in general.

Such judgments can be made in or out of a linguistic context. Many sentences that appear meaningless out of context become acceptable in an appropriate context or when interpreted metaphorically. For example, *Ideas breed papers* is literal nonsense but acceptable as a metaphor. Similarly assertions of connections between events may appear nonsensical (for example, *The notes were sour because the seams split*) until a contextual cue activates a relevant knowledge structure (for example, *bagpipe*). An example of judgments in context would have subjects rate the importance, centrality, or relatedness of a given text statement to the "overall meaning" of a short text or story. Subjects agree reasonably well in the ranking of text statements in their degree of importance; such judgments are determined by aspects of text structure (for example, degree of causal connectedness of the event described by the statement to the other events in the narrative) and are positively correlated with such performance indices as the likelihood of the statement's recall or inclusion in a summary of the text.

Production Tasks

Production tasks are used to study how people plan and execute their utterances. For instance, when we are planning our next sentence in a conversation, do we usually plan out the whole sentence before executing it, or do we just plan a phrase at time? Do people plan a unit of

speech in its totality before execution, or do planning and execution occur in parallel? How much does this process depend on global discourse knowledge, our beliefs about where the conversation is going, and the topic we are discussing? How do we establish convenient ways to refer to the entities we are discussing, especially in cases where these do not have familiar labels? Ideally one would like to answer questions like these with a computational theory that is able to plan and execute speech just as humans do.

In most production tasks subjects' vocal responses are either elicited on cue (and the reaction time to initiate speaking is recorded) or are tape recorded for later analysis of such prosodic features as speech rate, pauses, dysfluencies, intonation contours, and so on. Any way in which the subject departs from an "ideal delivery" (continuous, error-free execution) provides us with information about the production process. For instance, subjects usually pause during execution when they are heavily involved in planning a new constituent; therefore pauses can be used as an index to help researchers identify speakers' "units of planning." We should find that their performance on a secondary dual task is especially poor at just those points in composing their utterance.

Language-production tasks come in several varieties. One task has the subject say as rapidly as possible different word series or sentences in response to different cues. Such tasks have been used in research by Cooper and Paccia-Cooper (1980) and by Sternberg, Monsell, Knoll, and Wright (1978). Here the interest is in how the time to start producing speech increases with the length of the series to be produced. A second task, studied by Motley and Baars (1981), tries to create slips of the tongue in the laboratory. A subject silently reads a pair of words and is then cued to say the words aloud either in the order written or in the reverse order. When driven at a rapid pace, subjects make a small percentage (about 10 to 20 percent) of speech errors such as spoonerisms or perseverations. One can then investigate factors that promote more errors. For example, over a short series of trials one can create a temporary set for subjects to pronounce *ba-* then *fay-* by presenting a series such as *bat fake* and *bad fate;* thus when subjects are pressed to hurriedly pronounce the written pair *fag base,* they are likely to slip and say *bag face.* Phonetic switches that make real words are more likely to occur than are similar switches that do not make words.

In the tasks reviewed so far, the person's productions are strictly controlled. However, subjects can be given more liberty to generate speech at their "natural" rate. One useful task requires subjects to describe a known spatial layout; for instance, Linde and Labov (1975) asked apartment dwellers to describe their living quarters. Interest focused on the way subjects organized and planned their description. This was typically achieved by the speaker taking the listener on a verbal "walking tour" of the apartment, beginning with entering the

front door and proceeding through adjacent rooms with their contents and fixtures. Other investigators have studied recitals of everyday hassles and cooking recipes. The focus of these studies is on how people organize their recitals and how they pause as they finish generating one segment and plan the next segment.

Another class of studies investigates children's abilities to compose and generate coherent stories. Story generation requires considerable linguistic skill, conceptual knowledge, and familiarity with story conventions. The storyteller must introduce a central character (or group) who has a goal and then compose and relate a plan of action for the character to overcome obstacles and achieve his or her goal. Developmental psychologists track the way in which these component abilities improve as children develop greater levels of competence at the whole task (see, for example, Stein 1987).

A final class of speech production tasks studies social communication in controlled laboratory situations. One of the more fruitful tasks, devised by Krauss and Weinheimer (1966), studies how two subjects come to agree on a common ground of novel referring expressions. The subjects are separated by a screen; both have a collection of ten or so similar but unfamiliar nonsense figures on scrambled cards before them. One subject serves as the "sender," who is asked to verbally communicate a specified serial order of the ten figures to the "receiver" subject. Interest focuses on how the sender chooses to label or describe each of the figures; this is typically initiated by the sender describing what the figure resembles (for example, "the man with the top hat who's kicking out behind him"). Investigators study the "negotiating exchanges" by which the sender and receiver check, confirm, and eventually come to agree on a unique figure as the referent of a given description. The task is usually repeated many times, with the experimenter specifying on each trial a new random order of the same figures (cards) to be communicated by the sender. At issue is how subjects come to agree on an abbreviated shorthand to refer to each figure (for example, "The top hat" or "the kicker"), and how this substantially reduces the time they require to communicate each new order accurately. This communication game has been taken as a model of how two participants in a conversation search for common ground and effective referring expressions for their conversation. A variant of this task, devised by Wilkes-Gibbs (1986), has two subjects in separated cubicles planning a route through a city's streets by sharing knowledge gleaned from city maps that have large areas blacked out, different for the two subjects. At issue is how the subjects compare and coordinate their respective partial knowledge bases so as to compose a complete route through the patched-together maps. The performance is presumed to bring into prominence and model several *coordination* processes that are presumed to underlie social communication.

Reception Tasks

Reception tasks are used to study the processes by which people understand what they read or hear. Many aspects of comprehension have been investigated; these include recognizing individual words, constructing syntactic and semantic descriptions of an input sentence, and relating a statement to the ongoing discourse and to one's general knowledge about the world. In all such tasks subjects are presented with some sort of linguistic stimulus and are asked to respond in some way that will inform the investigator about either the process by which the material was understood or the resulting mental representation. Generally the investigator will vary some structural property of the stimulus material or its conditions of presentation. For instance, one might ask subjects to read sentences that vary in their grammatical complexity (according to a particular syntactic theory) and see if it takes subjects longer to understand more complex sentences (as indicated by the amount of time they spend reading them). We organize our review of reception tasks by first discussing various types of stimuli and stimulus manipulations and then describing several performance indices.

Language Stimuli Virtually every kind of language fragment has been used as stimuli in reception tasks. Discourse units vary in their length or inclusiveness—from single sounds or phonemes to syllables, words, phrases, sentences, and larger pieces of discourse such as complete stories or expository texts. People can be said to "comprehend" at each of these levels; they identify phonemes, retrieve word meanings, and interpret sentences and larger discourse units. Researchers have developed theories of how comprehension is achieved at each level. These theories specify relevant factors to investigate and the expected nature of their effects.

Words, for example, vary in their familiarity to the average subject, their length, number of meanings, visual or semantic confusability with other words, relation to the context in which they occur, and countless other dimensions. Any of these factors might affect comprehension. Research in this area attempts to identify those factors that do have an effect, and to establish where in the overall comprehension process these effects occur. For instance, longer words might be recognized more slowly because they take longer to read; on the other hand, infrequent words may not take longer to read than common words but they may require more time for their meanings to be looked up in our mental lexicon. Different performance indexes, such as reading versus paraphrasing, may be sensitive to these different types of effects.

At the sentence level investigators are usually interested in how people construct syntactic or semantic descriptions of the sentences they read or hear. One question that has occupied researchers is whether extraction of these two descriptions are really separate pro-

cesses that occur independently of one another (see, for example, Fodor 1983, Forster and Olbrei 1973). Do semantic or pragmatic factors affect the initial syntactic analysis, or is syntax a separate "module" in the language-understanding system? Given that it makes sense to sharply distinguish these processes, what is the "correct" description of each?

Although the questions at the sentence level are different from the major issues in word comprehension, the logic for answering them is essentially the same. Structural factors and their expected effects are specified by one's theory; the investigator then selects materials that differ on some factor of interest and tests whether this difference has the predicted effect. For instance, a syntactic theory might be used to rank order sentences in terms of their "syntactic ambiguity." One could then see whether this ranking accurately predicts how rapidly subjects read the sentences, how well they can paraphrase or recall them later, and so on.

The study of larger discourse units such as narratives or expository passages is currently a burgeoning research area. One practical aim of some investigators is to describe what makes texts easy to understand, easy to recall, or both. One of the stronger determinants of comprehensibility is the reader's familiarity with the ideas in the text, the words by which they are expressed, and the syntactic simplicity of the sentences. A good predictor of a word's familiarity is the frequency of its occurrence in English texts; Kucera and Francis (1967) have published tables of word-frequency counts that are often used in language research. The familiarity of "idea units" in text is often judged intuitively and reflects the "expertise" of the readers regarding the topics of the text. A text on organic chemistry is highly readable to professional chemists but not to elementary-school children. The most direct way to assess comprehensibility of texts is to have subjects from the target population read and rate a variety of experimental texts; investigators then select a sample of texts that have the desired comprehensibility level to use in further studies.

In studies of comprehension and recall of longer texts (say, 250 to 500 words), a critical variable is the organization of ideas in the text—how they are laid out, sequenced, interrelated, and signalled by rhetorical words and phrases. Texts can be designed that vary in logical coherence (for example, number of illustrations, amount of support for general assertions) and that vary in explicit signalling of the underlying organization of the ideas (see Grimes 1975, Meyer 1975, 1985). These factors influence how much a reader can comprehend and remember from the text. A volume edited by Britton and Black (1985) has many papers illustrating different analyses of expository prose. Mandler and Johnson (1977, Johnson and Mandler 1980) describe hypotheses regarding the structure of narratives (simple folk tales). Having in hand a hypothesis about what constituents (theme, goal, setting, embedded episodes, and the like) comprise a coherent story, the investigator may design stories

that either delete the constituents or rearrange the canonical order of the constituents in a story and then note how this influences people's comprehension and recall of the texts.

A different level of analysis of narratives is provided by identifying the several strands of causal connections that course through the events described by a story. Starting with work by Schank (1975), the analysis of texts for causal connections has been further developed by Black and Bower (1980), Trabasso and van den Broek (1985), and Trabasso and Sperry (1985). Because stories typically describe people's problems and their plans to solve their problems, successive events can be viewed as causing or enabling subsequent events, either by physical or psychological causation. Causal analysis is relevant because it is associated with several psychological indicators: for instance, story events that are multiply connected by causal linkages to other story events are more likely to be judged as "centrally important" events, more likely to be recalled, and more likely to be included in a summary of the story (see the Trabasso papers for details). An acknowledged weakness of this approach is that because texts are often cryptic, important cause-effect linkages within the story's actions may not be explicitly stated; thus a causal chain analysis either must infer unmentioned links or tolerate broken chains in the description of the narrative.

In investigating text understanding, researchers must remember that people can adopt alternative strategies for reading a text. They will choose a strategy appropriate to their own goal or to whatever goal has been set for them. For a competent adult each goal initiates a reading (or listening) plan that is designed to focus on the text information judged most relevant to the goal. Of course people best remember what they focus on and react to in a text. Other information that is given only passing attention will be quickly forgotten. Thus a person who proofreads a paragraph for typing errors may show relatively little understanding or memory for the content of the material; conversely a person who reads to summarize a text may not notice or recollect which words were misspelled.

A word of caution is needed concerning the use of language materials to test psychological hypotheses. The hypotheses are usually abstract in the sense of referring to the behavioral differences that should arise when people process linguistic inputs of type A versus those of type B, all other things being equal. The A-B inputs might be ambiguous versus nonambiguous words, abstract versus concrete words, causally ordered versus temporally ordered narratives, and so forth. To test the generality of such hypotheses, one should select sizeable samples of many language units that exemplify the theoretical variable and are otherwise roughly equivalent. Obviously the larger the sample, the more generalizable the result. To generalize a given result (based on a sample of linguistic units) to the population of all possible units of that type, the investigator should try to demonstrate that the effect holds

up consistently when comparing individual items of the same type. important methodological paper by Clark (1973) argues for the nee test the consistency of effects across language samples. Psycholing tics researchers now routinely follow Clark's injunctions regarding perimental design and statistical hypothesis testing with mult language samples.

Performance Indexes The index of linguistic performance chosen an investigator varies according to the purpose and object of study. studies of language reception investigators are usually interested in subject's comprehension and how it varies with properties of the t and the subject. The measures may be divided into those that trac serial reception task as it unfolds in real time, versus those that exami the resultant memory established by a reader or listener after havi comprehended some text. We call these "on-line" versus "memor measures of language reception.

On-Line Performance Measures

Repeating Back Perhaps the simplest reception task is one common used in studies of speech perception, namely, requiring subjects repeat the speech sounds they think they heard over noisy earphone The presentation and response units involved may be of any size, fro a phoneme to a syllable, word, or full sentence. It is well known th performance in such tasks is better the higher the intensity of the sign (speech sounds) relative to the noise, the fewer the confusable alte natives in the choice set, and the more probable (or expected) the sign item is in the given context (see, for example, Licklider and Miller 1951)

Reading Time Probably the most widely used method for studying on line comprehension is to measure the time a subject requires to read language unit such as a word, phrase, sentence, or fixed block of text With the widespread use of microcomputers that display text, the typ ical laboratory arrangement is to visually present words, phrases, o sentences on a CRT screen for a duration controlled by the reader. The reader presses an advance button on the keyboard to indicate when he or she has finished reading (and comprehending) the presented unit and is ready to view the next unit of text. One method, called rapid serial visual presentation (RSVP), presents only single words (from sentences) in this manner. Successive words may appear either at the same location on the CRT screen or may march across the screen as though one were reading a line of text from left to right.

The basic measure taken in RSVP is the time to read each word, indicated by the time between successive button presses. Slower times are taken to reflect greater underlying difficulty of "processing" the unit. Slower times are typically observed for longer words, for relatively unfamiliar words, semantically unexpected words, words involved in

unexpected syntactic constructions, and words that mark the end of major constituents.

The same reading-time measure can be used for whole phrases or sentences presented all at once on the CRT. Here the subjects' button presses control how long they take to read each phrase or sentence. Such reading times are sensitive to structural variables such as the length of a sentence in syllables, familiarity of the words or relations, and syntactic complexity of the sentence. Reading times are also affected by the expectedness of the sentence given the preceding linguistic context. The more easily a statement can be linked to the reader's current representation of the text, the more quickly it can be read and understood. The RSVP method is most often used in on-line psycholinguistic experiments because it is very sensitive to experimental variables and is also quite easy and inexpensive to set up.

Eye Tracking The RSVP method disrupts the normal reading process, introduces several problems, and fails to capture certain aspects of normal reading (such as going back over material read previously). More normal reading can be studied by recording readers' eye movements as they fixate on successive words or groups of words on a page (or screen) of text. Elaborate optical equipment can accurately determine and record exactly where a subject's eyes are looking at each moment. By superimposing this gaze location onto a copy of the text, the investigator can obtain a precise record of where the subject was looking and for how long. People read by successive fixations (of 20 to 5000 milliseconds), jumping across a standard line of text in 2 to 5 seconds. Actual reading occurs only for the word or two in the center of foveal vision; our ability to discriminate words more than a few degrees off center is very poor. To a first approximation, then, we may assume that the eyes fixate on a word (or cluster of words) until it has been integrated into the unfolding representation of the text (Just and Carpenter 1987). Thus fixation times provide a sensitive measure of how difficult particular words are to understand in context. As one might expect, poor readers have longer fixations, more fixations per sentence, and more eye regressions than do good readers.

Eye-movement records have provided a rich array of information about the "mechanics" of reading. They tell us that people look longer at content words than function words; longer at unfamiliar, unusual, or unexpected words; and longer at words that terminate and wrap up a phrase. The records also tell us when a reader's eyes "regress," or go back to reread difficult parts of the text. For example, when readers encounter that point in a "garden path" sentence where they realized they have been tricked, their eyes will regress to a previous part of the sentence, from which point they will start over with a different interpretation. Thus in reading the sentence *The boat floated down the river sank,* readers get into trouble when they get to *sank,* at which point they

stop, move their eyes back to the beginning, reread the sentence, and recategorize "floated" as the beginning of a complement phrase that modifies *boat*. Eye regressions are also frequent when readers encounter pronouns with indeterminate or ambiguous referents. The readers' eyes scan rapidly over the preceding text, looking for a proper antecedent for the problematic pronoun.

The main difficulty with eye-movement recordings is that the investigator is in danger of becoming buried in mounds of data as well as details of the technical apparatus. Therefore scientists considering the use of eye-movement recordings are advised to become familiar with the costs in time and money before they embrace such an expensive and data-rich source. Also they will need a set of practical data-reduction programs to help them deal with the huge volume of eye-fixation data generated by a few subjects reading just a few passages.

Probe Reaction Time A reasonable hypothesis is that readers' processing resources are more engaged (or used up) during their intake of certain difficult parts of a sentence. For instance, we might hypothesize that it takes extra mental effort (attentional "resources" or "capacity") for a person to understand syntactically complex or semantically ambiguous sentences or to find the referent of opaque pronouns in the sentence. Cognitive psychologists often use dual-task methods for measuring how much attentional capacity is engaged by such performances. In such procedures the subject performs a primary task (say, reading words by the RSVP method), while at the same time performing a second task. Of course frequent testing of subjects' comprehension of the primary-task material is required to ensure that they attend to it as much as required. A typical secondary task is to have subjects listen for an occasional soft tone (or look for a dim light) and to press a key as soon as they hear (or see) it. The probe stimulus appears only infrequently, in an unpredictable pattern. People's reaction time to a probe stimulus is presumed to be slower the more absorbing the primary task is at the moment the probe appears. In terms of an energy-reserve metaphor, the more processing resources are absorbed by the primary task at a given moment, the less capacity there is for performing the secondary task, resulting in a less efficient (slower) performance on the secondary task. These effects are usually quite small, however, relative to the baseline variability in the RT measure. Thus fairly large samples (about 50 observations per condition) are usually required to obtain reliable results.

Phoneme Monitoring A similar monitoring task asks the subjects to listen for a particular phoneme (such as *ba*) as they listen to a continuous stream of speech and to press a key as soon as they detect it. Thus hearing a sentence such as *The emperor went to the royal baths*, subjects should press the key upon hearing the initial *ba* in baths. A visual analog can be arranged wherein subjects look for a target pair of letters

(*ba*) in an RSVP reading task. The idea behind the phoneme-monitoring task is similar to that of the probe RT task: the more engrossing and difficult the primary task (of comprehension) just before the target phoneme is presented, the slower the subject should be in detecting and reacting to the target. Thus people will react more slowly to target probes that occur just after ambiguous words or just after words signalling a syntactically ambiguous construction (see Foss 1969, 1970, 1982). Unfortunately performance in phoneme monitoring itself is a complex affair and is affected by many "nuisance" variables such as the frequency and distribution of targets across sentences, the discriminability and voicing of the target phoneme, and the frequency (familiarity) of the word in which the target phoneme occurs and of the preceding word (see Newman and Dell 1978). Thus investigators must exercise some care in using the phoneme-monitoring task to reach conclusions about the mental effort involved in parsing and comprehending different parts of a sentence.

Interrupting Questions A simple way to catch some of the processes of comprehension in nearly real time is to interrupt subjects' reception of text and ask them a question. The question may or may not refer directly to elements in the antecedent text. An example is to ask subjects to name the referent of a pronoun in a sentence they are just hearing (or have just heard) and measure the time required to reply as well as the accuracy of the response. Another example is to have the reader answer who-what-where-when questions immediately after hearing or reading a sentence. Various hypotheses can be tested in this way—for example, that elements in syntactically complex constructions will be harder to retrieve as answers. Thus the agent can be shown to be more accessible immediately after presentation of an active rather than a passive sentence (Wright 1969).

Other types of interrupting questions are used to examine the way in which the preceding sentence or passage *primes* and speeds up the retrieval of semantically related information from long-term memory. An example is the so-called *lexical decision* task, in which the person decides whether or not a probe string of letters forms a word. Half the probes form words and half are near-miss nonwords (for example, *order* and *ordar*). The time to decide that an item is a real word depends on how much that word is expected, activated, or primed by the preceding context. Thus following a sentence describing a character entering a restaurant to eat, the word *order* will elicit a fast decision whereas a control word like *older* would be slower by about 50 milliseconds (Sharkey and Mitchell 1985). The lexical decision task has been used as a tool to collect information relevant to many issues in psycholinguistics. For example, it has been used to demonstrate that immediately after hearing or reading a polysemous word in context, it primes *several* of its related meanings, not just the one appropriate to that context. After hearing

John deposited his money in the bank, people are speeded in lexical decisions for words related to the contextually irrelevant meaning of *bank* (*river, water, ground*) as much as they are for words related to the contextually relevant meaning (*office, saving, teller*). This effect is time dependent, however; the irrelevant priming is measurable immediately following the polysemous word, but after a pause of a second or more, the relevant meaning becomes dominant and the activation of the irrelevant meaning decays to the control baseline (Swinney 1979).

A related interrupted technique is the Stroop interference task. In the Stroop task a probe word is presented in colored letters (such as red or green), and the subject is instructed to name the letters' color as quickly as possible. This requires inhibiting the strong tendency to read the word, which is harder the more the word is primed by the context. Thus a strongly primed word such as *lamb* in *Mary had a little* _____ would require more time to name the color of its letters than would the same word in an unprimed context. Thus the degree of Stroop interference is a derived measure of priming.

We mention a final probe test that allows the investigator to trace the effect of a text on the accessiblity of different parts of a knowledge structure in long-term memory. Morrow, Greenspan, and Bower (1987) had subjects first memorize the spatial layout of a building composed of four rooms, each containing four named objects. Subjects then read a story about a character carrying out a plan that required him to move from room to room. The question of interest was whether objects near the reader's focus of attention in the mental map would be activated and more accessible to retrieval than objects outside the focus of attention. To assess this, as subjects were reading one statement at a time on a CRT screen they were interrupted just after reading any statement describing the character moving between rooms, say, from room A to room B. The interruption consisted of a probe test of two objects from the building (for example, *projector, computer*), and subjects were to indicate as rapidly as possible whether the two objects were located in the same or different rooms. Morrow and colleagues found that access to objects' locations was very fast for objects in the current-focus room, somewhat slower for objects in the room just exited, and slowest for objects in other rooms of the building. The value of this technique is that it allows us to assess moment-by-moment changes in the activation of information in long-term memory as the story moves the reader's attention to different parts of the building and the corresponding mental map.

Verification Tests of Comprehension A simple test of comprehension of a declarative statement asks subjects to decide whether the statement is true or false with respect either to general knowledge or a specific visual display. Thus the subject might read and answer quickly statements such as *Six isn't an odd number* or *Four isn't larger than seven.* The

response times tell us something about the difficulty of understanding and untangling negatives (for example, *isn't odd* means *even*) as well as marked comparative adjectives. Such statements can also be tested against a continually changing display, such as (true or false) *The B doesn't precede the A . . . BA*. Clark (1974) has made ingenious use of this technique for testing a range of hypotheses about the comprehension of implicitly negative verbs. For example, subjects take longer to respond to a question such as *If John forgot to let the dog out, is the dog now in*? than a related question such as *If John remembered to let the dog out, is the dog now in*? The different response times presumably reflect the time required to decompose the implicitly negative verb, *forgot*.

Memory-Based Measures of Comprehension It is understood that comprehension is an aid to memory but distinct from it. A person with organic amnesia will fully understand a conversation, but will totally forget it within a few minutes if distracted. Similarly a person can memorize a passage verbatim (say, in a foreign language) yet understand none of it. Despite these caveats the correlation between comprehension and memory is normally so high that answering questions from memory is often taken as an index of comprehension of the material.

Question Answering The verification task noted previously asks a true-or-false question about information that is immediately available to the subject or is presumed to be common knowledge. This method can be extended to ask questions about a collection of statements that has recently been read for comprehension or for memorization. Lehnert (1978) enumerated the different categories of questions that might be asked about simple narratives and has provided a computer-simulation model of how people answer such questions from a memory representation of the text. In addition to true-or-false questions, there are fill-in-the-blank questions about events involving the who-why-when-where-how slots in a typical event frame. One can also ask for answers to *What happened after event X*? *Did event X precede Y*? *Why didn't such and so happen*? Lehnert's simulation program produces impressively human-like answers to such questions. The theory has not yet been tested in psychological laboratories, perhaps because it makes few if any predictions about the speed with which subjects could answer the different types of questions.

Recognition-Memory Tests The most commonly used tests of comprehension (in the "classroom" sense) require subjects to decide whether a test sentence is true or plausible in light of the sentences (text) they just read. The instructions may ask the subject to answer "yes" only if the test statement is an exact verbatim match to one of the studied sentences and to say "no" otherwise. In such cases the investigator focuses on the accuracy of verbatim memory in different conditions and

how often false-positive answers are given to paraphrases of the presented statements.

Speed of recognition-memory judgments is often used to investigate associative priming between two different statements from the text (see, for example, McKoon and Ratcliff 1980). In such experiments the test statements are presented singly on a CRT screen and the computer records how fast the subject presses the true or false key. In priming, when a given test statement queries (and thus activates) a particular episode from the studied narrative, then the very next test statement is answered more quickly if it also queries the same or an associatively related episode.

More commonly the subject is asked not for verbatim memory but for judgments of whether a test statement is plausible or probably true given the information in the text. The main interest in such cases is usually the availability in memory of different types of inferences about the text statements. Many kinds of inferences are plausible (for a classification of several, see Rieger 1975), and a few of these have been studied experimentally (see, for example, Singer 1986). Inferences about probable instruments, locations, consequences, and actions are common. For example, a *hammer* is commonly used to pound nails, a kiss is usually placed on the *lips*, people who fall out of sailboats usually get *wet*, and pythons who catch mice usually *eat* them. By checking the frequency and speed with which subjects assent to a particular test inference, investigators can test hypotheses about which inferences are normally made during initial comprehension and which are only derived as they are needed, in response to later queries.

Cued Recall Tests In contrast to true-or-false tests, cued recall tests require the subjects to fill in the blank or blanks in a series of test statements. The Cloze technique used in school tests is an example: subjects receive the original text they read previously except that a number of content words are missing and replaced by blanks; the subjects try to fill in the blanks with the correct word or one similar in meaning. Often in sentence-memory experiments the cue for recall is a single content word, such as the subject or object of the sentence. Scoring involves simply recording the percentage of content words filled in correctly.

Ordered Recall The most difficult memory test is to recall the original text (or sentences) in its original order and as close to verbatim as possible. The subjects either write or tape record their recall and are usually given sufficient time to reproduce the entire passage. The investigator may be interested in the temporal character of the subjects' recall; for example, people tend to recall a story in an initial burst, then pause while they retrieve the next episode, recall that episode rapidly, then pause again, and so on.

Scoring subjects' protocols for gist or substance recall presents for-

midable problems. One obviously needs a means of identifying idea units in the text, as well as some conventions for coding and scoring recall to give credit for paraphrases, confabulations or blends of two or more ideas, generalizations over several ideas, and plausible inferences from the text statements. With the recent advent of propositional scoring systems (see the papers in Britton and Black 1985), the way has been cleared for extensive studies of comprehension by means of recall. When applied to a medium-sized text (say, of more than 200 words), the full propositional scoring system is quite laborious and time consuming to apply. Its output is a recall score (of yes, no, or partial credit) for each of the text's propositional (idea) units for each subject, and this suffices to permit statistical analyses by subject conditions or type of items. Fortunately a far simpler measure, namely the number of content words recalled, turns out to be very highly correlated (in the 0.90s) with the more complex propositional scores (Paul 1959, Voss, Tyler, and Bisanz 1982). Thus for most purposes the number of content words recalled can be used as an easily obtainable measure of recall.

Summarization One of the most demanding tests of peoples' comprehension of a text is to have them summarize it, distilling and abstracting it into a few statements. A summary supposedly tells only the centrally important points of the text and shows their global interrelations. Summarization presumably reveals some kind of macrostructure that holds the various parts of the text together in memory.

Summarization is a difficult task. Typically it is done from memory, imposing an added burden on the performer; this can be relieved by letting people inspect the text as they compose their summary of it. To add to the difficulty, a summary is an ill-defined and fuzzy concept: for a given text there is not a well-specified "correct" summary that everyone will recognize. Thus investigators can ask their subjects to write summaries under different constraints; for example, subjects may be asked to write a summary of a text using no more than (say) 25 content words or to use only phrases from the text itself or to produce a distillation of the text's main points at a level more abstract than the text itself (for example, the moral of a folktale). Because subjects have differing notions of what comprises a summary, investigators would be well advised to explicitly instruct them on the kind of summary they want; otherwise subjects' summaries will hardly be comparable. Despite the fuzziness of the summary concept, people are fairly reliable in rank ordering a set of summaries of a text according to their quality.

Once summaries are produced within certain constraints, they are usually analyzed in terms of propositions, and if possible these are then identified with their source in the text. Often this identification requires the scorer to understand how several text propositions could have been combined to imply the summary statement. There are no mechanical rules for doing this, although Rumelhart (1977) and Lehnert (1981)

provide some techniques for analyzing narrative summaries, and the papers in the collection edited by Britton and Black (1985) provide analyses of summaries of expository text.

Perhaps because of these scoring difficulties, relatively few studies of summarization per se have been published. We know that propositions that are important to the text's global structure are likely to appear in summaries, as are narrative events that have high causal connectivity to other events in a story (Trabasso and van den Broek 1985). In narratives these elements are usually the main character's primary goal (or an event that instigated it), a penultimate successful plan of action, and the final upshot or outcome of the narrative. Any abnormal constraint or condition of the actor and his or her plan may also be mentioned in the summary. Although data collection on summarization is still relatively sparse, the simulation programs of Lehnert (1981) and DeJong (1979) provide useful starting theories of the on-line processes by which subjects might construct summaries of simple narratives. We may expect to see more empirical development of such theories in the future.

7.4 Concluding Comments

In our recital of specific techniques in experimental cognitive science we have been able to review in moderate detail only research methods for memory and language processing. Although these areas comprise a large portion of cognitive science research, we are fully aware of other important areas (including pattern recognition and problem-solving) whose procedures we have not been able to cover.

To recapitulate, we have touched briefly on the role of empirical investigation in justifying our beliefs and guiding our actions and reviewed several empirical methods including introspection, naturalistic observations, and correlational studies. But we argued that intuitive introspection is often a weak, uninformative, even biased, measuring instrument for rapid, nonconscious cognitive processes. And our powers of naturalistic observation are seriously limited; even if we could get beyond our blinding preconceptions to observe things carefully, nature rarely conveniently arranges the exact observational conditions we need to infer causal relationships.

The history of humankind is one of increasing use of tools for extending our power, expanding or amplifying our physical senses, and more recently for amplifying our intellectual senses, our mental powers. Experimentation is a *conceptual prosthetic,* an intellectual tool that allows us to create in the laboratory possible microworlds never seen before and then observe how specific cognitive subsystems operate in those microworlds.

Experimentation provides a generate-and-test heuristic for checking the validity of our causal theories, for testing theoretical predictions. We presented a few general methods used in experimental studies of

cognition and indicated how analogical theories can be placed in correspondence to observable, experimental events. We then reviewed some specific techniques used in studies of memory and language processing, trying in most cases to indicate a few substantive issues addressed by them.

Obviously these substantive theoretical issues are the meat and potatoes of the scientist's diet, whereas the experimental methods we have reviewed comprise only the metaphorical pots and pans used to prepare that intellectual feast. The later, substantive chapters in this book serve up various feats prepared with the help of these tools.

References

Anderson, J. A. 1983. Cognitive and psychological computation with neural models. *IEEE Transactions on Systems, Man, and Cybernetics* 5:799–815.

Anderson, J. R. 1987. Methodologies for studying human knowledge. *Behavioral and Brain Sciences* 10:467–505.

Anderson, J. R., and Bower, G. H. 1974. Interference in memory for multiple contexts. *Memory and Cognition* 2:509–514.

Anderson, J. R., and Singley, M. K. 1987. An identical-productions model of transfer. Paper given at the Ninth Annual Conference of the Cognitive Science Society, Seattle, WA.

Baddeley, A. D. 1976. *The Psychology of Memory.* New York: Basic Books.

Bartlett, F. C. 1932. *Remembering: A Study in Social Psychology.* Cambridge, Engl.: Cambridge University Press.

Bellezza, F. S., and Bower, G. H. 1986. *The Formation of Verbal Schemata: Mediation and Interference Processes.* Stanford University Psychology Department, Stanford, CA.

Black, J. B., and Bower, G. H. 1980. Story understanding as problem-solving. *Poetics* 9:223–250.

Bower, G. H. 1970. Analysis of a mnemonic device. *American Scientist* 58:496–510.

Bower, G. H. 1974. Selective facilitation and interference in retention of prose. *Journal of Educational Psychology* 66:1–8.

Bower, G. H. 1981. Mood and memory. *American Psychologist* 36:129–148.

Bower, G. H., Black, J. B., and Turner, T. J. 1979. Scripts and memory for text. *Cognitive Psychology* 11:177–220.

Brewer, W. F., and Treyens, J. C. 1981. The role of schemata in memory for places. *Cognitive Psychology* 13:207–230.

Britton, B. K., and Black, J. B. 1985. *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Explanatory Text.* Hillsdale, NJ: Erlbaum.

Chase, W. G., and Posner, M. I. 1965. *The Effect of Visual and Auditory Confusability on Visual and Memory Search Tasks.* Paper presented at the meetings of the Psychonomic Society, Chicago.

Chernoff, H., and Moses, L. E. 1959 *Elementary Decision Theory.* New York: Wiley.

Clark, H. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12:335–359.

Clark, H. H. 1974. Semantics and comprehension. In T. A. Sebeok, ed. *Current Trends in Linguistics, Volume 12: Linguistics and Adjacent Arts and Sciences.* The Hague: Mouton Publishers, pp. 1291–1498.

Cofer, C. N. 1971. Properties of verbal materials and verbal learning. In J. W. Kling and L. A. Riggs, eds. *Experimental Psychology.* 3rd ed. New York: Holt, Rinehart & Winston; pp. 847–904.

Cooper, W. E., and Paccia-Cooper, J. 1980. *Syntax and Speech.* Cambridge, MA: Harvard University Press.

Crowder, R. G. 1976. *Principles of Learning and Memory.* Hillsdale, NJ: Erlbaum.

DeJong, G. F. 1979. *Skimming Stories in Real-Time: An Experiment in Integrated Understanding.* Doctoral dissertation, Research Report No. 158, Department of Computer Science, Yale University, New Haven, CT.

Ericsson, K. A., and Simon, H. A. 1980. Verbal reports as data. *Psychological Review* 87:215–251.

Fodor, J. A. 1983. *The Modularity of Mind.* Cambridge, MA: MIT Press.

Forster, K., and Olbrei, J. 1973. Semantic heuristics and syntactic analysis. *Cognition* 2:319–347.

Foss, D. J. 1969. Decision processing during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior* 8:457–462.

Foss, D. J. 1970. Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior* 9:699–706.

Foss, D. J. 1982. Discourse on semantic priming. *Cognitive Psychology* 14:590–607.

Galambos, J. A., and Rips, L. J. 1982. Memory for routines. *Journal of Verbal Learning and Verbal Behavior* 21:260–281.

Gibson, E. J. 1969. *Principles of Perceptual Learning and Development.* New York: Appleton Publishers.

Gluck, M. A., and Bower, G. H. 1988. Evaluating an adaptive network model of human learning. *Journal of Memory and Language* 27:166–195.

Graesser, A. C., Woll, S. B., Kowalski, D. J., and Smith, D. A. 1980. Memory for typical

and atypical actions in scripted activities. *Journal of Experimental Psychology: Human Learning and Memory* 6:503–513.

Green, D. M., and Swets, J. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.

Grimes, J. E. 1975. *The Thread of Discourse*. The Hague: Mouton.

Hintzman, D. L. 1976. Repetition and memory. In G. H. Bower, ed. *The Psychology of Learning and Motivation* New York: Academic Press, pp. 47–93.

Hintzman, D. L. 1986. Schema abstraction in a multiple trace memory model. *Psychological Review* 93:411–428.

Johnson, N. S., and Mandler, J. M. 1980. A tale of two structures: Underlying and surface forms in stories. *Poetics* 9:51–86.

Just, M. A., and Carpenter, P. A. 1987. *The Psychology of Reading and Language Comprehension*. Boston, MA: Allyn and Bacon.

Kahneman, D. 1973. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kling, J. W., and Riggs, L. A., eds. 1971. *Experimental Psychology*. 3rd ed. New York: Holt, Rinehart & Winston.

Krauss, R. M., and Weinheimer, S. 1966. Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology* 4:343–346.

Kucera, H., and Francis, W. N. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Lehnert, W. G. 1978. *The Process of Question Answering*. Hillsdale, NJ: Erlbaum.

Lehnert, W. G. 1981. Plot units and narrative summarization. *Cognitive Science* 4:293–331.

Lewicki, P. 1986. *Nonconscious Social Information Processing*. Orlando, FL: Academic Press.

Licklider, J. C. R., and Miller, G. A. 1951. The perception of speech. In S. S. Stevens, ed. *Handbook of Experimental Psychology*. New York: Wiley, pp. 1040–1074.

Linde, C., and Labov, W. 1975. Spatial networks as a site for the study of language and thought. *Language* 51:924–939.

Loftus, E. S. 1979. *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Mandler, J. A., and Johnson, N. S. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive Psychology* 9:111–151.

Marr, D. 1982. *Vision*. San Francisco: W. H. Freeman.

McKoon, S., and Ratcliff, R. 1980. Priming in item recognition: The organization of propositions in memory for text. *Journal of Verbal Learning and Verbal Behavior* 18:369–386.

Medin, D. L., and Schaffer, M. M. 1978. Context theory of classification learning. *Psychological Review* 85:207–238.

Meyer, B. J. F. 1975. *The Organization of Prose and its Effects on Memory.* Amsterdam: North-Holland.

Meyer, B. J. F. 1985. Prose analysis: Purposes, procedures, and problems. In B. K. Black and J. B. Black, eds. *Understanding Expository Text* Hillsdale, NJ: Erlbaum, pp. 11–64.

Miller, J. O., and Pachella, R. G. 1973. Locus of the stimulus probability effect. *Journal of Experimental Psychology* 101:227–231.

Morrow, D. G., Greenspan, S. L., and Bower, G. H. 1987. Accessibility and situation models in narrative comprehension. *Journal of Memory and Language* 26:165–187.

Motley, M. T., and Baars, B. J. 1981. Syntactic criteria in pre-articulatory editing: Evidence from laboratory-induced slips of the tongue. *Journal of Psycholinguistic Research* 105:503–522.

Newell, A., and Simon, H. A. 1972. *Human Problem Solving.* Englewood Cliffs, NJ: Prentice Hall.

Newman, J. E., and Dell, G. S. 1978. The phonological nature of phoneme monitoring: A critique of some ambiguities studies. *Journal of Verbal Learning and Verbal Behavior* 17:359–374.

Nisbett, R. E., and Ross, L. 1980. *Human Inferences: Strategies and Shortcomings of Social Judgment.* Englewood Cliffs, NJ: Prentice Hall.

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

Parks, T. E. 1966. Signal detectibility theory of recognition memory performance. *Psychological Review* 73:44–58.

Paul, I. H. 1959. Studies in remembering: The reproduction of connected and extended verbal material. *Psychological Issues.* Monograph no. 2, vol. 1. New York: International Universities Press.

Posner, M. I. 1969. Abstraction and the process of recognition. In G. H. Bower and J. T. Spence, eds. *The Psychology of Learning and Motivation.* Vol. 3. New York: Academic Press, pp. 44–100.

Posner, M.I., and Keele, S. W. 1967. Decay of visual information from a single letter. *Science* 158:137–139.

Posner, M. I., and Mitchell, R. F. 1967. Chronometric analysis of classification. *Psychological Review* 74:392–409.

Postman, L. 1971. Transfer, interference, and forgetting. In J. W. Kling and L. A. Riggs, eds. *Experimental Psychology.* 3rd ed. New York: Holt, Rinehart & Winston, pp. 1019–1132.

Puff, C. R. 1982. *Handbook of Research Methods in Human Memory and Cognition.* New York: Academic Press.

Rieger, C. G. 1975. Conceptual memory and inference. In R. C. Schank, ed. *Conceptual Information Processing*. Amsterdam: North-Holland, pp. 157–288.

Rosch, E. 1973. On the internal structure of perceptual and semantic categories. In T. E. Moore, ed. *Cognitive Development and the Acquisition of Language*. New York: Academic Press, pp. 111–144.

Rosch, E. 1975. Cognitive representation of semantic categories. *Journal of Experimental Psychology: General* 104:192–233.

Rosch, E. 1977. Human categorization. In N. Warren, ed. *Advances in Cross Cultural Psychology*. Vol. 1. New York: Academic Press.

Rosch, E., and Mervis, C. B. 1975. Family resemblance studies in the internal structure of categories. *Cognitive Psychology* 7:573–605.

Rumelhart, D. E. 1977. Understanding and summarizing brief stories. In D. LaBerge and S. Samuels, eds. *Basic Processes in Reading, Perception, and Comprehension*. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., and Ortony, A. 1977. The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro, and W. E. Montague, eds. *Schooling and the Acquisition of Knowledge*. Hillsdale, NJ: Erlbaum.

Sattath, S., and Tversky, A. 1977. Additive similarity trees. *Psychometrika* 42:319–345.

Schank, R. C. 1975. *Conceptual Information Processing*. Amsterdam: North-Holland.

Schank, R. C., and Abelson, R. P. 1977. *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum.

Sekular, R., and Blake, R. 1985. *Perception*. New York: Random House.

Sharkey, N. E., and Mitchell, D. C. 1985. Word recognition in a functional context: The use of scripts in reading. *Journal of Memory and Language*. 84:253–270.

Shepard, R. N. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* 210:390–398.

Shepard, R. N., and Arabie, P. 1979. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* 30:87–123.

Singer, M. 1986. Answering wh____ questions about sentences and text. *Journal of Memory and Language*. 25:238–254.

Smith, E. E., and Medin, D. L. 1981. *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Sperling, G., and Dosher, B. 1986. Strategy and optimization in human information processing. In K. Boff, L. Kaufman, and J. Thomas, eds. *Handbook of Perception and Performance*. Vol. 1. New York: Wiley.

Stein, N. L. 1987. The development of children's story-telling skill. In M. B. Franklin and S. Bartan, eds. *Child Language: A Book of Readings*. New York: Oxford University Press.

Sternberg, S. 1966. High-speed scanning in human memory. *Science* 153:652–654.

Sternberg, S. 1967. Two operations in character recognition: Some evidence from reaction time experiments. *Perception and Psychophysics* 2:45–53.

Sternberg, S. 1969. Memory-scanning: Memory processes revealed by reaction-time experiments. *American Scientist* 57:421–457.

Sternberg, S., Monsell, G. S., Knoll, R. L., and Wright, C. E. 1978. The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach, ed. *Information Processing in Motor Control and Learning*. New York: Academic Press, pp. 117–152.

Stevens, S. S. 1957. On the psychophysical law. *Psychological Review* 64:153–181.

Swinney, D. A. 1979. Lexical access during sentence comprehension: Reconsideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18:645–659.

Thorndyke, P. W., and Hayes-Roth, B. 1979. The use of schemata in the acquisition and transfer of knowledge. *Cognitive Psychology* 11:82–106.

Trabasso, T., and van den Broek, P. 1985. Causal thinking and the representation of narrative events. *Journal of Memory and Language* 24:612–630.

Trabasso, T., and Sperry, L. L. 1985. Causal relatedness and importance of story events. *Journal of Memory and Language* 24:595–611.

Voss, J., Tyler, S. W., and Bisanz, G. L. 1982. Prose comprehension and memory. In C. R. Puff, ed. *Handbook of Recent Methods in Human Memory and Cognition*. New York: Academic Press, pp. 349–395.

Wald, A. 1950. *Statistical Decision Functions*. New York: Wiley.

Wickelgren, W. A. 1977. *Learning and Memory*. Englewood Cliffs, NJ: Prentice-Hall.

Wilkes-Gibbs, D. 1986. *Collaborative Processes of Language Use in Conversation*. Doctoral dissertation, Stanford University, Stanford, CA.

Winer, B. J. 1971. *Statistical Principles in Experimental Design*. 2nd ed. New York: McGraw-Hill.

Wittgenstein, L. 1953. *Philosophical Investigations*. New York: Macmillan.

Wright, P. 1969. Transformations and the understanding of sentences. *Language and Speech* 12:156–166.