

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 observ. 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,

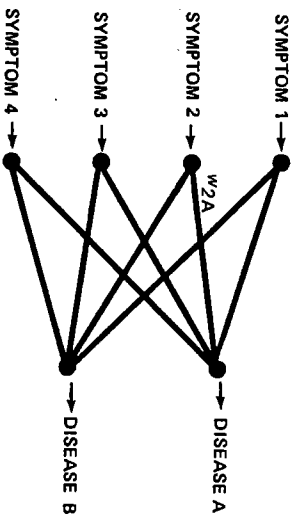


Figure 7.1 A simple connectionist network that learns to diagnose patterns of up to four symptoms as having one of two diseases. The specific w shown denotes the weight of evidence that presence of symptom 2 contributes toward disease A. (From Gluck and Bower 1988. Used with permission.)

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

