7

Processes in Evaluative Judgment

This chapter is concerned with the psychological processes that are likely to be involved in evaluating the performance of subordinates. We concentrate on cognitive processes, or on the way in which raters mentally process information about ratees, but will also discuss research on emotional or affective bases of evaluation. This chapter draws more heavily than others on experimental research conducted in laboratory settings. The relevance of this research for rating and evaluation in organizations has been widely debated; several of the issues in this debate will be discussed in the chapter.

It is useful to keep in mind that our focus here is on the evaluation that the rater arrives at, not necessarily on the rating he or she records. The model we presented in Chapter 1 suggests that this evaluation is partly context bound and partly context free. That is, the organizational context helps to define which behaviors or outcomes are valued and helps to define the meaning of performance. Concepts such as "good," "bad," "attractive,"

and so on exist independently of the organizational context, however, and judgments regarding performance will be heavily affected by these concepts. As we will discuss later, much of the research cited in this section is more relevant to the context-free aspects of judgment than to the context-bound aspects. This represents one of the potential barriers to generalizing from judgment research in the lab to actual judgments in organizations.

COGNITIVE PROCESSES

The first topic that must be examined in discussing cognitive processes in evaluative judgment is the meaning of cognitive. A quick survey of relevant research suggests that *cognitive* is a very broad term that includes almost any activity involving the mental manipulation or storage of information. Cognitive research might thus involve exploration of the basic processes in human memory (Craik & Lockhart, 1972; Tulving, 1974) or of person perception and social cognition (Wyer & Srull, 1986). Cognitive research also includes studies of the dynamics of halo error in rating (Cooper, 1981b), applications of personal construct theory (Borman, 1983), studies of attribution processes (Hogan, 1987; Kelley, 1971), research on the systematic distortion hypothesis (Shweder & D'Andrade, 1980), and examinations of the development and nature of tacit knowledge (Palermo, 1983; Wagner, 1986).

Given the wide range of things that might be included under the "cognitive" umbrella, it is a mistake to regard cognitive research in performance appraisal (PA) as a unitary field. Nevertheless, the majority of PA studies that would be labeled "cognitive" are concerned with the same basic issue—whether research in human information processing can be used to draw valid generalizations about the evaluation of job performance.

The rapid growth of research in cognitive processes in performance evaluation can be traced to several influential reviews that appeared between 1978 and 1982. Papers by Cooper (1981b), DeCotiis and Petit (1978), Feldman (1981), Landy and Farr (1980), and Wherry and Bartlett (1982) all called attention to the cognitive processes in appraisal and suggested ways of examining these processes and applying the results. Wherry and Bartlett (1982) described a model of rating developed by Wherry in the 1950s. Unfortunately, the paper describing the model was not published nor was the model widely known until it was described in Landy and Farr (1980).

More recently, two models of the judgment processes in appraisal have emerged, one presented by Feldman and colleagues (Feldman, 1981; Ilgen & Feldman, 1983) and another presented by DeNisi and colleagues (DeNisi et al., 1984; DeNisi & Williams, 1988). Much of the current cognitive research draws on these two models for hypotheses and interpretations of experimental findings.

Although the two models referred to above are in some ways different, they (as well as other cognitive approaches that have been applied to PA) can be described in terms of the five basic processes shown in Figure 7.1. First, raters must observe behavior, sorting relevant from irrelevant information (Banks & Murphy, 1985). Second, they must mentally represent, or encode, that information. This representation is not necessarily a totally faithful replication of what they have seen; it may lack many of the details that were observed and it may contain details that were not in fact present in the stimulus. This representation must then be stored in memory. It is likely that some information is lost in the transition from working memory, where immediate processing takes place, and long-term memory, where the memory traced is stored.

At some later time, raters must retrieve information from memory. Because performance appraisals are frequently conducted at annual or semi-annual intervals, demands on memory may be substantial and raters might find it difficult to retrieve all of the relevant information. Finally, raters must somehow integrate all of the information they have about each ratee. As discussed in Chapter 5, raters may have to integrate information from several different sources as well as from several different time periods. This last process, that of information integration, has received comparatively little systematic attention in cognitive models of performance appraisal, but has been a focus of considerable attention for judgment and decision researchers.

Lord and Maher (1991) show how the perspectives of cognitive science can be applied to understanding judgments in performance appraisal. Much of the cognitively oriented research in this area has drawn from the social-cognitive tradition of the 1960s and 1970s. In recent years, multidisciplinary research in cognitive science has taken a somewhat different direction, and it offers new perspectives on cognitive processes in appraisal.

The most fundamental difference between the social-cognitive approach that has dominated PA research and the cognitive science approach reviewed by Lord and Maher (1991) is the latters' emphasis on constructing

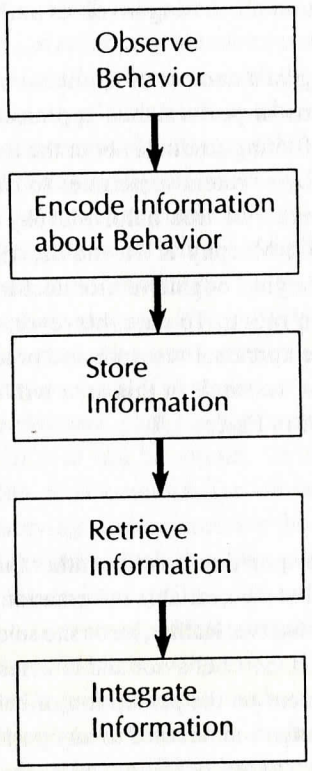

Figure 7.1. Basic Cognitive Model for Performance Evaluation

and testing models of fundamental processes involved in information processing. Cognitive science research suggests that the process of evaluative judgment probably proceeds simultaneously at several levels. Most generally, a distinction is drawn between connectivist models of cognition, which start from the neural level and attempt to explain cognitive processing in terms of neural networks, and symbolic models that take as their starting point the cognitive representation of concepts, events, and objects. Connectivist models emphasize "hard-wired" networks of neural associations, whereas symbolic models emphasize the representation and manipulation of knowledge. Cognitive science research suggests that *both* levels of analysis are useful for understanding tasks such as the evaluation of job

performance, and that information processing involves multiple simultaneous processes.

Lord and Maher (1991) suggest a number of applications of a cognitive science perspective to problems in performance appraisal. For example, they show how apparently conflicting findings about the dynamics of halo error (in particular, the tendency of raters sometimes to overestimate and sometimes to underestimate levels of true halo: Murphy & Jako, 1989; Murphy, Jako, & Anhalt, 1993; Murphy & Reynolds, 1988a) might be explained in terms of the underlying cognitive architecture (i.e., connectivist vs. symbolic) involved in rating. To date, however, applications of this perspective to performance appraisal research and practice have been limited, and our presentation of research in this area will conform to the more traditional model outlined in Figure 7.1.

Information Acquisition

The process of information acquisition is active rather than passive. The rater does not simply bring in all of the available information from whatever he or she has an opportunity to observe. Rather, he or she selectively attends to some features of the ratees and their behavior and devotes little attention to others. Focusing for the moment on the perception of behaviors, cognitive research suggests that the attention devoted to any particular behavior is a function of three variables: (a) the behavior itself, (b) the context of observation, and (c) the purpose of observation. That is, some behaviors are likely to attract more attention, regardless of the context or purpose of observation.

Organizational norms and standards (see Chapter 6) define some behaviors as important, desirable, unacceptable, and so on. Behaviors that carry strong evaluative implications will probably receive attention, almost regardless of the local context (Murphy, 1982b; Wegner & Vallecher, 1977). Beyond this simple principle, cognitive research has had relatively little to say about the behaviors that are most and least likely to attract attention. Research in this area has concentrated most on the second two influences, the context of observation and the purpose of observation.

Research on context suggests two conclusions. First, the salience of most behaviors (i.e., the likelihood that they will be the focus of attention) varies across situations (McArthur, 1980; Taylor & Fiske, 1978). This is due in part to differences in the evaluative implications of specific behaviors in

different situations. For example, a loud, verbally aggressive style of conversation may not attract any attention among a group of heavy-equipment salespeople but might be very noticeable in a receptionist; the behavior is appropriate in one situation but not in others.

Second, distinctive, novel features of the ratee or his or her behavior will be highly salient (Langer, Taylor, Fiske, & Chantowitz, 1976). Gender thus might be a very salient characteristic in an office where there is only one female but may not be at all salient if half of the employees are female (Cleveland, Festa, & Montgomery, 1988). Behaviors that are infrequent might become salient through their novelty; behaviors that are important but commonplace may attract less attention.

Research on the purpose of observation, sometimes referred to under the heading of *observational goals,* is highly relevant for understanding what behaviors will or will not be attended to by raters in organizations. In laboratory studies it is common for subjects to concentrate all of their attention on observing performance for the sole purpose of evaluating the performance of the individuals they have observed. In organizations this is rarely the case. Supervisors typically face multiple task demands and rarely have the luxury to devote all of their attention to behavior observation and evaluation (Balzer, 1986; Murphy et al., 1989). Thus, raters in organizations are likely to acquire information about ratees' performance while they are concentrating on tasks other than evaluation.

Murphy and associates (1989) demonstrated that the purpose of observation can have a direct impact on the accuracy of behavior ratings. In particular, they showed that raters who concentrate primarily on observing and evaluating ratees have a short-term advantage over raters for whom observation and evaluation are secondary tasks. This advantage disappears over the long term, however. It may be that raters who concentrate on observation take in more information, both relevant and irrelevant, than raters for whom evaluation is a secondary task. In most organizational settings raters may attend to relatively little information, but it is likely that what they do attend to will be relevant. A small amount of relevant information may lead to more accurate evaluations than a large amount of information that includes both irrelevant and relevant observations (Murphy & Balzer, 1986).

When raters *do* consciously seek out information about ratees, the purpose of observation will affect their information-acquisition activities. Raters will seek different information if they want information to find out

ut people, tasks, or characteristics of the group (Murphy, Garcia et al., 1982; Williams et al., 1985). For example, Cafferty et al. (1986) found that acquiring information in a person-blocked pattern (i.e., all information about Ratee #1 before moving on to Ratee #2), a task-blocked pattern (i.e., all information organized by task), or a mixed pattern had a direct effect on several measures of the accuracy of evaluations.

Most evaluations are concerned with assessing overall differences between persons—a purpose that might be most consistent with a person-blocked strategy (see Srull & Brand, 1983). This, however, is not always the case. Raters who are trying to assess the training needs of their work-group might focus on the overall strengths and weaknesses of the group; raters who are trying to decide which subordinate to recommend for which position might want to know about individual patterns of strength and weakness (Murphy, Garcia et al., 1982). The appropriate information search and acquisition strategy will depend on the demands of the task.

Before concluding this section, it is important to note that information-acquisition strategies have both conscious and unconscious aspects. That is, the rater might consciously develop a goal and might plan steps to achieve that goal (see Chapter 8). However, features of the stimulus and the context of observation may also unconsciously shape the rater's information-acquisition activities. Ilgen and Feldman (1983) note that many important information-processing activities are done in an automatic mode, where little if any conscious deliberation is carried out. Lord and Maher (1991) suggest that what appear to be mode-of-processing effects (i.e., effects of automatic vs. controlled processing) may in fact reflect differences in the level (e.g., neural networks vs. symbolic manipulation) at which various judgment tasks are carried out. Both the traditional social-cognitive and the cognitive science perspectives agree that information-acquisition activities can be affected without the rater's conscious knowledge.

One implication of this fact is that interventions to help direct the rater's attention to relevant and valid cues cannot concentrate solely on the conscious strategy the rater uses in searching for information but must also attempt to deal with differences in the natural salience, the novelty, and the distinctiveness of relevant and irrelevant behaviors. In order to efficiently redirect the rater's attention to the most relevant or valid behavioral indicators, changes may be needed in both search and acquisition strategies *and* the context in which behavior is observed (Cleveland & Hollmann, 1991).

Encoding and Mental Representation

At one time, a movie camera served as a reasonable analogy for the way in which perception and the mental representation of information was assumed to work. That is, the mental representation of information was thought to be relatively automatic and faithful. The more current thinking in cognitive psychology, and in particular in social cognition, is that the mental representation of what one has observed involves a complex process of categorization that may proceed automatically but often requires careful attention and mental effort. Furthermore, mental representations are not necessarily snapshots of what one has observed. Rather, some information may be lost when a stimulus is categorized.

In general, it is likely that one remembers the *category* rather than the *stimulus*. An example of this process is illustrated in Table 7.1. One mental category of people might represent persons who have the properties of (a) tall, (b) dark, and (c) handsome. An individual who is tall, dark, and average in appearance might be sufficiently similar to that category to be perceived as a member. If he or she is categorized in that manner, the information that will be later retrieved from memory is that the individual was tall, dark, and handsome (i.e., the person is now seen as having the properties of the category).

It is useful to distinguish among three related terms that are frequently encountered in research on encoding and categorization: category, prototype, and schema. A *category* is a group of related objects that takes the form of a "fuzzy set" (Rosch, 1977; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). That is, categories do not have rigid boundaries but rather represent potentially unstable groups of objects that are held together by similarity or "family resemblance." A *prototype* is an exemplar of that category. It is an image that summarizes the typical and distinguishing features of a category. Thus, if the categories are "dependable economy car" and "American sports car," a Saturn and a Corvette might serve as prototypes of their respective categories. Finally, *schema* (or *schemata*) refers to higher-level memory structures that contain verbal or propositional information (Feldman, 1986; Ilgen & Feldman, 1983). Schema can be used to represent the self and/or well-known others in familiar situations (Ilgen & Feldman, 1983). One type of schema, referred to as a *script* (Abelson, 1976), pertains to a mentally stored representation of a familiar event (e.g., visiting a restaurant).

TABLE 7.1 Effects of Categorization on Memory

Properties of the Stimulus	*Properties of the Category*	*What Is Remembered*
Tall	Tall	Tall
Dark	Dark	Dark
Average Appearance	**Handsome**	**Handsome**

In the most general sense, categorization depends on the similarity between a target and each of the categories that are available to the rater (Rosch, 1978; Tversky & Gati, 1978). Thus, to understand the process of categorization one must address two separate issues: how similarity judgments are made and how categories become more or less available to the rater. Research on the similarity judgment process suggests that the process of assigning persons (or objects) to categories involves a comparison of the features of a person with the features of the prototype for each category. (Research in this area has not addressed *how* to compare stimuli and prototypes. The literature on similarity judgment—see Gregson, 1975—suggests that the similarity-matching process is not a simple one.)

Returning to our earlier example, a car will be classified as a "dependable economy car" if it shares many critical features with a Saturn. It will be classified as an "American sports car" if it shares many critical features with a Corvette. The features that are critical are those that best represent each category and that best distinguish between categories.

The concept of category accessibility is an essential one in this line of research. It is assumed that the set of all possible categories is extremely large and that individuals can work with only a subset of all the possible categories at any given point in time. The model of social cognition presented by Wyer and Srull (1986) suggests that *using* categories makes them more salient (see also Srull & Wyer, 1980). That is, if a person uses the category "American sports car" to classify the first car he or she sees, the likelihood that the person will use the same category for the next few cars is increased. Using a category also increases the accessibility of related categories. Thus, if the person does not categorize the second car he or she sees as an "American sports car," the likelihood that the person will use a related category is increased.

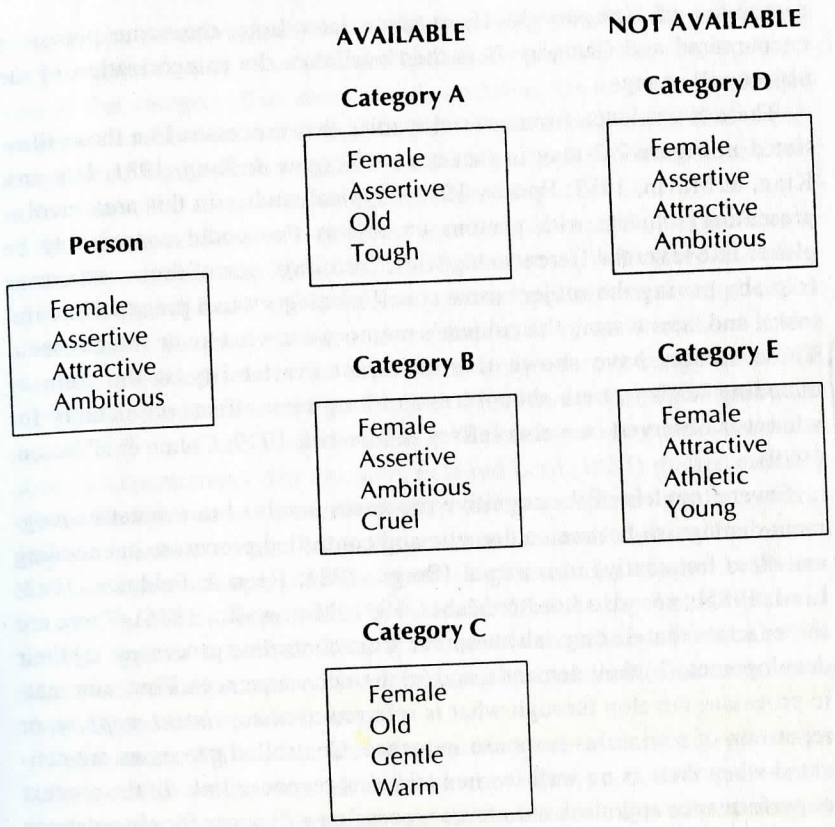

Figure 7.2. The Category Matching Process

NOTE: Properties of the person and the categories are listed in boxes.

The accessibility or availability of different categories may vary over time or situations. It follows that the same stimulus will not always be assigned to the same category. Figure 7.2 illustrates how this might occur. A person and five different categories are each described in terms of four properties. The match between the person and the category is a function of the number of properties the person and the category have in common. Of the five categories, the person fits Category D the best, but that category is not currently available. As a consequence, the person will be classified as

a member of Category B. If, at some later time, the same person is encountered and Category D is then available, the categorization of the object will change.

There is evidence from several studies that processes like those illustrated in Figure 7.2 may in fact occur (Higgins & King, 1981; Higgins, King, & Mavin, 1982; Posner, 1978). Typical studies in this area involve presenting subjects with persons or objects that could conceivably be placed into several different categories, "priming" one of those categories (e.g., by having the subject use a specific category in a preexperimental task), and later testing the subject's memory for what he or she has seen. These studies have shown that construct availability at the time of encoding leads to both short-term and long-term effects on memory for what was observed (see also Jeffrey & Mischel, 1979; Cohen & Ebbeson, 1979).

Several models of the cognitive processes involved in evaluative judgment distinguish between automatic and controlled processes in encoding and (less frequently) in retrieval (Bargh, 1984; Ilgen & Feldman, 1983; Lord, 1985b; see also Lord & Maher, 1991; Motowidlo, 1986). There are three factors that distinguish automatic from controlled processes: (a) their development, (b) their demands, and (c) their consequences. First, automatic processes develop through what is referred to as *consistent mapping* or repetition of a stimulus-response sequence. Controlled processes are activated when there is no well-learned stimulus-response link. In the context of performance appraisal, automatic processing will occur for observations that are repetitive, frequent, or simple, whereas controlled processing will occur for novel or unusual observations. Second, automatic processing makes only minimal demands on processing capacity, making it possible to perform on several well-learned activities simultaneously (e.g., driving a car while carrying on a conversation). Controlled processes demand conscious attention and involve serial rather than parallel processing. That is, the controlled process of encoding occurs when raters consciously attend to observations and effectively screen out surrounding activities.

Finally, and most important, the consequences of automatic versus controlled processes in encoding can be substantial. Automatic processing involves a simple, unconscious matching of objects and categories and involves the loss of information about the object that is not shared by the category. Thus, in the example shown in Figure 7.2, automatic processing would involve the loss of information about one of the properties of the

person (i.e., attractive); subjects would remember the object as possessing the properties female, assertive, ambitious, and confident (i.e., the properties of the category that most closely matches the object). Controlled processes involve an active search for information. The faithful representation of detailed information about the stimulus is probably more likely under controlled than under automatic processing.

The concepts of schema and prototypes have been applied to the area of leadership, as well as to PA (Lord, 1985b; Lord et al., 1984; Lord & Maher, 1991; Phillips & Lord, 1982). In fact, it is our opinion that the application of schematic theories has been more successful in the domain of leader perception and less successful in the domain of performance appraisal. The most obvious reason is that prototypes of leaders are more likely to exist and be similar across situations than will prototypes of good workers or successful performers. For example, Foti and Lord (1987) studied memory for leader behavior at a board meeting and were able to predict several important outcome variables using script and schema theory. One reason for this is that most people have at least a general idea of what behaviors are appropriate and expected from the leader of a meeting (i.e., scripts exist for this type of behavior). Thus it is possible to determine which of the behaviors actually observed are or are not consistent with a schema and to make fine-grained predictions.

In contrast, PA studies are often carried out in contexts where the subjects will have no clear idea of what represents good, average, or poor performance. In the language of schema theories, studies of performance appraisal are often carried out in situations where schema, scripts, and even performance categories are not well defined. As a result, it is difficult to use schematic theories to make specific predictions about appraisal outcomes. One of the few studies that was successful in applying this approach in studying appraisal led to very mixed results (Nathan & Lord, 1983).

Content of Categories. The outstanding weakness of schema and categorization theories as applied to PA has been a lack of attention to defining what performance-related categories look like. For example, is "good worker" a category? Is "average worker" another? Researchers in this area have argued both sides of this question but there has been little empirical work that addresses the content of job- or work-related categories (see, however, Blencoe, 1984; Ostroff & Ilgen, 1986). Cardy, Bernardin, Abbott, Senderak, and Taylor (1987) have presented evidence suggesting that raters do have

Why is this not tie into the Thorndike's ultimate criter

prototypes of good and poor workers but their research has not examined the content question in any depth.

Although Borman (1987) was not directly concerned with the content of the categories that are used in encoding performance-related behaviors, his research on the personal constructs, or "folk theories," that raters use to describe job performance is highly relevant. He found that general dimensions such as maturity/responsibility, organization, technical proficiency, and assertive leadership were among the personal constructs most often used by raters in their descriptions of work. One possibility is that categories are characterized by different levels of particular dimensions or by different profiles of levels on several dimensions. That is, the dimensions identified by Borman (1987) might define the properties of each category, in that categories might be defined in terms of the presence or absence of each property. For example, one category might be people who are (a) high on maturity, (b) low on organization, (c) average on technical proficiency, and (d) very high on assertive leadership. Other categories might be defined in terms of relative and absolute strengths on different dimensions.

It is likely that the categories used in defining job performance change over time, especially when new raters are compared to experienced raters. In general, experience with the task and/or with evaluating performance on the task leads to an increase in the number and sophistication of the categories that are used in evaluation (Cardy et al., 1987; Kozlowski & Kirsch, 1986). In addition to the number and the detail of job-related categories, the content of these categories is likely to change with experience. One of the several aspects of organizational socialization may be to develop a category system that is comparable to the one used by other members of the organization. As new managers learn their jobs, they are also likely to learn categories that are useful for mentally representing the work of their subordinates.

One dimension that is likely to be highly central in defining categories is the general evaluative dimension (Murphy, 1982b; Murphy & Balzer, 1986). That is, categorization is very likely to involve placing each person along a good-bad, preferred-nonpreferred continuum. Cognitive models have generally assumed that evaluation was the end product of the cognitive process, but there is evidence that evaluation is primitive and universal and that evaluation may occur at the same time as or even precede other information-processing activities (Kim & Rosenberg, 1980; Osgood, 1962; Wegner & Vallecher, 1977; Zajonc, 1980).

Research by Hastie and Park (1986) suggests that evaluations are stored at the time that behavior is observed, and that subsequent memory may be for evaluations rather than for behaviors. (Hastie & Park's distinction between on-line vs. memory-based judgments is similar to the distinction Ilgen & Feldman [1983] draw between stimulus-based and memory-based judgments.) One implication of this is that overall evaluations of individuals may not be greatly affected by the encoding-storage-retrieval process, whereas perceptions of specific behaviors may be directly affected. Raters may be highly accurate in remembering their evaluations but may have a great deal of difficulty in remembering supportive detail. Indeed, there is evidence that raters may *infer* behavioral details from their evaluations, rather than (as is generally assumed) basing their evaluations on the total set of behaviors they have observed (Murphy, Martin, & Garcia, 1982).

One final point about categories deserves attention. It is widely assumed that categorization simplifies perception and encoding. Ilgen and Feldman (1983) assert that "categorization is necessary to cognitive economy; it reduces the amount of information that must be processed and stored" (p. 155). The assumption that categorization is more economical in terms of information-processing resources depends on the untested assumption that the number of categories is small relative to the number of stimuli that are encountered. In fact, the number of categories that might be used is potentially larger than the number of stimuli, because it is possible to construct categories that represent null sets. One such category of persons might be "female presidents of the United States" (hopefully, this category will not always be a null set).

It seems critical for proponents of the view that categorization is cognitively efficient and economical to demonstrate that this is in fact so. One way would be to show that stable categories exist only at a general level. For example, categorization will be economical if "good worker" and "bad worker" represent typical categories, but will not be economical if "good worker who sits at this desk and who is sometimes absent on Thursdays" represents a typical category.

Storage and Retrieval

Tulving (1983) noted that more than 100 years of experimental research on human memory have failed to produce any clear consensus on how memory works or on how many different types of memory exist. Memory

researchers generally agree that it is useful to distinguish between short-term working memory and long-term memory, but several researchers have proposed different models for each type of memory. Thus, it is important to realize that the models we will describe here are not universally accepted, and that alternative theories can lead to different predictions about what behaviors will or will not be remembered at the time of appraisal.

There are two principal distinctions between working memory and long-term memory: duration and capacity. Working memory is assumed to have a limited capacity, both in terms of the number of items that can be held in memory and in terms of the length of time they can be held. It is assumed that some information enters working memory, where immediate information-processing activities occur, but is not transferred to long-term memory. Information that is not transferred to long-term memory within a very short period (often a few seconds) may be permanently lost.

Long-term memory is assumed to have effectively infinite capacity, both in terms of the number of items that can be stored and in terms of the length of time that they can be held. Although it is possible that the passage of time, as well as activities that intervene between storage and retrieval, may interfere with the ability to *retrieve* information from memory, it is assumed that information that is committed to long-term memory is not subsequently lost (Tulving, 1983).

One theory suggests that two separate memory systems exist: semantic and episodic memory (Tulving, 1983). More recent research suggests that transitional forms of memory that share features of both semantic and episodic memory may also exist. Semantic memory provides storage for verbal, factual, and propositional information, whereas episodic memory provides storage for actions, occurrences, and experiences (both direct and vicarious). Information about a subordinate's performance may be stored in *both* semantic and episodic memory, which may have implications for the way in which this information is used. Integration of information from two separate memory systems may not be as easy as integration of several pieces of information from the same memory system.

Wyer and Srull (1986) present an alternative model—one that has had a significant impact on cognitive research in PA. Some of the aspects of this model are illustrated in Figure 7.3. A storage bin is used as an analogy for human memory. It is assumed that there are several bins, each of which is designed to hold different types of information. Each bin has a label (or a prototype) that describes its contents. Information is stored in a bin in the

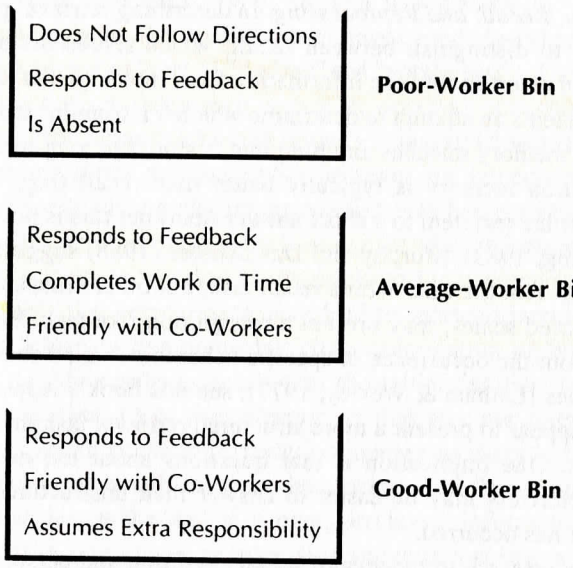

Figure 7.3. A Storage-Bin Model of Memory

order of receipt; the same information may be stored in several different bins. Information that is at the top of each bin is most accessible, and information is retrieved by first locating the appropriate bin and then by searching through the bin. Bins in this model are similar to categories. One implication of this model is that a memory search will not locate an individual piece of information unless the search includes the appropriate bin(s).

The model illustrated in Figure 7.3 suggests two major conclusions. First, some items (e.g., responds to feedback) can be found regardless of which bin is searched. Thus, categorization will not affect the *possibility* of retrieving this item, although it will be less accessible (i.e., the probability of locating the item will be lower) in a search of the poor-worker bin than in a search of the other two bins. Second, failure to search the correct bin(s) will result in a memory failure. Thus, if you search the average-worker or good-worker bins for the item "absent," you will report that you do not remember it; a search of the poor-worker bin is needed to locate that item.

Recognition, Recall, and Reprocessing. In describing retrieval procedures, it is useful to distinguish between *recall,* which represents a relatively unstructured effort to retrieve information from memory, and *recognition,* which represents an attempt to determine whether a stimulus that is present in working memory matches anything that is stored in long-term memory.

Recognition memory is typically better than recall (e.g., compare a multiple-choice test item to a short-answer item) but this is not always the case (Tulving, 1983). Murphy and Davidshofer (1988) suggest that some PA scales present the rater with a recall task, but other scales, particularly behavior-based scales, may present a recognition-like task. Rating scales that ask about the occurrence of specific behaviors (e.g., Behavior Observation Scales [Latham & Wexley, 1977]; see this book's Appendix for an example) appear to present a more structured retrieval task than do global trait scales. One implication is that questions about the occurrence of specific behaviors may be easier to answer than unstructured questions about what has occurred.

Memory research in the area of social cognition and person perception has been primarily concerned with the effects of categorization on memory. Several studies have shown that subjects recall traits that are consistent with the category used to represent persons mentally, even when those traits were not in fact present (Cantor & Mischel, 1977; Hastie & Kumar, 1979; Wyer & Srull, 1986). Other studies have established that manipulating category accessibility at the time of encoding will systematically affect long-term memory (Cohen & Ebbeson, 1979; Higgins & King, 1981; Higgins et al., 1982; Jeffrey & Mischel, 1979). At one point this research seemed to suggest that initial categorization determined the eventual contents of long-term memory. Recent studies of reprocessing suggest that this is not necessarily so.

DeNisi and Williams (1988) reviewed several studies of the effects of reprocessing on memory-related performance evaluation outcomes. Reprocessing occurs when information is obtained (and presumably encoded) for one purpose and subsequently used for another purpose. For example, Williams, DeNisi, Meglino, and Cafferty (1986) had subjects acquire information either for the purpose of making a designation decision (i.e., pick the best worker out of a group) or a deservedness decision (i.e., decide how well each worker performed), and then asked them, several days later, to rate all workers from memory. As expected, subjects in the deservedness condition were more accurate than those in the designation condition.

Williams and associates (1986) explain this finding by noting that subjects in the designation condition may have encoded workers' performance into two categories (such as "the best" and "not the best"). The subsequent PA task asked them to distinguish among all workers—something that was difficult to do for subjects who had initially categorized workers into two gross categories. It was not *impossible,* however, for subjects in the designation condition to distinguish among workers they had previously categorized into the same group, which suggests that information about individuals' performance was not completely lost in categorization.

It is possible that reprocessing may lead to recategorization. The evidence is not clear on this point, but close examination of Williams and colleagues' (1986) results suggests that the initial decision task did not cause raters to lose all of the information that was not relevant to the category (e.g., differences in performance among workers in the "not the best" category). Although most models incorporate the assumption that categorization entails the loss of category-irrelevant information (Ilgen & Feldman, 1983), research on reprocessing suggests that this is not always the case. DeNisi and Williams (1988) suggest that the initial categorization of each person may involve at least two separate codes: a behavior code and an evaluation code. Recategorization may involve changing the emphasis on the behavioral versus evaluative information as a result of the demands of different tasks.

Memory Aids. More than 40 years ago, Wherry proposed that raters use behavioral diaries as memory aids (see Wherry & Bartlett [1982] for a description of a model of rating developed by Wherry in the early 1950s); diaries are still advocated as a useful tool for appraisal (Balzer, 1986; Bernardin & Walter, 1977). The rationale here is that it is very difficult for raters to remember all of the relevant behaviors they observe, and that consulting a behavior diary before rating could lead to more accurate evaluations. Although empirical evaluations of the effects of keeping a diary are somewhat mixed, there are some aspects of memory research that suggest that diaries may indeed be useful.

In order to understand the role of diaries, it is important to assess the precise nature of the memory problems that are most frequently encountered in rating. The problem is not that the information that is needed is not in the system; anything that is stored in long-term memory could, in principle, be retrieved at some later date. Rather, the problem is typically

one of retrieval. The entries in behavioral diaries probably serve as retrieval cues. One implication is that a diary entry such as, "Worker submitted a report that contained many typographical errors" will help to retrieve incidents of that behavior and will also help in retrieving incidents of conceptually similar behaviors (e.g., behaviors that are stored in the same memory bin). Diary-keeping not only aids the recall of behavioral information, it may also help the raters organize that information into coherent categories (DeNisi, Robbins, & Cafferty, 1989).

One unanticipated consequence of using behavior diaries may be a change in the method of categorization. Level-of-processing theories (e.g., Craik & Lockhart, 1972) suggest that consciously attending to and recording behaviors will shift categorization from an automatic to a controlled mode. This will imply more extensive searches for information and a greater likelihood of preserving specific details of what has been observed. It will also mean that raters may now process behavioral information in sequence rather than engaging in parallel processing. In the automatic mode raters can process many pieces of information simultaneously, but in the controlled mode they will process only one piece at a time and will not encode information about other stimuli that are present at the time. It was argued elsewhere (e.g., Murphy & Balzer, 1986) that preserving all of the behavioral details one observes does not necessarily lead to more accurate evaluations and may even detract from accuracy. Thus, the shift from automatic to controlled processing that may result from the use of behavioral diaries could be a mixed blessing.

Rater Training. Rater training might range from simple efforts to acquaint raters with the performance rating scales and the mechanics of the system to elaborate methods of developing skills in observation, rating, and delivering feedback; the variability in scope and the focus of rater training efforts reflect large differences in the purpose of training across organizations. Many training programs focus on developing skills relevant to observing and evaluating ratee behavior.

The purpose of many rater training programs is to increase the validity, fairness, and accuracy of ratings. Over the years the emphasis in training research and application seems to have shifted away from training designed to eliminate specific problems with rating data (e.g., training to reduce rater errors; Bernardin & Pence, 1980; Pulakos, 1984), and toward training methods that increase raters' ability to accurately observe, recall, and/or

classify ratees' behavior. There is some consensus that one of the more successful training methods for that purpose is some variation of frame-of-reference (FOR) training (Bernardin & Buckley, 1981; Bernardin & Pence, 1980; McIntyre et al., 1984; Pulakos, 1984, 1986). FOR training is designed to provide raters with a common set of standards for evaluating their subordinates, and it typically involves training raters to match behavioral exemplars to performance dimensions and to map behavioral exemplars to a common evaluative rating scale. If successful, FOR training helps to "calibrate" raters so that particular scores on individual rating dimensions have at least roughly equivalent meanings for all raters.

Another promising method of rater training utilized cognitive modeling principles. Here, trainers encourage raters to "think aloud" or to make their judgment processes explicit, provide feedback about the way in which judgments are being made, and demonstrate judgment processes that are likely to be effective (Luthans & Kreitner, 1985; McIntyre, 1986). In this method of training, raters attempt to model their own judgment processes on those processes that are known to be effective.

FOR and cognitive modeling methods share a number of components that are likely to contribute to their effectiveness, in particular the use of practice and feedback (McIntyre et al., 1984; Pulakos, 1984, 1986). The apparently critical role of practice and feedback in rater training mirrors its role in general learning theory. There is evidence from a wide range of sources that practice and feedback enhance learning and retention (e.g., Bandura, 1986), and it is possible that the use of practice and feedback is more critical than the specific training method (e.g., FOR vs. cognitive modeling) that is used.

Integration

Both the DeNisi and the Feldman models imply that integration of different pieces of information to form an overall judgment occurs after encoding, storage, and retrieval (see also Cooper, 1981b), although both models allow for the fact that previous information may effect subsequent encoding. This possibility has been examined in a series of studies by Murphy et al. (1985) and by Murphy, Gannett et al. (1986). The main results of these studies are summarized in Table 7.2.

These studies showed that both previous and subsequent performance can affect evaluations of present performance. Also, the integration of

TABLE 7.2 Effects of Previous and Subsequent Performance on Evaluations of Present Performance

Delay Between Observation and Rating	*Previous Performance*	*Subsequent Performance*
Short	Strong contrast effect	Weak assimilation effect
Long	Weak contrast effect	Strong assimilation effect

SOURCES: K. R. Murphy, W. K. Balzer, M. Lockhart, & E. Eisenman, "Effects of Previous Performance on Evaluations of Present Performance," *Journal of Applied Psychology, 70* (1985) pp. 72-84. Also: K. R. Murphy, B. A. Gannett, B. M. Herr, & J. A. Chen, "Effects of Subsequent Performance on Evaluations of Previous Performance," *Journal of Applied Psychology, 71* (1986) pp. 427-431.
NOTES: Contrast: Ratings are biased in the opposite direction from that of previous performance. Assimilation: Ratings are biased in the same direction as that of subsequent performance.

information about present performance with information about previous and subsequent performance can lead to either assimilation or contrast effects. By itself, demonstration of assimilation and contrast effects in evaluation is neither new nor impressive but the overall *pattern* of results shown in Table 7.2 is highly informative.

First, previous performance sets up expectations for the future, and when present performance is different from previous performance, contrast effects occur. For example, if a worker usually does a very good job on a task, but this time does only an average job, this present performance will be perceived as poor rather than average. As memory demands increase, this contrast effect grows weaker. Second, when subsequent performance differs from present performance, assimilation effects occur. That is, if a worker is now performing at an above-average level it is likely that one will remember that his or her previous performance was also above average, even though it was in fact not that good. Assimilation effects grow *stronger* as demands on memory are increased.

A study by Steiner and Raine (1989) suggests some boundary conditions for the conclusions reached by Murphy and colleagues (1985, 1986). They noted that the likelihood of assimilation versus contrast effects in evaluations of performance depended in part on the average performance level of the ratee. In particular, an employee whose performance is typically average but whose present performance is discrepant with the norm must be performing either very well or very poorly during the appraisal period. An employee whose typical performance is either very good or very poor and whose present performance is discrepant from the norm may be an

average performer during the appraisal period. Contrast effects are more likely when the performance being evaluated is average; assimilation effects are more likely when the performance being evaluated is extreme (Murphy et al., 1985).

The studies summarized in Table 7.2 suggest an ongoing integration process, in that previous and subsequent performance affect two distinct information-processing activities. First, the expectations that are generated by previous performance direct one's attention to instances of present performance that differ from those expectations. In terms of the Feldman model, present performance that is different from previous performance may be processed in a controlled mode, whereas performance that is consistent with expectations may be processed in an automatic mode. Increasing one's attention to behaviors that violate expectations will lead raters to overestimate the differences between present and past performance, which results in a contrast effect.

Second, when subsequent performance is different from previous performance, one's *memory* for previous performance is biased; here, assimilation rather than contrast effects occur. The effects of subsequent performance on ratings of present performance cannot be due to biases in attention and encoding (Murphy et al., 1986), because present performance is encoded before subsequent performance even occurs. The fact that attention and encoding are ruled out, together with the pattern of results shown in Table 7.2, suggest that subsequent performance affects the retrieval of information about present performance from long-term memory.

The results summarized in Table 7.2 have clear implications for performance evaluations in organizations. Because performance appraisals are typically conducted at infrequent intervals and are generally done from memory rather than using behavior diaries or other memory aids, results obtained when there is a relatively long delay between observations are most relevant. These results suggest that previous performance will have only a weak effect on evaluations, but that subsequent performance may lead to strong assimilation effects in rating performance that has occurred in the past.

In all, the results shown in Table 7.2 suggest that performance evaluations may lead to underestimates of the variability (over time) of a worker's job performance. If assimilation effects occur, raters will remember the past as being more similar to the present than is actually the case. Therefore, raters may have a very hard time accurately evaluating changes in their subordinates' performance.

Judgment Research. Although many cognitive models allow for the possibility of ongoing integration of information about performance (as described above), most models also assume that some summary integration is carried out at the time the performance is formally evaluated. Research on judgment and decision making has attempted to address processes that might be involved when judges are asked to integrate several pieces of information to arrive at a judgment.

Table 7.3 illustrates the type of task that is often used in studying judgment. Subjects are given multiple pieces of information about each person or object (in this case, about each bank) and are asked to make judgments on the basis of that information. Regression models, subjectively weighted models, and Bayesian opinion revision models have all been used in studying judgments of the type illustrated in Table 7.3 (Hammond, McClelland, & Mumpower, 1980; Slovic & Lichtenstein, 1971). Anderson and colleagues have applied more complex designs in studying algebraic models of human judgment (Anderson, 1971; Anderson & Alexander, 1971).

The approach illustrated in Table 7.3 uses an algebraic analogy to explain and explore information integration. That is, a judgment about a person or object is assumed to be the result of a weighting and averaging process. It is *not* assumed that people actually calculate the average of their evaluation of each property or cue when judging a person or object, but it is assumed that determining the weights assigned to properties will help to tell how much influence each property had on the final judgment. The term *policy capturing* is often used to refer to an approach in which a mathematical model of the judgment is used to infer the importance of each cue or property in judgment.

Although policy capturing has been used to study information integration in performance appraisal (Zedeck & Cascio, 1982), this approach is no longer as popular as it was in the 1960s and 1970s. The principal problem with policy capturing is that the mathematical models employed frequently lead to ambiguous or unclear results. This can be illustrated with a simple example. Returning to the type of study illustrated in Table 7.3, the policy-capturing approach would entail the following:

1. Ask each subject to judge a large number of different banks (typically 50 to 100), each described in terms of the four properties included in Table 7.3.

TABLE 7.3 Typical Task Used in Studying Information Integration in Judgment

The properties of several banks are described below. Rate each bank on a scale from 1 (not very desirable) to 7 (very desirable) in terms of how desirable each bank might be as a place to do business.

	Rating
Bank #1	
Interest on savings—very high	
Charges for checking—average	
Accessibility—very easy to get to	
Evening and weekend hours—very limited	
Bank #2	
Interest on savings—very low	
Charges for checking—very low	
Accessibility—average	
Evening and weekend hours—very often open	

NOTE: This task was used by K. R. Murphy, "Assessing the Discriminant Validity of Regression Models and Subjectively Weighted Models of Judgments," *Multivariate Behavioral Research, 17* (1982a) pp. 354-370.

2. Transform the verbal information about each property into scaled values (e.g., Charges for Checking—average would have a scale value of 4 on a 7-point scale).
3. Use the scores on the four cues as predictors and use the judgment about each bank as a criterion in a regression equation.

The regression weights in this equation are assumed to provide information about the importance of each cue in judgment (Lane, Murphy, & Marques, 1982). Thus, if the regression equation computed to predict a subject's judgments yielded regression weights of 1.0, 4.3, 0.20, and 0.33, respectively, for the four cues, you might conclude that (a) Charges for Checking are very important in your evaluation of a bank and (b) you also pay some attention to Interest on Savings but do not really care about Accessibility or Evening and Weekend Hours.

The problem with the conclusions drawn above is that regression equations capture the *outcomes* of judgments, without necessarily capturing the *process.* That is, the regression equation will accurately predict the actual judgments, but that does not mean that the weights that define the equation were actually applied. Many studies have compared subjective weights,

which are obtained by asking the subject how important each cue or property is in his or her judgment, with regression weights (see Hobson & Gibson, 1983; Murphy, 1979). There are two general findings in this literature. First, regression weights are often quite different from subjective weights. Second, subjective weights predict judgments just about as well as do regression weights. In fact, any weights are likely to work about as well as regression weights.

As long as cues are to some extent correlated, the choice of weights in a prediction equation has virtually no effect on the accuracy of predictions (Dawes & Corrigan, 1974; Wainer, 1976). It is therefore hard to argue that one particular set of weights captures the judgment process (i.e., the regression weights) when any other set of weights would do just as well in predicting judgments.

The relevance of judgment research for information integration in performance appraisal is doubtful. (For a different evaluation of policy capturing in PA research, see Hobson & Gibson, 1983.) This research is useful for describing situations in which the decision maker is given several pieces of information at the same time and is required to make judgments solely on the basis of that information. The process of graduate admissions is an example of this. Here, each member of the admissions committee reads a folder describing each applicant in terms of a number of common properties (e.g., grade-point average, etc.) and makes a judgment about each one.

Performance appraisal does not involve a single summary evaluation, but rather probably involves an ongoing process of evaluation and opinion revision. Some decision research is relevant to this sort of ongoing evaluation process (Slovic & Lichtenstein, 1971) but policy-capturing research probably is not. Our conclusion is that we still know relatively little about the information integration processes involved in appraisal, and that policy capturing and related approaches are unlikely to provide an accurate description of these processes.

Applications of the Cognitive Approach

One of the most frequently encountered questions in cognitive research in performance appraisal is whether this approach has led or is likely to lead to advances in the *practice* of PA. On the whole, this approach has not yet led to new advances in application, in part because researchers and practitioners have not cooperated fully (Banks & Murphy, 1985). Some of

the areas where application has been tried have turned out to be blind alleys, whereas others have not yet been pursued on a sufficient scale to determine whether they will work.

In the areas where cognitive research *has* made a contribution, the contribution has often been negative. For example, behavioral anchors on rating scales have traditionally been regarded as useful guides for observing and evaluating ratee behavior (Bernardin & Smith, 1981). Murphy and Constans (1987) showed that behavioral anchors can be a source of bias rather than a source of validity in rating, and that anchors can misdirect attention and retrieval processes (but see Murphy & Pardaffy, 1989).

Other cognitive research (Ilgen & Feldman, 1983; Murphy, Martin, & Garcia, 1982) has served to further undercut the rationale that is typically put forth in arguing for behavior-based scales rather than for trait-based scales. This research suggests that the categories used in encoding and storage are more traitlike than behavioral, and the trait inferences cannot be avoided by simply phrasing scales in terms of behaviors. Rather, it is likely that people infer behavioral information from subjective trait judgments and that behavior-based scales are not as objective as they appear (Murphy, Martin, & Garcia, 1982).

At one time it appeared that cognitive complexity would provide one key to improving the appraisal process. It was argued that individuals differ in their cognitive complexity—or their ability to sort persons and behaviors into many rather than few categories—and that appraisal would be most effective when the dimensional complexity of the rating form matched the level of cognitive complexity of the rater (Schneier, 1977). For example, if all work behavior is mentally sorted into two categories (e.g., task accomplishment and interpersonal relations), it will be difficult to rate subordinates using a form that requires performance to be separated into 15 dimensions.

Studies by Bernardin, Cardy, and Carlyle (1982) and Lahey and Saal (1981) cast doubt on the cognitive complexity hypothesis. One major problem with this line of thinking is the construct validity of the measures that are typically used to assess cognitive complexity. In our opinion it is doubtful that there are stable individual differences in cognitive complexity (i.e., there may be no such thing as cognitive complexity). If there are, the construct is very difficult to measure.

Although the examples cited above suggest that, to date, applications of cognitive research have not advanced the practice of performance appraisal,

there are reasons to be optimistic for the future. There are two areas where progress in cognitive research is likely to lead to progress in application.

The first area is in rating scale formats. Landy and Farr (1980) noted that the great majority of the studies of performance appraisal over the past 35 years have been concerned with determining whether one scale format was better than another. The contribution of this research has been so minimal (no one format is consistently better or worse than the others) that Landy and Farr (1980) called for an end to scale format research. Feldman (1986) and Murphy and Constans (1988), however, noted that previous research on scale formats has largely ignored cognitive issues. It is quite likely that different scales may involve different cognitive processes, which could have a definite impact on rating processes. Although no one scale is likely to be best in all situations, selecting the scale according to the cognitive demands imposed may lead to significant improvements in rating. For example, if the purpose of rating is to identify candidates for promotion, scale formats that concentrate on between-person differences (e.g., ranking rather than rating) may lead to more valid decisions than scales that concentrate on individual strengths and weaknesses.

A second area in which cognitive research has clear potential for improving practice is in the area of rater training. Feldman (1986) has suggested that training should provide raters with uniform and valid schema and prototypes. If all raters knew what good performance looked like and agreed in their definitions of good, average, and poor performance, it is likely that the quality of rating data would improve dramatically. Feldman's (1986) suggestions are similar to what is already done in frame-of-reference training (Bernardin & Beatty, 1984) and reflect ideas that were suggested in research on interviewing and on performance appraisals carried out in the 1950s (see Wherry & Bartlett [1982] and Webster [1982] for reviews). The advantage of the cognitive approach is that it provides methods of assessing the categories and prototypes that are currently being used (Lord, 1985b).

Assessment of the raters' current category systems represents one step in training needs assessment; if raters are already using valid schema and prototypes, training of this type may not be needed. However, in situations where prototypes vary over raters, or where the category of "good performance" is defined in ways that are inconsistent with organizational values and norms, this sort of training might be very useful.

Limitations of the Cognitive Approach

There are two very different criticisms of cognitive research in PA: (a) it is not good science, and (b) it does not lead to good practice. The scientific criticism of this approach centers on shortcomings of schema-based and categorization-based theories. It is important to note that *direct* evidence for categorization processes of the sort featured in the DeNisi and the Feldman models is hard to come by (Heneman et al., 1987; Ilgen & Feldman, 1983). Although research in this area has shown that categorization-induced errors *can* occur, there is little evidence that these errors are frequent or important (Funder, 1987). It is even more important to note that schematic theories themselves are open to attack. Alba and Hasher (1983) noted that schematic theories are not needed to account for many of the phenomena that are currently thought to result from categorization, and that some of the critical predictions of schema-based theories of memory have not been supported. Alba and Hasher's (1983) article asked the question, Is memory schematic? There is still no clear answer.

Social cognition research represents only one small section of the broader literature on human information processing (Lord & Maher, 1991). It is likely that cognitive research in performance appraisal will at some future point establish contact with that broader literature but this has not yet happened. The concepts that drive much of the current cognitive research described here are useful but they do not exhaust the possibilities. For example, an exchange of ideas with researchers in artificial intelligence (AI) would probably be fruitful for both groups. Much of the cognitively oriented research in AI is concerned with understanding the way in which experts make judgments and decisions. It is not known if raters in organizations could be considered experts in evaluation (Philbin, 1988), but even if they do not meet the standard definition of *expert,* many of the concepts from expert systems research may be useful for understanding performance appraisal.

Many other critics suggest that cognitively oriented research is too concerned with theory and not sufficiently concerned with application (Banks & Murphy, 1985; Ilgen & Favero, 1985; Latham, 1986; Wexley & Klimoski, 1984). Part of the problem is that the issues that are most often examined in cognitive research are not highly relevant to appraisal in organizations (Banks & Murphy, 1985).

There is even greater concern, however, over the issue of external generalizability. Cognitive research is almost always carried out in the laboratory, using college students as subjects. Although it is recognized that laboratory research can be used to make valid generalizations to the field (Locke, 1986), there is justifiable concern over the extent to which the typical lab study captures the essential features of appraisal in organizations. Bernardin and Villanova (1986) have described several critical features that are present in performance appraisals in organizations but are typically lacking in laboratory studies. These include: (a) the use of multiple ratees, (b) real consequences for the rater and the ratee, (c) time pressures, and (d) ratees who know the rater and who sign off on the rating form. The absence of these features may seriously restrict the generalizability of laboratory studies.

In our opinion, the debate over the external generalizability of lab studies of cognitive processes in appraisal has been somewhat misguided because of a failure to distinguish between performance *evaluation* (a private judgment) and performance *appraisal* (a public, written form that is submitted to the organization). Consequences for the rater and the ratee are the result of appraisals, not judgments. This suggests that Points b and d above may not be relevant when laboratory studies are used to make inferences about *judgments* in organizations. On the other hand, Point c is probably relevant to both judgments and appraisals.

For the most part, however, laboratory research is concerned with judgments rather than with decisions that have real consequences. It is incumbent on cognitive researchers to make this distinction and to indicate clearly what aspects of the overall appraisal process are or are not being examined in a particular study. It is incumbent on critics of cognitive research to recognize that laboratory studies do not have to mimic all of the features of performance appraisals in the field to provide useful information about the way in which raters transform their observations of ratee behavior to judgments about their effectiveness.

Research focusing on the cognitive processes involved in performance appraisal has made both theoretical and practical contributions (Ilgen, Barnes-Farrell, & McKellin, 1993). For example, it has advanced our understanding of such things as the role of attention, storage, and retrieval in complex judgment tasks. It has also suggested that early assumptions about the importance of behavioral-based as opposed to trait-based appraisal systems were unrealistic; it seems clear that incorporating behavioral information into appraisals does not solve the problem of subjectivity. However, it is hard to avoid the conclusion that, given the volume of research in this area, the overall contributions have been relatively minor.

AFFECT AND PERFORMANCE EVALUATION

Dipboye (1985) noted that recent research on the interview has addressed cognitive variables at length but has largely ignored several other important variables. In particular, both this literature and the PA literature have paid little attention to the issue of affect. Cognitive models portray evaluation as a "cold" process that does not explicitly involve emotion. It is more accurate to think of performance evaluation as an instance of "hot cognition" or of judgment that involves both cognitive and affective processes. Raters typically feel strongly about several of their subordinates and these feelings are likely to influence their evaluations (see Cardy & Dobbins [1994] for a review of research on the role of affect in appraisal).

Unfortunately, *affect* is a fairly broad term. PA research that has examined affect has generally concentrated on liking (Cardy & Dobbins, 1986; Dobbins & Russell, 1986). *Liking* can be thought of as directed affect—it represents an emotional reaction to a specific person. Not all affect, however, is directed toward a specific person. Rather, there are at least two categories of undirected affect: (a) mood—which represents transient undirected affect, and (b) temperament—which represents chronic undirected affect. It is likely that both mood and temperament influence evaluations. A rater who is in a good mood at the time of observation and/or evaluation may give more positive evaluations than one who is in a sour mood. Raters who have a very positive, upbeat disposition may evaluate performance differently than raters whose temperament is surly or mean.

It is likely that raters' reactions to demographic cues (e.g., race, gender), physical attractiveness, and nonverbal behaviors are more affective than cognitive (Demuse, 1987). That is, a rater may have a "gut reaction" to a person of a different race or to a very attractive (or unattractive) person; this reaction will color the rater's evaluation. Zajonc (1980) has argued that affective reactions may, in some instances, occur independently of cogni-

tive reactions. Thus, if a rater "knows" that an attractive person does not necessarily make a better secretary than an unattractive one, this may not prevent the rater from reacting positively to the attractive person and negatively to the unattractive one. Affective biases may be very difficult to detect and remove.

Affect and Cognition

Affect can influence cognition in many ways (Robbins & DeNisi, 1994). First, one's affective reaction to a ratee could serve as one piece of information to be integrated with other pieces (e.g., observations of behavior, reports from customers). It is thus possible that affect directly influences evaluations (Isen, Shalker, Clark, & Karp, 1978). It is likely, however, that the influence of affect on evaluations is at least partially indirect. That is, affect may change the cognitive processes themselves, as well as serving as a cue to be cognitively processed.

Studies by Clark and Isen (1982) and Isen and Daubman (1984) have shown that mood can affect encoding and retrieval processes. There is some evidence for mood congruency effects (Bower, 1981; Teasdale & Fogarty, 1979), which occur when the similarity between one's mood at the time of encoding and one's mood at the time of retrieval is positively correlated with the likelihood of retrieving specific pieces of information. The reasoning here is that mood can serve as a retrieval cue. Although it has been shown that having the same cues available at both encoding and retrieval enhances memory (Tulving, 1983), it is still not clear whether mood congruency effects really occur, or if they do, whether they are strong enough to make any practical difference.

At one time the existence of mood congruency effects led to the suggestion that learning might be state-dependent, in that material learned under one mood (e.g., a positive mood) would be more easily remembered while the learner was in that same mood than when the mood at learning was different from the mood at the time of testing (Bower, 1981). This suggestion is no longer considered a viable one; if congruency effects exist at all, they are likely to be too weak to substantially affect learning.

Affect influences the categorization process in two ways. First, affect increases the salience of some categories (Tajfel, 1969). It is possible that some categories are conceptually associated with affective states (e.g., a

category of "things you detest"); the association between affect and categories may even result from classical conditioning. If a category of stimuli is repeatedly paired with aversive (or desired) outcomes, that category may develop a strong affective connotation. Affect also influences category breadth (Sinclair, 1988). There is evidence that positive affect leads to an increase in the breadth of categories, in such a way that persons in a positive affective state will classify more persons or objects into a given category than will persons in a negative affective state.

Although the precise explanation is not yet clear, there is evidence that interrater agreement is higher among raters with similar affect than among raters with dissimilar affect (Tsui & Barry, 1986; Zajonc, 1980). This may be because affect itself is a piece of data that is integrated into the evaluation, but may also reflect the indirect influence of affect. That is, raters with similar affect may employ similar categorization strategies and may have similar retrieval cues available at the time of evaluation.

Research suggests that the effects of affect are not symmetric, in the sense that positive and negative affect do not necessarily lead to opposite outcomes (Isen, 1984). For example, positive affect is more strongly linked to halo than is negative affect (Williams & Keating, 1987). In one area, however—leniency—there is evidence of symmetric effects (Williams, Allinger, & Pulliam, 1988). Positive affect leads to lenient ratings, whereas negative affect leads to ratings that are unduly severe.

The current research on affective bases of performance evaluation and appraisal has barely scratched the surface of an important question. It seems reasonable that the influence of affect on evaluations can be as strong as or even stronger than the influence of cognitive variables. Managers like some of their subordinates, dislike others, and have no strong feelings about still others. Some managers have happy dispositions, whereas some focus on the negative sides of life. Managers are sometimes in good moods and sometimes in bad moods. All of these variables are likely to influence managers' judgments regarding the performance of their subordinates. Furthermore, affect effects the judgments of some raters but not others (Harris & Sackett, 1988). Current understanding of affective processes in performance evaluation and appraisal is not sufficiently advanced to allow us to predict the strength or direction of all of these effects, or to determine the circumstances under which affect will have a large or a small effect on evaluations. There is clearly need for more work in this area.

SUMMARY

Information processing has become a central concern in many areas of psychology. During the 1980s, performance appraisal research was swept up in the so-called cognitive revolution. Most current PA studies are carried out in the lab, using college students as subjects and using a variety of tasks that appear to mimic features of appraisals in organizations. This research has advanced our understanding of evaluative judgment, but it is not clear whether it has contributed to our understanding of appraisals in organizations.

Cognitive research has dealt with the acquisition, encoding, storage, and retrieval processes involved in judgment; less research has been devoted to the integration of information. One general conclusion that can be reached on the basis of this research is that the way in which information is initially categorized and encoded can have a substantial and perhaps decisive impact on the subsequent use of that information. This suggests that training raters to develop consistent and accurate categorization schemes (as in frame-of-reference training) are probably worthwhile.

Research on the influence of affect on appraisals suggests that affect is an important component of appraisal. The rater's mood, temperament, and like or dislike of individual ratees is almost certain to influence ratings.

The most significant challenge to cognitive and affective researchers is to develop applications for their findings. As we noted in this chapter, most of the applications of this research to date have been negative (i.e., demonstrations that existing techniques do not work). The field needs to move beyond this stage, which will require the active cooperation of researchers and practitioners. It may be several years before it is possible to determine whether research on cognition and affect will have an impact on performance appraisal in organizations.

8

Rater Goals

Performance appraisal (PA) research has traditionally treated appraisal as a measurement process in which the principal goals were to provide reliable and valid measures of ratee performance. As a result, many of the deficiencies of appraisal systems are blamed on inadequate measurement instruments (i.e., rating scales), inadequate training, inappropriate schema and cognitive structures, and so on.

As discussed in Chapter 1, it may be more useful to treat appraisal as a goal-directed *communication* process in which the rater attempts to use PA to advance his or her interests. Raters are not passive measurement instruments, and it is unwise to assume that they are always trying to provide accurate measures of each ratee's performance. This is not to say that the rater always, or even often, engages in cynical manipulation of appraisals. Rather, it is a simple recognition of the fact that raters are likely to consider the implications of the ratings they give and may sometimes conclude that