Performance Appraisal Process Research in the 1980s: What Has It Contributed to Appraisals in Use?

Daniel R. Ilgen
Michigan State University

Janet L. Barnes-Farrell
University of Connecticut

and

David B. McKellin
Michigan State University

In the early 1980s, Landy and Farr (1980) and Feldman (1981) redirected performance appraisal research from issues related to the development of psychometrically sound rating scales to those involving the cognitive processes of raters. Since that time, several reviews have attempted to translate principles from social cognition and cognitive psychology to the specific conditions of formal appraisal systems in work-oriented organizations. In addition, a number of empirical studies have been conducted on this topic. This article reviews empirical research during the 1980s that focused on performance appraisal processes, particularly the research that has focused upon rating accuracy. The review is structured around a three-stage process model of gathering, storing, and retrieving information about social stimuli for the purposes of rating performance. Factors affecting this process are clustered into four categories: appraisal settings, ratees, raters, and the nature of the scales used for the appraisal. Once reviewed, the research is evaluated in terms of its contributions to improving the quality of appraisal systems as they are used in organizations. © 1993 Academic Press, Inc.

In 1980, Landy and Farr published a watershed review and evaluation of the performance appraisal literature that both stimulated and redirected research on the topic. Prior to that time, most research was directed toward developing rating scales that were valid and reliable, training

Portions of this paper were presented at the 21st International Congress of Applied Psychology, Jerusalem, Israel, July 1986. This paper was supported, in part, by grants from the Office of Naval Research, Organizational Effectiveness Unit (Grant N00014-83-K-0756, Daniel R. Ilgen, Principal Investigator, and Grant N00014-83-K-0757, Janet L. Barnes-Farrell, Principal Investigator). Although the support is greatly appreciated, the ideas expressed herein are those of the authors and not necessarily endorsed by the supporting agency.
ratees to reduce rating errors and improve their observational skills, and developing performance appraisal practices. The Landy and Farr (1980) article, along with another by Feldman (1981), shifted the focus of performance appraisal research from scales and rater training to understanding the rater as a decision-maker who processes social cues. The rater was construed as an information processor confronted with a judgment task involving a social stimulus, the ratee. From this perspective, research focused on the accuracy with which raters make such judgments and the application of knowledge about judgment processes to the development of appraisal systems.

The focus on raters has dominated the literature on performance appraisal since 1980. Several theoretical reviews and extensions of the rater perspective have been published (see for example, DeNisi, Cafferty, & Meglino, 1984; Ilgen & Feldman, 1983; Wexley & Klimoski, 1984). Our purpose here is to review only the empirical research on appraisals that was done in the 1980s which has as its focus the rater’s rating process. In doing so, we shall first present a conceptual framework for performance appraisal from the appraiser’s perspective and then evaluate empirical research as it relates to that framework. Our primary concern is to describe the research and then comment on the contribution of process-focused research for developing effective performance appraisal systems.

The issue of contribution deserves some elaboration. We believe that research on the cognitive processes of raters in organizational settings has the potential of making two primary contributions to the knowledge base. The first contribution is to basic research directed at understanding fundamental cognitive processes. Rater cognitive process research, in our opinion, has a great deal of potential for contributing to the understanding of social cognition and the way it functions in judgment processes. A second way in which rater process research contributes is by providing information that suggests ways to improve performance appraisal systems in organizations. This second contribution is more indirect than the first. That is to say, knowledge is gained about the effects of various conditions on raters’ ability to make judgments about others, and this knowledge can then be used to create organizational policies and practices that take advantage of what is known about rater capabilities. Although we consider the first of these two areas of contribution legitimate and important, our focus here is on the second.

**PERFORMANCE APPRAISAL: A RATER PROCESS PERSPECTIVE**

**A Process Model**

Appraising performance from the appraiser’s perspective is construed as a process of cognitively processing information in order to make judg-
ments/evaluations that are associated with others. As such, three critical sets of operations have been identified. These are: (1) acquisition of information about those to be evaluated, (2) organization and storage of this information in memory, and (3) retrieval and integration of the information in a fashion that leads to the recording of an evaluation of the person being appraised (DeNisi et al., 1984; Feldman, 1981; Ilgen & Feldman, 1983). According to this view, the process begins with appraisers obtaining information about the persons appraised. The information acquisition process varies both within raters and between raters in the degree to which acquisition is systematic, conscious, and guided by valid predispositions about appraisees' performance. Organization and storage of the information in memory follows its acquisition. Here again, the amount of attention paid to the nature of the dimensions on which information is stored and the nature of the dimensions themselves vary as a function of many factors. The last set of processes are recall and evaluation of information stored in memory. Recall involves the appraiser's ability to remember information stored in memory, and evaluation combines that information in some fashion in order to form an evaluative judgment about the appraisee's performance. Often recall and evaluation are treated as separate functions. However, although these two cognitive activities can be separated theoretically, they are inextricably paired in practice. For that reason, we have chosen to treat them together.

The three phases of the rating process are also assumed to be loosely temporally related. The process begins with the acquisition of information, which is then stored in memory and finally recalled at some time in the future (Feldman, 1981). This temporal ordering has been demonstrated in a number of settings (Ilgen & Feldman, 1983), but the order is not without exception. For example, the nature of the cognitive systems needed to store information about appraisees also influences the types of behaviors that are observed—or acquired—in the information storage phase (DeNisi et al., 1984). Likewise, evaluations affect the prior two phases (observation and storage) by directing attention and other mechanisms. This point will be raised later when each phase is discussed. These evaluations, in turn, may affect the nature of cognitive systems employed in future observations. But, despite reversals and oversimplification, this general order has been found to be a useful way to construe the process (Feldman, 1981).

A great deal of research in cognitive psychology and, in particular, social cognition has addressed theoretical issues in each of the three process domains. Authors have borrowed heavily from the basic psychological literature to develop theories of the performance appraisal process (see, for example, DeNisi et al., 1984; Feldman, 1981; Ilgen & Feldman, 1983; Wexley & Klimoski, 1984). However, much of the basic research
from cognitive and social psychology has failed to develop constructs and operational conditions essential for transfer of knowledge specifically to performance appraisal settings (Ilgen & Favero, 1985). Therefore, in order to judge the relevance of these literatures to the appraisal setting, it is often necessary to conduct research with the explicit purpose of focusing on performance appraisal. The current review will be limited to empirical studies explicitly investigating the rating process as it relates to performance appraisal.

Process Approaches in Perspective

To evaluate the objectives of performance appraisal process research in general and those of this review in particular, consider a simple pie chart where the area of the circle represents the total set of variables that affect the effectiveness of performance appraisal systems in practice. Let one slice of the pie represent the research on scale construction and training that dominated performance appraisal research prior to 1980, let a second slice represent the performance process research of the 1980s, and let the remainder of the area represent unlabeled sources of variance. For the time being, assume that the unlabeled portion of the total area far exceeds that of the other two. Within this framework, our review will first address what has been found with respect to the contents of the slice that we have labeled the process approach. We will then comment on the size of the slice relative to the total space and to the space represented by scale and training research. Finally, we will comment on whether or not it appears as if future research is better directed at understanding better rater processes in order to push the boundary of the process area farther into the unlabeled area or whether it appears that the process boundary has been pushed about as far as it will go in terms of leading to a greater understanding of performance appraisal systems in use.

Ratings and Judgments

The end product of a performance appraisal is a rating of one individual that is produced by another. There are a wide variety of factors that influence the rating recorded on an appraisal form, some of which are related primarily to the rater’s best judgment about the behavior of the person that is rated and some of which are related to other factors such as the rater’s willingness to give a low rating to the person even if he or she believes the ratee deserves it. Research on the appraisal process, by choice, focuses upon the judgment component of rating. It is the judgment process that dominates the area of the circle described in the previous section. At the same time, some of the research uses ratings as indications of rater judgments. As will be pointed out, such measures are sometimes
appropriate for measuring judgments and at other times they are too contaminated to provide very good measures of judgment.

The direct result of the role of judgment in the performance appraisal process is that empirical research in this area is dominated by concerns for performance appraisal accuracy as the primary criterion of interest. The appropriateness of this criterion can be considered from two perspectives. The first of these is from the standpoint of the scientific study of the appraisal process as it relates to raters’ capability for rating others. In this case, accuracy is the central concern since the goal of the research is to understand how and under what conditions peoples’ perceptions of others are or are not veridical with the cues that are available to them regarding the others. For this research, accuracy is a very appropriate criterion.

With respect to the appropriateness of rating accuracy as a criterion for appraisal systems in use, the case is less strong. From the standpoint of rating validity, accuracy is a very legitimate criterion. If ratings cannot be shown to have any relationship to the characteristics of the persons being rated, it is difficult to justify the use of such ratings for most any purpose. Thus, some level of accuracy is a necessary but not sufficient criterion for ratings. Other criteria, such as specificity or perceived fairness may be as important or more so when it comes to the impact of the rating in practice. There may even be times when communicating accurate ratings is detrimental to the operation of the unit in question. For example, accurate ratings may point out differences in performance among people when teamwork and working closely together is the more desirable state for the group. Nevertheless, this limitation with respect to the use of accurate ratings does not limit accuracy as a criterion for understanding the rating process. If one is interested in the rating component of appraisals and what the rater can do, then accuracy can be used to understand this. Whether accurate ratings should be used is a question of application that can be addressed after we understand what the rater can do. It is in the understanding mode that process research has operated, and it should be judged in that mode before the application mode is applied. In this review, we will first address what has been learned in the understanding mode and then address application. In doing that, much of the research we review will focus on performance accuracy. Since the measurement of rating accuracy is complex and, at times, confusing, we will first discuss several issues related to accuracy measures before turning to the review.

**Performance Rating Accuracy Measures**

**Accuracy criteria.** One of the major impediments to the study of rating accuracy in the context of performance appraisals has been difficulty identifying and developing appropriate criteria for accuracy. Latham (Latham 1986; Latham & Wexley, 1981) echoed the early advice of
Thorndike (1949) and argued that good appraisals should be reliable, valid, practical, and accepted by users. Underlying the first two of these is the notion that appraisals should provide an accurate representation of the domain they purport to measure. For a number of years, accuracy was indexed indirectly by assuming that the presence of a variety of common psychometric biases, termed rating errors (e.g., leniency, central tendency, and halo), implied a lack of accuracy. It was assumed that decreasing the biases increased accuracy. The fallacy of this argument was that bias-free ratings (as operationally defined) were not necessarily accurate; rater biases were weak surrogates for more direct indices of the accuracy construct (Hulin, 1982; Murphy & Balzer, 1989). The need for better measures of rating accuracy was quite clear, but their development was far less straightforward.

The concern for accuracy shifted from rater errors to the discrepancy between ratings and some known standard of performance. Conceptually, the discrepancy view was far more appealing, but practically it too was plagued with problems. Some of the problems resulted from the difficulty of knowing the standard or "true" level of performance to which to compare ratings (cf. Smither, Barry, & Reilly, 1989; Sulsky & Balzer, 1988; Weekley & Gier, 1989). A partial list of other problems are (1) the reliability of the accuracy index itself, (2) extraneous factors that contribute to variance in the index, and (3) the extent to which the measure is sensitive to errors of exclusion or inclusion (Cronbach 1958; Lord, 1985; Sulsky & Balzer, 1988; Padgett & Ilgen, 1989).

Addressing all or even most of the issues surrounding the assessment of performance rating accuracy is beyond the scope of this review. However, because this review focuses on research that has used as its criterion some measure of rating accuracy, and in recognition of the fact that accuracy is not a unitary construct, it will be necessary for us to describe or refer to the nature of the accuracy measures relevant to each study. Therefore, we shall briefly comment on the measures of accuracy that will be used in the research reviewed.

*Cronbach's discrepancy scores.* A number of researchers have turned to work on the measurement of veridicality in person perception as the basis for performance appraisal accuracy indices. Of these positions, Cronbach's (1955) multivariate conception of perceptual accuracy dominated the literature for some time. His indices of accuracy were based on the assumption that a rater evaluated a number of ratees on a finite set of attributes or dimensions and that there existed some assessments of each ratee's performance on every dimension from a source other than the rater's judgment. Raters' judgments could then be compared to the other assessments and the discrepancy between the two could serve as an inverse index of accuracy. Cronbach's accuracy measures were simply
sums of discrepancy scores from the rows, columns, or within cells of a ratee-by-dimension matrix containing entries based on the differences between a rater's judgment and the standards. *Elevation accuracy* was defined as the average of a rater's ratings across all behaviors and ratees and reflected how the rater interpreted the rating scale. *Differential elevation* reflected a rater's ability to judge deviations in a ratee's performance from his or her average or normal performance level across all job dimensions. *Stereotype accuracy* described how well the rater's judgments agreed with the mean rating for all raters summed over all dimensions. *Differential accuracy* indicated how well a rater discriminated among dimensions across ratees and among ratees. Finally, *overall accuracy* was simply the comparison of the grand mean judgment to the grand mean standard. These components reflected fundamentally different aspects of perceptual accuracy, and the relative importance of each type of accuracy in the performance rating context depends on how those judgments will be employed (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982).

*Other accuracy indices*. In the last few years, a growing appreciation for the distinction between accuracy in the tasks prerequisite to performance appraisal and accuracy of the appraisal itself has led to other indices of rating accuracy. Lord (1985) proposed three accuracy indices. First, when a rater observed, accurately recalled, and correctly recognized particular subordinate behaviors (e.g., communication skills), Lord labeled this *behavioral accuracy*. On the other hand, if the rater correctly recognized that the employee behaved consistently with a performance level for a particular category (e.g., highly effective communication skills) and correctly recalled that category, but the specific behavior was not necessarily observed, Lord called this *classification accuracy*. Finally, if ratings were lenient but also consistent with the ratee's actual performance level, Lord argued that these ratings were accurate based on a rater's *decision criteria*. It is the first of these indices, behavioral accuracy, which Lord felt was the most appropriate index of rating accuracy (but not necessarily most desirable for ratings in use), and he suggested that signal detection theory (SDT) provided a useful framework for distinguishing it from other accuracy indices.

A different perspective on what comprises a rater's essential task was proposed by Nathan and Alexander (1985). They pointed out that, although appraisers must observe and recall behavioral information, the key to their effectiveness is the extent to which they make accurate inferences about the information they observed. (See also Dickinson, 1987, and Funder, 1987, for discussions of related issues.) This position suggests an increased focus on a rater's ability to accurately classify ratees' performance level (e.g., low, average, high) and to correctly perceive true
levels of covariation in performance across dimensions, rather than conceptualizing rater accuracy in terms of behavioral accuracy.

Yet another approach to indexing performance rating accuracy has focused on the problem of halo bias in ratings. Although this perspective emphasizes the role of intercorrelations among performance dimensions instead of ratees' absolute performance levels, its logic is similar to that proposed by Cronbach. A major requirement of this approach is the ability, using one of several procedures (e.g. Kenny & Berman, 1980; Landy, Vance, & Barnes-Farrell, 1982; Murphy, 1982), to estimate the true level of intercorrelation among the dimensions being rated (cf. Cooper, 1981). It is important to remember that accurate performance measurement implies that both the intercorrelations among ratings on different performance dimensions and the ratings within those dimensions correspond to the true intercorrelation and performance levels present, respectively. This fact is sometimes obscured in discussions of "true halo" and has often been ignored in operational definitions of halo bias (Pulakos, Schmitt, & Ostroff, 1986).

Summary

Recent empirical studies by Lord (1985) and by Padgett and Ilgen (1989) demonstrated that the effect of such variables as rater performance and rating conditions on accuracy depends on the manner in which accuracy is conceptualized. This raises the question, "Is there a 'best' index of accuracy?" Our answer is no. Rather, we believe that the heart of the issue emerges most clearly from Nathan and Alexander's (1985) contrast between current interest in an appraiser's role as an accurate "observer/recorder" and his or her role as an accurate "evaluator/judge." Both of these roles are real, important, and scientifically interesting. However, the tasks and requirements for effectively carrying out the two roles are not the same. At the same time, the roles cannot be completely separated from one another. The measurement of performance is essentially a two-step process requiring two different types of scaling tasks for the rater. First, the rater must map events—ratee actions—onto a frequency scale reflecting the "amount" of the behavior that occurred. Second, this information must be mapped onto an evaluative scale reflecting the quality of the behavior. The problem of distinguishing between these two steps is complex because they overlap chronologically, and because they mutually affect one another. Since some accuracy indices relate more clearly to one than to the other, decisions about the most appropriate criterion measures require both an awareness that accuracy is not a unitary concept and a clear understanding of the particular aspects of the appraiser's role one wishes to study. This makes it impossible to summarize comprehensively all of the indices which have been used in research-
ing performance rating accuracy. Therefore, it will be necessary to state how accuracy was measured in studies that are reviewed.

AN ORGANIZING MODEL

In the review that follows, research related to each one of the three major processing phases in the appraisal model will be addressed. Within each phase, the literature will be divided into four sources, the characteristic impacts of which on the rating process have been studied. First, the effects of ratee characteristics, such as race, gender, likability, and the amount of true halo exhibited by the ratee's performance, will be considered. Next, rater attributes, including race, cognitive style, and knowledge of the job to be rated, will be discussed. Third, the impact of the rating scale on rating accuracy will be examined. Finally, characteristics of the setting within which rating occurs, such as the purpose of the appraisal, whether or not rater training has occurred, and the length of the delay between observing the ratee and rating his or her performance, will be considered.

The four classes of sources were selected, in part, because they represent major topic areas addressed in the literature and, in part, for practical reasons. It is our position that the process focus on performance appraisal arose out of frustration with the limitations of earlier attempts to improve performance appraisal systems. The process approach was predicated on the assumption that, of the factors that affected raters' ability to provide accurate ratings, this knowledge could be applied to design and carry out better appraisals. Yet, in order to "engineer" or construct better appraisal systems, it seems likely that our attempts to intervene will involve the same general mechanisms that have been used in the past. These are: (1) changing raters through training, (2) changing characteristics in the appraisal settings regarding policies and practices, and (3) changing the rating task through the design of the rating instrument. We selected the sources because of their fit to the need to make inferences about ways to improve appraisal systems. Table 1 presents a summary of the research reviewed, classified according to process phases and the sources that impact on the process.

Attention and Observation

While rater cognitive processing has dominated the performance appraisal literature since 1980, all stages of the process have not received equal attention. The stage most neglected is attention and observation. This is unfortunate since job behaviors must first be noticed before they can be stored in memory for later recall and evaluation.

Ratee. In investigating the effects of ratee characteristics on attention and observation, Favero and Ilgen (1989) found that raters spent signifi-
<table>
<thead>
<tr>
<th>Source</th>
<th>Study</th>
<th>Variable(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1: Attention and Observation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratee</td>
<td>Favero &amp; Ilgen (1989)</td>
<td>Prototypicality</td>
</tr>
<tr>
<td></td>
<td>Balzer (1986)</td>
<td>Initial impression</td>
</tr>
<tr>
<td>Rater</td>
<td>Cardy &amp; Kehoe (1984)</td>
<td>Selective attention</td>
</tr>
<tr>
<td></td>
<td>Foti &amp; Lord (1987)</td>
<td>Knowledge of task</td>
</tr>
<tr>
<td>Scale</td>
<td>DeNisi &amp; Summers (1986)</td>
<td>Scale presentation (before/after obs.)</td>
</tr>
<tr>
<td></td>
<td>McDonald (in press)</td>
<td>Dimension content</td>
</tr>
<tr>
<td></td>
<td>Murphy &amp; Constans (1987)</td>
<td>Behavioral anchor presence</td>
</tr>
<tr>
<td></td>
<td>Murphy &amp; Pardafy (1989)</td>
<td>Behavioral anchor presence</td>
</tr>
<tr>
<td></td>
<td>Piotrowski, Barnes-Farrell, &amp; Esrig (1989)</td>
<td>Behavioral anchor presence</td>
</tr>
<tr>
<td>Setting</td>
<td>Foti &amp; Lord (1987)</td>
<td>Purpose</td>
</tr>
<tr>
<td></td>
<td>Balzer (1986)</td>
<td>Centrality of task</td>
</tr>
<tr>
<td></td>
<td>Murphy, Philbin, &amp; Adams (1989)</td>
<td>Purpose</td>
</tr>
<tr>
<td></td>
<td>Williams, DeNisi, Blencoe, &amp; Cafferty (1985)</td>
<td>Purpose</td>
</tr>
<tr>
<td><strong>Phase 2: Storage/Memory</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratee</td>
<td>Foti &amp; Lord (1987)</td>
<td>Leader prototypes</td>
</tr>
<tr>
<td></td>
<td>Frank-Major &amp; Foti (1985)</td>
<td>Behavioral consistency</td>
</tr>
<tr>
<td></td>
<td>Foti &amp; Lord (1987)</td>
<td>Information form</td>
</tr>
<tr>
<td></td>
<td>Ostroff &amp; Ilgen (1985a)</td>
<td>Category match</td>
</tr>
<tr>
<td>Scale</td>
<td>DeNisi &amp; Summers (1986)</td>
<td>Task- vs trait-based scales</td>
</tr>
<tr>
<td>Setting</td>
<td>Foti &amp; Lord (1987)</td>
<td>Purpose</td>
</tr>
<tr>
<td></td>
<td>Lauterbach &amp; Barnes-Farrell (1989)</td>
<td>Information amount</td>
</tr>
<tr>
<td></td>
<td>Williams, Cafferty, &amp; DeNisi (1990)</td>
<td>Salience</td>
</tr>
<tr>
<td></td>
<td>DeNisi, Robbins, &amp; Cafferty (1989)</td>
<td>Salience</td>
</tr>
<tr>
<td></td>
<td>Ostroff &amp; Ilgen (1985b)</td>
<td>Diary-keeping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td><strong>Phase 3: Recall/Evaluation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratee</td>
<td>Schmitt &amp; Lappin (1980)</td>
<td>Race</td>
</tr>
<tr>
<td></td>
<td>Pulakos, White, Oppler, &amp; Borman (1989)</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Dobbins, Cardy, &amp; Truxillo (1986)</td>
<td>Race</td>
</tr>
<tr>
<td></td>
<td>Cardy &amp; Dobbins (1986)</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Murphy &amp; Jako (1989)</td>
<td>Likability</td>
</tr>
<tr>
<td></td>
<td>Murphy &amp; Reynolds (1988)</td>
<td>True halo</td>
</tr>
<tr>
<td></td>
<td>Padgett &amp; Ilgen (1989)</td>
<td>True halo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Performance consistency</td>
</tr>
<tr>
<td>Source</td>
<td>Study</td>
<td>Variable(s)</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>--------------------------------------------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>Smither &amp; Reilly (1987)</td>
<td>True halo</td>
<td></td>
</tr>
<tr>
<td>Schmitt &amp; Lappin (1980)</td>
<td>Race</td>
<td></td>
</tr>
<tr>
<td>Wexley &amp; Youtz (1985)</td>
<td>Beliefs about human nature</td>
<td>Cognitive complexity</td>
</tr>
<tr>
<td>Smither &amp; Reilly (1987)</td>
<td>Intelligence</td>
<td></td>
</tr>
<tr>
<td>Cardy &amp; Dobbins (1986)</td>
<td>Liking</td>
<td></td>
</tr>
<tr>
<td>Sinclair (1988)</td>
<td>Mood</td>
<td></td>
</tr>
<tr>
<td>Zalesny (1990)</td>
<td>Confidence</td>
<td></td>
</tr>
<tr>
<td>Bernardin, Cardy, &amp; Abbott (1982)</td>
<td>Motivation</td>
<td></td>
</tr>
<tr>
<td>Kozlowski &amp; Kirsch (1987)</td>
<td>Job knowledge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ratee familiarity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Conceptual similarity</td>
<td></td>
</tr>
<tr>
<td>Fisicaro (1988)</td>
<td>Halo</td>
<td></td>
</tr>
<tr>
<td>Becker &amp; Cardy (1986)</td>
<td>Halo</td>
<td></td>
</tr>
<tr>
<td>Fox, Ben-Nahum, &amp; Yinon (1989)</td>
<td>Peer-perceived similarity</td>
<td></td>
</tr>
<tr>
<td>Farh &amp; Dobbins (1989)</td>
<td>Self and supervisor CPI</td>
<td></td>
</tr>
<tr>
<td>Mount &amp; Thompson (1987)</td>
<td>Subord.-perceived congruence</td>
<td></td>
</tr>
<tr>
<td>Scale</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fay &amp; Latham (1982)</td>
<td>BOS, BARS, Trait</td>
<td></td>
</tr>
<tr>
<td>Pulkos (1984)</td>
<td>BOS, BARS</td>
<td></td>
</tr>
<tr>
<td>Murphy, Martin, &amp; Garcia (1982)</td>
<td>BOS</td>
<td></td>
</tr>
<tr>
<td>Murphy &amp; Constans (1987)</td>
<td>BARS</td>
<td></td>
</tr>
<tr>
<td>Murphy &amp; Pardaffy (1989)</td>
<td>BARS</td>
<td></td>
</tr>
<tr>
<td>Piotrowski, Barnes-Farrell, &amp; Esrig (1989)</td>
<td>BARS</td>
<td></td>
</tr>
<tr>
<td>Setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heneman &amp; Wexley (1983)</td>
<td>Delay</td>
<td></td>
</tr>
<tr>
<td>Smither &amp; Reilly (1987)</td>
<td>Delay</td>
<td></td>
</tr>
<tr>
<td>Murphy &amp; Balzer (1986)</td>
<td>Delay</td>
<td></td>
</tr>
<tr>
<td>Zedeck &amp; Cascio (1982)</td>
<td>Purpose</td>
<td></td>
</tr>
<tr>
<td>Murphy, Balzer, Keilam, &amp; Armstrong (1984)</td>
<td>Purpose</td>
<td></td>
</tr>
<tr>
<td>Murphy, Philbin, &amp; Adams (1989)</td>
<td>Purpose</td>
<td></td>
</tr>
<tr>
<td>Farh &amp; Werbel (1986)</td>
<td>Purpose</td>
<td>Expectation of validation</td>
</tr>
<tr>
<td>Pulkos (1984)</td>
<td>Training × scale</td>
<td></td>
</tr>
<tr>
<td>Pulkos (1986)</td>
<td>Training × task</td>
<td></td>
</tr>
<tr>
<td>McIntyre, Smith, &amp; Hassett (1984)</td>
<td>Purpose</td>
<td></td>
</tr>
<tr>
<td>Hedge &amp; Kavanaugh (1988)</td>
<td>Training</td>
<td></td>
</tr>
<tr>
<td>Athey &amp; McIntyre (1987)</td>
<td>Training</td>
<td></td>
</tr>
</tbody>
</table>
cantly less time observing ratees who were more easily stereotyped or
classified than those who were less easily classified. In a simulation of the
job of a nursing supervisor, raters were asked to observe and rate the
videotaped performance of a nurse who worked in their unit. Background
information about the nurse was manipulated prior to observation. For
half the raters role-playing the supervisor, the nurse was described as a
"social activist," and for the remainder she was described by several
attributes that did not fit a general category or label. Over three experi-
mental sessions raters were given a number of supervisor activities, one
of which was observing the nurse by watching her videotaped per-
formance. It was found that raters spent more time observing subordinates
described by traits that did not fit a general label than subordinates who
fit a prototypical pattern. These results were congruent with the social
perception literature, which has found that the confirmation of a proto-
type leads raters to conclude that they possess adequate information
about the person being rated, thereby making it unnecessary to collect
additional information (Feldman, 1981).

Favero and Ilgen (1989) also found that as time spent observing the
ratee increased, the accuracy of ratings increased (using Cronbach's mea-
sures of overall accuracy, elevation, and stereotypic accuracy). Similarly,
Heneman and Wexley (1983) found that raters who observed more of a
ratee's performance were more accurate. Finally, Balzer (1986) examined
the effects of initial impressions on the recording of performance infor-
mation. Initial impressions (positive or negative) produced a significant
contrast effect, thus biasing later ratings. Note, here is an example of the
absence of a clear ordering from information gathering to evaluations;
evaluations formed prior to gathering information affect what is observed.

Beyond the few studies cited here, little other empirical work has linked
ratee characteristics to raters' observations of ratees. Theoretically, race,
job title, ratee competence, and other attributes have been suggested as
variables that potentially may influence observations (DeNisi et al., 1984;
Ilgen & Feldman, 1983), but research on these variables is lacking.

Rater. Just as ratee attributes affect the accuracy of performance rat-
ings, rater characteristics may also assume an important role in the ap-
praisal process. Individual difference characteristics that the rater brings
to the evaluation may influence the cues that he or she perceives to be
present when observing ratee behavior. Cardy and Kehoe (1984) pro-
posed that a rater's ability to selectively attend to relevant information
influences rating accuracy. Using the Hidden Figures Test (Jackson, Mes-
sick, & Myers, 1964), raters high or low on field dependence–
independence rated the performance of four hypothetical instructors de-
picted in written vignettes. Raters high in selective attention were more
accurately differentiated among ratees (differential accuracy) than those low on selective attention ability.

Observations may also be influenced by raters' preconceived ideas about what ratees should do (i.e., rater's knowledge of the ratee's task). Although we are aware of no research directly investigating raters' prior knowledge of the ratees' tasks, Foti and Lord's (1987) research with groups implies that prior knowledge is important. In their study, subjects viewed a 15-min videotape of a mock school board meeting. Prior to watching the videotape, half of the raters read instructions outlining the purpose of the group meeting, while the other half received no such information. After watching the tape, raters attempted to recall the behaviors which had occurred. Results showed that raters given explicit information about the purpose of the group meeting tended to recall behaviors relevant to group goals better than those who lacked such information. However, raters from whom this information was withheld tended to organize their recall around general characteristics of the group members. This suggested to Foti and Lord that implicit personality theories guided what was observed and recalled. While it is not clear from Foti and Lord's data, we suspect that knowledge about an individuals' goals would similarly organize individual ratings.

Research to date indicates that rater characteristics influence what ratee behaviors are observed. Lacking, however, is a systematic treatment of which characteristics impact observation, a set of hypotheses about what effects are likely to be observed, and empirical evaluations of these hypotheses.

Scale. To our knowledge, five studies have directly examined rating scale effects on observational behaviors as they relate to accuracy. DeNisi et al. (1984) proposed that exposure to the rating instrument prior to observing ratee performance might alert raters to the kinds of traits or behaviors to which they should attend. DeNisi and Summers (1986) tested this hypothesis.

Following subjects' observation of three ratees via videotape, DeNisi and Summers measured how well raters recalled each ratee's behavior and how accurately they rated performance. One-half of the raters rated performance using trait-based rating scales, while the other half used task-based ones. Rating scales were presented to raters at three stages in the experimental session: one-third received the rating instrument prior to observing the videotape, one-third after watching the tape but prior to completing the recall task, and one-third after finishing the recall task. Results demonstrated that raters given the rating instrument prior to observing ratee behaviors were more accurate at recalling and rating performance. DeNisi and Summers concluded that the introduction of the
rating instrument prior to observation served to "prime" raters' use of a category system for organizing ratees' behaviors (cf. Wyer & Srull, 1981), and that raters who used such a system rated performance more accurately.

McDonald (1991) provided raters with either correct or incorrect information regarding the dimensions of performance that the raters were about to observe and compared each to a group that received no prior information about performance dimensions. The results implied that the prior information activated different schemas which, in turn, influenced observational behaviors. Using a procedure by which participants observed tapes of a classroom instructor and pushed a button when they believed a unit of behavior had changed, she found that cueing persons as to the important performance dimensions led to attending to segments of behavior consistent with expectations.

Murphy and Constans (1987) argued the Behaviorally Anchored Rating Scales (BARS) may bias ratings both by cuing raters who are familiar with the scales into the behaviors to which attention should be paid when observing behavior and by biasing memory after the observations are made. By making the rating instrument available to the raters either before observation or after observation but before rating performance, Murphy and Constans were able to separate observation from memory effects. They found that knowledge of the BARS dimensions and items did bias ratings, but the absence of any difference between those who were made aware of the scale before vs after observing a videotape of a college professor's lecture implied that the effects of the scale were due to memory rather than observation. These effects were replicated by Piotrowski, Barnes-Farrell, and Esrig (1989) in a study which examined the effect of retention interval on such biases. However, the presence of similar levels of scale bias in both immediate and delayed ratings was interpreted as demonstrating that the scale effects were not memory-related. Another recent study by Murphy and Pardaffy (1989) examined the extent to which scale bias generalizes to related performance dimensions. They concluded that BARS anchor bias effects were largely scale-specific. All three of these studies suggest that behavioral anchors affect observation and possibly storage/retrieval processes.

Setting. Research into the effects of setting characteristics on observation has concentrated on observation and rating purpose. In their investigation into the characteristics of raters' cognitive schemas when rating school board meetings, Foti and Lord (1987) found that accuracy was affected by observational purpose. Specifically, they discovered that: (1) those instructed to remember events in the videotaped school board meeting recalled actual events more accurately than those who were either asked to form a general impression of the board's chairperson or those
who were given no memory-based instructions; (2) advice to form impressions of the chairperson created more well-developed impressions; (3) a memory set led to recall of a greater proportion of behaviors consistent with a behavioral set presented before observation; and (4) impression formation instructions led to recalling a greater proportion of information about the school board chairperson than other observational instructions. Overall, these data supported the conclusion that people recall most accurately the behaviors to which they are cued to attend. However, since the purposes given for observations may have also influenced memory/storage process, caution must be taken in interpreting these effects as due solely to observation.

A study by Williams, DeNisi, Blencoe, and Cafferty (1985) more clearly demonstrated the effects of rating purpose on observation. They discovered that raters differentially searched for performance information used to make appraisal decisions for different purposes and outcomes. Subjects were told to conduct a performance appraisal for one of three purposes: (1) salary increase, (2) promotion recommendation, or (3) referral for remedial training. The decision focused on one of two outcome dimensions: (1) the extent to which an individual deserved recognition or some other action, or (2) the actual designation of an individual for a specific treatment or outcome. For all three rating purposes, information about how the ratee performed on other tasks within his or her job scope was sought most often, followed by information about how well all workers performed the given task and how well the ratee performed the same task on other occasions. Furthermore, there was an increased number of search requests for information concerning how others behave in the same setting when making salary and remedial training decisions. The authors concluded that the search for performance information is sensitive to the demands of the situation, and that raters may not only be differentially motivated by having to make appraisal decisions for different purposes, but they may actually require different information to reach those decisions.

Murphy, Philbin, and Adams (1989) obtained ratings of a videotaped college professor under conditions where performance appraisal was the central purpose or was secondary to the purpose of evaluating the content of the lectures. Observed differences in purpose were explained using Hastie and Park's (1986) notions of memory-based and on-line processing. Murphy et al. (1989) argued that a purpose known prior to observation acts to influence information gathering in an on-line fashion as observations are made. They argued that since evaluations of others have been shown to be part of all phases of the performance appraisal process, stating the purpose of the ratings as part of the performance appraisal process or stating the purpose of the ratings as part of performance eval-
uation did not alter what would have occurred if no purpose had been given. However, cueing the raters into the need to evaluate the content of the lectures did affect the observation process because it added an effect on observation that would not have been present if the purpose had not been stated. Thus, in their research, purposes did affect observation but only to the extent that the purpose drew attention to issues that would not be considered if the purpose had not been provided.

Finally, in the study described earlier, Balzer (1986) manipulated the centrality of the rating task. He observed that, although there was no main effect for the centrality of the rating task on the number of performance incidents collected by subjects, there was an interactive effect with initial impressions. When rating task centrality was low, appraisers who collected information about a worker who presented a negative first impression collected more performance incidents than did appraisers who collected information about a worker who created a positive first impression. On the other hand, when the centrality of the appraisal task was high, those appraisers exposed to ratees who presented a positive first impression collected more incidents than appraisers exposed to ratees who created a negative first impression. If the amount of performance information collected influences the accuracy of performance judgments, then the extent to which prior expectations about worker performance levels bias performance ratings may depend on the degree to which appraisers focus on the appraisal task when they collect performance information.

Summary. Interest in the attention and observational stage of the appraisal process is predicated on the basic assumption that what is observed affects appraisal accuracy; research indicates that this assumption is a valid one (Favero & Ilgen, 1989; Heneman & Wexley, 1983). The more interesting questions involve the nature of conditions that affect attention and observation. Here a single mechanism with two levels of abstraction appears to be operating. This mechanism is one of creating a "set" that influences the amount and nature of what is observed about ratees. At a more abstract level, cues regarding the general uses to which ratings will be put influence the amount and the level of detail of behaviors observed (DeNisi et al., 1984; Foti & Lord, 1987). More concretely, the set appears to prime the rater regarding specific behaviors that need to be observed. The source of the set can be from all four of the major categories reviewed. In particular, it has been shown to be cued by stereotypic information about the ratee (Favero & Ilgen, 1989), the rater knowledge about the rateee's task (Foti & Lord, 1987), familiarity with the rating scale (DeNisi & Summers, 1986), and the setting characteristic of rating purpose (DeNisi et al., 1984).

Although the observation and attention literature to date collapses onto
a single process, it hardly provides a solid basis for understanding observation. Within each class of variables, there is little or no systematic sampling of the set of variables that may affect observation. In addition, the sizes of the effects on observations that have been observed are not very large. Finally, observation effects often are not directly measured but are inferred from recall. This inference is always open to the alternative explanation that recall was affected by memory rather than observation. A useful extension of the present literature would be a more systematic mapping of the domain of observational conditions with a better understanding of the elements within ratees, raters, scales, and settings that are most critical for priming the observational process. Such information would be useful for the next step, that of addressing directly the critical conditions that enhance observational quality.

**Categorization and Storage Processes**

Because performance appraisals are rarely executed immediately following the observation of performance, it is necessary for the appraiser to categorize and store observations in memory over periods of time which may range up to a year or more. Categorization and storage processes result in some information being lost, discarded, or distorted before it is recalled for the appraisal. Therefore, to the extent that information recalled does not reflect what was actually observed, the accuracy of subsequent evaluations is hindered. For this reason, an understanding of how observations are organized and stored is critical to understanding the conditions likely to encourage accurate appraisal. In this section, we will consider research that has addressed the relationship between categorization and storage processes and the accuracy of subsequent evaluations of employee performance.

**Ratee.** The previously described study by Foti and Lord (1987) focused on information processing strategies used to evaluate leader performance. In this case, the leader was the ratee. As part of the study, Foti and Lord examined how one ratee characteristic, the extent to which a leader’s behavior was similar to behaviors generally accepted as typical or expected, affected how accurately episodes of performance were recalled. Using a signal detection criterion of rating accuracy, they found that leader behaviors similar to those considered the prototype of effective leadership were less accurately recalled than behaviors and events that were dissimilar to the prototype. This suggests that using prototypes to store observations may interfere with accurate recall, particularly for those behaviors and events directly related to the prototype.

**Rater.** Several studies have examined the premise that raters differ in the organizational processes they use to store information and that these processes affect how accurately observed behaviors are recalled. Nathan
and Lord (1983) conducted a laboratory study which assessed how appropriately two different cognitive process models described appraisers' organization of performance information. In particular, the study directly compared the so-called traditional model of Borman (1978) and the cognitive categorization one described by Feldman (1981). According to the traditional model, raters store information in relatively independent dimensions. This implies that, under favorable rating conditions, raters should be able to discriminate among performance dimensions, and, therefore, halo bias should be low. The cognitive categorization model, on the other hand, proposes that information is stored more globally as a general impression (e.g., good worker, poor worker), and that the prototypical features of the general descriptive category, based on an overall impression, guide recall. Therefore, a large general impression halo effect would be expected even under favorable rating conditions. Nathan and Lord found that the traditional model described their data fairly well. However, they also found support for the cognitive categorization model in the form of a large halo effect. The authors suggested that support for both models is not unreasonable, since these results may simply reflect aggregation of their data over many subjects.

In an attempt to extend Nathan and Lord's (1983) work, Frank-Major and Foti (1985) examined the effect of behavioral inconsistency on performance ratings. They hypothesized that a primacy effect would support the categorization model, while an averaging effect would support the traditional model. Again, there was mixed support for the two models. Although performance ratings were consistent with the traditional model, raters' inability to recall accurately the behaviors they observed and a large halo effect were consistent with the cognitive categorization model.

If, as Nathan and Lord (1983) suggested, cognitive style is a rater variable important in understanding storage processes, we might ask what aspects of cognitive style have the largest impact on accuracy in appraisal. The study by Foti and Lord (1987) discussed earlier bears on this issue. Foti and Lord proposed that appraisers generally process information in one of two ways: using prototypes or using scripts. According to the authors, in prototype processing, the appraiser focuses on forming an impression which will allow him or her to appropriately categorize the ratee. In this case, category membership, rather than the individual behaviors leading to categorization, is retained. During recall, characteristics of the prototype with which the ratee was associated are likely to be recalled, whether or not they were actually observed. In script processing, on the other hand, the appraiser focuses on remembering sequences of events. When the accuracy criterion is memory for the occurrence of specific behaviors/events, script processing should be superior to prototype processing. Consistent with this hypothesis, Foti and Lord found
that when subjects were experimentally induced to use either prototype processing or script processing as they observed leader behavior, script processing was equal or superior to organization around a prototype.

Finally, Ostroff and Ilgen (1985b) examined the cognitive categories that raters use, rather than the mode of processing employed. They found that rating accuracy, defined primarily as overall accuracy, was enhanced to the extent that raters' cognitive categories were consistent with the performance categories required by the rating scale. Prior rating experience also had a positive influence on rating accuracy, as did raters' ability to differentiate among rating dimensions. On the other hand, no support was found for the notion that the match between the orientation of the rating scale (trait vs behavior) and the overall tendency of raters to classify observations in terms of behaviors or traits affected rating accuracy.

Scale. Apart from the Ostroff and Ilgen study described above (Ostroff & Ilgen, 1985b), the only research located that directly investigated the effects of the appraisal instrument on the organization and storage of information or on the relationship between categorization processes and accuracy of evaluation was the DeNisi and Summers (1986) study previously described. In addition to studying how the timing of rating scale introduction affected rating accuracy, DeNisi and Summers identified several categories that raters used to cluster ratee behaviors. Using the Adjusted Ratio of Clustering (ARC) index (Roenker, Thompson, & Brown, 1971), the information raters recalled about the performance of each of three ratees was classified by whether the information referred to elements of the task, the ratee him- or herself, or the ratee's performance. Participants were then divided into those who consistently used the same classification schemes and those whose organizational pattern was inconsistent so it could not be identified.

Results indicated that raters using task-based rating scales and those given the rating scale prior to observation were more likely to have identifiable classification schemes. In addition, raters who used some consistent pattern to organize information were more accurate both in recalling ratees' behaviors and in rating performance. Among raters who used consistent patterns, those whose patterns were predominantly person-oriented were highest in recall accuracy, raters with task-oriented patterns were the second most accurate subgroup, and raters with performance-oriented patterns were least accurate. With respect to rating accuracy, raters using person-oriented patterns were again the most accurate. Raters using performance-oriented patterns were next most accurate. Raters using task-oriented patterns were no more accurate than subjects without identifiable classification patterns. Overall, while research is sparse, rating instrument characteristics do seem to cue how information is categorized and stored in memory.
Setting. How might the demands and characteristics of the appraisal setting impinge upon information storage processes and their relative effectiveness? Foti and Lord (1987) found that both the intended purpose of raters' observations (memory vs impression-formation) and the amount of information they possessed about the goals of an event (such as a committee meeting) influenced how raters categorized their observations. Specifically, script schemas were more likely to be used by raters who observed for recall purposes and by those who were knowledgeable about the goals of the event being observed. On the other hand, raters attempting to form impressions and raters unaware of the basic goals of the event were more likely to use prototype schemas to categorize and store observations. As noted earlier, these differences are important because they affect raters' ability to recall accurately the presence or absence of behaviors and events representing ratee performance.

Similarly, a study by Lauterbach and Barnes-Farrell (1989) examined the degree to which congruence between encoding and recall tasks enhanced the performance of the recall task itself. Subjects watched videotaped vignettes of a waiter performing six different tasks and were instructed to observe for the purpose of: (a) generally recalling behavior; (b) recalling behaviors from three particular performance areas; (c) evaluating performance in those three task areas; or (d) forming an overall evaluation. One to two days later, subjects recalled behaviors and evaluated overall performance and performance in six task areas. Rating accuracy was defined by (1) the mean squared difference between performance dimension ratings and expert raters' estimates of the true performance level exhibited on each performance dimension and by (2) a recognition task, with the percentage of true positive responses adjusted for the percentage of responses that were false alarms. Lauterbach and Barnes-Farrell found that rating accuracy was positively affected by one aspect of congruence between the encoding task and the recall task: the cueing of relevant categories. Subjects who anticipated that they would recall behaviors or evaluate performance in particular performance areas provided more accurate ratings on those dimensions than both subjects who anticipated that they would be required to recall specific behaviors that they observed and subjects who anticipated that they would provide evaluations of overall performance. Furthermore, raters who were provided with directions about which three performance areas would be evaluated judged those dimensions more accurately than they did the other three performance dimensions, which were ostensibly irrelevant at the time of encoding. However, unlike Foti and Lord (1987), no congruence or "cueing" effect was observed for the recall of work behaviors or for the accuracy of overall evaluations of work performance. The study suggests that the categories raters believe are important at the time of encoding are
more important than raters' expectations about whether the purpose of observation is to accurately recall worker behaviors or to draw accurate inferences about worker performance levels.

The impact of category priming reported by Lauterbach and Barnes-Farrell (1989) was replicated and explored in more detail in a recent study by Barnes-Farrell, L'Heureux-Barrett, and Holzworth (1990). They hypothesized that increasing the salience of specific task dimensions at the time of encoding should increase the availability of information relevant to those categories at the time of recall. This should have two implications. First, less cognitive effort (as reflected by reduced response latencies) should be required to complete tasks that require information retrieved from categories "primed" at the time of information acquisition. Second, tasks that require information retrieved from primed categories should be completed more accurately. They used the same stimulus materials as those used by Lauterbach and Barnes-Farrell (1989) and a computerized process tracing technique that allowed them to record response latencies for behavior recognition and performance rating tasks. Category priming increased the accuracy of memory-based performance ratings, although it did not affect the number of behaviors correctly recognized. Contrary to expectations, response latencies to both behavior recall and performance rating tasks were significantly longer for performance areas that had been primed at the time of information acquisition. The authors suggested that priming or cueing particular performance areas may serve the function of enhancing the perceived importance of those task areas, thus increasing the amount of cognitive effort that raters invest in recall and rating tasks for these performance areas.

In yet another study, Barnes-Farrell and Couture (1984) hypothesized that if the rating task was salient to raters while observing ratee performance, they should be more likely to both observe and store appropriate information and to complete tasks that are highly salient more accurately than tasks low in salience. Using videotapes developed by Borman (1979), they found that when making memory-based evaluations, raters made more accurate ratings if the rating task was made salient prior to observing ratees. Although salience did not affect rating accuracy when ratings were completed immediately after observing performance, it did improve accuracy when the rating task was delayed. This suggests that rating task salience most likely affected either the content or the structure of what was stored, rather than the information which subjects attended to during observation.

Williams, Cafferty, and DeNisi (1990) carried out a pair of laboratory studies which examined the impact of appraisal salience on the organization and recall of performance information and the accuracy of performance ratings. As hypothesized, they found that raters tend to cluster
performance information in person categories when appraisal salience is high at the time of observation and tend to cluster information in task categories when appraisal salience is low. Furthermore, they found that high appraisal salience was associated with greater recall of performance information, although rating accuracy was only slightly affected by appraisal salience. Williams et al. suggested that these findings are indicative of the use of on-line information processing when appraisal salience is high and memory-based processing is low (using the distinction outlined by Hastie & Park, 1986).

DeNisi, Robbins, and Cafferty (1989) have also reported research which sheds light on the circumstances which affect the organization of performance information and its relationship to recall and rating accuracy. In a laboratory study using videotaped stimulus materials similar to those used by Williams et al. (1990), subjects who kept diaries stored information in a pattern consistent with the manner in which the diary was organized (by ratee or by task). Subjects given a choice preferred to organize their diaries by ratee, and subjects who organized their diaries by ratee recalled performance information more accurately and provided more accurate performance ratings than subjects using other organization patterns. As with the work by Williams et al. (1990) described above and an earlier study by Williams, DeNisi, Meglino, and Cafferty (1986), person organization of performance information was associated with increased accuracy in performance ratings.

One final study which investigated the effects of setting characteristics on organizational processes in appraisal is a field study by Ostroff and Ilgen (1985a). Here, the effects of two types of rater training on the cognitive categories appraisers used to rate nurses' performance and on the accuracy of these ratings was examined. Ostroff and Ilgen designed two training programs, one specifically intended to provide feedback on the convergence between rater categories and scale categories, and a second designed to increase accuracy. It was found that both training programs increased rating accuracy and, of particular interest here, also resulted in increased convergence between rater cognitive categories and scale categories. Furthermore, it was found that feedback training tended to produce larger changes in both cognitive categories and in accuracy than did the generalized accuracy training, but these differences were quite small and only observed under limited conditions.

Summary. Research on the encoding process for performance appraisals has been limited primarily to comparing two information processing models and to exploring the role of consistency. At one level, the conclusions regarding the models are equivocal. Neither the traditional model, based upon the assumption that raters have relatively independent performance dimensions into which they sort information about ratees,
nor the impression formation model receives clear support. The most parsimonious conclusion is that both models possess some explanatory value. However, the lack of clear support for either does not mean that the research has made no contribution. The contribution comes from a weakening of zealous commitment to the traditional model. Prior to the last 10 years, most recommendations for performance appraisal practices followed the traditional model. For example, it was advocated that raters act as reporters without evaluating alternatives rather than evaluating as they go (Latham & Wexley, 1981). Recognizing that the very process of storing information in memory is affected by evaluations tempers the kinds of conclusions that are drawn from performance ratings on job dimensions.

Conclusions about consistency are more involved primarily because there are several forms of consistency. First, there is the degree of consistency between the level and nature of ratee performance compared to prototypical behavior for the role being performed. Foti and Lord (1987) found that those who perform consistently with the prototype are less accurately evaluated than those who perform inconsistently. Presumably, this occurs because the nonprototypical stands out and gets special attention. It should be pointed out that this affect was observed with a measure of accuracy that treats false positive errors the same as false negative ones, thus, penalizing raters for falsely identifying a prototypical behavior that did not occur as being present.

Consistency is also of concern between the dimensions of performance used to judge ratees and the belief structures that the raters appear to be using to judge performance. Most of the evidence regarding this form of consistency is indirect and indicates that the greater the consistency the better the ratings (Barnes-Farrell & Couture, 1984; DeNisi & Summers, 1986; Lauterbach & Barnes-Farrell, 1989). However, when attempts were made to measure directly the cognitive categories that raters used, there was a great deal of difficulty measuring the categories and showing a consistency effect (Ostroff & Ilgen, 1985b).

Recall and Evaluation

Ultimately, all performance appraisal systems result in a rating of one person by another. Since such ratings rarely involve judging a ratee’s performance at the time he or she is performing, performance ratings usually involve recalling and evaluating information from memory. Evaluation and memory are so intertwined that it is difficult to separate them under the typical conditions in which performance ratings are performed.

Given the primary interest in evaluations that result from ratings, it is not surprising that the vast majority of research conducted on the performance appraisal process has addressed issues of recall and evaluation.
This section discusses those studies which have investigated how the recall and evaluation of performance information influences rating accuracy.

*Ratee.* Process concerns for performance appraisal have devoted surprisingly little attention to trait-like characteristics of ratees, especially considering the fact that much of the early work on performance appraisals focused on such variables. In a laboratory study, Schmitt and Lappin (1980) investigated the effects of rater and ratee race and gender on rater accuracy. Selecting raters on the basis of race and sex, subjects viewed one of four videotapes of a stimulus person stacking books in a library. While the task performed and the level of performance were controlled and constant across the four stimulus tapes, performers' race and gender varied. Operationalizing accuracy as the correlation between subjects' and experts' ratings of the ratee's performance ("true scores"), a number of statistical interactions were observed. One of the most interesting effects was that raters rated same-race ratees' performance more accurately. At first glance, this finding seems inconsistent with earlier field studies which found that raters tended to rate ratees of their same race more favorably than those of another race (Landy & Farr, 1980). In fact, since their accuracy measure was the correlation between performance ratings and true scores, the two conditions could coexist in the Schmitt and Lappin data. Correlation coefficients are not sensitive to mean differences between the correlated variables.

Other interactions between raters' and ratees' race and gender were also observed by Schmitt and Lappin. In particular, black females were rated more accurately by white raters of both sexes than by black males, and white males were rated more accurately than white females by all subgroups. More importantly, race and sex variables accounted for far less variance in performance ratings (12%) than did ratee performance, which accounted for 70% of rating variance. A field study by Pulakos, White, Oppler, and Borman (1989) echoed this finding. Pulakos et al. were able to remove the effects of ratee true performance level before examining the effects of ratee and rater race and sex. They found that ratee race accounted for no more than 1% of the variance in overall performance ratings, and ratee sex produced no significant bias in performance ratings when the effect of performance level was held constant.

Another study dealing with ratee sex effects on rating accuracy was reported by Dobbins, Cardy, and Truxillo (1986). Using written performance descriptions of high-performing and low-performing male or female college professors, students evaluated teaching performance with the expectation that their evaluations would be used either for research, instructor feedback, or pay and promotion decisions. Results indicated that ratee sex affected the direction of inaccuracies in performance rating.
such that ratings of males were inflated relative to their true performance levels, while females were underrated relative to their true performance.

In a study involving the appraisal of vignettes of college professor performance, Cardy and Dobbins (1986) manipulated the likability of target ratees independently of performance level. They found that ratee likability significantly affected the accuracy (not just the level) of performance ratings. More specifically, this study indicated that affect operates as an integral dimension such that raters evaluated ratees' performance less accurately when liking varied orthogonally to performance levels than when liking was constant.

Focusing on predicting rating accuracy from variability in ratee performance, Padgett and Ilgen (1989) constructed videotapes of clerical workers who performed either well or poorly and consistently or inconsistently over time. Both behavioral and classification accuracy derived according to Lord (1985) were used as criteria. In the former case, accuracy was the extent to which raters were able to correctly identify behaviors actually displayed by the ratees. In the latter case, behaviors consistent with ratees' general performance level were considered accurate even if such behaviors had not been observed. Padgett and Ilgen found that classification accuracy was higher when evaluating consistent performers than when judging inconsistent performers. However, within consistent performers, raters were more likely to incorrectly report observing behaviors consistent with the ratee's general performance level than behaviors which were inconsistent. Finally, raters were more accurate when rating good performers than when rating poor performers. The reason for this is not clear. One explanation offered was that most people perform reasonably well on most jobs. As a result, observers of others' performance may have more practice dealing with good performance behaviors than bad and are therefore better able to judge such behaviors.

In a similar vein, two studies examined the consistency of ratee performance across performance dimensions (i.e., "true halo") and its effect on accuracy. First, Murphy and Reynolds (1988) examined the extent to which raters are sensitive to true levels of halo. They found that the level of true halo in ratees' performance only affected the intercorrelations among performance dimension ratings when the true level of intercorrelation was extremely high or extremely low. Particularly interesting is their finding that raters exposed to ratees exhibiting low true halo overestimated the intercorrelation among performance dimensions, but raters exposed to ratees who exhibited high true halo underestimated the true intercorrelation. This effect was replicated in a second study by Murphy and Jako (1989). Similarly, Smither and Reilly (1987) used a magnitude estimation task to examine the effect of true halo on individual rating accuracy. Asking raters to view videotaped customer sales representative
performance, they found that both overall accuracy and differential accuracy were positively correlated with the level of true intercorrelation among job components.

In sum, ratee characteristics have been demonstrated to affect rating accuracy over a wide range of constructs, varying from specific individual difference characteristics to consistency with stereotypes and the pattern of behavior over time. In none of these cases, however, was the variance accounted for by the ratee characteristics very high. We believe this is because the sum total of ratee effects is likely to be underestimated based on the current literature since none of the studies investigated more than one or two ratee characteristics at a time. Given the large number of such characteristics that may impact on accuracy, the total effect should be larger than currently evident despite the probable degree of multicollinearity among individual difference variables. Therefore, although research to date provides little basis for confidence that the effects of ratee characteristics on performance accuracy are well known, diversity in the types of variables investigated points to the need to look more closely and systematically at ratee effects in the future.

Rater. Two general approaches have been taken to address the effect of rater characteristics on the accuracy of performance evaluations. The more traditional approach has identified characteristics attributable to raters, such as personal traits, level of knowledge about the job, or characteristics associated with positions held by the rater. These have then been related to accuracy. The second approach has looked at the cognitive processes believed to be employed by the rater when rating. Although the latter approach is less typical in the industrial–organizational psychology arena, it is far more common in other areas of psychology that study rater effects on accuracy.

A number of studies have looked at the effects of rater traits on rating accuracy. In the Schmitt and Lappin (1980) study reported earlier, rater race interacted with the race of the ratee such that ratings were more accurate when the race of the two was similar than when it was not. In addition, Wexley and Youetz (1985) showed that raters’ beliefs about human nature were associated with rating accuracy. Using behavioral observation scales, they investigated the relationships among several rater characteristics and indices of rater error and rater accuracy. Wexley and Youetz found that rater accuracy, defined as the difference between rater behavior frequency ratings and the actual frequency of behaviors exhibited in an hour-long videotape of worker performance, was negatively correlated with rater beliefs about the variability of human nature.

Cognitive skills and traits, such as cognitive complexity, have been suggested as characteristics which should influence raters’ ability to accurately differentiate among workers and among employee strengths and
weaknesses. However, neither a study by Bernardin, Cardy, and Carlyle (1982) using student ratings of instructor effectiveness nor Wexley and Youtz (1985) found evidence that cognitive complexity influences rater accuracy. On the other hand, Smith and Reilly (1987) did find that rater intelligence was positively related to overall accuracy, stereotype accuracy, and differential elevation.

Rater affective states have also been shown to influence rating accuracy. In the study by Cardy and Dobbins (1986) described in the previous section, it was discovered that a rater’s liking for the ratee affects the accuracy of performance ratings. In addition, Sinclair (1988) recently reported that a rater’s mood impacts the performance appraisal processes. Because elevated mood states are associated with the use of broader categories and increased attention to positive information, Sinclair expected that “performance judgments of those in depressed states may be characterized by less halo, greater accuracy, and more validity” (Sinclair, 1988, p. 25). Sinclair presented student raters with a list of instructor behaviors. Before evaluating performance and recalling behaviors, subjects were exposed to depression or elation mood inductions. Using the within-subject correlation between the number of positive behaviors in a particular performance category and the subjects’ ratings of the teacher on that category as the accuracy index, results showed that depressed mood subjects exhibited significantly less halo than elated subjects (i.e., narrower category width). Also, more discrete categories were predictive of the overall evaluation, and the overall evaluation was more predictable for depressed subjects. Finally, as expected, ratings from subjects with depressed moods were significantly more accurate than were those of elated mood subjects.

In a study of the influence of rater confidence on social influence in multiple-source performance appraisals, Zalesny (1990) observed that rater confidence was negatively related to rating accuracy (indexed as discrepancy from expert ratings) for student teachers evaluating a videotaped class lecture. This finding is particularly interesting in light of the fact that she also found confidence to influence the impact that a rater had on combined evaluations based on multiple sources.

In a different vein, Bernardin, Cardy, and Abbott (1982) experimentally manipulated motivation to rate accurately and combined this with both raters’ knowledge of the job to be rated and the compatibility between raters’ information and actual job conditions. Prior to rating, all subjects (undergraduates) were asked to describe what they believed were important dimensions of an interviewer’s job. Then, raters watched videotapes of interviewers’ performance and were either given or not given a bonus for rating accurately. Using the free response items, Bernardin et al., calculated scores reflecting the degree to which raters’ a priori dimen-
sions were compatible with experts' judgments about the jobs. Raters' knowledge about the rating scale prior to conducting the rating was assessed through an objective examination of the rating scale. Using an index of differential accuracy based on the correlation between each rater's ratings and experts' ratings, they found that accuracy was higher for more highly motivated raters and for raters whose prior beliefs about the job were more similar to the actual job. Also, knowledge of the rating scale interacted with rater motivation; subjects with greater knowledge about the rating scale who were more highly motivated produced more accurate ratings.

Rater familiarity with the job was also studied by Kozlowski and Kirsch (1987), who examined the evaluation of baseball player performance under difficult memory conditions. Three weeks after the end of the baseball season, the authors found that both raters' job knowledge and familiarity with the ratee were positively related to Cronbach's (1955) component indices of rating accuracy. This study also supported the position that individuals' perceptions of the conceptual similarity among performance dimensions is associated with the systematic distortion of the covariation between ratings of performance on those dimensions, but the distortion depended upon an interaction of rater knowledge of the job and the familiarity of the ratees. Systematic distortion occurred for low job knowledge raters rating familiar ratees and vice versa. This may be important in understanding both those rater differences which contribute to halo bias and the conditions under which they influence ratings.

The second approach to studying rater effects on accuracy has involved attempts to identify differences among raters in cognitive style, rather than in traditional trait-like constructs. If we think of the propensity to exhibit halo as an appraisal "style" variable on which raters systematically differ, it is useful to consider the relationship between halo and rating accuracy. The relationship between so-called "illusory halo" and accuracy in measuring performance levels is not clear. Rather surprisingly, weak positive correlations between halo bias and rating accuracy have been reported (cf. Borman, 1979; Cooper, 1981; Murphy & Balzer, 1981). Recent studies, however, have suggested that the true relationship between halo error and rating error may have been obscured (a) by a failure to recognize the conceptual and empirical implications of different procedures for indexing halo and accuracy (Becker & Cardy, 1986) and (b) by the tendency to operationally define illusory halo in terms of rater overestimation of the intercorrelation among various aspects of performance. The study by Murphy and Reynolds (1988) discussed earlier noted that raters sometimes underestimate the degree of relationship between performance dimensions. Ficicaro (1988) demonstrated that when data are reanalyzed allowing for negative halo error as well as positive halo error, the relationship between halo error and rating accuracy is negative.
On the basis of social comparison theory, Fox, Ben-Nahum, and Yinon (1989) posited that peer assessors use those that they perceive as similar to themselves as a reference for self-evaluations. Therefore, they will be more sensitive to the behaviors of similar peers and will be able to evaluate their performance more accurately. In a direct test of this hypothesis, Fox et al. measured the relationship between military trainees' perceptions of peer similarity and the accuracy of their estimates of peer success in the training program. They found that accuracy in evaluating peer performance was substantially lower for dissimilar peers.

A second study based on a social comparison perspective examined the extent to which social comparison performance information (CPI) affects the accuracy of self-ratings and agreement between self-ratings and supervisor ratings (Farh & Dobbins, 1989). Farh and Dobbins hypothesized that the frequently observed disagreements between self-evaluations and supervisor evaluations are largely due to systematic differences in the amount of CPI available to raters in these two roles. In a laboratory experiment, subjects carrying out a proofreading task, either as an editor or a head editor, were asked to provide self-ratings or supervisor ratings, respectively. As hypothesized, receiving the same CPI that was available to supervisors increased the correlation between self-ratings and supervisor ratings of performance and increased the accuracy of self-ratings of performance (as indicated by the correlation between self-ratings and objective performance indicators).

Mount and Thompson (1987) were concerned with the effect that rater cognitive categorization of the appraisee (i.e., congruent, incongruent) has on the quality of performance ratings. Using a true performance score calculated as the mean rating assigned by the manager's supervisor and the manager's subordinates, they conducted a field study in which managers were evaluated by their subordinates. This study demonstrated that congruence between subordinates' expectations of appropriate supervisor behavior and their perceptions of the way their own supervisors behaved affected the accuracy, leniency, and halo exhibited in ratings of their supervisors' performance.

Nathan and Lord's (1983) distinction between traditional information processing and prototype processing suggests that, to the extent that individuals use one style of processing more than the other, rater characteristics may indirectly influence rating accuracy through the choice of processing mode. In a similar fashion, Foti and Lord's (1987) research on prototype and schema processing and that of Ostroff and Ilgen (1985b) on rater cognitive categories are relevant here if there are, in fact, predictable individual differences that cannot be ruled out. However, there is also no support for this position, primarily because it has not been explored in empirical studies. Therefore, we would not encourage researchers to litter the field with a large number of trait-focused studies corre-
lating traits or other individual difference measures with accuracy in a rather atheoretical fashion. Rather, there are good theoretical reasons for expecting individual differences to affect accuracy, and there is still room for well-reasoned individual research in this area.

Scale. The primary rating scale characteristics which have been studied relate to the types of ratee and/or work task dimensions that raters are asked to rate. These dimensions can be classified into two primary clusters: ratee traits and ratee behaviors. In addition, the latter is often subdivided again depending whether behaviors are assessed on trait scales, Behaviorally Anchored Rating Scales (BARS) or on Behavior Observation Scales (BOS). Both of the latter use critical incidents to generate lists of job behaviors. BARS further refines these critical incidents by clustering them into coherent dimensions, and specific behaviors are selected through a particular scaling process to represent discrete levels of performance on these dimensions. Thus, the rater's task is to consider the rating dimensions and use the incidents as examples of performance levels but not to recall specific behaviors observed. The BOS technique, on the other hand, does not attempt to scale responses but simply selects representative behavioral descriptions and asks the rater to recall the frequency with which these behaviors occurred.

Using only BARS, a number of studies have shown that knowledge of the scale after observation of performance but before rating does bias recall in the direction of increasing the likelihood that behaviors described on the scale will be recalled. Murphy and Constans (1987) first observed this effect when comparing observation and memory effects of BARS. In their case, the ratings were done immediately after familiarization with the rating scale. As noted in our discussion of observation processes, Piotrowski, Barnes-Farrell, and Esrig (1989) replicated Murphy and Constans (1987) and extended it by showing that the bias also held when the rating occurred several days later. Murphy and Pardaffy (1989) showed that the bias was limited to specific dimensions of the BARS on which some behaviors actually were shown by the ratee and did not generalize to unobserved dimensions.

In a laboratory study, Fay and Latham (1982) assessed the relative accuracy resulting from using three types of rating scales (trait, BARS, and BOS). Judging the overall accuracy with which raters rated videotaped interviewers' performance, they found that both of the behaviorally oriented scales, BARS and BOS, were superior to trait scales, but they were not significantly different from each other.

Similarly, Pulakos (1984) compared the two behaviorally based rating scales in terms of their cognitive demands. According to Latham and Wexley (Latham, Fay, & Saari, 1979; Latham & Wexley, 1977), an advantage of behavioral observation scales is that the rater does not have to
evaluate the behavior; he or she needs only to report the extent to which the behavior in question did or did not occur. Pulakos (1984) reasoned that if the rating task is construed as one involving gathering information, storing that information in memory, and recalling and evaluating the information stored in memory, then BOS scales focus on the information gathering process. In contrast to BOS, BARS present raters with a few job dimensions defined by a number of behavioral descriptors. These are generated by job incumbents so that the descriptors are in the vernacular of the job holders. Finally, the behaviors on each dimension are scaled and ordered from good to poor performance. Pulakos reasoned that because they involve attaching ratee behaviors to discrete rating dimensions, BARS place their heaviest demands on the encoding or information storing abilities of the raters.

To test these hypotheses, Pulakos used the same interviewing stimulus materials described earlier (Borman, 1979) to determine if training focusing either on developing information gathering skills or information storing skills affected rating accuracy when using the corresponding rating scale. She found that raters did, in fact, rate more accurately using scales corresponding to the cognitive processing skills on which the training focused.

The work of Pulakos (1984) accepted the cognitive processing orientation of BOS scales offered by their advocates (e.g., Fay & Latham, 1982; Latham et al., 1979); Murphy, Martin, and Garcia (1982) questioned it. Murphy, Martin, and Garcia, arguing from a cognitive information processing position, suggested that raters using BOS would not be able to recall accurately what ratee behaviors they had seen. They reasoned that these behaviors, once observed, were distorted during the encoding process. If this were true, then information recalled from memory would already include distorted data. Thus, ratings would be biased regardless of the type of scale used. Comparing students’ ratings of classroom lecturers with and without delay (representing two levels of memory requirements), using either BOS or more general types of rating scales, yielded results consistent with this memory distortion hypothesis; BOS scales produced ratings that were affected by storage processes.

In summary, Nathan and Alexander (1985) provide an excellent discussion of the cognitive processes involved in BARS and BOS. They agree with Murphy, Martin, and Garcia (1982) and others (e.g., Bernardin & Kane, 1980) that BOS are no less subject to cognitive distortions than are any other scales. With respect to BARS, Nathan and Alexander (1985) suggest that the process of developing these scales has no effect on how individual raters’ cognitive structures impact the information that is stored and recalled, but it does increase the probability that the raters will share common cognitive systems on which to conduct the ratings.
Setting. Setting characteristics represent those aspects of the rating context that influence rating accuracy. Given the stimulus–response thinking that drives a great deal of research on performance appraisals, most studies have addressed at least one setting/context variable. In addition, these variables have tended to mirror the types of setting variables researched prior to the concern for accuracy. In particular, this research has focused on the effects of the amount of delay between observing and rating performance, rating purpose, and rater training. Each of these will be addressed below.

Not too surprisingly, most research has shown that rating accuracy is affected by the delay between when a ratee is observed performing a task and when that performance is rated. Specifically, ratings made immediately after observation tend to be more accurate than those made after a delay of one day or more (Heneman & Wexley, 1983). However, since most research investigating delay effects has compared no delay to only one delay of some specified time period, it is not clear whether the loss in rating accuracy changes linearly or even monotonically over time.

The effects of delays on rating accuracy also are not completely stable. Smither and Reilly (1987) found that a 24-h delay between observation and rating did not affect the accuracy of ratings. Barnes-Farrell and Couture (1984) found that, while a one week delay between observation and rating significantly reduced the proportion of performance behaviors correctly recognized by subjects, it did not affect the overall accuracy of performance ratings.

Murphy and Balzer (1986) replicated the delay effect for behaviors but not for the relative rankings of individuals. Using videotapes of graduate students delivering class lectures, undergraduates rated multiple lecturers’ performance on several specific behavioral dimensions and on overall performance. Measures of differential accuracy showed the expected decrease in accuracy with respect to the behavioral dimensions over time. Accuracy in recreating the rankings among lecturers on overall performance, however, was better after a delay. Murphy and Balzer suggested that memory demands lead raters to form overall evaluations of ratees and to store information about ratees using those prototypes. They surmised that prototype matching aids recall of overall impressions as long as the ratee matches the prototypic categories reasonably well. The authors also suggested that an observed interaction between rating accuracy and the ability to correctly rank-order people may help explain the initially unexpected, but frequently observed, positive correlations between accuracy and the size of halo errors (e.g., Berman & Kenney, 1976; Borman, 1979; Cooper, 1981; Murphy & Balzer, 1981). In particular, forming a general impression of the ratee may aid in recalling his or her general performance level. On the other hand, a general impression is also likely
to decrease the amount of variance present in ratings among several performance dimensions when a general impression exists. Such an interpretation is consistent with Foti and Lord’s data (1987), demonstrating that prototypically inconsistent behaviors were recalled more accurately than prototypically consistent ones when consistent and inconsistent behaviors occurred with equal frequency. The latter effect was due to the tendency to falsely report the presence of prototypically consistent behaviors when they were not present. The authors suggested that when raters are reasonably accurate at classifying ratees into prototypic categories and ratees actually display prototypically consistent behaviors in uncontrolled situations, accurate ratings will tend to result from lucky guessing rather than an astute memory. Furthermore, combining the distribution of prototypically consistent ratee behaviors with raters’ information processing errors would lead to a similar pattern of results.

A second setting characteristic believed to affect rating accuracy is raters’ perceptions about the purpose for which the appraisal is being conducted. Zedeck and Cascio (1982) found that appraisers told to rate performance in order to assign merit raises produced less accurate ratings than those who rated the same materials in order to either provide developmental feedback or remember as much as possible about performance. McIntyre, Smith, and Hassett (1984) also investigated the effects of appraisal purpose, but they were unable to find such effects among college students’ ratings of videotaped lecturers. However, since both the nature of the purposes and the manner in which accuracy was defined differed between these two studies, a straightforward explanation of why the purpose effect was observed in the former case but not the latter is not available. One possibility is that the Zedeck and Cascio study used the ability to predict a known criterion from raters’ performance ratings as their measure of accuracy, a procedure not typically used in performance accuracy research.

In a study more directly comparable to the work reported by McIntyre and his colleagues, Murphy, Balzer, Kellam, and Armstrong (1984) compared ratings of videotaped teaching performance made for research purposes with ratings that students expected to be used in making important decisions. They found that rating purpose affected neither the accuracy with which students reported the frequency of critical teacher behaviors nor the accuracy of performance ratings (indexed in terms of Cronbach’s multiple accuracy components). However, purpose did affect the relationship between indices of observational accuracy and indices of evaluation accuracy. Therefore, Murphy et al. (1984) suggested that rating purpose primarily affects the manner in which behavioral information is handled rather than the effectiveness with which it is utilized.

Farh and Werbel (1986) examined the impact of rating purpose (in this
case research vs assignment of grades) on the accuracy of students’ self-rated performance. A major criticism of self-appraisals has been their tendency to suffer from positive leniency bias. Therefore, Farh and Werbel’s main concern was to understand how purpose and expectations about how self-ratings would be validated would affect leniency errors. After manipulating their expectations that self-ratings would be validated against other performance measures, subjects’ self-ratings were compared with externally observed performance. As expected, leniency was significantly greater when students expected that grades would be based on their self-appraisals, but it was significantly reduced when they expected external validation.

Although there are good reasons to expect that the purpose under which ratings are obtained affects rating accuracy, research on the purpose has been limited by a general tendency to treat the purpose as if it were a unitary construct in some meaningful way. In fact, there are a wide variety of purposes that are likely to affect ratings; purpose, per se, has little meaning. The more central constructs of purpose must be addressed for purpose to have much meaning.

One way to look at purposes is to view them in light of processes affected by purposes. Murphy et al. (1989) suggested that the purpose of a rating if known before the rater observes the ratee could affect attentional processes during observation and memory processes after observation. Their data showed that purposes did have an effect on ratings, but they found no evidence for an observation effect different from that of memory. On the other hand, the impression formation research (Lingle & Ostrom, 1979; Hamilton, Katz, & Leiner, 1980) has shown that the purpose (either to form general impressions of others or recall detailed information about them) affects both observations and memory. Finally, field data that finds a difference in leniency between ratings obtained for purposes of research versus those for purposes of personnel decisions (Pritchard, Peters, & Harris, 1973) when the issue of purpose is raised immediately before ratings take place implies that ratings can vary based on purpose when neither observation nor memory mechanisms have varied. In the latter case, the most likely explanation for the difference is motivational. It has been documented that a common organizational practice is to use performance ratings for multiple purposes once they are gathered (Cleveland, Murphy, & Williams, 1989). Given the potential importance of purpose effects, more clearly developed theoretical and empirical work would be helpful.

Another setting feature thought to influence the accuracy of performance ratings, the sex-typing of job tasks, was addressed in a study by Barnes-Farrell, L’Heureux-Barrett, and Conway (1991). In this research, subjects read written behavioral descriptions of ratees performing gender-
typed occupations, and then they evaluated ratees' task and overall performance. A significant interaction between the gender type of the occupation and the gender type of the task was observed such that performance information for sex-typed tasks was evaluated more accurately when it was presented in the context of an occupation with a congruent sex type. The authors suggested that such effects may be due to differences in the accessibility of relevant categories when appraising task performance for a male-typed or a female-typed occupation. These results imply that accuracy in performance measurement may not be independent of job and task characteristics, and therefore direct comparisons of task performance ratings gathered in different occupational contexts may be risky.

Rater training represents the final frequently investigated setting characteristic. With respect to performance appraisal accuracy, the controversy driving training research involves whether training aimed at reducing errors also reduces accuracy. Rater error training has been used for a number of years (Wexley & Latham, 1981). Its aim is to make raters aware of typical rating biases, including primarily halo, leniency, similarity-to-me, and contrast errors. Once trained to avoid such errors, there is strong evidence to support the conclusion that raters can reduce the presence of rating errors.

At first glance, the reduction of rater errors would seem to be beneficial for improving rater accuracy, and therefore rater error training should be beneficial. This conclusion overlooks the fact that a consistent, low, negative correlation has been reported between one rating error, halo, and rater accuracy (cf. McIntyre et al., 1984). This negative correlation has led to closer evaluations of training's influence on rating accuracy and to the development of training targeted directly at accuracy rather than at rating errors.

Two studies found that training focused on accuracy was superior to rater error training for improving accuracy. Pulakos (1984) compared the two and found that rater accuracy training led to higher rating accuracy when accuracy was defined by Cronbach’s (1955) differential accuracy. McIntyre et al. (1984) found a similar difference for two other accuracy measures: Cronbach’s (1955) correlational accuracy and distance accuracy. The latter study also included training that combined both forms of training (error and accuracy), but the combination was still not as beneficial as accuracy training alone. One note of caution must be added concerning the McIntyre et al. study. A problem with heterogeneity of variance across treatment conditions necessitated a conservative test to control for the problem. This reduced the effect size such that the results for only one of the two accuracy measures were statistically significant.

Hedge and Kavanaugh (1988) compared error training with observation
training and decision-making training programs. Using halo, leniency, range restriction, and rating accuracy (the correlation between rating and true score for each subject for each dimension) as dependent variables, they found that rater error training reduced leniency and halo, but it also reduced rating accuracy. On the other hand, both observation training and decision-making training produced significantly higher rating accuracy but had little impact on the rater error indices. These results, in combination with those of Pulakos (1984) and McIntyre et al. (1984), indicate that rater error training alone may reduce the accuracy of ratings and that training geared directly to accuracy can be beneficial.

Pulakos (1986) focused on the importance of training which meets the needs of the evaluation task. She suggested that different rating tasks have different cognitive requirements for collecting and using information. Specifically, some rating tasks have an evaluative orientation; others have an observational orientation. She hypothesized that BARS focuses on encoding performance information by presenting raters with both sets of categories into which performance information should be stored and examples of how certain behaviors should be evaluated in terms of performance effectiveness units. In contrast, BOS scales suggest the exact behaviors that the person should be attempting to observe. Therefore, Pulakos surmised that, in order to facilitate accuracy, training programs should be congruent with the requirements of the rating task. To test this, students provided with evaluative training, observational training, or no training rated videotaped vignettes of managerial performance. Cronbach’s four accuracy components were calculated for the evaluative rating task and for the observational rating task and, in both cases, rating accuracy was enhanced when training was congruent with the evaluation task, although the effect was not strong.

Finally, Athey and McIntyre (1987) explored the effectiveness of rater training programs from the perspective of levels-of-processing theory. Levels-of-processing theory would predict that training which requires raters to process information more deeply should enhance the accurate recall and evaluation of that information. Frame-of-reference training requires ratees to observe, evaluate, justify evaluations, and then respond to feedback on those evaluations. This should encourage deeper processing of relevant performance categories and behaviors than informational training, which in turn should result in deeper processing of appraisal-relevant information than no training. Athey and McIntyre provided frame-of-reference training, information training, or no training to groups of trainees who were then required to rate the performance of a videotaped lecturer. Distance accuracy and correlation accuracy indices were calculated for all subjects. They found that the frame-of-reference groups did exhibit significantly higher distance accuracy, higher retention of
training content, and lower halo than the information and no training groups, although correlation accuracy was not affected by training method. In addition, the information training groups exhibited higher knowledge of training content than the no training groups. They interpreted their results as consistent with levels-of-processing theory.

Summary. General conclusions about recall and evaluation are more complex and less clear than was the case with the first two processes. This may be due, in part, to the fact that the recall and evaluation processes have received a great deal more attention and, therefore, require fitting more points.

With respect to ratees, the data are generally consistent with the conclusion that ratee characteristics provide cues that may or may not facilitate recall and evaluation. Unfortunately, there does not seem to be any consistent theoretical position that would provide a priori reasons for expecting accuracy to be facilitated or inhibited. In part, this may be due to the paucity of studies available at this time. It may also be due to the diversity of the set of possible cues ratees may provide.

Rater effects cluster into those due to rater characteristics and those due to cognitive processing factors. Although the individual difference approach is somewhat out of favor among cognitive psychologists today, research from an individual perspective did provide some interesting observations in both effects observed and not observed. First, with the latter, the expectation that more cognitively complex raters would be more accurate and create fewer halo errors was not observed but more intelligent raters were better (Smither & Reilly, 1987). On the other hand, raters experienced with the job but relatively unfamiliar with the ratees had greater distortion in intercorrelations among dimensions (Kozlowski & Kirsch, 1987). Mixed results were also observed with affective variables. Raters motivated to be accurate actually were more accurate (Bernardin, Cardy, & Abbott, 1982) whereas, in one case, a more depressed mood reduced rating errors (Sinclair, 1988).

Research on rating scale formats and their effects on recall and evaluation has provided clear support for the fact that raters do not simply report what they have observed. Thus, rating scales, in particular the BOS, may be useful for feedback and evaluation but using them should not be construed as simply a task of reporting what was observed. Ratings on these scales, like any other recalled information, are subject to cognitive processing distortions. At the same time, it has been shown that scale formats do stress different phases in the appraisal process and that training targeted to the process central to the scale is beneficial (Pulakos, 1984).

The conditions external to the persons involved in ratings and the scales that have received most attention are: (1) the amount of delay
between observation and rating, (2) the purpose of the rating, (3) sex type of the job, and (4) training for errors or accuracy. In all cases, these conditions impacted rating accuracy. Consistent with predictions, accuracy decreased over time except when the accuracy measure captured the general rather than specific performance. Over time, overall impressions remained stable, although recall of specific performance decreased (Barnes-Farrell & Couture, 1984; Murphy & Balzer, 1986). Purpose and job sex type effects are consistent with the conclusion that these cues provide bases for establishing performance-linked schemas. Thus, recall follows the typical schema effects. Finally, the research looking at rater training revealed the complexities of dealing with rating errors and rating accuracy simultaneously (Hedge & Kavanaugh, 1988).

DISCUSSION

Contributions

The appraisal process approach has made at least four major contributions to performance appraisals in organizations. The first of these is a heightened awareness of the importance of observation in performance appraisals. Research clearly shows that raters use cues from knowledge about the ratees themselves, the jobs, the uses to which ratings will be placed, and other factors to form impressions of ratees, and these impressions influence the information to which the raters attend. Prior to the research on process, some attention was paid to the observation phase of ratings [e.g., suggesting that the raters keep diaries on the ratees (Bernardin & Walters, 1977)], but the importance of observation was underestimated. This led to the tendency to ignore the observational phase. It is clear at this time that it is an important phase that cannot be ignored. More careful attention needs to be paid to methods of ensuring the systematic sampling of information about ratees.

Other contributions of the process approach were primarily in the form of correcting invalid assumptions and practices operating prior to the most recent research. The first issue needing correction was the belief that rating errors (as commonly defined) were evidence for rating inaccuracy. It is now less clear that rating errors, particularly halo assessed by covariation among performance dimensions, are necessarily evidence for inaccuracy in rating. Obviously, no one is advocating that we desire rating errors. At the same time, when dimensions are intercorrelated and performance is based on general abilities crossing several performance dimensions, the presence of what appears to be halo error does not necessarily indicate that there is inaccuracy in the ratings. In fact, it has been demonstrated that accuracy may, in some situations, covary positively with the presence of halo (Borman, 1979; Cooper, 1981; Murphy & Balzer, 1981, 1989; Smither & Reilly, 1987).
A second modification of earlier beliefs involves the rating task itself. In comparing two scales, the BARS to the BOS, Wexley and Latham (1981) argued that for the latter, the rater need only function as a reporter of what was observed and did not need to get involved in evaluation. It is now clear that the process of social judgment involves forming an impression as observations occur and that the impressions affect what the rater believes he or she observed. As a result, regardless of the instructions presented on the scale, raters do not simply report what has occurred; reports already include evaluations indirectly as they, the evaluations, influence what is observed and what is recalled as having occurred (Pulakos, 1986).

The third correction to conventional wisdom relates to the use of archival performance appraisal ratings. Such ratings have typically been used as criteria for a variety of purposes, such as validating selection procedures and assessing the effectiveness of training programs and assessment centers (Cleveland et al., 1989). Most frequently, the ratings on file are ones that have been obtained from raters who provided the ratings as part of a formal appraisal system. Raters generated those ratings under a particular set of expectations about the purposes for which the ratings would be applied yet these purposes may be quite different from those of the researchers who access the files at a later date. Previously, these differences have received little attention. However, recent research showing the strong influence of rater expectations on ratings implies that the purposes under which ratings are obtained will have a strong influence on the ratings themselves. This inference is substantiated by research showing that rating purpose does indeed bias ratings (e.g., Farh & Werbel, 1986; McIntyre et al., 1984; Murphy et al., 1984; Murphy et al., 1989; Zedeck & Cascio, 1982). The implication is clear; archival performance data should be used as criteria for purposes other than those for which the ratings were obtained only with extreme caution. At the very least, the purposes under which the ratings were made must be clearly understood and a judgment made by the researcher regarding the consistency of that purpose with the later purpose.

Future Directions

Future research needs can be considered at two levels. The first of these is from within the process perspective. The question asked from this perspective is: What rating process research is needed to improve performance appraisals in practice? Here, three needs seem to emerge. The first of these is a need for a more systematic sampling of the observation, memory, and recall and evaluation processes across the domains of ratees, raters, rating instruments and settings. It is clear from this review that research was abundant in some areas and lacking in others. Although
a reasonable rationale existed for studying every particular variable in these studies, little concern has been paid to the total domain.

Second, measurement procedures that more accurately summarize the nature of rater cognitive processes need to be used more extensively. For example, Martin and Klimoski (1990) have advocated the analysis of verbal protocols ("thinking aloud") as a means of tracing the kinds of cognitions reported by workers during performance evaluations. Others have explored the uses of process tracing techniques that record the point at which raters notice relevant information (cf. Banks, 1979) and unobtrusive measures of cognitive effort such as response latency (cf. Barnes-Farrell et al., 1990). Modelling the appraisal process requires that researchers be somewhat more imaginative in borrowing and developing techniques that will help us to describe the content and process of the cognitive aspects of rating behavior more precisely.

Third, up to this point, the research has tended to focus on the rating process with much less attention placed on the content or elements important for rater accuracy. It is assumed that the process is a universal one, operating in the same fashion regardless of content. However, from a practical standpoint content is important. Consider, for example, the issue of rating purpose. At the process level, it is of interest to know if raters' perceptions regarding the rating purpose affect evaluations. The literature reviewed here indicates that this is clearly the case. However, from an applied standpoint, the real interest is in the effects of particular purposes on ratings, such as when ratings are used for salary increases or for developmental purposes. The introduction of content makes research more difficult because it requires that the content domain be relatively well known. For purpose, this problem is minimal because there tend to be a relatively small number of purposes for which appraisals are used. However, if the issue is the information about ratees stored in raters' schemas, the domain contains a vast number of elements that are far less well understood.

The extension to content also raises issues related to the research paradigm. When the focus of research is on a general psychological process, laboratory research often possesses sufficient external validity to generalize to field settings. However, if the focus is on content, then the content dimensions must possess sufficient fidelity to generalize to the field. The generalization criteria in this case are much more stringent. As a result, such stringent conditions may make it impossible to do research in the laboratory for many types of content issues. Nevertheless, if performance process research is to continue contributing to the implementation and operation of performance appraisal systems in ongoing organizations, we believe that it is going to have to understand better the content of the processes identified.
Our discussion of future directions thus far has been limited to ways in which research within the appraisal process perspective could contribute to appraisal practices. Another way to look at contribution is to focus more broadly on appraisal systems in general and examine the nature of research that will contribute the most to performance appraisal practices. Banks and Murphy (1985) criticized process approaches to performance appraisal for failing to narrow the gap between research and practice. They argued, as we have, that the process approach was initiated primarily out of dissatisfaction with the extent to which early research on rating errors, rater training, and scale construction was able to meet the needs of practice, and that the turn to process has offered far too few contributions to the problems of appraisals in practice. Although we share some of Banks and Murphy’s concerns about the practice of performance appraisals, we disagree with their conclusions regarding performance process research. First, we disagree with the extent to which progress has been made, and we have already mentioned a number of contributions. (Some examples of progress are a better understanding of the advantages and limitations of behaviorally referenced appraisals, recognition of the fact that observed halo may not be a rating error, and an increased awareness of the need to clearly state the purpose of ratings.) More importantly, we disagree with the attribution of failures in performance appraisal practices to the process approach. We contend that the process focus did not set out to address all (or even most) of the problems facing performance appraisal systems in practice. Neither did the earlier work on rating scales. Rather, each approach defined a domain in which to address concerns that would appear to be part of the set of problems facing appraisals. Within those limited domains, progress relevant to practice has been made, and this progress should not be overlooked. For rater process research, the domain was that of factors affecting raters’ ability to rate others accurately.

At the same time, it is legitimate to ask if this progress has been sufficient. Here we believe Banks and Murphy’s (1985) answer would be a clear no; we would strongly agree. Many of the problems that have plagued these systems still remain. However, rather than attribute them to the failure of either the process approach or a focus on rating errors, it would be better to explore other domains where greater relative contributions may be made in the future. We feel that it is far more likely that the major problems facing performance appraisals at this time lie neither in the cognitive process domain nor in that of rating scale construction. Consider a case in point. Padgett (1988) argued that the values and expectations generated by the social milieu in which a rater finds himself or herself controls the ratings that are produced over-and-above the raters’ ability to produce accurate ratings. Using causal modeling, her data
showed that inflation in ratings could be predicted from the extent to which raters felt that they could be open and honest when rating their subordinates. Beliefs about freedom to be open and honest were a function of a number of conditions in the rating environment, such as the rating purposes, the extent to which ratees discussed their ratings with others, the extent to which raters could document their ratings with other measures of ratee performance, and the extent to which raters desired to be liked by the ratees. None of these factors are issues of cognitive processing nor do they involve rating errors likely to be controlled by the nature of the rating scale. On the other hand, they are variables that frequently are mentioned by raters in the field as being important for ratings. We suggest that expanding the types of variables included in our models of the rating task beyond simply cognitive processing variables may advance the field more quickly. At the same time, we are not implying that the processing view was incorrect or misleading. Rather, we suggest that the amount of variance that cognitive processes account for in appraisal ratings is limited. It appears it is time to redirect attention toward other possible influences on ratings if the primary objective of the research is to improve the quality of performance appraisal systems.

**Conclusions**

The seminal review of performance appraisal by Landy and Farr (1980) abruptly shifted the direction of performance appraisal research from a focus on rating errors, rater training, and rating scale development to that of understanding the rating process. In the years that followed, sufficient empirical research has been conducted to evaluate the contribution of rater appraisal process research to the design of performance appraisal systems. In our opinion, there has been a number of contributions, but they have been limited. Furthermore, if the thrust of the research continues on its present path, we predict these limits will become even more apparent. That it is reasonable to construe the appraisal task as a specific case of social cognition in which others' behavior is observed, stored in memory, later recalled, and evaluated has clearly been demonstrated. This demonstration has been useful for guiding some of our thinking about the design of rating tasks, and about rater characteristics that may facilitate more accurate ratings. However, in our opinion, this body of research has reached a point of diminishing returns. Additional simple demonstrations of cognitive effects are not needed. Rather, contributions to the design and operation of appraisal systems are likely to come from a shift toward (1) the investigation of the content of cognitive variables, (2) the identification of work group and organizational factors that influence these variables, and (3) the design of appraisal systems that incorporate cognitive principles.
REFERENCES


Pritchard, R. D., Peters, L. H., & Harris, A. (1973). The effects of confidentiality on the
distribution of naval performance appraisals (Tech. Rep. No. 1). West Lafayette, IN: Purdue University, Department of Psychological Sciences.


Youz, M. A., & Ilsen, D. R. (1986). *The impact of performance level on alternative mea-
sures of rater accuracy (Tech. Rep. No. 86-1). East Lansing: Michigan State University, Department of Psychology.

Received: December 1, 1989