Organization of Information in Memory and the Performance Appraisal Process: Evidence From the Field

Angelo S. DeNisi
Rutgers University

Lawrence H. Peters
Texas Christian University

No studies dealing with cognitive processes in performance appraisal have been conducted in field settings, raising questions about the usefulness of this research for practice. The field experiments described here, conducted in 2 organizations, were designed to evaluate interventions that laboratory research has suggested enable raters to better organize performance information in memory: structured diary keeping and structured recall. After these interventions, raters had more positive reactions to the appraisal process, were better able to recall performance information, and produced ratings that were less elevated and better able to discriminate between and within rates. The implications of these results for practice and for cognitive research in performance appraisal are discussed, along with the limitations of these studies and the problems with criteria for evaluating ratings in the field.

In their watershed review of the performance appraisal literature, Landy and Farr (1980) suggested that future research efforts concentrate on understanding rater decision-making processes underlying appraisals. With this, research began in earnest on what has come to be called the “cognitive approach” to performance appraisal. Initially, this new approach manifested itself primarily in proposed cognitive models of the appraisal process. Models by Feldman (1981), Ilgen and Feldman (1983), and DeNisi, Cafferty, and Meiglin (1984), among others, tended to draw heavily on the research on person perception and social cognition (e.g., Srull, 1983; Wyer & Srull, 1981) and tended to emphasize such processes as information acquisition, storage, and retrieval. Subsequent research empirically examined implications from these newer cognitive models (see DeNisi & Williams, 1988, for a review of research in this domain).

However, almost before these empirical studies were published, criticism of the approach began to appear. Ilgen and Favero (1985) were among the earliest to suggest that the social psychological models that formed the basis for most of the cognitive appraisal models were not readily applicable to performance appraisal settings. These authors also argued that cognitive appraisal studies were primarily conducted in laboratory settings, where important process issues were not considered, and this criticism persisted even after a number of empirical studies had appeared (cf. Murphy & Cleveland, 1991, chap. 7; but, see Ilgen, Barnes-Farrell, & McKellin, 1993, for a more positive commentary). Clearly, the utility of the cognitive approach must be demonstrated in field settings, where real supervisors rate real subordinates on relevant dimensions of their work.

The very notion of moving from the laboratory to the field, however, surfaces a whole set of issues regarding the appropriate criteria for appraisal researchers. Traditionally, the primary focus of cognitive appraisal research has been on determining the factors that might lead to appraisal accuracy (or inaccuracy). But what if accuracy is not the “right” criterion for applied settings? Ilgen (1993), for one, has recently suggested that it may not be, and Murphy and Cleveland’s (1991) emphasis on contextual factors also implies that research aimed at improving rating accuracy will indicate only what raters are capable of doing. To understand how appraisals actually work in the field, one may need to turn to alternative criterion measures.

In this article, we present findings from two field experiments that provide the first test (to our knowledge)
of the effectiveness of organizing information in memory in a field setting. However, in moving this research from the lab into the field, we faced a number of issues not present in more typical laboratory experiments. In particular, we realized we would not be able to rule out some alternative explanations for findings, because we would not be able to establish the same degree of experimental control possible in the lab. Furthermore, we would not be able to use rating accuracy as a criterion measure, even though this has been widely used in lab studies. Instead, we used three alternative outcome measures that, we believe, are indicative of the potential usefulness of appraisal information as typically used by organizations. Our goal was to provide meaningful data on the potential usefulness of interventions designed to improve information organization in memory.

Organization of Information in Memory

Because formal appraisals in most organizations are infrequent events, rating accuracy would seem to be dependent, at least in part, on memory accuracy. Research both in social cognition (e.g., Srull, 1983; Wyer & Srull, 1981) and in performance appraisal (see reviews by DeNisi & Williams, 1988; Feldman, 1994) has suggested that a rater's ability to accurately recall information is largely dependent on how the information was organized in memory during the encoding process.

Two experiments reported by Cafferty, DeNisi, and Williams (1986) illustrate this point. In the first study, the authors examined how raters acquired information they would use in performance appraisal decisions and found that, for the most part, raters acquired information either blocked by persons (acquiring data for one ratee at a time on different tasks) or blocked by tasks (acquiring data for multiple ratees on one task at a time). In their second study, Cafferty et al. found that person-blocked acquisition led to information being organized in memory by persons, whereas task-blocked information led to information being organized in memory by tasks. Consistent with research in social cognition, these authors also found that organizing information either way resulted in more accurate recall and more accurate ratings than when the raters did not organize information in memory (cf. Bousfield, 1953; Mandler, 1967).

But Cafferty et al. (1986) also found that some raters acquired information in an unblocked (moving across both ratees and tasks rather haphazardly) fashion. This acquisition pattern did not allow organization in memory in any recognizable format and, more critically, resulted in poor recall and inaccurate ratings. Subsequent research by Williams and his colleagues (Williams, Cafferty, & DeNisi, 1990; Williams, DeNisi, Meglino, & Cafferty, 1986) suggested that raters who acquire information for a nonappraisal purpose (e.g., evaluating the difficulty of the task) organize the information in a way that facilitates making the nonappraisal decision at hand. These raters must later "reprocess" that information before they can use it for making appraisal decisions. Because a fair amount of information will be acquired or encountered in nonappraisal contexts, this information will, therefore, not be organized in a fashion conducive for recall and subsequent ratings. In such cases, a rater, even if he or she desired to be accurate (see Murphy & Cleveland, 1991, chaps. 8 and 9, for an excellent discussion of rater motivation and reasons why raters may decide not to be accurate), would find it difficult to do so.

The practical implication of this research is that, if raters are to be able to provide accurate ratings (assuming they are motivated to do so), they must be able to access performance information stored in memory. This, in turn, requires that the information be stored in memory in some organized or structured fashion. Yet, because a great deal of the information raters acquire is inherently unorganized or encountered in nonappraisal contexts, there is a need for interventions designed to provide those raters with the structure they need. Three potential interventions, designed to impose structure on memory, were subsequently investigated, although only two are relevant for the present research.

One possibility, not investigated in this research, would be to provide raters, before they acquire information, with performance categories that could be used to organize that information. Feldman (1981) emphasized the importance of such categories, and several studies (e.g., Balzer, 1986; DeNisi & Summers, 1986; Favero & Ilgen, 1989; Foti & Lord, 1987) provided support for the fact that, when raters were presented with such categories, they relied on them to organize information.

A second way of imposing structure on memory, as information is being encountered, involves the use of performance diaries. DeNisi, Robbins, and Cafferty (1989) reported that raters who kept diaries organized by persons also appeared to have stored "unorganized" performance information by persons, whereas raters who kept diaries organized by tasks (or performance dimensions) appeared to have stored information by tasks. Raters in either of these conditions were better able to recall information and provided more accurate ratings than raters who did not keep diaries. These authors also reported that raters who kept diaries organized by persons provided the most accurate ratings of any raters in the study. This latter finding is consistent with a great deal of research in social cognition (e.g., Hamilton, Katz, & Leirer, 1980; Ostrom, Pryor, & Simpson, 1981; Sherman, Judd, & Park, 1986). It is also worth mentioning, at this
point, that DeNisi, Robbins, and Cafferty (1989) included a condition in which raters were free to organize their diaries in any manner they saw fit. Almost every rater in this condition organized his or her diary according to persons and appeared to organize information in memory in the same fashion. Thus, structured diaries may do more than provide raters with a "hard copy" of performance information. Diary keeping is a potential intervention that could help raters in the field to organize information they encounter as it is being encoded, although it should be noted that Balzer (1986) reported that diaries might also be influenced by performance expectations.

Finally, it might be possible to organize information after it has been encountered and encoded but before it is actually used in rating decisions. Williams and his associates (Williams et al., 1986, 1990) found that having raters engage in a structured recall task before making ratings (i.e., having raters record, from memory, all critical incidents of performance they observed according to a given format) resulted in both more accurate recall and more accurate ratings, even though information had not been previously organized. That is, raters could actually impose structure on information in memory after the fact. This appears to facilitate the recall process by making performance information more accessible for decision making.

Thus, a series of laboratory experiments have shown that (a) organization of information in memory is important for recall and rating accuracy and (b) it might be possible to impose structure even if the information is acquired or encountered in an unorganized fashion (which is likely to often be the case). The present research was designed to test whether two interventions—structured diaries and structured recall—would be effective in field settings.

Role of Memory in Appraisal

The research described earlier clearly suggests a direct relationship between memory and recall and, by extension, subsequent ratings. Although such a suggestion may seem reasonable, it reflects only one perspective on the role of cognitive processes in appraisal decisions. A different perspective suggests that raters do not actually store specific performance incidents in memory. Instead, they store, and update, general impressions about ratees. When they are asked to make appraisal decisions, rather than asked to recall the incidents and combine them to form a judgment, they simply recall the readily accessible general impression they have formed about the rateee and use that impression for making their judgments (cf. Nathan & Lord, 1982). This position is probably best represented in the work of Hastie and Park (1986) and their distinction between what they termed on-line and memory-based processing.

On-line processing is assumed to take place when raters know the purpose for which they are acquiring information from the outset and so can form categories that can then be updated and used to make judgments. In such cases, the actual information involved in forming the impression is lost, and, as a result, there would be no relationship between memory and ratings. Memory-based processing is assumed to take place when raters are not aware of the ultimate purpose for which they are acquiring information or when that purpose is changed (e.g., when raters acquire information for one purpose but then are asked to use it for a different purpose). As a result, the rater must rely on the information in memory because no categories or impressions are readily available. In this instance, there would be a relationship between memory and ratings.

Evidence suggests that both processes can occur in performance appraisals, and, in fact, on-line decisions appear to be more accurate than memory-based ratings (Williams et al., 1990). Nonetheless, memory-based processing seems to be extremely important in appraisal settings, because raters have other demands on their time and energy in addition to appraisal tasks and so would not be able to easily form and maintain the categories needed for on-line processing. In fact, DeNisi, Robbins, and Williams (1989) found that raters in laboratory settings who were given additional, competing tasks to perform were more likely to exhibit memory-based processing (i.e., there were significant correlations between recall and ratings). At least to some extent, raters must rely on information stored in memory; If they cannot access that information, they cannot provide ratings that are useful to the organization. Thus, it seems clear that memory is important for ratings in the field, at least under some circumstances, and the present research was designed to help determine what role memory plays in performance appraisal decisions made in the field. It remains, however, to develop measures that reflect the potential usefulness of appraisal information to an organization. We turn now to a consideration of that problem.

Criteria for Evaluation and Outcome Measures

Evaluating appraisal systems in field settings is not a simple matter. Ideally, one might want to evaluate appraisal interventions against a criterion of rating accuracy, as is typically done in lab studies. However, even in lab settings, there is disagreement over how rating accuracy should be operationalized (e.g., Becker & Cardy,
1986; Sulsky & Balzer, 1988). In the field, where “true scores” (required for computing accuracy) are unavailable or even nonexistent (cf. Feldman, 1981), the problem of assessing accuracy becomes much more complicated. Furthermore, there is at least some reason to believe that an exclusive interest in accuracy may be misplaced and that there are other potential outcome measures that are also relevant to organizational goals (Ilgen, 1993). Consistent with these latter views, we developed and used, in the present studies, three types of criterion data that we believe address potential relevant organizational goals and, thus, help address the potential usefulness of any appraisal intervention in field settings. These three types of data were reactions to the appraisal system, indexes related to the recall of performance information, and rating distribution indexes.

Ideally, reactions to the appraisal system would involve both rater and ratee reactions, because it would be useful to know whether raters felt more confident in the ratings they provide or feel they can better defend ratings under one system relative to another (see Dickinson, 1993, for a review of these issues). In a similar fashion, it would be useful to know whether raters have more confidence in ratings they receive under one system as compared with another, or whether they are more accepting of ratings under one system. In both cases, these reactions should be related to continued use of the appraisal system by the raters and the credibility of the feedback resulting from the appraisal system as viewed by the ratee, factors that would seem to be important to an organization engaging in performance appraisal.

However, in the present studies, we were able to obtain data only from the raters. This is less than optimal (although we should note that all of our raters were also ratees in a different context, and so it seems reasonable to assume that their reactions were due, in part, to their views of the appraisal process from the perspective of ratees as well) but still provides insights into the reactions of one of the parties to the evaluation process.

In laboratory studies, recall accuracy has generally been associated with rating accuracy, but this relationship may simply indicate that memory-based processing was taking place in these settings. Even if recall accuracy is not related to rating accuracy, however, the content of recall should be related to the amount and specificity of feedback, as well as to the ability of the raters to substantiate their ratings with specific behaviors. Just as one cannot assess the accuracy of ratings in the field, one cannot assess the accuracy of recall. One can, however, examine the number of incidents recalled and the nature of those incidents as indexes of recall quality. We argue that, because it is important for raters to recall specific performance incidents for the purposes of both documentation and feedback, assessing the quality of recall should represent a second important approach to evaluating appraisal research in the field.

Indexes that are related to the distribution of ratings may also provide important information for judging the usefulness of appraisal interventions (DeNisi, in press). For example, within-rater discriminability indexes (DeNisi & Peters, 1992) reflect the extent to which raters differentiate in their ratings of different aspects of a ratee’s performance. These indexes may indicate how useful the ratings can be for feedback purposes, because raters who do not discriminate across performance dimensions within ratees will not be able to provide useful feedback about ratees’ strengths and weaknesses. Elevation indexes (DeNisi & Peters, 1992), which reflect the mean rating level in a set of ratings, may also provide useful information about ratings. Hauenstein (1992) suggested that ratings tend to be erroneously elevated when raters lack clear standards on which to base their evaluations. Thus, elevation indexes can help suggest whether or not such standards are present, and Buckley and Bernardin (1980) have suggested that less elevated ratings are more useful to organizations. Furthermore, as recently noted by Kane, Bernardin, Villanova, and Peyrefitte (1995), elevated ratings can undermine incentive pay systems and pose legal problems for organizations seeking to terminate poor performers. Finally, indexes of between-ratees discriminability tell whether raters discriminate across ratees on different performance dimensions. In the absence of this discriminability, it is less likely that raters will be able to correctly rank order ratees for decision-making purposes, reducing what Murphy (1991) has called classification accuracy.

Thus, we believe that rater reactions to ratings, measures of recall quality, and distribution indexes are all outcome measures that might be useful to an organization and might therefore serve as reasonable criterion measures. We further believe that interventions designed to improve the organization of information in memory should make performance information more accessible to raters. This, in turn, should allow raters to recall more information about ratee performance and so help them be more comfortable with their ratings. Finally, findings from laboratory experiments (e.g., DeNisi, Robbins, & Cafferty, 1989; Williams et al., 1990) have demonstrated that interventions designed to produce organization of information in memory also result in ratings that are better able to discriminate among levels of performance both within and between raters. We believe that similar results will be obtained in the field. Furthermore, because it is unlikely that everyone in an organization is performing at high levels, we also expect the organization of information to reduce the overall elevation of ratings.
by making more complete (i.e., both positive and negative) performance information accessible to raters before they make judgments.

Focus of the Present Research

The present research involved two field experiments designed to replicate the findings reported by DeNisi, Robbins, and Cafferty (1989) and those reported by Williams et al. (1990). Specifically, the present studies were designed to determine whether structured diary keeping (Study 1) and structured recall (Study 2), two interventions designed to influence the extent to which information is organized in memory, affect rater reactions to the appraisal system, the recall of performance information, and distributional indexes of performance ratings. To our knowledge, these studies represent the first attempts to extend the study of rater cognitive processes in performance appraisal to a field setting.

Study 1: Structured Diary Keeping

Raters acquire a great deal of the performance information they use for making appraisal decisions in an unstructured or unorganized manner. Unorganized information is inherently difficult to store in a meaningful pattern in memory (cf. Cafferty et al., 1986); thus, it is also difficult to accurately recall that information, which would make it difficult for raters to accurately rate subordinates even if they wished to do so. The question driving the present study, then, was whether using diaries could help raters bring organization to this information that would otherwise remain unorganized. Wherry (1952) was the first to propose that raters should keep incident diaries when observing performance. He argued that because raters would tend to forget some of the information they observed, such diaries should increase rating objectivity. Similar suggestions were made by Guion (1965) concerning the importance of diary keeping, and both Bernardin and Worthy (1977) and Buckley and Bernardin (1980) provided empirical support for the role of diary keeping in improving ratings while also reporting that raters who kept diaries indicated more positive reactions to the appraisal process.

The DeNisi, Robbins, and Cafferty (1989) study, however, suggested that diary keeping could be viewed in an additional light. Rather than focusing on diaries as simply hard copy records of observed performance, DeNisi, Robbins, and Cafferty suggested that raters, by keeping structured diaries, could impose organization on inherently unorganized information. As discussed earlier, these authors found that structuring diaries led to better organization in memory, which improved both recall and ratings. Furthermore, they reported that structuring dia-

ries according to ratees (persons) produced the most accurate recall and ratings and that, given a choice of formats, raters preferred to organize their diaries according to persons. Finally, they found that structuring diaries according to either persons or performance dimensions improved the accuracy of ratings of overall performance (i.e., improved classification accuracy) but that organizing diaries according to persons also improved raters' ability to recall and differentiate examples of effective and ineffective performance exhibited by a given ratee (i.e., improved behavioral accuracy).

DeNisi, Robbins, and Cafferty's (1989) results suggest a potentially critical role for diary keeping in the appraisal process, regardless of the criterion measure under consideration. However, their study was conducted in a laboratory where raters had no ongoing relationships with or real impressions of ratees, there were no consequences associated with ratings, and the raters had little else to compete for their attention other than the appraisal task (although participants did also complete an in-basket task); thus, there is some question about the external validity of the results.

Study 1 was designed to replicate and extend the DeNisi, Robbins, and Cafferty (1989) study in a field setting and extend that work by investigating how diary-keeping activities relate to recall, rating distributional indexes, and rater reactions to the appraisal process. Several hypotheses consistent with extending the findings of that earlier study were tested.

Raters who keep diaries organized by persons will organize information in memory according to persons; raters who keep diaries organized by performance dimensions will organize information in memory according to those dimensions; and raters who do not keep diaries will exhibit no discernible pattern of organization of information in memory. (Hypothesis 1)

Raters given a choice of how to organize their diaries will choose to organize them according to persons. (Hypothesis 2)

Both of these hypotheses were tested and supported in the earlier laboratory study. The first relates to the proposed role for structured diary keeping, and the second stems from both the earlier laboratory study and related research in social cognition (cf. Srull & Brand, 1983).

Raters who keep structured diaries will exhibit better recall of performance information, and will provide ratings that are less elevated and discriminate better both within and between ratees, than will raters who do not keep diaries. (Hypothesis 3)

Raters who keep diaries organized by persons will do "better" (in terms of the ratings and recall criteria described earlier), than will raters who keep diaries organized in any other way. (Hypothesis 4)
These two hypotheses also follow from DeNisi, Robbins, and Cafferty (1989), although they deal with distributions of ratings rather than rating accuracy. The first reflects the notion that structured diary keeping will also have an effect on both recall and ratings. The second suggests that, as in the laboratory, organizing information by persons in the field will prove to be superior to any other organization scheme. The final hypothesis extended this work further to consider rater reactions.

Raters who organize diaries according to persons will report the most positive reactions to the appraisal process, followed by those who organize their diaries in any other way; those raters who do not keep diaries will report the least favorable reactions. (Hypothesis 5)

Method

Sample

The sample included 112 first-line supervisors (71% men, 77% Caucasian, average age of 41 years, and average tenure of 8.3 years) from various manufacturing plants of a large multinational electronics firm. Supervisors were randomly selected to participate in the study and were randomly assigned to one of the experimental conditions described later. Participation was voluntary, and the training sessions were conducted on supervisors' personal time (i.e., on a Saturday morning; it should be noted, however, that it was fairly typical of these supervisors to come to work on Saturday, even if they were not scheduled for work).

Procedure

Raters reported to a corporate training facility in large groups for a half-day training program designed to improve performance management skills. Raters who were assigned to the no-diary condition were trained separately from those assigned to one of the diary-keeping conditions. Their training focused on performance coaching and the role of critical incidents; no mention was made of diary keeping. The training program for raters in the diary-keeping conditions stressed the role of memory in the appraisal process, noted the nature and role of critical incidents, and discussed how these incidents could be kept in a performance diary or incident file. Specifically, raters were instructed on the importance of and how to record descriptive rather than evaluative statements and to include positive as well as negative incidents. Raters practiced these skills after watching a videotape depicting a work scene. After each practice session, there was feedback and a group discussion aimed at reinforcing the importance of writing descriptive statements and recording both positive and negative incidents. At no time during the training session was there any mention of rating errors or rating accuracy.

At the end of this training session, raters were given instructions for the diary-keeping portion of the study. Raters were asked to write down the names of all subordinates (up to eight) for whom they had supervisory responsibilities. Raters who had more than eight subordinates were instructed to select eight representing the full range of performance. The actual number of subordinates evaluated ranged from two to eight, with an average of six. There were no differences for any dependent variable as a function of the number of raters, and there were no interactions involving the number of raters. Raters were then given loose-leaf notebooks with section dividers and blank tabs for each section. They were instructed to use the notebooks to maintain performance diaries for the next 3 months on the subordinates whose names they had supplied earlier. They were instructed to record all critical incidents in their diaries, as they occurred, according to the procedures they had practiced in the training session.

All participants were told they would return in 3 months, at which time they would rate their subordinates using a rating instrument distributed to them. This instrument was developed for the study, and ratings made for the study were to be retained in employee files. Raters who kept diaries were also told that their diaries would be collected on their return.

Approximately 3 months later, the organization began scheduling follow-up sessions. Eighty-eight raters returned for the second session. All diaries were collected, after which raters (including those in the control group) were given blank tablets and asked to recall all of the critical performance incidents they could remember over the preceding 3 months for the subordinates under consideration. They were told to write down these incidents in any format they chose. These written incidents provided the raw data from which recall indexes, described later, were developed. After 30 min of recall, all raters were instructed to rate each of their subordinates on the appraisal forms provided. Raters with diaries were told they could refer to them, as needed, when making their ratings. After completing all ratings, raters were asked to complete a 13-item general reactions measure in which they described their reactions to the training, the diary keeping, and the appraisals they made. After these materials had been collected, all participants were thanked, questions were answered, and the session was brought to a close.

Experimental Conditions

Raters were randomly assigned to four experimental conditions, three diary-keeping conditions and one no-diary control condition. The major experimental manipulations involved instructions to the raters as to how to organize their diaries.

Raters in the person (ratee) condition (n = 29) were instructed to write down the name of one of the subordinates they had submitted earlier on each tab in their binder. They were instructed to list critical incidents, as they occurred, in the section now designated for each ratee, according to the training guidelines they had received for writing critical incidents. Raters in the performance dimension condition (n = 37) were given binders with each tab filled in with one of the performance dimensions from the appraisal form. Raters were instructed to record incidents in the section for the relevant performance dimension according to the guidelines from the training program. Raters in the free-recall condition (n = 22) were given binders with the divider tabs still blank and were told that they could
organize their diaries in any way they wished. A number of examples for organizing their diaries were provided (e.g., by person, by performance dimension, by performance level, or by week), and raters were instructed, once they decided on an organizing scheme, to label the separator tabs accordingly. Raters in the control condition (n = 24) did not keep diaries.

**Dependent Variables**

**Recall data.** Recall indexes were based on the critical incidents written during the 30-min recall period. These data were independently coded by two individuals (neither author was involved in this task) who were unaware of the condition to which raters had been assigned. After receiving instructions about how to code the recall data, the coders practiced this task, and their work was reviewed by the second author. The coders achieved agreement levels of more than 85% for each of the following indexes: (a) total number of incidents recorded, (b) number of positive incidents recorded, (c) number of negative incidents recorded, and (d) number of descriptive (non evaluative) incidents recorded. Only incidents for which there was agreement between coders were used in computing indexes.

As a means of controlling for the number of subordinates being evaluated (which ranged from two to eight, with an average of six), the numbers of positive, negative, and descriptive incidents were converted to percentages of the total number of incidents recalled (note that the total number of incidents was the sum of the total number of positive evaluative incidents, total number of negative evaluative incidents, and total number of descriptive incidents). The results reported later were based on these percentages.

In addition, adjusted ratio of clustering (ARC) scores were computed for each rater for both persons (i.e., rates) and performance dimensions. The ARC index is commonly used to assess how data are stored in memory (Roenker, Thompson, & Brown, 1971), even though it actually reflects only the structure of responses to a recall task (i.e., a task likely to be influenced by the structure of information in memory). Scores could range from 0 (no clustering) to 1.00 (perfect clustering). In the present study, we assessed the extent to which recall was structured along four dimensions: persons, performance dimensions, time, and performance level. These were the patterns of organization found in the earlier lab setting (DeNisi, Robbins, & Cafferty, 1989), and each pattern was patterned as a possible way of organizing diary information to the raters in the free-recall condition. The ARC indexes for performance level and for time were essentially zero in all conditions. Therefore, we report only the ARC scores for the person and performance dimension conditions, and all comparisons were based on these two scores alone.

**Ratings data.** The appraisal instrument, developed specifically for this research, consisted of eight specific dimensions of performance (reaction to pressure, communication skills, job knowledge, interpersonal skills, timeliness, problem solving, adaptability, and initiative), as well as overall performance. We did not base these dimensions on any particular job analysis, because we needed to use the same instrument for a variety of related manufacturing jobs from several different manufacturing facilities. Instead, we adapted the dimensions from other appraisal forms. These eight dimensions were chosen from a larger set of possible dimensions by management officials of the participating organization as most relevant for use in their company. All ratings were made on 5-point rating scales (1 = poor performance and 5 = exceptional performance). There were neither main effects nor interactions as a function of the number of raters on any dependent variable involving ratings.

Data from these ratings were used to compute several rating distribution indexes (as discussed earlier). The first index rating was elevation, which measured the extent to which ratings were elevated (or, in more traditional psychometric terms, were more lenient). In the present study, elevation was operationalized simply as the mean rating given by a rater for each rating dimension, as well as for overall performance, across all raters. Thus, the sample size for analyses using this index was the number of raters.

The second index was within-ratee discriminability, which measured the extent to which raters exhibited variance across the eight dimensions for each ratee. Within-ratee discriminability was operationalized as the average within-ratee standard deviation across all rating dimensions, standardized within each performance dimension (consistent with the suggestion of Pulakos, Schmitt, & Ostroff, 1986). Greater within-rater discriminability was defined as having a higher within-ratee standard deviation across rating dimensions. One within-ratee index was computed for each rater.

The final index was between-ratee discriminability, which measured the extent to which raters discriminated across ratees. Between-rates discriminability was operationalized as the average standard deviation across rates for each dimension, as well as for overall performance; thus, there was a separate index for each performance dimension, as well as one for overall performance.

**Reactions.** Finally, all raters completed a 13-item measure of reactions to the rating process; this measure was developed for the present study but was based on available measures of reactions to ratings (e.g., Landy, Barnes, & Murphy, 1978; Landy, Barnes-Farell, & Cleveland, 1980). All items were rated on 7-point scales, higher scores reflecting more positive reactions.

The items were subjected to a principal-components analysis with varimax rotation of the resulting factors. Four dimensions emerged, explaining roughly 70% of the variance: clear understanding of what to attend to, fair appraisal program, accurate and defensible ratings, and ratings would not be surprising to raters. Scores were computed by summing the responses to relevant items. The four dimensions constituting each dimension are provided in the Appendix, along with reliability coefficients for each factor (specific factor loadings are available from the first author). As can be seen in the Appendix, the reliabilities for the last two dimensions were somewhat disappointing; to fully examine the effect of this intervention on rater reactions, however, we included all dimensions in subsequent analyses. In addition, raters in the diary-keeping conditions responded to five additional items (using the same rating scale) that all loaded on a single dimension: keeping a diary helped me in the appraisal process.
Results

The first hypothesis dealt with the structure and nature of recall. Means for recall measures are presented in Table 1 for each experimental condition. Maximum sample sizes for all analyses reflect the actual numbers of raters in each condition who returned for the second session (ns = 26, 24, 16, and 22 in the person, performance dimension, free-recall, and control conditions, respectively) rather than the number assigned to each condition (differential rates of return by condition are discussed later). Sample sizes differed for various analyses because of missing data.

The first hypothesis predicted that raters would organize information in memory consistent with the way they organized their diaries. This hypothesis was tested by comparing ARC performance dimension and ARC person scores in different experimental conditions. One-way analyses of variance (ANOVA) indicated significant effects for both ARC performance dimension scores, $F(3, 68) = 3.24, p < .05$, $\eta^2 = .05$, and ARC person scores, $F(3, 68) = 10.76, p < .001$, $\eta^2 = .12$.

Planned comparisons (because the specific nature of differences was hypothesized) indicated that ARC performance dimension scores were significantly ($p < .01$) greater in the performance dimension condition than in all other conditions; however, these scores did not differ from each other across the other three conditions. Regarding the ARC person scores, planned comparisons indicated that raters in the person condition did not differ from those in the free-recall condition (but see the results for Hypothesis 2 described subsequently). Both groups, however, had significantly higher scores ($p < .01$) than the control group, which, in turn, had significantly higher scores ($p < .01$) than the performance dimension group. These results (especially in light of the results for the next hypothesis) provide reasonably strong support for the hypothesized consistency between diary and memory structure, although there was clearly a tendency for raters who had not kept diaries to also organize their recall (and presumably memory) according to rates.

The second hypothesis predicted that raters who were given a choice would prefer to organize their diaries according to persons. Results indicated that all raters in the free-recall condition chose to organize their diary by persons, providing strong support for the second hypothesis. Furthermore, in light of these findings, the results reported earlier provide very strong support for the first hypothesis, that raters in the free-recall condition should have been expected to behave like those in the person condition.

The third and fourth hypotheses addressed differences in recall and the distributions of ratings across conditions. The relevant recall indexes are shown in Table 1. Table 2 presents elevation indexes (i.e., mean ratings) for each rating dimension as well as for overall performance in each experimental condition. Table 3 presents between-rates discriminability indexes (i.e., standard deviations) for each dimension as well as overall performance, along with the mean within-ratee discriminability indexes for each condition.

Beginning with the recall data, one-way ANOVAs indicated no differences in the percentage of negative incidents recalled across conditions, but significant effects were observed for percentage of positive incidents, $F(3, 63) = 4.63, p < .005$, $\eta^2 = .13$; percentage of descriptive incidents, $F(3, 63) = 39.63, p < .001$, $\eta^2 = .35$; and total number of incidents recalled, $F(3, 63) = 8.29, p < .001$; $\eta^2 = .19$. Planned comparisons contrasting raters who kept diaries with those in the control condition (for Hypothesis 3) indicated that the former recalled more total incidents, a greater percentage of descriptive incidents, and a smaller percentage of positive incidents (all $ps < .01$) than the latter. These findings provide moderate support for the third hypothesis relative to recall (but see later discussion).

In terms of the ratings data, the first set of analyses involved the index of elevation. An omnibus multivariate analysis of variance (MANOVA) considering mean ratings (elevation indexes) for all eight performance dimension ratings and overall performance was significant, $F(8, 75) = 3.81, p < .05$. Univariate tests indicated that elevation indexes differed across conditions for each of the eight dimensional ratings (all $ps < .05$), and that elevation differed for ratings of overall performance as well, $F(3, 80) = 10.56, p < .01, \eta^2 = .08$. In each case, planned comparisons indicated that ratings made by control group raters were significantly ($p < .05$) higher than those provided by raters who kept a diary.

The second set of analyses dealt with within-ratee dis-

---

Table 1

<table>
<thead>
<tr>
<th>Index</th>
<th>Person</th>
<th>Performance Dimension</th>
<th>Free-recall</th>
<th>No diary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total recall</td>
<td>27.8a</td>
<td>16.1b</td>
<td>26.8</td>
<td>9.3b</td>
</tr>
<tr>
<td>Positive (%)</td>
<td>17.0a</td>
<td>41.2b</td>
<td>21.5</td>
<td>62.3b</td>
</tr>
<tr>
<td>Negative (%)</td>
<td>20.5</td>
<td>25.4</td>
<td>27.2</td>
<td>22.8</td>
</tr>
<tr>
<td>Descriptive (%)</td>
<td>62.7a</td>
<td>33.0b</td>
<td>51.2</td>
<td>14.7b</td>
</tr>
<tr>
<td>Specific raters (%)</td>
<td>0.91a</td>
<td>0.29a</td>
<td>0.99</td>
<td>0.75b</td>
</tr>
<tr>
<td>ARC performance dimension</td>
<td>0.08a</td>
<td>0.45b</td>
<td>0.06</td>
<td>0.07a</td>
</tr>
</tbody>
</table>

Note. Means for the same dependent variable with different subscripts differ from each other at $p < .05$. Person, $n = 26$; performance dimension; $n = 24$; free recall, $n = 16$; no diary, $n = 22$. Sample sizes for specific analyses may differ as a result of missing data. ARC = adjusted ratio of clustering.
Table 2

Mean Elevation Indexes in Study 1 According to Diary Organization Experimental Condition

<table>
<thead>
<tr>
<th>Rating dimension</th>
<th>Person</th>
<th>Performance dimension</th>
<th>Free-recall</th>
<th>No diary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reaction to pressure</td>
<td>3.56_a</td>
<td>3.63_a</td>
<td>3.67_a</td>
<td>3.87_b</td>
</tr>
<tr>
<td>Communication skills</td>
<td>3.77_a</td>
<td>3.81_a</td>
<td>3.53_a</td>
<td>4.04_a</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>3.84_a</td>
<td>4.01_a</td>
<td>4.00_a</td>
<td>4.31_a</td>
</tr>
<tr>
<td>Interpersonal skills</td>
<td>3.48_a</td>
<td>3.56_a</td>
<td>3.58_a</td>
<td>4.08_a</td>
</tr>
<tr>
<td>Timeliness</td>
<td>3.60_a</td>
<td>3.69_a</td>
<td>3.70_a</td>
<td>3.96_b</td>
</tr>
<tr>
<td>Problem solving</td>
<td>3.41_a</td>
<td>3.50_a</td>
<td>3.59_a</td>
<td>3.71_b</td>
</tr>
<tr>
<td>Adaptability</td>
<td>3.43_a</td>
<td>3.39_a</td>
<td>3.50_a</td>
<td>3.67_a</td>
</tr>
<tr>
<td>Initiative</td>
<td>3.57_a</td>
<td>3.46_a</td>
<td>3.53_a</td>
<td>3.89_a</td>
</tr>
<tr>
<td>Overall performance</td>
<td>3.31_a</td>
<td>3.42_a</td>
<td>3.43_a</td>
<td>3.67_b</td>
</tr>
</tbody>
</table>

Note. Rating elevation was computed as the mean rating given by a rater for the dimension across raters. Means for the same analyses with different subscripts differ from each other at \( p < .05 \). Person, \( n = 26 \); performance dimension, \( n = 24 \); free recall, \( n = 16 \); and no diary, \( n = 22 \). Sample sizes for specific analyses may differ as a result of missing data.

The third set of analyses dealt with between-raters discriminability, and, as above, an overall MANOVA indicated a significant effect across conditions, \( F(8, 75) = 3.05, p < .05 \). More specifically, between-raters discriminability differed significantly across conditions for seven of the eight performance dimensions (all except timeliness), as well as for overall performance (all \( ps < .05 \)). Planned comparisons indicated that, in six of the seven cases, the control condition raters exhibited significantly (\( p < .05 \)) less discriminability than the raters who kept diaries (the pattern did not hold for communication skills). Thus, there was considerable support for the third hypothesis in terms of both the recall and the rating data.

The fourth hypothesis specified that raters who organized their diaries by persons (i.e., those in both the person and free-recall conditions) would exhibit better recall

Table 3

Mean Discriminability Indexes in Study 1 According to Diary Organization Experimental Condition

<table>
<thead>
<tr>
<th>Rating dimension</th>
<th>Person</th>
<th>Performance dimension</th>
<th>Free-recall</th>
<th>No diary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between-rater discriminability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reaction to pressure</td>
<td>.96_a</td>
<td>.80_a</td>
<td>.87_a</td>
<td>.57_a</td>
</tr>
<tr>
<td>Communication skills</td>
<td>.97_a</td>
<td>.73_a</td>
<td>.92_a</td>
<td>.69_a</td>
</tr>
<tr>
<td>Job knowledge</td>
<td>.73_a</td>
<td>.75_a</td>
<td>.68_a</td>
<td>.46_a</td>
</tr>
<tr>
<td>Interpersonal skills</td>
<td>.99_a</td>
<td>.70_b</td>
<td>.90_a</td>
<td>.46_a</td>
</tr>
<tr>
<td>Timeliness</td>
<td>.84_a</td>
<td>.70</td>
<td>.76</td>
<td>.76</td>
</tr>
<tr>
<td>Problem solving</td>
<td>.92_a</td>
<td>.74_a</td>
<td>.90_a</td>
<td>.61_a</td>
</tr>
<tr>
<td>Adaptability</td>
<td>.68_a</td>
<td>.76_a</td>
<td>.90_a</td>
<td>.50_a</td>
</tr>
<tr>
<td>Initiative</td>
<td>.85_a</td>
<td>.75_a</td>
<td>.89_a</td>
<td>.50_a</td>
</tr>
<tr>
<td>Overall performance</td>
<td>.65_a</td>
<td>.66_a</td>
<td>.67_a</td>
<td>.45_b</td>
</tr>
</tbody>
</table>

| Within-rater discriminability | | |
| Average | .81_a  | .71_b | .86_a | .60_b |

Note. Between-rates discriminability was computed as the standard deviation of the ratings for a certain rater on a certain dimension. Means for the same analyses with different subscripts differ from each other at \( p < .05 \). Person, \( n = 26 \); performance dimension, \( n = 24 \); free recall, \( n = 16 \); and no diary, \( n = 22 \).
and provide better ratings on all three distributional indexes than those who organized their diaries according to performance dimension. Planned comparisons of the recall data indicated that raters keeping diaries organized by persons had greater total recall ($p < .01$), recalled a smaller percentage of positive incidents ($p < .01$), and recalled a greater percentage of descriptive incidents ($p < .01$) than did raters organizing diaries by performance dimensions. Planned comparisons between these groups on rating elevation indicated no significant differences. Other planned comparisons indicated that the two groups of raters differed significantly on between-ratees discriminability only for reaction to pressure ($p < .001$), communication skills ($p < .01$), interpersonal skills ($p < .01$), and problem solving ($p < .01$). Finally, the planned comparison for within-ratee discriminability was significant ($p < .05$) and indicated that raters keeping diaries organized by performance dimensions exhibited less within-ratee discriminability than did raters who kept diaries organized by persons. Thus, these results provide strong support for the fourth hypothesis relative to recall but only moderate support relative to ratings.

Finally, the fifth hypothesis dealt with differences in reactions to the rating process. The means for each reactions dimension, by experimental condition, are presented in Table 4. An omnibus MANOVA considering four of the five reactions dimensions (diary keeping helped me in the appraisal process was not included because raters in the control condition were not asked to respond to these items) was significant, $F(4, 71) = 3.05$, $p < .05$, but follow-up one-way ANOVAs revealed significant differences across conditions only for the third dimension (accurate and defensible ratings), $F(3, 78) = 4.21$, $p < .01$, $\eta^2 = .09$. Planned comparisons indicated that raters who organized their diaries by persons believed their ratings were more accurate and defensible than did raters in either the task condition or control condition, who in turn did not differ from each other. These results provided only limited support for the final hypothesis; however, this particular dimension would seem to assess a particularly important aspect of raters’ reactions to the appraisal process, even though it was one of the two dimensions that exhibited rather low reliability.

In addition, although no hypothesis was stated, there was a significant difference, $F(2, 56) = 9.50$, $p < .001$, $\eta^2 = .14$, across diary-keeping conditions for the fifth dimension (diary keeping helped me in the appraisal process). This effect was due to the raters in the free-recall and person conditions more strongly endorsing statements reflecting that the use of diaries was helpful than did raters in the performance dimension condition.

**Discussion**

Although diary keeping has previously been shown to be an effective aid for raters in field settings, the present study focused on the role of diaries in organizing information in memory. Beyond replicating prior laboratory research, the present study is the first to demonstrate the effect of cognitive models of performance appraisals in field settings and, as a result, to address a major concern regarding the applicability of this approach to applied settings.

In the laboratory, raters who kept diaries produced ratings that were more accurate than those who did not. In previous field studies (e.g., Bernardin & Walter, 1977), raters who kept diaries were less prone to rating errors than those who did not. In the present study, raters who kept diaries produced ratings that were less elevated and that discriminated more both within and between rates than those of raters who did not keep diaries. The specific format (person vs. performance dimension) in which raters kept their diaries had an effect on actual ratings and recall, as well as (to a lesser extent) rater reactions. Specifically, raters who organized diaries by persons recalled more information overall (as well as recalling a lower percentage of positive and a greater percentage of descriptive information), appeared to have somewhat more confidence in their ratings, and believed that diary keeping was more useful than raters who organized diaries by performance dimension or did not keep diaries at all. Although the differences between the two diary-keeping formats were not as strong as those that had been obtained in the laboratory (DeNisi, Robbins, & Cafferty, 1989), the artificiality of the laboratory setting may well have accentuated differences between these formats. Thus, it may well be that, in the field, the format of the diary is less important than whether a rater keeps a structured diary at all. Nonetheless, the present results do suggest some advan-

---

**Table 4**

Mean Scores for Reactions Dimensions in Study 1 According to Diary Organization Experimental Condition

<table>
<thead>
<tr>
<th>Reaction dimension</th>
<th>Person</th>
<th>Performance dimension</th>
<th>Free-recall</th>
<th>No diary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear understanding of what to attend to</td>
<td>23.28</td>
<td>22.13</td>
<td>22.69</td>
<td>23.12</td>
</tr>
<tr>
<td>Fair appraisal program</td>
<td>26.16</td>
<td>25.17</td>
<td>25.67</td>
<td>25.89</td>
</tr>
<tr>
<td>Accurate and defensible ratings</td>
<td>18.08a</td>
<td>16.68a</td>
<td>18.48a</td>
<td>16.69a</td>
</tr>
<tr>
<td>Ratings would not be surprising to rates</td>
<td>11.60</td>
<td>12.52</td>
<td>12.07</td>
<td>12.04</td>
</tr>
<tr>
<td>Diary keeping helped me in the</td>
<td>16.75a</td>
<td>13.36a</td>
<td>18.01a</td>
<td></td>
</tr>
<tr>
<td>appraisal process</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* Means for the same reactions dimension with different subscripts differ from each other at $p < .05$. Person, $n = 26$; performance dimension, $n = 24$; free recall, $n = 16$; and no diary, $n = 22$. Sample sizes for specific analyses may differ as a result of missing data.
tages to person-organized diaries, and, in general, they support the external validity of those laboratory findings.

Thus, it would appear that diary keeping is an effective aid to raters, not only because it provides a hard copy of performance information but also because it helps raters to organize information in memory. These results suggest that, even when raters did not or could not refer to the diaries they were keeping, they were better able to access specific incidents of performance about raters. This would allow raters to give better feedback on a day-to-day basis without having to refer to their written records. This, in and of itself, is an important implication of the present results.

In addition, by extending this work to consider rater reactions, the results of the present study also revealed that raters keeping diaries organized by persons perceived their ratings to be more accurate and defensible. This would probably result in these raters being more confident of the feedback they would give as well as the personnel decisions they would make. In fact, the percentage of raters who kept diaries organized by performance dimensions and who returned for the second session (65%) was considerably smaller than the percentage of raters who kept diaries organized by person (82%) or those who did not keep diaries at all (92%). We see this as yet another indicator that organizing diaries by performance dimensions is not preferred by raters who keep diaries.

Much has been made of the various pressures operating on raters during the appraisal process (see Murphy & Cleveland, 1991, chaps. 8 and 9, for an in-depth discussion of these pressures), and anything that would make the raters' task seem easier to them should be considered. Furthermore, as noted earlier, rater (as well as ratee) reactions to the appraisal process may represent an important criterion measure for appraisals in the field, and, although there were not many differences in reactions as a function of diary keeping, increased rater confidence may be an extremely important determinant of the success of any appraisal system.

Finally, although the effects of diary-keeping format on actual ratings were not strong, some differences were obtained. These differences were found even though raters could refer to their diaries when making their ratings. Therefore, any effects found were over and above those attributable to the simple fact that raters could refer to a hard copy of performance information. Clearly, the format in which a diary is kept does affect ratings (even if not dramatically), which further underscores the importance of diaries as an information organization intervention in addition to their role as a record-keeping intervention.

Indeed, if rating accuracy were to continue to be a primary concern for evaluating appraisals in the field, the results from the present field study, together with the results of the prior laboratory study by DeNisi, Robbins, and Cafferty (1989), support the possibility that having raters keep diaries organized by ratees may result in both improved behavioral accuracy and classification accuracy (Murphy, 1991). The fact that raters in the present study who kept diaries organized by persons had increased within-ratee discriminability and exhibited better recall suggests that person-organized diaries might be able to improve behavioral accuracy. Results from DeNisi, Robbins, and Cafferty (1989) indicated that this diary-keeping format improved classification accuracy. Thus, use of diaries may facilitate making “correct decisions” and providing meaningful feedback (cf., Meyer, Kay, & French, 1965). The present results therefore, suggest that the cognitive role for diary keeping, based on laboratory results only, may well generalize to field settings.

Study 2: Memory Reorganization

Results from Study 1 suggest that having raters organize information as it is being acquired, through structured diary keeping, has a positive effect on recall, ratings, and reactions. As noted earlier, it might also be possible for raters to reorganize existing information to form relevant cognitive structures before using that information to form evaluations. Williams and his associates (Williams et al., 1986, 1990) found that the use of a structured recall task (asking raters to recall performance incidents and record them according to a given format) produced results similar to those obtained when the information was initially obtained in a structured manner. Because this approach does not require that raters maintain diaries throughout the observation period, it is less likely to face resistance. It therefore offers real promise as a means of introducing organization to unorganized performance information.

As with structured diary keeping, however, all support for this intervention has come exclusively from laboratory settings. The present study was designed to determine whether the effects obtained in the lab could generalize to the field. The prior laboratory studies assigned raters to one of three structured recall conditions. The same conditions were used here. All raters received a brief training program designed to help them identify and record descriptive critical incidents. Then, before making their ratings, raters were asked to record all critical incidents they could recall according to one of these three sets of instructions (conditions are described later).

The previous lab studies dealing with structured recall (as well as the DeNisi, Robbins, & Cafferty, 1989, diary-keeping study) found that structuring information ac-
cording to any format resulted in better ratings and recall than those produced by a control group. These studies also found that organizing information according to persons was better for both ratings and recall than organizing according to task or performance dimensions and that, given a choice, raters would prefer to organize information in memory according to persons. As in Study 1, we considered rater reactions to the appraisal process; we evaluated the intervention relative to the recall and rating distribution indexes used in the previous study as well. Because Study 1 indicated that an intervention designed to enhance memory organization, especially one designed to produce organization according to persons, also resulted in somewhat more favorable reactions of raters to the appraisal process and their ratings, we expected a similar reaction to the present intervention. On the basis of these findings, then, we tested three hypotheses.

Raters given a choice of how to organize their recall (free recall) will choose to organize information by persons. (Hypothesis 1)

Raters completing a structured recall task will provide ratings that are less elevated and that discriminate better both within and between raters than those of raters who do not complete a structured recall task, and they will report more positive reactions to the appraisal process (Hypothesis 2)

Raters who organize their recall according to persons will provide less elevated and more discriminating ratings, and will report more positive reactions and exhibit better recall, than will raters who organize their recall according to any other pattern. (Hypothesis 3)

Method

Sample and Setting

The study involved 118 middle-level managers from various departments of a large international transportation organization. Managers were randomly chosen to participate in the study and were randomly assigned to training sessions in which the experimental conditions were induced. The sample included 90 men (3 non-White, average age of 43 years, average tenure with organization of 20 years, and average tenure in present position of 7 years) and 28 women (all White, average age of 40 years, average tenure with organization of 16 years, and average tenure in present position of 6 years). Thus, the raters were experienced and were probably experts on the jobs to be evaluated.

All raters participated in a training program that focused on the observation and acquisition aspects of the appraisal process. The program involved a 3-hr workshop in which the importance of memory in appraisal was discussed, critical incidents were defined and problems associated with them (e.g., focusing on negative and evaluative incidents) discussed, and participants practiced identifying and writing critical incidents. There was no mention of rating errors or rating accuracy.

Experimental Conditions and Procedure

There were three experimental conditions involving different instructions on how to structure the recall task, as well as a no-recall control condition. Participants were asked to write down all of the critical incidents they could recall for up to eight of their subordinates. Participants with more than eight direct reports were instructed to select eight subordinates representing the full range of performance. The names of all raters were listed before the recall task was initiated. The actual number of subordinates rated in the study ranged from two to eight. There were no main effects or significant interactions involving number of subordinates rated. The recall task lasted 30 min.

Participants in the person condition were instructed to recall all of the incidents they could recall for one subordinate and then go on to the next subordinate (and so on until they had done this for up to eight subordinates). Participants in the performance dimension condition were given the same instructions except that they were told to organize the recall data according to key performance dimensions included in the organization's appraisal instrument. Specifically, 8 of the 18 dimensions currently in use were randomly selected before the start of the study to make this condition comparable with the person condition (see list in Table 5). Participants in the free-recall condition were told to organize their recall data in any fashion they chose. Participants in the control condition did not complete a structured recall task.

After completing the recall task, all managers rated the eight (or fewer) subordinates for whom they had provided recall data using the organization's standard appraisal instrument (the control group raters did so without the benefit of the recall task). After completing the ratings, all participants completed a "reactions" measure assessing their views of the appraisal task, the specific information they could recall, and the training program itself.

The first training session was originally designed to be a pilot person condition session allowing practice of the training procedures with target employees. This session went so smoothly that we chose to include data from the 27 participants in this session in the study. This was done without consideration of how inclusion of these data would influence the results. In fact, the data were analyzed both with and without these pilot data with no differences in any of the results. The pilot data are included here, resulting in a total of 48 participants in the person condition, 21 in the performance dimension condition, 28 in the free-recall condition, and 21 in the control condition.

Dependent Variables

Recall data. After the completion of the training program, raters in the recall conditions took 30 min to write down all critical incidents of performance they could recall from the previous 3 months for their eight (or fewer) subordinates. These data were transcribed and independently coded by the same two independent judges who had coded the recall data from Study 1. There was, again, a high degree of agreement between the judges (almost 90%) for each of the following indexes: (a) total number of incidents recorded, (b) number of positive incidents
Table 5
Means for Distributional Indexes in Study 2 According to Diary Organization Experimental Condition

<table>
<thead>
<tr>
<th>Rating dimension</th>
<th>Person + free recall</th>
<th>Performance dimension</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leadership</td>
<td>2.93oked</td>
<td>3.29oked</td>
<td>3.34oked</td>
</tr>
<tr>
<td>Reaction to pressure</td>
<td>3.3oked</td>
<td>3.73oked</td>
<td>3.9oked</td>
</tr>
<tr>
<td>Development of people</td>
<td>3.14oked</td>
<td>3.33oked</td>
<td>3.9oked</td>
</tr>
<tr>
<td>Written communication skills</td>
<td>3.12oked</td>
<td>3.20oked</td>
<td>3.52oked</td>
</tr>
<tr>
<td>Interpersonal skills</td>
<td>3.18oked</td>
<td>3.45oked</td>
<td>3.50oked</td>
</tr>
<tr>
<td>Decisiveness</td>
<td>3.11oked</td>
<td>3.36oked</td>
<td>3.44oked</td>
</tr>
<tr>
<td>Adaptability</td>
<td>3.40oked</td>
<td>3.41oked</td>
<td>3.89oked</td>
</tr>
<tr>
<td>Initiative</td>
<td>3.46oked</td>
<td>3.72oked</td>
<td>3.90oked</td>
</tr>
<tr>
<td>Overall performance</td>
<td>3.25oked</td>
<td>3.64oked</td>
<td>3.89oked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Between-raters discriminability</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall performance</td>
<td>1.02oked</td>
<td>0.85oked</td>
<td>0.59oked</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Within-raters discriminability</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1.00oked</td>
<td>0.77oked</td>
<td>0.50oked</td>
</tr>
</tbody>
</table>

Note: Means for the same dependent variable with different subscripts differ from each other at p ≤ .05.
Person and free recall, n = 76; performance dimension, n = 21; and control, n = 21. Sample sizes for specific analyses may differ because of missing data.

Recorded, (c) number of negative incidents recorded, and (d) number of descriptive (nonevaluative) incidents recorded. To control for the number of subordinates being evaluated, we converted the number of positive, negative, and descriptive incidents to percentages of the total number of incidents recalled, as was done in Study 1 (again, the percentages of positive, negative, and descriptive incidents summed to 100%). In addition, ARC scores were computed for each rater for both persons (i.e., rates) and performance dimensions.

Ratings data. On the basis of the company’s appraisal instrument, each manager rated all subordinates on 18 specific performance dimensions, as well as on overall performance. All ratings were made on 5-point scales ranging from poor performance (1) to exceptional performance (5). Only data relative to the eight performance dimensions randomly chosen for inclusion in the performance dimension condition were used in the analyses, because these were the dimensions those raters used to structure their recall. Data from these ratings were used to compute three indexes reflecting rating distributions.

The first index, rating elevation, measured the extent to which ratings were elevated. This index was operationalized as in Study 1, and the sample size was the number of raters. The second index, within-rater discriminability, measured the extent to which raters exhibited variance across the eight dimensions for each rater. Again, this index was operationalized as in Study 1 such that greater within-rater discriminability was defined as having a higher within-rater standard deviation across rating dimensions. One within-rater index was computed for each rater.

The final index, between-raters discriminability, measured the extent to which raters discriminated across raters separately for each rating dimension. Once more, this index was operationalized as in Study 1, and there was a separate index for each of the eight performance dimensions considered, as well as one for overall performance.

Reactions data. Finally, all raters completed a somewhat modified (to meet the needs of this organization) version of the 13-item measure of reactions to the rating process developed for Study 1. This version of the instrument contained only 11 items, items were worded somewhat differently than were the items used in the prior instrument, and, of course, the raters did not respond to the additional items from Study 1 dealing with diary keeping. All items were rated on 7-point scales (higher scores reflecting more positive reactions), and the items were intended to assess reactions similar to those measured in Study 1.

The items were subjected to a principal-components analysis with varimax rotation of the resulting factors. Four dimensions, explaining 65% of the variance and paralleling the dimensions obtained in Study 1, emerged: clear understanding of what to attend to, fair and accurate appraisal program, defensible ratings, and ratings would not be surprising to raters. Scores for each dimension were again computed by summing the responses to relevant items. The actual items used for each factor and the resulting reliability coefficients are included in the Appendix. Reliability coefficients were quite acceptable for the first three dimensions, but the reliability for the fourth dimension was low. Nonetheless, all four dimensions were again retained for analyses.

Results

Table 6 contains the means for the three recall conditions on the various indexes computed for the recall data.
The first hypothesis predicted that raters in the free-recall task would choose to organize their recall according to persons. The mean ARC person scores in the three conditions suggest that raters in the free-recall condition organized memory according to persons almost to the same extent as raters in the person condition. In fact, only one rater in the free-recall condition failed to organize recall according to persons. A one-way ANOVA for ARC person scores indicated a main effect for condition, $F(2, 93) = 12.56, p < .001, \eta^2 = .21$. Post hoc (Scheffé) tests indicated that raters in the person and free-recall conditions did not differ from each other but that these raters did differ significantly ($p < .01$) from those in the performance dimension condition. As a further check, a one-way ANOVA for ARC performance dimension scores also indicated a main effect for condition, $F(2, 93) = 24.67, p < .001, \eta^2 = .27$. Post hoc tests again indicated that raters in the person and free-recall conditions did not differ from each other but that these raters did differ significantly ($p < .001$) from those in the performance dimension condition. Thus, there was strong support for the first hypothesis.

The remaining hypotheses involved both the rating data and the reactions data, and the third hypothesis involved the recall data as well. The mean rating elevation indexes for all eight performance dimensions, as well as for overall performance, are presented in Table 5. These data are presented for the person and free-recall conditions combined because, as just noted, all but one of the raters in the free-recall condition organized their recall by persons. (It should be noted, however, that we also analyzed the data keeping the two conditions separate, with no difference in the interpretation of results.) In addition, Table 5 presents the average within-ratee discriminability index for each condition and the between-ratees discriminability index for overall performance only. The between-ratees data for the eight separate rating dimensions are not presented in Table 5 because, in no case were the differences across conditions significant. Table 7 presents the means for the four reactions factors for each condition.

Hypothesis 2 predicted that raters who participated in the structured recall task (regardless of specific condition) would provide less elevated and more discriminating ratings and report more positive reactions than would raters from the control condition. A MANOVA considering means (elevation indexes) for the eight performance dimension ratings and the ratings for overall performance was significant, $F(8, 89) = 6.28, p < .01$. Univariate tests indicated that elevation indexes differed across conditions for each individual performance dimension, as well as for overall performance (all $ps < .05$). Furthermore, in every case, planned comparisons indicated that elevation was significantly ($ps < .05$) lower for raters who participated in the structured recall task. As noted earlier, there were no differences for any individual performance dimension on between-ratees discriminability. However, this index for overall performance did differ across conditions, $F(2, 94) = 16.22, p < .01, \eta^2 = .12$, and planned comparisons indicated that raters in the structured recall conditions exhibited significantly ($p < .05$) greater discriminability across raters than did those in the control condition. Finally, the means for within-ratee discriminability differed across conditions, $F(2, 94) = 11.52, p < .01, \eta^2 = .16$, planned comparisons again indicating that the raters in the structured recall conditions exhibited significantly ($p < .05$) more within-ratee discriminability than did those in the control condition.

Significant effects for condition were found for two reactions dimensions: fair and accurate appraisal program, $F(2, 93) = 7.62, p < .05, \eta^2 = .07$, and ratings would not be surprising to raters, $F(2, 93) = 8.56, p < .05, \eta^2 = .08$. In each case, planned comparisons revealed that the raters

<table>
<thead>
<tr>
<th>Table 7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Scores for Reactions Dimensions in Study 2 According to Diary Organization Experimental Condition</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reaction dimension</th>
<th>Person</th>
<th>Free-recall</th>
<th>Performance dimension</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clear understanding of what to attend</td>
<td>11.35</td>
<td>7.79</td>
<td>11.17</td>
<td>10.21</td>
</tr>
<tr>
<td>Fair and accurate appraisal program</td>
<td>15.34</td>
<td>15.61</td>
<td>15.50</td>
<td>11.24</td>
</tr>
<tr>
<td>Defensible ratings</td>
<td>16.37</td>
<td>16.43</td>
<td>15.89</td>
<td>16.18</td>
</tr>
<tr>
<td>Ratings would not be surprising to raters</td>
<td>11.21</td>
<td>11.40</td>
<td>10.21</td>
<td>9.56</td>
</tr>
</tbody>
</table>

*Note*: Means for the same reactions dimension with different subscripts differ from each other at $p < .05$. Person, $n = 48$; free recall, $n = 28$; performance dimension, $n = 21$. Sample sizes for specific analyses may differ because of missing data. ARC = adjusted ratio of clustering.

<table>
<thead>
<tr>
<th>Table 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Means for Recall Indices in Study 2 According to Diary Organization Experimental Condition</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>Person</th>
<th>Free-recall</th>
<th>Performance dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total recall</td>
<td>38.5</td>
<td>39.2</td>
<td>41.5</td>
</tr>
<tr>
<td>Positive (%)</td>
<td>30.6</td>
<td>31.3</td>
<td>38.5</td>
</tr>
<tr>
<td>Negative (%)</td>
<td>26.0</td>
<td>25.8</td>
<td>24.9</td>
</tr>
<tr>
<td>Descriptive (%)</td>
<td>43.4</td>
<td>42.9</td>
<td>36.6</td>
</tr>
<tr>
<td>ARC person</td>
<td>0.99</td>
<td>0.95</td>
<td>0.24</td>
</tr>
<tr>
<td>ARC performance dimension</td>
<td>0.10</td>
<td>0.06</td>
<td>0.93</td>
</tr>
</tbody>
</table>

*Note*: Means for the same dependent variable with different subscripts differ from each other at $p < .05$. Person, $n = 48$; free recall, $n = 28$; and performance dimension, $n = 21$. Sample sizes for specific analyses may differ because of missing data. ARC = adjusted ratio of clustering.
in the structured recall conditions had significantly \( p < .05 \) more positive reactions in these areas than did raters in the control condition. Thus, overall, there was moderate support for the second hypothesis. Predicted differences were not found for between-ratee discriminability, except for overall performance, and reactions differed for only two (albeit critical) reactions factors.

The final hypothesis predicted that the ratings with the lowest elevation and the most discriminability, the best recall, and the most positive reactions would be observed for those who structured recall according to persons. To test this hypothesis, we compared raters in the person condition (including those initially assigned to the free-recall condition) with those in the performance dimension condition. Planned comparisons revealed that raters who structured recall according to persons had significantly \( p < .05 \) less elevated ratings for six of the eight performance dimensions (all except written communications skills and adaptability), as well as for overall performance. Furthermore, these raters exhibited significantly \( p < .05 \) greater between-ratee discriminability on overall performance, and significantly \( p < .05 \) greater within-ratee discriminability, than raters who structured their recall by performance dimensions.

In terms of recall (the relevant data are presented in Table 6), there were no differences in the number of incidents recalled, but there were differences in the nature of the incidents recalled. Specifically, raters structuring recall according to persons recalled a significantly \( p < .05 \) greater percentage of descriptive incidents and a smaller percentage of positive incidents. Finally, raters structuring recall by persons reported a significantly \( p < .05 \) more positive reaction to the fair and accurate appraisal program dimension than did raters organizing recall according to performance dimensions. Thus, moderate support was found for Hypothesis 3 as well, and the results for Hypotheses 2 and 3 were roughly parallel. That is, to the extent that structured recall seems to have an effect on cognitive processes, structuring recall according to persons appears better, relative to these outcome measures, than structuring recall according to performance dimensions.

**Discussion**

The present results suggest that a structured recall task, studied previously only in the laboratory, may be able to help raters recall and organize previously acquired performance information in a field setting as well. This intervention appeared to have a fairly consistent effect on both ratings and reactions, although the effects obtained here were somewhat less powerful than those obtained in the laboratory.

The present results further suggest that interventions aimed at having raters organize information in memory according to persons (rates) will be especially effective. In the present study, person organization of recall resulted in better ratings relative to the outcome variables used here. Also, the "quality" of the information recalled by raters who structured recall this way seemed to be better; there were more descriptive than evaluative incidents recalled, and a smaller percentage of positive versus negative incidents were recalled. Furthermore, these raters also reported believing more strongly that their ratings were fair and accurate, and, as noted earlier, such reactions may be as important as any other criteria for evaluating appraisal systems. These results are especially encouraging in that raters, given a choice of how to organize information, almost unanimously chose to do so by person. Thus, it seems unlikely that there would be much resistance to interventions aimed at inducing person organization. Person organization also seems to be the natural way to organize information for appraisal decisions (because, in such cases, persons rather than tasks must be appraised), consistent with the findings from studies in social cognition (e.g., Srull, 1983).

In conclusion, the present results further support the usefulness of interventions aimed at affecting raters cognitive processes in field settings. The fact that the present intervention occurs after performance information has been collected and so requires less effort on the part of the raters makes it especially promising.

**General Discussion**

The results of these two field experiments provide consistent support for the importance of interventions designed to impose structure on unstructured information for recall, ratings, and, to a lesser extent, reactions to the appraisal process. These results suggest that the same processes that seemed important to decisions in the laboratory may well be important to decisions in the field as well.

**Limitations**

Before further discussion of the implications of these findings, it is useful to discuss the limitations of the present studies (the limitations of cognitive approaches to performance appraisal have been widely discussed). First, the ratings provided in these studies, although they became part of subordinates' personnel files (and in one case were used in decision making), were not, to our knowledge, shared with subordinates. Thus, raters should have been more motivated to be accurate here than they might have been under conditions in which ratings would be shared with ratees (cf. Bernardin & Beatty,
The present studies, then, did not fully address all sources of rater motivation that might come into play (see Murphy & Cleveland, 1991, chap. 9). This also means that the present studies could not address issues arising from the political pressures involved in raters attempting to balance different goals in the appraisal process; thus, the focus may still have been more on judgments than on ratings (see Murphy & Cleveland, 1991, p. 211, for a discussion of this distinction). Nonetheless, these studies involved real managers rating the actual performance of real subordinates; thus, results reflect many contextual variables not considered in laboratory studies. As such, the present studies represent a meaningful extension of prior laboratory-based research.

A second limitation involves the fact that the ratings provided were made on scales assessing mostly nonbehavioral dimensions of performance. Although most scholars in the field have traditionally suggested that such rating dimensions be avoided (e.g., Bernardin & Beatty, 1984; Murphy & Cleveland, 1991), these are the types of dimensions typically found in most organizations (cf. Bernardin & Villanova, 1986). Furthermore, more recent work on performance models (cf. Borman, 1991) suggests that nonbehavioral ratings may be more relevant for most raters' models of effective performance, especially in circumstances in which raters need to provide ratings on which to base personnel decisions. Nonetheless, it remains to be determined whether different results would be obtained with different rating scales.

Third, the interventions tested here were short-run, experimental interventions. In the diary-keeping study, for example, raters were asked to keep diaries for only 3 months. If the intervention were to last much longer, more raters might cease keeping diaries. In this regard, it is interesting to note that Flanagan and Burns (1955), in discussing their critical incident appraisal system at General Motors, reported that raters preferred to keep diaries on a daily rather than a weekly basis. Nonetheless, J. C. Flanagan (personal communication, June 1992) reported that there was general resistance on the part of supervisors to keeping diaries once the critical incident system was about to evolve from a pilot program to an ongoing program.

It is likewise possible that, when raters are asked to keep diaries over an extended period of time, they will lose interest and become less vigilant in their diary keeping. Thus, the success of the diary-keeping approach in the present study may not generalize well to programs that extend over longer periods of time. On the positive side, however, both theory and empirical findings suggest that there is value in keeping an incident diary, even if its use is sporadic or not continued. Evidence supporting this view was provided by DeNisi, Robbins, and Cafferty (1988). They found that even after raters who kept diaries had ceased to do so, there was still a substantial residual tendency to organize information in memory consistent with the format of the diary. These results are also consistent with Feldman's (1986) suggestion that cognitive structures, once formed, prime the encoding of subsequent information. Nonetheless, one cannot be sure how a prolonged structured recall program might affect raters' motivation to be diligent.

Related to this problem, we cannot be sure that the information raters included in their diaries actually reflected ratee performance. That is, the raters may have made up the information they included in their diaries just to provide us with information we requested. We see no incentive for raters to do this, and there was no indication that this was actually happening, but such a possibility reflects yet another area in which experimental control is lost in the move to field settings.

Next, there is the issue of the criteria used for evaluating these interventions. These studies did not involve measures of rating accuracy, and for some this may be seen as a serious shortcoming. We disagree. Earlier in the article, we alluded to some of the controversy over the best criterion measure to use in field research on performance appraisal. It is becoming increasingly clear that although rating accuracy reflects an important concern, it may not be the best measure from an organizational perspective. Clearly, it should not be the sole criterion. Instead, rater and ratee reactions may make as much if not more sense, because they have more to do with the perceived usefulness of any appraisal system. In fact, it is interesting to note that, in the present studies, effects found for rater reactions mirrored those for rating distribution indexes (although they were less powerful). Furthermore, these results paralleled those reported in earlier laboratory studies in which accuracy was used as the criterion measure. This raises the intriguing possibility that, at least under some circumstances, the same kinds of programs that lead to increased accuracy also lead to more discrimination both between and within rates and to improved rater reactions to the appraisal process. Future research should investigate the relationships among these criteria, but we have more to say about that later. In any case, these findings support the use of multiple criterion measures to help one better understand the effect of appraisal interventions.

Finally, we must consider the extent to which the data we collected actually reflect raters' cognitive processes. Although the variables considered here have typically been considered in both cognitive appraisal studies and social cognition studies, one might argue that they do not directly assess cognitive processes. ARC scores, for example, may simply reflect how information is recalled.
rather than how it is stored in memory. Nonetheless, the literature recognizes that ARC scores do allow one to understand how information is stored in memory. Furthermore, although we cannot be sure when raters reviewed their diaries, or even when they made their diary entries, one can never be sure of such things in field studies. We must, therefore, accept that raters behaved as they were instructed because the present results are consistent overall with predictions based on extensive literature from social cognition and with the notion that we accessed the cognitive processes of interest.

**Implications**

Given these potential limitations, there are still a number of important things we have learned from the present studies. First, raters seem to appreciate and benefit from interventions designed to impose structure on unstructured information. Past research suggests that most information acquired by raters to make appraisal decisions is unorganized (which makes it less accessible to raters) and that this is problematic for raters attempting to make appraisal decisions. The present results suggest that anything that helps raters to structure information results in more positive reactions and better recall of that information. In addition, these interventions enable raters to better discriminate both between and within raters.

Second, interventions designed to organize unstructured information by persons are both preferred and seem to be more effective in the field, as has been found to be the case in the laboratory. It would seem that organizing information by persons would allow raters to more easily form impressions about persons, which should make ratings easier in either setting. The present results suggest, however, that the relative advantage of organization by person may not be as great in field settings as in laboratory experiments. Perhaps, in the field, the fact that there is some cognitive organization is relatively more important than the basis for that organization, even if organization by person might be preferable. It should be noted here that this advantage is surely due, in part, to the fact that ratings are made of persons, making person-based categories the most useful. We surely would not suggest that person organization is best for all types of evaluations; we do suggest, however, that it is better for appraisal decisions.

Third, there can be convergence between findings from the laboratory and the field. Critics of laboratory research in performance appraisal have either focused on the differences between lab and field settings (cf. Bernardin & Villanova, 1986) or focused on those aspects of field settings that are missing from laboratory settings and assumed to play a major role in determining results. Yet, the present results substantially replicate those found in the lab, and the fact that the lab studies did consider accuracy actually enhances the credibility of our findings. As noted earlier, however, there were areas of differences. Effects were not as strong in the field, and differences attributable to organizational format were not as pronounced. Furthermore, in both studies, effects for between-rates discriminability were generally weaker than those for revelation or within-rate discriminability. These somewhat weaker effects, however, may well have been due to the nature of the interventions, which tended to focus on effective and ineffective behaviors exhibited by a given ratee. Such interventions may well produce stronger effects on within-ratee indexes, with some corresponding trade-off against between-ratee indexes. This, of course, is quite consistent with the trade-off alluded to earlier in Murphy's (1991) distinction between classification and behavioral accuracy. In any case, future studies might be better served by specifically examining sources of differences in effects between field and lab studies rather than assuming that such differences exist in the first place.

**Future Research Considerations**

There are several issues for future research. One important direction, we believe, concerns the use of reaction data, such as those used here, to evaluate appraisal systems. Anecdotal accounts of rater discomfort and failures to effectively implement an appraisal system are widespread, and these reactions data may actually point to a key issue underlying such reports. Raters who feel unskilled or uncomfortable in making appraisals (and the subsequent decisions that are often based on those appraisals) may choose to avoid doing them or taking them seriously. The present findings suggest that raters who use interventions such as diary keeping and structured recall believe their ratings to be more accurate, more defensible, and less surprising to ratees than do those who do not have access to these interventions. If such feelings and beliefs translate into more effort and care in performing appraisal-related duties, efforts to improve memory organization may go beyond short-term rating and recall effects and extend to those longer term motivational issues that are so elusive. This, in turn, suggests the possibility of a link between cognitive and motivational underpinnings of the appraisal process that remain unexplored despite repeated calls for conducting research that brings these two perspectives together (e.g., Murphy & Cleveland, 1991, pp. 167–168; Robbins & DeNisi, 1994). It seems clear that future research should concentrate on criterion measures that assess affective reactions such as these.
We also believe that more research is needed in the area of distributional indexes. In this we agree with calls for moratoriums on studies in which attempts are made to minimize error indexes in which these indexes are used as proxies for accuracy (cf. Balzer & Sulsky, 1992; Murphy, 1991). However, we suggest that this research needs to explore different directions. It is certainly not clear whether indexes such as these are the best measures for evaluating appraisal systems; again, however, they may provide useful information as part of an overall system of evaluation. Although we have been clear not to suggest that any of these indexes reflected rating errors, many of the criticisms leveled against error indexes can be applied to our measures as well. Nonetheless, we believe that distributional indexes provide information on the limits of appraisal systems. Consider what it would mean, for example, if a set of ratings did not demonstrate within-rater discriminability. In such a case, it would not be possible to provide ratings with meaningful feedback about their strengths and weaknesses. Likewise, if ratings did not exhibit between-rater discriminability, it would not be possible to distinguish between the best and worst raters, and those who simply do well enough. Of course, the fact that ratings do discriminate within and between raters does not guarantee that the distinctions inherent in the ratings are valid. These indexes simply suggest that if the differences are real, organizations will find the ratings more useful for feedback and decision making. In the present study, it would seem difficult to argue against the usefulness of interventions that made more critical incidents accessible to raters, who, when using that information, gave ratings that were less elevated and reflected greater variability both within and between raters.

Furthermore, these distributional indexes may also be relevant to rater reactions to ratings (see Dickinson, 1993, for an outstanding review of the research on determinants of rater reactions). Specifically, it would seem that rating distributions may be related to the acceptability of ratings. In the absence of information about accuracy, acceptability may well be the most critical criterion for evaluation from the organization's perspective, because raters may respond (i.e., change behavior to improve performance) only to ratings they perceive as fair and legitimate. Kavanagh and his associates (Kavanagh & Hedge, 1983; Kavanagh, Hedge, Ree, Earles, & DiBiasi, 1985) found that positive attitudes toward an appraisal system depended not only on whether or not a rater received a positive evaluation but on factors such as the ability of the system to distinguish among ratee proficiencies and provide useful feedback. An index of within-rater discriminability might, therefore, affect ratee acceptability. Furthermore, Hedge and Teachout (1992) suggested that rating acceptability was a function of how well ratings could discriminate across raters (between-raters discriminability) and instill raters with confidence in their ratings. These arguments are also related to concerns about the importance of perceptions of procedural justice in determining the perceived fairness of appraisal systems (cf. Folger, Konovsky, & Cropanzano, 1992). Thus, given the preceding discussion about the potential importance of ratee reactions, it may be too soon to completely discard indexes of rating distributions. Instead, the research agenda regarding these indexes might meaningfully be extended to examine their relationships with reactions and appraisal system acceptability.

In the end, we return to the notion that cognitive models of the appraisal process can yield practical as well as interesting information. The recent Ilgen et al. (1993) review suggested several concrete contributions to practice stemming from this approach, but the present results suggest that the focus on memory issues, typical of cognitive models, may also be useful. Furthermore, the disagreements over the relationship between ratings and memory (cf. Nathan & Lord, 1983) can be traced to the fact that this approach has made memory issues more salient in the appraisal area. Of course, it has been noted that cognitive approaches, as a rule, tend not to consider motivational issues (cf. Murphy & Cleveland, 1991). But some recent cognitive researchers have begun integrating motivational variables in their research (e.g., Varma, DeNisi, & Peters, 1993), and, here again, it is the emphasis of these cognitive models on information processing that, in part, made the need to consider motivational forces more obvious. We are not suggesting that cognitive approaches to performance appraisal are the answer to all of an organization's appraisal problems, but we are suggesting that this approach has advanced understanding of the appraisal process. We believe the present research adds something more to that understanding.

References


Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the task is on-line or memory-based. Psychological Review, 93, 256-268.


Kane, J. S., Bernardin, H. J., Villanova, P., & Peyrefitte, J.


Appendix

Reactions Items and Resulting Reactions Dimensions From Study 1 and Study 2

Study 1

Dimension 1: Clear Understanding of What to Attend To ($\alpha = .77$)
I know what performance indicators to attend to when making these ratings.
I know where my subordinates stand on all relevant performance indicators.
I have a well organized picture of how my subordinates contribute to the company.
My subordinates and I agree upon the duties and responsibilities for evaluations.

Dimension 2: Fair Appraisal Program ($\alpha = .71$)
Performance appraisal policies and responsibilities are adequately explained.
My subordinates know what they are being evaluated on.
Performance appraisal here at ___ is a fair process.
My subordinates would regard these evaluations as fair.

Dimension 3: Accurate and Defensible Ratings ($\alpha = .58$)
These performance ratings are accurate evaluations of my subordinates' performance.
I can easily recall specific examples from work upon which to base my evaluations.
I could easily defend these evaluations to a subordinate who might disagree with them.

Dimension 4: Ratings Would Not Be Surprising to Ratees ($\alpha = .59$)
My subordinates would not be surprised to receive these ratings.
I know enough about my subordinates' behavior to accurately rate their performance today.

Dimension 5: Keeping a Diary Helped Me in the Appraisal Process ($\alpha = .81$)
I tended to notice critical noteworthy behaviors.
I wrote down critical noteworthy behaviors in an incident diary.
I am better able to remember critical noteworthy behaviors.
I found it helpful to have an incident diary.
I will continue to keep an incident diary.

Study 2

Dimension 1: Clear Understanding of What to Attend To ($\alpha = .64$)
I know what performance indicators to attend to when making these ratings.
I have a well organized picture of how my subordinates contribute to the company.

Dimension 2: Fair and Accurate Appraisal Program ($\alpha = .83$)
The ratings I give are an accurate evaluation of my subordinates' performance.
The appraisals are fair reflections of my subordinates' performance.
I am confident that my subordinates' appraisals are fair.

Dimension 3: Defensible Ratings ($\alpha = .74$)
I can remember enough about my subordinates' behavior and results to accurately rate their performance.
I can cite specific instances at work to support my ratings.
I can easily recall specific work examples upon which to base my evaluations.
I could easily defend these evaluations to subordinates who might disagree with them.

Dimension 4: Ratings Would Not Be Surprising to Ratees ($\alpha = .56$)
My subordinates would not be surprised to receive these evaluations.
My subordinates would view these ratings as fair.

Received December 1, 1994
Revision received June 20, 1996
Accepted June 24, 1996