APPRAISAL PURPOSE

The Role of Appraisal Purpose: Effects of Purpose on Information Acquisition and Utilization

KEVIN J. WILLIAMS
Department of Psychology, University of South Carolina

ANGELO S. DE NISI AND ALLYN G. BLENCOE
College of Business Administration, University of South Carolina

AND

THOMAS P. CAFFERTY
Department of Psychology, University of South Carolina

Past research on the role of appraisal purpose in the appraisal decision-making process has concentrated on the motivational role of purpose. Research has found that raters are less willing to give poor ratings when appraisals are to be used for some purposes rather than others. The present paper describes two experiments which explore how appraisal purpose might affect rater cognitive activities as well. The first experiment investigated how appraisal purpose and outcomes affect how raters differentially utilize information to make appraisal decisions. Few differences were found. The second experiment investigated how raters differentially search for performance information to make appraisal decisions for different purposes and outcomes. Raters were found to search for more comparative information when they had to select one of several raters for some treatment. The results also indicated a discrepancy between how information is collected and how it is used. Implications for defining the role of purpose in the appraisal process, as well as for recent process approaches to performance appraisal, are discussed. © 1985 Academic Press, Inc.

Performance appraisal information is widely used as the basis for making a number of organizational decisions such as who should get promoted or how large a salary increase a worker deserves, as well as for providing feedback to workers. Since the consequences of giving a worker a poor (or good) rating vary as a function of the purposes for the appraisal, it is not surprising that a number of researchers have suggested that the purpose for an appraisal should be considered as part of any model of the appraisal process (DeCotiis & Petit, 1978; Wherry, 1952).

The projected role for appraisal purpose, and subsequently the research on appraisal purpose, has been rather limited, however. Research has focused mainly upon the motivational consequences of conducting appraisals for different purposes. Studies have shown that appraisals conducted for feedback or developmental purposes are less prone to rating bias (such as leniency) than are appraisals conducted for administrative decision-making purposes, such as determining which worker should be promoted (e.g., Meyer, Kay, & French, 1965; Zedeck & Cascio, 1982). Ratings are also dependent on the cooperation of the persons making the ratings and on the perceived use to which ratings are to be put (Rothe, 1949). Many researchers have also focused exclusively upon the consequences of ratings, moving further away from the effects of actual appraisal purpose. The resulting evidence suggests that raters dislike giving poor ratings in general (Fisher, 1979; McGregor, 1957), but that this aversion is heightened when ratings are to be made known to the ratees (Fusilier, 1980; Sharon & Bartlett, 1969).

Although this research has proven fruitful, its usefulness is limited. Simply determining that raters are likely to make different types of errors under different sets of circumstances is not as valuable as determining how raters make appraisal decisions for different purposes. Understanding the process by which raters reach appraisal decisions becomes more significant when one considers how essential performance appraisal is to the human resource system of any organization. Bernardin and Beatty (1984) have identified the many interdependent purposes for performance appraisals: appraisal may be used to improve the utilization of staff resources (identifying training needs, providing intermittent feedback, etc.) or as a basis for personnel actions (recognition of performance, promotions, merit raises, terminations, reassignments, etc.). A supervisor trying to identify the training needs of subordinates may process relevant information differently from a supervisor allocating merit raises. Indeed, rating subordinates’ deservedness of merit raises may be a different process from allocating merit raises.

Recent cognitive models of the appraisal process (DeNisi, Cafferty, & Meglino, 1984; Ilgen & Feldman, 1983) address this issue by viewing the rater as a decision maker. Accordingly, the purpose for which an appraisal is conducted may influence how information is used to make appraisal decisions. The purpose for an appraisal may cue raters to search for, weigh, and combine information in a certain way, with raters utilizing different types of information for different appraisals. This is not to suggest that appraisal purpose does not have a motivational function, but
simply that it may also have a cognitive function within the appraisal process.

Zedeck and Cascio (1982) have recently conducted one of the few studies that directly addresses this potentially important issue. These authors used policy-capturing procedures to determine whether raters weighed, combined, and integrated performance information about supermarket checkers differently when ratings were to be used for making decisions about merit raises, the need for development, and retention. The performance information (contained in essays about each worker) was simply rated on performance on various dimensions of the checker's job. Although this is the kind of information encountered by a rater, and although the results indicated that the dimensions receiving the greatest weight in each case seemed to be appropriate for the purpose of the rating, this approach to presenting performance information is theoretically unsatisfying. There will be different dimensions of performance associated with different jobs, and thus it would be difficult to formulate any theory or predictions about how raters will utilize information in any given situation. Unfortunately, since there is not presently available any theoretical basis for characterizing different types of performance information that will apply to all jobs, any such predictions are not now possible. We suggest such a theoretical basis, but first there is another dimension to appraisal purpose that we feel should be considered.

Studies such as that by Zedeck and Cascio (1982) have considered purpose to vary on only one dimension—the purpose for which a decision is made. In their study, Zedeck and Cascio required ratings of recommendations for a merit raise (anchored with no raise recommended, average raise recommended, and highest raise recommended), need for development, and recommendations for retention. We call these deservedness decisions, because each requires the rater to decide how deserving each rater is of each type of action. But, this is not the only type of decision made on the basis of appraisal information. A supervisor might be asked to rate how "promotable" is each rater, or the rater might instead have to promote one rater; one could determine the extent to which each rater needs development, or one might have to send one rater to a remedial training program; one could rate how deserving each rater is of a fixed pay increase, or one might have to allocate a certain salary increase for a rater. We refer to these latter types of decisions as nomination or designation decisions, because each requires the rater to select one (or more) of the ratees for a particular treatment, or to designate a specific outcome. Designation and deservedness appraisal decisions may be treated differently in the appraisal process. First, different motivational forces may be generated upon the rater. The act of designation, for example, has potentially more personal consequences for the rater. Having to promote one person, as opposed to rating deservedness, heightens both organizational and personal risk involved with the decision. These perceived consequences may affect subsequent ratings. It is also possible that the distinction in the type of decision to be made leads to different information-processing strategies. Specifically, different types of performance information may be attended to and different representations of ratee behavior established.

It should be useful, thus, for research to (1) consider the rater as a decision maker who collects and utilizes information about performance and (2) investigate whether different appraisal purposes result in different, stable, predictable rater cognitive activities. Model building concerned with studying this cognitive function, however, requires some theoretical basis for categorizing and describing, a priori, the different "types" of information available. Unfortunately, as noted earlier, no adequate theoretical basis presently exists within the field of performance appraisal; the simple labels of good and poor are probably of limited usefulness in any larger context.

Kelley's (1971) covariation principles in attribution theory may provide a basis for establishing evaluation standards. Kelley proposed a cognitive mechanism whereby observers gather and organize information in an attempt to attribute observed behavior to the person exhibiting the behavior, the characteristics of the tasks, or the circumstances surrounding the behavior. Kelley suggested that observers attempt to assign causality for events by attending to

1 how the target person behaves in the same setting on other occasions (consistency information),
2 how the target person behaves in other similar settings (distinctiveness information), and
3 how other people behave in the same setting (consensus information).

While performance appraisals should not be treated as special cases of attributions—the determinations of causality, for example, may not be pertinent to performance appraisal—the two decision-making processes share certain elements. First, the social context for both attribution and appraisal tasks are characterized by a high degree of complexity. The observer often receives, and must integrate, information about multiple targets from multiple sources. Second, the decision to be made in both cases is based on imperfect, or incomplete, data. What is seen or read is only a subset of possible behaviors. As Bazerman and Atkin (1982) have previously outlined, this partial data is also characterized by inconsisten-
cies and noise. Response variation is intrinsic in both situations; a worker's performance level varies as does, say, a person's reasons for laughing (a common question used in attribution research is to figure out why a person is laughing at a comedian). Consistency, distinctiveness, and consensus information provide the attribute with an effective strategy for using incomplete data; similar covariation information should help reduce the uncertainty surrounding performance levels. Knowing (1) how the ratee performs on the same task on other occasions, (2) how the ratee performs on other tasks within his or her job scope, and (3) how all workers perform on the given task should enable the rater to effectively reach performance decisions. We hesitate to use Kelley's terms of consistency, distinctiveness, and consensus information to define these three dimensions of performance data because of the semantic confusion which may arise given the context in which the evaluation is made. For example, when comparing one or more persons on one particular performance dimension, distinctiveness, rather than consensus, information may best describe the relevant information. For the sake of uniformity, however, we will define the proposed types of performance information in the same manner as attribution researchers. It should be noted that Bazerman and Atkin (1982) have previously used attribution theory as a basis for defining subsets of performance information. They also suggested, as we have, that the specific dimension of information used by the rater is a function of the type of decision to be made.

Decisions made by subjects in attribution studies suggest that the type of decision does influence information processing. When making attribution decisions, observers do differentially weight the different types of information they receive (e.g., McArthur, 1972). There is also evidence that different types of information are more important for making different types of attribution decisions. Observers attempting to make attributions to a person rather than a task or situation rely primarily upon distinctiveness and consistency information while observers attempting to make situational attributions rely upon consensus information (Garland, Hardy, & Stephensen, 1975; Hansen, 1980; Major, 1980). Nonetheless, across a wide variety of situations, consensus information seems to be given the least weight (cf. McArthur, 1972).

The decisions reached by attribution studies, however, do not neatly correspond to those made for different appraisal purposes. Rather than deciding if the comedian is funny or if John is easily entertained, raters conducting performance appraisals must decide if John is the worker to be promoted, or if he deserves a 6% pay increase. Although we would argue that the rater in an appraisal setting is always trying to make a person attribution, different appraisal decisions may require different covariation information, much like person versus situational attribution decisions. Appraisals requiring designation decisions (e.g., what one person should be promoted, or how much of a raise should this person get) may require more comparative judgments, and thus necessitate the use of consensus information, than deservedness decisions (e.g., how promotable is this person, or how deserving of a raise is she/he).

Using this framework, it may be possible to test the effects of appraisal purpose and decision outcomes on both passive, information integrative and active, information acquisition processes. If the nature of the appraisal decision affects the cognitive processes involved in information integration, the way in which the available information is combined should vary with appraisal purpose and outcome. If only the rater's motivation is affected, the use of covariation data should stay constant, although elevation of ratings might vary across different types of decisions. When performance information is not readily available, however, appraisal outcomes may also affect active search processes. Raters in certain situations may prefer specific covariation information and actively search for it (Crocker, 1981). If this were true, conducting different performance appraisals without access to complete information would lead to gathering different kinds of information, which, in turn, may result in different ratings.

The present research investigated the effects of performance appraisal purpose and outcome on both rater's information integration and acquisition processes. Experiment 1 was designed to see if subjects do in fact use information differently to yield different ratings for different appraisal purposes. Subjects were given covariation data on ratees and instructed to evaluate the ratees for one of three purposes, with one of two possible outcomes. The ratings were analyzed with respect to differences in the use of covariation information. Experiment 2 was designed to test the hypothesis that raters will search for different performance information depending on the nature of the performance appraisal. Subjects were instructed to conduct appraisals for one of the same six experimental conditions used in Experiment 1. No presearch information regarding the ratees was given to the subjects. Preferences for the types of information selected were analyzed within each experimental condition.

### EXPERIMENT 1

Based on attribution theory research, it seems that concentrating on the performance of individuals will facilitate a prevailing tendency to use distinctiveness and consistency information for person attribution decisions; concentrating on performance relative to others should increase
the saliency of often disregarded consensus data. Thus, the following hypotheses were tested:

1. Subjects will utilize consensus, consistency, and distinctiveness information similar to the way observers have been found to utilize covariation information for attribution decisions. Variations in consistency and distinctiveness information should be reflected in performance ratings; variation in consensus information should have negligible effects on ratings (cf. McArthur, 1972). Specifically, good performance will not be rated as highly when it is inconsistent and distinctive as when it is consistent and indistinctive. Poor performance will be rated lower when it is consistent and indistinctive than when it is inconsistent and distinctive. Low or high consensus information should not affect either set of ratings.

2. Raters will utilize both consistency and distinctiveness information for appraisal decisions regardless of the appraisal’s purpose or outcome.

3. Raters will utilize consensus information more for appraisal decisions made for designation outcomes (i.e., salary allotments or single promotions and referrals) than for deservedness outcomes. This hypothesis is largely exploratory, since no direct empirical evidence has addressed this issue. It is assumed that the comparative nature of designation type decisions will increase the salience of often ignored normative data.

Method

Subjects. One hundred twenty undergraduates (78 males and 42 females) enrolled in introductory psychology and business administration courses at the University of South Carolina participated in the experiment in partial fulfillment of course requirements. (Preliminary tests indicated no differences by sex or course, nor any interactions involving these variables.) Twenty subjects were randomly assigned to each of the six experimental groups in the 3 (Purpose) × 2 (Outcome) design. In all six conditions, subjects were told they would be evaluating the performance of eight hypothetical administrative assistants from different departments in the same company.

Materials. Performance information was contained in vignettes which described the performances of target workers as good or poor. These vignettes had been pilot tested with a separate sample of undergraduates and results indicated that performance in each vignette was interpreted in the appropriate direction. The design used to present the information to subjects was similar to that used by McArthur (1972). For each employee, target performance on a certain task (budget preparation) was followed by concrete and specific consistency (how the employee did on the previous month’s budget), consensus (how three co-workers in her department performed on the budget task), and distinctiveness information (how the employee did on a different task—answering customer inquiries). 2

Two sets of performance information were constructed. The first set, used for the promotion and pay increase purpose conditions, described the performance of all eight ratees on the budget task was good. For example, subjects read that “Betty prepared a financial report on office expenses. A supervisor checked the report and all of her figures were found to be accurate.” Each ratee represented one cell in a 2 (high vs low consistency) × 2 (high vs low consensus) × 2 (high vs low distinctiveness) within-subjects covariation matrix. High consistency was indicated by good performance on the previous month’s budget while low consistency was indicated by poor performance the previous month. High consensus information was indicated by good performance by the three co-workers on the budget task; low consensus was indicated by poor co-worker performance. High distinctiveness was indicated by poor performance on the customer inquiry task; low distinctiveness was indicated by good performance on that task. The second information set was used for the remedial training purpose conditions and described the performance of all eight ratees on the budget preparation task as poor. For example, “Cathy prepared next month’s budget for office expenses. A supervisor checked the report and a number of errors were found.” Indications of high and low consistency, consensus, and distinctiveness information were reversed accordingly.

Procedure. Subjects reported to the experiment and received initial instructions in groups (n = 8–15) but worked on the experimental task individually at separate tables. Brief descriptions of the tasks involved in the job of an administrative assistant initially were given to each subject. Subjects were told to conduct a performance appraisal for one of three purposes: (1) salary increases, (2) promotion recommendation, or (3) remedial training referrals. The decision to be reached focused on one of two outcome dimensions: (1) individual recognition or deservedness, or (2) designation or specific indication. Subjects in the deservedness condition were instructed to indicate how deserving each employee was of a promotion, a remedial training referral, or a 6% raise. Subjects in the designation condition were instructed to select one employee for promotion or remedial training, or determine how large a pay raise each employee should receive. Subjects were then given the target performance and covariation information and were allowed 20 min to study the conditions.

A pilot study was conducted to investigate possible order effects in the presentation of the covariation information (see Ruble & Feldman, 1975; Sherman & Titus, 1982). There were no significant differences in appraisal ratings for the various presentation orders tested. The consistency–consensus–distinctiveness order was chosen because it offered the most fluent method of presentation.
information. After 20 min the information was taken away, and all subjects were subsequently asked to rate all the rates on 7-point Likert scales. The ratings were then collected, and subjects were debriefed and dismissed.

Dependent measures. Since each worker represented one cell in the $2 \times 2 \times 2$ configuration of covariation information, raters were required to make judgments for each rater. For subjects in the deservedness outcome condition, those assigned to the promotion condition were asked, "How deserving is each employee of being promoted?" Each employee was rated on a 7-point Likert type scale ranging from very undeserving to very deserving. Those assigned to the salary increase condition responded to the question, "How deserving is each employee of a 6% pay raise?" on the same scale. Those assigned to the remedial training condition were asked, "How strongly do you feel each employee should be sent to a remedial training program?" and responded on a scale anchored at definitely should send and definitely should not send.

The rating scales used for the designation conditions had to reflect the intended comparative nature of the appraisal task while still providing a performance judgment for each worker. Thus, rather than just selecting one person for the treatment, subjects making designation decisions for promotion and remedial training purposes indicated how they felt about each employee being the one person to be promoted or referred on a 7-point Likert type scale anchored at definitely not the one to be promoted (referred) and definitely the one to be promoted (referred). Subjects assigned to the salary increase purpose responded to the question, "How much of a salary increase would you give each employee?" on a 7-point scale ranging from 0 to 12% increase.

**Results**

The mean ratings for the eight configurations of covariation information in each experimental condition are presented in Table 1. A 2 (Consistency) $\times$ 2 (Consensus) $\times$ 2 (Distinctiveness) $\times$ 3 (Purpose) $\times$ 2 (Outcome) mixed-factor analysis of variance (ANOVA) was conducted on the mean ratings, with appraisal purpose and outcome as between-subjects factors. The results of this analysis are contained in Table 2. Significant main effects were found for appraisal outcome, $F(1,114) = 16.70, p < .001$, and appraisal purpose, $F(2,114) = 4.43, p < .01$, but no Purpose $\times$ Outcome interaction resulted, $F(1,114) = 1.86$, n.s. Subjects in the deservedness outcome condition made overall higher ratings ($M = 4.43$) than those in the designation condition ($M = 3.91$). Post hoc Scheffe tests for the significant purpose main effect revealed that the ratings given in the remedial training purpose condition ($M = 4.44$, reverse scored) differed from those given in the salary increase ($M = 4.01$) and promotion

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>MEAN PERFORMANCE RATINGS AND STANDARD DEVIATIONS AS A FUNCTION OF COVARIATION INFORMATION, APPRAISAL PURPOSE, AND OUTCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Purpose</strong></td>
<td><strong>Consensus</strong></td>
</tr>
<tr>
<td><strong>Outcome</strong></td>
<td><strong>Low Consist.</strong></td>
</tr>
<tr>
<td>Promotion</td>
<td>6.15 (1.27)</td>
</tr>
<tr>
<td>Salary</td>
<td>6.40 (1.16)</td>
</tr>
<tr>
<td>Training</td>
<td>6.70 (1.16)</td>
</tr>
</tbody>
</table>

Note: Consistency = consensus information; Distinctiveness = distinctiveness information; Consist. = consistency information. Standard deviations are in parentheses.
(M = 4.06) conditions, F(2, 114) = 3.82, p < .05, while the latter two did not differ from each other.

Significant main effects were also found for consistency, F(1, 114) = 493.07, p < .001, and distinctiveness information, F(1, 114) = 464.82, p < .001. Consistent performance (M = 4.89) was rated higher than inconsistent performance (M = 3.45) and low distinctiveness was rated higher than high distinctiveness (M's = 4.87 and 3.47, respectively). Although low consensus information resulted in higher ratings (M = 4.29) than did high consensus (M = 4.05), F(1, 114) = 6.54, p < .05, it accounted for less than 1% of within-subjects variance and the F value failed to reach significance when adjustments for higher order factorial designs were made. No other effects were meaningful.

**Discussion**

These results suggest that raters are sensitive to covariation information in a similar manner to observers making attribution decisions. Varying the levels of consistency and distinctiveness information affected performance ratings; varying consensus data did not influence ratings. In addition, similar effects existed for all appraisal outcomes and purposes. Individual levels of performance across time and tasks seem to be considered optimal for evaluation decisions. No support, however, was found for the exploratory hypothesis that appraisal tasks requiring comparative judgments (i.e., designation outcomes) would lead to greater use of consensus information than other appraisal decisions. The pattern of information utilization was fairly consistent across all conditions. This seems puzzling since it suggests that raters were willing to make comparative judgments without really utilizing comparative data. It is possible, however, that raters (even naive student raters) have sufficiently internalized standards of performance against which any ratee's performance can be compared. Thus, except for situations where no such standards exist (and we can only speculate on what those situations are), raters may not feel the need to use comparative information, even to make comparative judgments. Such a phenomenon would help explain the virtually universal underutilization of consensus information.

The significant main effects for appraisal purpose and outcome indicate that evaluation standards may change with perceptions of the rating's consequences. Appraisals focusing on deficiencies (i.e., remedial training referrals) displayed slightly higher ratings than those used for other purposes. Any interpretation of this effect would be highly speculative given the extremely low effect size associated with this finding. The outcome main effect was due to ratings of deservedness being generally higher than ratings of designation. This suggests that situations where all ratees are capable of receiving benefits may yield higher global ratings than

### TABLE 2

<table>
<thead>
<tr>
<th>Source</th>
<th>MS</th>
<th>df</th>
<th>F</th>
<th>Est. $\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Purpose (A)</td>
<td></td>
<td></td>
<td></td>
<td>.109</td>
</tr>
<tr>
<td>Outcome (B)</td>
<td></td>
<td></td>
<td></td>
<td>.048</td>
</tr>
<tr>
<td>A × B</td>
<td></td>
<td></td>
<td></td>
<td>.012</td>
</tr>
<tr>
<td>Subject (A, B)</td>
<td></td>
<td>114</td>
<td>17.54</td>
<td></td>
</tr>
<tr>
<td>Within subjects</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus (C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distinctiveness (D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C × D × E</td>
<td></td>
<td>114</td>
<td>158.93***</td>
<td>1.38</td>
</tr>
<tr>
<td>Subject × D</td>
<td></td>
<td>114</td>
<td>171.37***</td>
<td>5.34*</td>
</tr>
<tr>
<td>Subject × E</td>
<td></td>
<td>114</td>
<td>6.54*</td>
<td></td>
</tr>
<tr>
<td>Subject × C × E</td>
<td></td>
<td>114</td>
<td>3.96*</td>
<td></td>
</tr>
<tr>
<td>Subject × C × D × E</td>
<td></td>
<td>114</td>
<td>244.40</td>
<td></td>
</tr>
</tbody>
</table>

Note. Estimated $\omega^2$ values for the between-subjects effects were computed relative to total between-subjects variation, $\eta_p^2 < .05$, $p < .05$. **$p < .001$**.
those where equal distribution of benefits is impossible. Raters may have a tendency toward leniency in that they do not want to deny rates any possible benefits they may be eligible for. On the other hand, situations where all individuals cannot share in outcomes may decrease ratings for two reasons. First, raters may undergo a more critical review of ratee performance. Concern for selecting the most appropriate individual may initiate this extensive review and reduce leniency bias. This would imply, however, that the rater would attend more to consensus data in making his or her decision, and the results from Experiment 1 do not support this implication. A second explanation for lower ratings in situations with unequal recognition may be that raters want to make sure that the one individual singled out for the outcome is sufficiently distinct from the other rates. Designation outcomes may impose justification demands upon the rater and he or she may want to ensure that the decision made is clearly perceptible to all. Obviously, this needs more investigative attention, but, based upon these results, it would seem that the way in which outcomes are to be distributed among rates is another important dimension of appraisal purpose that should be considered in future research.

The data from Experiment 1 suggest that raters use certain information while making appraisal decisions. This knowledge, in and of itself, may be limited in its implications for the appraisal process. Before analyzing a subordinate's performance, the rater must acquire relevant information. If the rater does not have sufficient information to reach an evaluative decision, he or she may first determine what requisite information is missing and then attempt to collect it (Crocker, 1981). The question that needs addressing is what information is preferred by raters when complete information is not available? Do raters search for distinctiveness and consistency information, and do they show a preference for different types of information for different appraisal purposes, an effect that is absent when information is available?

**EXPERIMENT 2**

The results of Experiment 1 suggest that appraisal purpose and outcome have a limited effect on how raters utilize information. Purpose and outcome may, however, have additional effects when raters do not possess sufficient performance data. One implication of the recent social-cognitive approaches to performance appraisal is that raters might make decisions about the kinds of information they need and then commence to collect that information. Hansen's (1980) propositions about the economical decision maker are relevant to this issue. Hansen (1980) suggested that a decision maker might not need all types of covariation information to make his or her decision; an exhaustive search for consisten-

cy, consensus, and distinctiveness information may often result in redundant information. If instead the rater could determine the ratee's job ability by learning how the ratee performs on related tasks within the job scope (i.e., gather distinctiveness information), he or she could infer a great deal about the ratee's overall performance without the need for temporal consistency information. Hansen also suggested that raters could determine the difficulty of the task by learning how other workers perform on the same task (i.e., gather consensus information) and then infer a ratee's overall performance from his or her performance on a difficult or easy task. This strategy, however, involves establishing facilitative or inhibitive forces (Kelley, 1971) and is therefore less efficient than a simple distinctiveness search. Thus, Experiment 2 tested the following hypothesis:

1. Raters will seek primarily distinctiveness information for making appraisal decisions regardless of purpose or outcome. This information preference, however, may be tempered by the demand of the appraisal context. Specifically, designation decisions should require more comparative judgments, and therefore consensus information should be sought more.

The following hypotheses were, therefore, also tested:

2. Requiring raters to make designation decisions will lead to greater search for consensus information than merit decisions.

3. Requiring raters to make deservedness decisions will lead to greater search for consistency information than designation decisions.

**Method**

**Subjects.** One hundred twenty undergraduates (53 males and 42 females) enrolled in introductory psychology courses at the University of South Carolina participated in the experiment as partial fulfillment of a course requirement. Since no effects for sex nor any interactions involving sex were found, all data were analyzed for the entire sample.

**Overview.** Subjects were randomly assigned to one of the six experimental conditions in the same 3 (Appraisal Purpose) × 2 (Outcome) design used in Experiment 1. Twenty subjects were assigned to each condition. Subjects were given their specific appraisal purpose and outcome and then allowed, unlimited access to information describing four hypothetical female administrative assistants performing four separate tasks. After completing their search, subjects in the merit outcome condition made ratings for each employee on 7-point Likert type scales; those in the designation condition indicated their specific selections. The main dependent variable of interest was the search strategy used by each subject to collect performance information.

**Apparatus.** The information search phase of the experiment was con-
ducted on an Apple II Plus microcomputer terminal. Information was presented to the subject on a CRT screen and the subject's requests for information were entered on the attached typewriter keyboard. The performance information about the four fictitious employees was contained in a sequential textfile dimensioned into a $4 \times 4 \times 4$ array. The 64 items of information portrayed four employees (Anne, Betty, Cathy, Diane) performing four different tasks (budget preparation, memo typing, customer service assistance, and file clerk supervision) on four separate occasions. For each employee there were two satisfactory and two unsatisfactory performances on each task. The order of these incidents was alternated over the four employees, thus producing a different sequence of good and poor performance for each employee.

Procedure. Each subject was given a description of the experiment and the specific tasks involved. They were told the experimenters were interested in the strategies used by raters to gather information while making performance appraisal decisions. Subjects were briefed about the nature of the job of an administrative assistant and the tasks involved. The specific appraisal purposes were randomly assigned to subjects prior to the search phase of the experiment. Subjects were run individually in one of four rooms each of which had a computer terminal.

Subjects were asked to take a seat in front of the computer console and were given operating instructions. They activated the program by striking one of the typewriter keys and each was presented with the same following initial item of information:

Anne. Budget. Anne prepared next month's budget for departmental expenses. When it was checked by a supervisor, no errors were detected.

Following this, and every performance, subjects made their requests for information. Requests were made from one of four choices: (1) the same person performing the same task, (2) the same person performing a different task, (3) a different person performing the same task, or (4) a different person performing a different task. Consistent with traditional attribution research, the first three choices corresponded to consistency, distinctiveness, and consensus information requests, respectively, while the last choice was considered a request for noncovariation information. The computer recorded the subjects' requests in sequential order. A counter was set up by the computer program that moved through the three-dimensional array locating the information requested. If the requested piece of information had already been seen, the counter was incremented until a novel piece of information was encountered, thus guaranteeing no duplicate presentations. For example, if Betty's performance on typing memos was presented, and the subject asked to see another person doing the same task, the counter would go to Cathy's

first performance on typing memos. If the subject had already seen this, the counter would go to Cathy's second performance on that task, and so on. Prior to receiving the fourth piece of information for any one worker on any one task, a message was presented on the CRT screen to the subject indicating that this was the last time he or she could ask for that type of information. Any time a subject requested to see one person's performance on one task more than four times, a message appeared on the screen telling them to make another selection. The search phase of the experiment could be terminated in one of two ways: (1) if the subject had seen all 64 items, i.e., the counter could not be incremented further, the program terminated itself, or (2) the subject terminated the search whenever he or she felt enough information had been obtained to reach their appraisal decision(s). Subjects were left alone in the room and allowed to work at their own rate. They were told that there was no limit on the time it took to complete their search.

After completing their search, subjects were administered a rating scale corresponding to the condition they were assigned. These scales were identical to those used in Experiment 1, with the exception that those in the designation outcome, promotion and referral purpose conditions, selected one person to promote or one person in need of training. These ratings were not the main dependent variables in Experiment 2 (the focus was upon search strategies) and were included mainly for experimental task closure for the subjects. Unfortunately, these measures could not be subjected to the same analysis as the measures in Experiment 1 because each subject's ratings were based on different amounts and types of information. After making their ratings, subjects were debriefed and the experiment concluded.

Results

The mean number of requests for information made by raters, collapsed across purpose and outcome, was 38.43 (out of a possible 64). A 3 (Purpose) $\times$ 2 (Outcome) ANOVA was conducted on the mean number of requests per condition. There was no significant main effect for appraisal purpose, $F < 1$, or for appraisal outcome, $F < 1$. A significant Outcome $\times$ Purpose interaction was found, $F(2,114) = 6.67, p < .01$. A simple effects test revealed that the total number of requests made for salary increase purposes was greater for the designation outcome (salary allocation) than for the deservedness outcome ($M's = 43.5$ and $30.95$, respectively). $F(1,38) = 5.89, p < .05$.

The types of information sought by a rater were calculated as percentages of the total number of choices made by a rater, to control for differences in the total number of choices. Overall, distinctiveness information was sought 46.7% of the time, consensus information was sought
The results of Experiment 2 support the first hypothesis: Naive raters prefer distinctiveness information when systematically collecting performance data.
performance information about ratees. This preference may be attributed to the principle of cognitive economy. Ideally, the rater's decision would be based on consistency, consensus, and distinctiveness information. Since the rater possesses none of this information initially, he or she must actively search for it. To search for complete covariation information on each individual would strain one's cognitive processing abilities. The most parsimonious strategy, therefore, is to seek out distinctiveness information because it conveys information about the targets' propensity to perform at a certain level across a variety of tasks. Although the rater may be sacrificing temporal consistency data, a measure of each target's performance stability is obtained with limited cognitive effort. In situations similar to those constructed here—where the component tasks are not similar and performance across the tasks does not reflect the same abilities—distinctiveness information may be especially informative.

The Purpose × Outcome interaction, however, indicates that the search for performance information is sensitive to the demands of the situation. In partial support of the second hypothesis, the increased number of consensus requests evident for salary increase and remedial training purposes with designation outcomes may result from the greater need for knowledge of normative data in such situations. When a rater must not only determine if a worker deserves a raise, but also designate a certain amount in relation to other employees, distinctiveness information by itself is no longer the most parsimonious strategy. Normative data are required to adjust initial ability estimates relative to co-worker performance. Distinctiveness augmented by consensus data, or as is evident for salary allocation, consensus data alone may now be the preferred information.

There is no supporting evidence for the third hypothesis. No strong preference for consistency information exists within any condition. While consistency requests are greater in the merit condition than in the designation condition, this effect is relatively weak. This apparent disregard for consistency information has implications for recent discussions advocating the value of performance variability for the accurate assessment of ratee performance (Kane & Lawler, 1979; Latham & Wexley, 1977). Kane and Lawler (1979) argue that focusing on the average level of performance exhibited by a worker ignores motivational and situational antecedents to performance. They suggest, instead, to measure the distribution of the rates at which ratees exhibit various levels of performance goodness. Access to the variability of a worker's performance over time and settings is necessary for such an approach. Likewise, although reflecting different theoretical assumptions than Kane and Lawler, behavior observation scales (e.g., Latham & Wexley, 1977) require raters to report the frequency with which individuals engage in critical incidents. The present results suggest that the value of behavioral consistency may not be immediately salient to raters.

There may be a number of reasons why consistency information was not sought by raters in this experiment, all of which should be considered by performance appraisal models emphasizing performance variability. First, the immediacy of the decision to be reached may have facilitated the tendency for raters to concentrate on typical or differential levels of performance. Second, no assumptions were made regarding the importance of the separate tasks. Subjects may have weighted performance equally across tasks. If this were so, distinctiveness information would have supplied sufficient information about performance variability. Finally, effort-reducing heuristics may have been employed such that initial observations of a person's performance on a task were taken as indicative of average performance by that person on that task. This may be of special concern when raters are tracking good or poor behavior. For example, raters making remedial training referrals may search for nega-
tive instances of performance. When good performance is encountered, the rater may ignore further instances of that worker performing that task and continue to search for episodes of poor performance elsewhere. Only when poor performance is identified will consistency of performance receive attention. Even then, raters may eschew temporal variability for variability across job dimensions. The same may occur for raters seeking to promote someone; they may forgo instances of poor performance, and the variability surrounding such performances, in favor of successful performances. This suggests a number of actions required to ensure the use of performance consistency in ratings. Raters should be taught how to continuously monitor performance variability, and how to easily recall it, so that they will have ready access to this information when decisions are periodically required. As Experiment 1 demonstrates, raters will use consistency information when it is available to them. The different job dimensions for which consistency of performance is important should also be made clear to observers and any distinctions between temporal, intradimensional variability and interdimensional variability made explicit. Finally, steps should be taken to make sure raters understand the usefulness of consistency information for all levels of performance.

The above discussion also suggests that a closer comparison between searches conducted for positive consequences and searches conducted for negative consequences may be useful. While the types of information requested did not vary across purpose, the distribution of requests over workers and tasks may have. It may be that search for information about any one individual is initiated and terminated by different cues.

**GENERAL DISCUSSION**

The results of these two experiments offer information-processing explanations for rater search strategies under varying appraisal conditions. First, the purpose and outcome of an appraisal decision may serve a cognitive function in addition to the motivational function usually proposed. Raters making appraisal decisions for different reasons seem to differentially search for, if not use, performance information. Although information integration does not appear to differ substantially for different types of decisions, raters requiring additional information appear to do what Crocker (1981) suggests observers do when making covariation decisions. They first decide what kinds of information are needed, and then search for it. Thus, raters may not only be differentially motivated by having to make appraisal decisions for different purposes, they may actually require different information to reach those decisions. This has potentially serious implications for issues of accuracy in appraisals, since it may be necessary to consider what purpose appraisal decisions will be used for and then to ensure that the rater has access to the necessary information. Along these lines, the present research may help define the role of purpose in the performance appraisal process. Traditionally, the effects of an appraisal's purpose have been investigated along one dimension—the use to which the ratings are to be put. The designation/deservedness outcome dimension investigated in the present study did show some important effects, especially in terms of search strategies. Multidimensional conceptions of appraisal purpose should be considered in future research on cognitive and motivational processes involved in performance appraisal.

Another major implication of these results relates to defining the dimensions of performance data used in the appraisal process. Covariation information, of the type used in making attribution decisions, appeared to provide a useful basis for understanding the appraisal process. Raters were found both to use covariation information in a predictable manner and to search for the kinds of information that recent research would predict when relevant information was not available.

It is interesting to note, however, that a lack of correspondence existed between the integration of available information and information search. No strong preference for a consistency search strategy was found, yet consistency information was heavily weighted when subjects were presented with it. Also, consensus information was largely ignored when presented along with consistency and distinctiveness information, but was requested for certain appraisal decisions. A cognitive economy explanation has been offered for this discrepancy. When the rater does not already possess performance information, he or she must search for it. To search for complete covariation information on each individual would strain cognitive processing abilities. The rater must therefore decide when the costs of obtaining additional information outweigh the benefits provided by such information (Stigler, 1961). A parsimonious strategy would be to seek out distinctiveness information since it conveys information about the targets' propensity to perform at a certain level across a variety of similar tasks. Such concerns of parsimony are not relevant when the rater is given free access to complete covariation information. In such situations, no penalty is incurred for using all of the information. This explanation implies that conditions where the benefits of additional information are obvious (e.g., when extensive information regarding variability of performance is desired) should lead to a more complete search for information. Likewise, situations where normative data are needed should lead to the use of consensus search strategies because such strategies would be most cognitively simple. Future research should examine whether raters adapt their search strategies to reflect the demands of the situation or whether they display one preferred strategy across different situations.
The discrepancy between the way in which raters integrated information available to them and what kinds of information raters sought when they required information also suggests there may be a critical distinction between the artificially constrained decision-making processes usually studied in lab settings and more realistic processes (cf. Hogarth, 1981). Laboratory research on decision making typically concentrates on the logical analyses used by raters while often overlooking selective processes that operate in more realistic settings. Presenting complete information to raters may lead to logical analyses atypical of realistic decision-making tasks. Selective processes, such as information salience (Taylor & Fiske, 1978) and primary and recency effects (DeNisi & Stevens, 1981), as well as unintentional, or incidental, information processing, must also be considered.

It is noteworthy that raters in the present study altered information search strategies for an experimental task where the perceived consequences of decisions made for different purposes were not salient. That is, the cost of making an error, even relative to who was to be promoted, was nonexistent. As such, this study attempts to capture some of the basic underlying cognitive processes involved in reaching appraisal decisions. Obviously, there are interactions between raters and ratees that we have not addressed. The consequences associated with different types of decisions are clearly more prominent in organizational settings. One can assume that the effects obtained in the present study would be more pronounced in such settings. It is likely that raters who have had continuous interchanges with employees have developed implicit theories for each individual. Not only may such conceptualizations color performance ratings, they may also bias any further search for information in such a way as to support the existing theory. Additionally, experienced raters may have well-developed habits for conducting performance appraisals. They may have their own convention for assembling and integrating performance information and may not be susceptible to contextual factors. Attempts to induce changes in the way information is processed may either be met with resistance or, if initially successful, their effects may not be maintained over an extended period of time (cf. Ivancevich, 1979). Clearly, research on cognitive processes in appraisal decision must move to field settings. Only then can we fully assess interactions between motivational and cognitive effects of decision purpose and perhaps determine how these interactions relate to the accuracy of appraisal decisions.

The approach here is exploratory, yet it has applicability for a potentially wide range of current performance appraisal issues. Landy and Farr’s (1980) paper initiated a shift away from summation concerns in performance appraisal toward a more process-oriented approach. Before a rater can integrate and analyze performance data, those data must be acquired. Accuracy becomes dependent upon attention to relevant behaviors, the organization of information in memory, and the accessibility of the information in memory. Feldman (1983) has effectively pointed out that training raters to recognize relevant behaviors should improve accuracy. We suggest that such training be extended one step further—raters should receive instructions on how to search for relevant behavior when it is not present. The relevancy of behavior will, in turn, depend on the purpose and outcome dimensions of the appraisal. Obviously, the development of any such training procedures must attend to the search biases inherent in person perception processes (see Snyder & Gangstad, 1982).

Suggestions for new avenues of research also emerge from this discussion. Differences in information search strategies have been demonstrated. Closer attention should now be paid to the outcomes of these strategies. What differences in the level of information abstraction, encoding, organization in memory, retrieval, and/or affective judgments result from various patterns in which information is acquired. Different search strategies may facilitate different organization of information in memory and result in differences in the nature of the recalled information (e.g., global impressions versus performance detail). The present study was not able to address these issues because the observed performance information was not controlled by the experimenters. Research that varies information acquisition patterns and controls observed performance is thus encouraged.

REFERENCES
applied to real organizations: Narrowing the gap. Symposium presented at the 91st meeting of the American Psychological Association, Anaheim, CA.


RECEIVED: December 5, 1983