Journal of Accounting Literature
Vol. 14, 1995, pp. 107–139

A REVIEW AND SYNTHESIS OF RESEARCH IN PERFORMANCE EVALUATION IN PUBLIC ACCOUNTING

Steven C. Hunt
Assistant Professor
University of North Texas

1.0 INTRODUCTION

The performance evaluation process is important to virtually all organizations since an entity's success often largely depends on recognizing, retaining, and rewarding the best employees. Performance evaluation information can be used to validate a firm's hiring procedures, motivate employees, establish how well its training programs work, and provide feedback to employees in order to direct effort towards job behaviors viewed as most important by management.

A number of unique characteristics of the public accounting environment make it essential that the performance evaluation system be well-designed and monitored. First, large CPA firms rate performance frequently. Generally an auditor is rated at the end of each engagement.[1] Second, there is not a continuous work relationship between a superior and subordinate auditor since an auditor may work for and be evaluated by a number of supervisors during the year. Third, the environment has typically been "up-or-out," with both employees and management depending on the evaluation system to provide accurate performance information that can be used to determine promotions. Finally, the superior knows that he or she may not work with the ratee in the near future. These factors may reduce the auditor's effort to perform proper performance evaluation procedures.

Performance evaluation is an important part of the firm's control system since it can be used to determine how firm goals are met. Since performance evaluations are important determinants of rewards for achieving goals, the appraisal system affects job effort and performance [Jiambalvo, 1979]. If subordinates misperceive the importance to the rater of a particular dimension, they may apply effort toward unimportant components of a task and receive low performance evaluations. This can lead to ratee dissatisfaction and excessive turnover. Staff accountants leaving public accounting have cited the performance evaluation process as a major cause of dissatisfaction [Rhode et al., 1977; Hellreigel and White, 1973].[2]

The author wishes to thank Bill Messier, Alan Mayper, and two anonymous reviewers for their helpful comments on earlier versions of this paper.

[1] Performance is evaluated less frequently in smaller CPA firms [Reinstein and Smith, 1983].

[2] Similarly, Albrecht et al. [1983] found the three highest sources of job dissatisfaction among CPAs to be factors related to the performance evaluation system: firm policy and administration, feedback on performance, and recognition for a job well done.

To be useful in accomplishing a firm's objectives, the performance evaluation ratings must be accepted by both ratees and management as valid indicators of actual performance. Landy et al. [1978] found that subordinates more readily accepted ratings as fair when performed by supervisors with high perceived evaluation skills. Similarly, a formal performance evaluation system that the firm considers to be of poor quality may be circumvented when promotion and retention decisions are considered [Ferris and Larcker, 1983]. Research into performance evaluation in public accounting may identify improvements that can increase acceptance of the rating system.

This paper provides a discussion of existing research in performance evaluation in public accounting. The remainder of the paper is organized as follows. The next section describes a model used to organize the discussion. The following two sections discuss prior research studies and provide suggestions for further research. The final section provides some concluding remarks.

2.0 MODEL

The review of prior studies in this paper is organized around a process model of performance evaluation developed from psychological research [Landy and Farr, 1980; Feldman, 1981; Ilgen and Feldman, 1983; DeNisi et al., 1984].[3] This model is presented in Figure 1.

In the model, performance evaluation is viewed as a process consisting of four steps. The first step is the search for and attention to relevant information. According to the model, a rater actively searches for information about the ratee. Even when the rater comes across information about the subordinate by chance, he or she must determine whether to devote attention to it for future evaluation purposes. The second step is categorization (organization of information in memory). Categorization involves identifying a particular individual as a member of a certain class of persons based on how well his or her attributes are perceived to match those of a category prototype [Rush and Russell, 1988]. For example, a plumber more closely resembles the prototypical skilled worker than does a paramedic [Feldman, 1981]. A prototypical good auditor might be well-organized, even-tempered, and assertive without being belligerent. Categorization promotes cognitive efficiency since the rater can more easily store an impression or categorization of a subordinate in memory rather than a list of behaviors. The third and fourth steps, recall and rating judgment, are closely related. The rating judgment involves the recall of category prototypes or previous overall judgments, not individual behaviors. This leads to dimensional ratings being affected by the overall rating (since behaviors on various dimensions may not be remembered) instead of ratings on various dimensions being combined into an overall rating.

The first three steps are not strictly sequential. For example, the rater's category system can affect what information the rater chooses to examine, instead

[3] Looking at raters' cognitive processes has been the major focus in the psychological literature on performance evaluation since the early 1980's. Support for the Feldman [1981] model, whose basic features comprise much of the model used in this paper, was found in Padgett and Ilgen [1989] and Lance et al. [1991].

Figure 1

MODEL OF THE PERFORMANCE EVALUATION PROCESS

FACTORS AFFECTING PERFORMANCE EVALUATION	STEPS IN THE PERFORMANCE EVALUATION PROCESS: Information Search and Observation	Categorization and Storage in Memory	Recall from Memory	Rating Judgment
Ratee Behavior	X			
Ratee Characteristics	X			
Rater Characteristics	X	X		X
Context 1. Purpose 2. Preconceived Notions 3. Time pressure	X	X		X
Rating Instrument	X	X	X	X

An "X" in a box means that that factor affects a particular step in the performance evaluation process directly.

of just later recall. Determining where in the evaluation process problems occur facilitates developing corrective action. This approach also helps avoid difficulties involved with attempting to find optimal performance on each step separately. As Libby and Luft [1993] note with regard to audit judgments, it may be more efficient to allow poor performance in a relatively costly area if it can be offset by good performance in a less costly portion of the task. However, most prior research in accounting has looked at the last step in the performance evaluation process, the final rating judgment.

Several factors are believed to affect various steps of the process (see Figure 1). One factor is the behavior of the subordinate. Another factor is the characteristics of the ratee (experience, personality, age, and sex) and how they affect performance or the way the rater observes and later categorizes that performance.

The third major factor affecting the performance evaluation process is rater characteristics. Individual differences such as personality can affect category selection [Cantor, 1976]. Additionally, categorization appears to vary with experience [Ilgen and Feldman, 1983]. Experienced evaluators can be expected to have developed prototypes for good, poor, and possibly average subordinates [Feldman, 1986]. This may lead to less information search required for categorization. Rater characteristics may also affect the final rating.

The fourth factor is the rating context. The rating context involves several important aspects of the CPA firm environment. One is the purpose for which the evaluation will be used. Performance ratings in a public accounting firm are used for many purposes (promotion and salary decisions, scheduling personnel to provide an optimal mix of talent on an engagement, providing feedback to the ratee, and the determination of the effectiveness of hiring and training programs). In psychology research (Williams et al., [1985]), purpose has been found not only to affect final ratings but also the amount and type of information for which the rater searches. If information is obtained for one purpose and later used for another purpose, inaccuracies may result.

Preconceived notions are a second aspect of the rating context in the performance evaluation process. As previously noted, there is usually not a continuous working relationship in public accounting between superior and subordinate. This can lead to reliance on preconceived notions about the subordinate either from the superior's own previous work with the subordinate or from some other superior's interaction with the subordinate. Such information may be necessary in order for the supervising auditor to determine what tasks to assign to the subordinate, how much supervision is likely to be necessary, and so forth. Preconceived notions can also affect how much additional information the supervisor believes is needed to make an evaluation, how the ratee is categorized, and the final evaluation itself.

A third aspect of the rating context is the effect of time pressure. Evaluating performance is an important control procedure in a firm [McNair, 1991], but in auditing there is an inherent tradeoff between profitability and quality. Seniors frequently complain of being unable to complete necessary auditing procedures without underreporting time [McNair, 1991]. Pressure to complete the audit in a timely manner may cause auditors to reduce time spent on performance evaluation. As a result, information search may be reduced (perhaps by reliance on preconceived notions to categorize ratees) and the final rating may be done in a perfunctory manner.

The final factor affecting the performance evaluation process is the rating instrument. Various suggestions have been made for improving the rating instrument to avoid certain problems (incomplete memory of subordinate behaviors and uncertainty about what specific ratee behaviors correspond to particular numerical ratings) in recall and final rating. The instrument may also guide information search and categorization by stressing certain performance dimensions.

Three of these major factors (rater characteristics, rating context, and rating instrument) are particularly important because they correspond to suggested ways for improving performance evaluations [Ilgen et al. 1993]. These measures include: (1) improving the skills and sensitivity of raters through training, (2) changing the evaluation setting, perhaps through changes in firm policies, and (3) changing the rating task by designing a new rating instrument.

Final rating is examined first, followed by categorization, information search, and memory issues. While not following the sequence indicated by the model, this order maps the chronological progression of research regarding performance evaluation in CPA firms.

3.0 ISSUES IN PERFORMANCE EVALUATION IN PUBLIC ACCOUNTING

3.1 Auditors' Final Rating Judgments

The step that has attracted the most interest in prior studies is the rating judgment. Most accounting performance evaluation research has viewed this step as being the entire rating process. The main area of research in the final rating judgment step has been raters' cue weightings, both self-perceived and actual. This section addresses several key issues in cue weighting: (1) cue weighting among performance dimensions, (2) raters' self-insight into cue weights, (3) accuracy, (4) consensus, (5) congruence, and (6) consistency. Table 1 summarizes research in each of these areas. Following the discussion of research looking directly at cue weighting in final judgments are sections describing research which has examined the effect on cue weightings or directly on final rating judgments of each of the four factors in the performance evaluation model (Figure 1).

3.1.1 *Cue Weighting Among Performance Dimensions*

Cue weighting is important because it can determine the importance auditors attach to various aspects of ratee performance. If weights differ from the ratee's expectation, the overall evaluation may be confusing and frustrating to the ratee. If weights differ from firm policy, the overall evaluation may be improperly used to award raises and promotions.[4] CPA firms typically require that a number of performance dimensions be considered in making an overall rating. Wright [1982] found the objective weights of senior auditors overwhelmingly favored the staff auditor's

[4] Note that firm policy typically does not require a specific weighting of the various factors, therefore firm policy in this regard has been inferred through the weightings of partners. See section 3.1.3.

Table 1

CUE WEIGHTING IN FINAL RATING JUDGMENTS
AND EFFECT OF RATER CHARACTERISTICS ON CUE WEIGHTING

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Ramanathan et al. [1976]	Congruence, Rater rank	Survey	233 Professional staff (all levels) of Big 8 firms	Moderate congruence, increased with rank
Maher et al. [1979]	Congruence, Accuracy, Rater rank	Survey	234 professional staff (all levels) in 8 offices of firms from local to national	Moderate congruence; accuracy increased with rank
Jiambalvo [1982]	Accuracy, Congruence	Experiment (By Mail)	49 subjects from three offices of one large firm	Low accuracy; moderate congruence
Wright [1982]	Consensus, Self-insight	Field Study	110 seniors from seven Big 8 and three large non-Big 8 firms	High consensus; low self-insight
Jiambalvo et al. [1983]	Cue weighting, Self-insight, Consensus, Consistency, Rater rank	Experiment (By Mail)	152 partners, managers, and seniors in audit, tax, MAS of one large firm	High consistency; moderate consensus; low self-insight; cue weights did not sig. differ with rank
Kida [1984]	Cue weighting, Consensus, Consistency, Rater rank.	Experiment (By Mail)	32 seniors and 40 managers in Big 8 firms	High consistency; low consensus; technical performance most highly weighted; no significant differences in cue weights by rank
Wright [1985]	Accuracy	Survey	78 seniors from various firms	High accuracy

Table 1 (Continued)

CUE WEIGHTING IN FINAL RATING JUDGMENTS
AND EFFECT OF RATER CHARACTERISTICS ON CUE WEIGHTING

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Hassell and Arrington [1989]	Cue weighting, Self-insight	Experiment (By Mail)	8 personnel partners (five national, three regional)	Cue weighting and self-insight depended on the modeling technique (AHP vs. regression) used
Regel and Murray [1989]	Accuracy, Training, Consensus	Experiment (Mail)	49 Big 8 staff accountants at one firm	High consensus, not sig. improved by training; feedback improved accuracy
Luckett and Hirst [1989]	Consensus, Self-insight, Accuracy	Experiment	48 Australian supervisors, seniors, and senior assistants from one large firm	Feedback improved consensus and accuracy; high self-insight
Hassell et al. [1992]	Cue Weighting	Experiment	14 Big 8 managers, partners, and principals	Cue weighting depends on modeling technique used (AHP vs. regression)
Hirst and Luckett [1992]	Accuracy	Experiment	48 Australian supervisors, seniors, and senior assistants from one large firm	Initial accuracy higher with task properties feedback; over time, accuracy higher with outcome feedback

technical ability to the virtual exclusion of most other factors. However, Wright [1980] noted that firm administrators ranked motivation level and oral communication skills as being equally important as technical skills in evaluating staff auditors.[5] This suggests that raters are not using or weighting cues in accordance with firm objectives.

Most research in cue weighting has examined evaluations of seniors. Jiambalvo et al. [1983] and Kida [1984] performed experiments using virtually the same methodology. CPA firm personnel were given ratings of 24 hypothetical seniors on several dimensions and asked to combine them into an overall rating for each senior. The relative criterion scores were obtained by regressing the individual criteria scores on the global performance score. Jiambalvo et al. [1983] found that willingness to accept responsibility, ability to identify and develop practical workable standards, and technical ability were the three most important factors, when objective weights were determined.Kida [1984] found technical competence was the most important factor. Practice development was second most important, which was surprising since discussions with managers and seniors indicated it was of lesser importance at the senior level (see section 3.1.2).

Cue weightings differed in Wright [1982], Jiambalvo et al. [1983], and Kida [1984]. This may be due to differences in the number of categories and how each category was defined, as well as differences in the rank of subjects (see Table 1). Technical performance, though important, was less dominant in Jiambalvo et al. [1983] and Kida [1984] than in Wright [1982]. This would be expected, since factors other than technical ability, such as practice development, are appropriate at higher staff and management levels.

In interpreting the above results, it is helpful to determine the extent to which they depend on the method used to obtain cue weights. Two articles compared the results of Jiambalvo et al. [1983] and Kida [1984]'s linear model to those from an eigenvector-scaling technique called the Analytic Hierarchy Process (AHP). In Hassell and Arrington [1989] each subject performed the AHP procedure and then the procedures used in Jiambalvo et al. [1983]. The two sets of rankings were significantly correlated for only one of the eight expert subjects.[6] It appeared that the objective weights were sensitive to the elicitation technique.

Hassell et al. [1992] evaluated the different, though generally overlapping, criteria sets of Jiambalvo et al. [1983] and Kida [1984] using AHP. Subjects performed AHP procedures on either the Jiambalvo et al. [1983] or Kida [1984] criteria set. For both criteria sets, objective weights for participants were highest for technical competence, consistent with Kida [1984], but not Jiambalvo et al. [1983].

---

[5] Surprisingly, ability to meet time budgets was regarded as an insignificant criterion in Wright [1982] despite prior studies (e.g., Hellreigel and White [1973]) which found that accountants felt considerable pressure to meet time budgets.

[6] Subjects were large firm national or regional personnel partners. The small sample size was considered appropriate because "the (AHP) technique requires considerable time and effort and does not lend itself to aggregating decision models across individuals in the interest of inference" [Hassell and Arrington, 1989, p. 531]. On the other hand, other accounting studies using AHP (e.g., Apostolou and Hassell [1993]) have used as many as 126 subjects.

This type of research can identify suboptimal cue weighting and it can be used to design training programs to help auditors weigh cues in accordance with firm policy. However, as long as research results vary by different levels of ratees and different elicitation methods, it is difficult to use the research findings as a basis for improving practice. Additionally, this research used "paper people," (written descriptions of subordinates' behavior) instead of actual subordinates thus omitting many factors such as the rater's like or dislike of a real subordinate that can affect overall evaluations in practice.

3.1.2 *Auditors' Self-Insight in Performance Evaluation*

If a rater lacks self-insight he or she will be unable to communicate objective weights to subordinates and thus may misdirect the subordinate's effort. This may in turn lead to ratee dissatisfaction. Jiambalvo et al. [1983] found low self-insight among raters. Although the raters saw themselves as using a variety of cues, they primarily rated subordinates on a few major categories. Wright [1982] found similar results and he suggested that this might reflect seniors' difficulty in evaluating subjective nontechnical performance [Wright, 1980]. Subjects in these two studies and in Luckett and Hirst [1989] overestimated the importance of minor categories and underestimated the importance of major ones. Luckett and Hirst [1989] found high self-insight which did not increase with feedback. They indicated, however, that this result may have been due to the extreme cases in the design leading to an overstatement of self-insight.

Self-insight, like cue weighting, may depend on the modeling technique used to obtain objective weights. Hassell and Arrington [1989] found that for 5 of their 8 subjects, self-insight was correlated with different models (regression or AHP). The use of AHP instead of regression did not greatly increase self-insight. What changed, however, was *which* subjects exhibited poor self-insight.

None of these studies required subjects to write justifications of their ratings, as is frequently required in practice. This practice increases self-insight. Thus the experimental studies described above may have shown lower self-insight than exists in practice. Judgment studies in other contexts generally show higher self-insight for auditors.

3.1.3 *Accuracy in Cue Weighting*

There is no objective criterion for accuracy in performance evaluation. Researchers have generally used a surrogate, the agreement between seniors' and partners' perceptions of partners' weightings of various evaluation dimensions. Several studies have been performed, with differing results. In a nonexperimental questionnaire study, Wright [1985] found that seniors and partners agreed on the relative ranking of most evaluation criteria with quality of technical work and level of motivation viewed equal in importance. Jiambalvo [1982], however, found low accuracy. A lack of accuracy can be due to a failure of partners to adequately communicate weights, possibly due to their own poor self-insight or failure of subordinates to understand communicated weights. Differences in the results of the two studies may be due to differences in sample size, level of persons evaluated, number of performance dimensions, or type of partner used as the base.

One would expect there to be a correlation between a subordinate's performance and accuracy in knowing raters' cue weights since knowledge of such weights should lead to proper direction of one's activities. However, raters identified by their firms as "high performers" in Wright [1985], Maher et al. [1979], Ramanathan et al. [1976] and Jiambalvo [1982] did not significantly differ from others in their subjective cue weights.

Two studies examined whether raters could be trained to use "official" cue weights. Previous research assumed that simply communicating this information to raters would cause raters to use the official cue weights. Regel and Murray [1989] had audit staff members given ratings on six dimensions for each of 33 hypothetical subordinates and then asked to produce final ratings.[7] Task properties feedback (information about cue weights reflecting the firm's policy) increased accuracy by 19%. Similar results were found in Hirst and Luckett [1992]. Luckett and Hirst [1989] and Hirst and Luckett [1992] also found that judgment performance increased over time when information about the correct response (outcome feedback) was provided after the subject's rating.

Accuracy in the above studies was determined by comparing auditors' perceptions of partners' cue weightings with partners' perceptions of their own ratings. Such a surrogate for accuracy is reasonable only if partners have high self-insight into their rating schema. However, as noted in section 3.1.2, several studies found low self-insight. This limits the usefulness of this research and suggests the need for other ways to determine accuracy.

Ilgen et al. [1993] pointed out that determining raters' accuracy should be a major goal of research in performance evaluation. On the other hand, accuracy is a necessary, but not sufficient, criterion for evaluation of existing rating systems. Other criteria, such as ratees' perceptions of fairness, are very important as well. Perceived fairness has been inferred by congruence.

3.1.4 *Congruence in Cue Weighting*

Congruence (agreement between auditors' perceptions of partners' weights on various performance dimensions and the auditors' desired weights) is useful to examine in order to determine the perceived fairness of firm policy. Additionally, congruence may affect performance [Jiambalvo, 1982]. If employees concentrate on job dimensions they feel are important rather than the ones they believe actually are considered important, lack of effort, frustration, and poor evaluations may result. Ramanathan et al. [1976], Maher et al. [1979] and Jiambalvo [1982] reported considerable differences in congruence indicating some dissatisfaction with the rating systems used in the sampled firms.[8] For example, Ramanathan et al. [1976] found auditors preferred more emphasis on the quality of technical work and supervision of staff and less on quantity of billable hours. Congruence did not

[7] The authors believed that training in performance evaluation should begin early and therefore looked at staff auditors' ability to respond to training in performance evaluation. However, staff auditors rarely prepare performance evaluation reports and also do not suffer potential anchoring effects of relying on improper dimension weightings based on past experience.

[8] Ramanathan et al. [1979] looked at company goals and criteria for promotion rather than performance evaluation weights per se.

significantly differ between high performers and other auditors. Surprisingly, congruence was not significantly related to job satisfaction in Jiambalvo [1982]. Perhaps auditors have sufficient commitment, either to their firms or to the profession, to perform their best even when disagreeing with the firm's cue weighting in performance evaluation.

3.1.5 *Consensus in Cue Weighting*

Consensus appears useful as a means of determining to what extent ratings are rater-specific. If supervisors vary greatly in evaluating the same individual for performing similar tasks on similar audit engagements, ratees may be confused about their actual level of performance and thus not obtain feedback from evaluations. Several experimental studies examined this issue with differing results. Wright [1982] found a high level (.85) of consensus among senior raters. However, the use of only extremely high and extremely low performance levels may have contributed to this result. Regel and Murray [1989] also found a high (.752) level of consensus, while Jiambalvo et al. [1983] found lower levels of consensus (0.61, 0.64, and 0.49 in auditing, management services, and tax, respectively). This difference may have been due to Jiambalvo et al.'s (1983) broader range of subjects (both seniors and managers). Kida [1984] also found low consensus. Hassell et al. [1992] found low interrater reliability (consensus) using AHP. Subjects used different weighting schemes except that technical competence was judged the most important criterion.

Two studies have examined the issue of whether auditors' consensus can be improved with training. Regel and Murray [1989] found that providing task properties feedback was of little value in increasing consensus because of high initial consensus. Training actually decreased consensus among more experienced auditors, who had higher consensus initially but may have been more reluctant than less experienced auditors to switch to the "correct" weights. Luckett and Hirst [1989], however, found both task properties and outcome feedback improved consensus among Australian auditors.

None of the above studies claimed that consensus was a surrogate for accuracy. Instead, high levels of consensus may indicate that raters are uniformly evaluating performance in ways that deviate from stated criteria. This possibility was supported by the results of several studies [Wright, 1982; Kida, 1984] which found that raters focused on technical competence.

The methodology of the above studies may have limited their practical usefulness. The subjects were told that they had prepared the dimensional ratings they were given; an experimental task was to combine them into a final rating. Many current psychological models of performance evaluation [e.g., Feldman, 1981] stress that the rater prepares overall ratings by comparing a target person to a prototype (such as a good, bad, or average staff accountant), instead of rating the individual on various categories and then combining them into an overall judgment.

3.1.6 *Consistency in Cue Weighting*

A sixth line of research has examined consistency of raters' cue weighting across ratees. High consistency would indicate that cue weighting is not random.

Consistency is also useful in determining possible reasons for interrater differences. Jiambalvo et al. [1983] and Kida [1984] found very high consistency among raters in cue weighting when they evaluated 24 hypothetical subordinates. Thus, cue weighting rather than inconsistent application of a rating policy appeared to have caused the relatively low consensus found in those two studies.

3.1.7 *Ratee Characteristics*

Ratee characteristics may affect either ratee performance or the way that performance is perceived by raters.[9] Blocher [1979] found no change in ratings for auditors on repeat assignments with a client. This might have occurred because the auditor worked with the same supervisor, who formed an initial opinion of the ratee that influenced the rating on the repeat assignment. Alternatively the ratee's improvement was matched by the expectations of the rater. A third possibility was that there was no improvement. There was no change in ratings for auditors whose successive clients were in different industries, but auditors in consecutive assignments with different clients in the same industry showed a slight decline at first, followed by improvement.

Two other ratee characteristics have been examined. Blocher [1980] and Ferris and Larcker [1980] found that the academic degree (bachelor's or master's) held by the ratee did not affect ratings. Ratings were affected by whether an auditor had an industry specialization. However, it is unclear whether actual performance was affected by industry specialization. Alternatively, higher performing auditors may have been selected to specialize in a given industry or raters may have been more lenient rating subordinates known to be industry specialists. Ratee factors used in Blocher [1980] appear to have been selected on an ad hoc basis. Research on the effects of ratee characteristics on final ratings is summarized in Table 2.

3.1.8 *Effects of Rater Characteristics on Cue Weighting*

Few rater characteristics have been examined in accounting performance evaluation research. This is unfortunate since rater characteristics such as experience in performance evaluation or personality or cognitive style may affect how information about a subordinate is processed. Research on rater characteristics is summarized in Table 1.

Looking at auditors' experience in performance evaluation is important because of the need to determine whether repeated performance of the rating task leads to improvement in cue weighting. Both accuracy [Maher et al., 1979] and congruence [Maher et al., 1979; Ramanathan et al., 1976] increased with higher rank in the firm, a surrogate for experience in performance evaluation. As Maher et al. [1979] pointed out, the fact that congruence increases with experience is not surprising. Over time, those auditors who object to a firm's policies either resolve

[9] In the model, ratee and rater characteristics affect information search and not the final judgment directly. However, since the studies described in this section and in the section on rater characteristics were not conducted using a process model and instead focused only on final judgments, they are included in the final judgment section of this review.

Table 2

EFFECTS OF RATEE CHARACTERISTICS ON FINAL RATINGS

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Blocher [1979]	Ratee experience	Field study	58 seniors from two large firms	No change in ratings for consecutive assignments with same clients or with clients in different industries; decline at first, followed by subsequent improvement in consecutive assignments in the same industry
Blocher [1980]	Ratee experience and education	Field study	58 seniors from two large firms	Ratee academic degree and repeat audit did not affect ratings; industry specialty did affect ratings
Ferris and Larcker [1983]	Ratee education	Field study	90 staff auditors from one firm	Ratee academic degree did not affect ratings

their differences or leave. Jiambalvo et al. [1983] and Kida [1984] found no significant differences between managers and seniors in their objective weighting schemes.

The above studies did not focus on the effect of experience on cue weighting and the reported results merely were a by-product of examining other topics. Future research looking directly at this issue should consider two important factors. First, determining what knowledge is needed and how it is likely to be acquired is necessary in order to predict the effect of increased experience on performance evaluation. Improvement in all tasks does not necessarily come with greater experience [Libby and Luft, 1993]. Second, Hunt and Messier [1995] found that persons of higher rank may not necessarily have done more performance evaluations than persons of lower rank. Bonner [1990] has pointed out the importance of using task-specific measures of experience. Number of evaluations performed appears to be a more appropriate experience measure. Looking at quality, as well as quantity, of evaluations performed could lead to development of an even finer measure of experience.

In Kida [1984], the rater's leadership style affected cue weighting. Those scoring higher on consideration for other people favored client relations and communication skills, while those higher on initiating structure put more emphasis on technical skills. Overall ratings, however, were not significantly different between subjects with different leadership styles.

3.1.9 *Contextual Factors Affecting Final Ratings*

Various aspects of the performance evaluation context may affect ratings. Research on contextual factors is summarized in Table 3.

One line of research has examined how the performance evaluation environment affects ratee motivation, performance, and evaluation. Ideally, knowledge that performance evaluations will be used in important decisions such as salary increases should lead to greater ratee effort. This should occur if the ratee has confidence that he or she understands the factors on which the evaluation is based and believes that appraisals are closely related to allocation of rewards. Four accounting studies have used expectancy theory to predict performance.[10]

In the expectancy model, effort is expected to lead to attainment of a high level of performance, which then should lead to a desired outcome (reward). Ferris [1977] found that expectancy models were weak predictors of staff accountant performance, but did predict employee job satisfaction. Jiambalvo [1979] expanded the expectancy model to include the performance evaluation system as a mechanism which linked job effort and reward. The model accounted for self-rated performance better than manager-rated performance. Jiambalvo [1979] found a stronger relationship between motivation and performance than Ferris [1977]. In

[10] Dillard and Ferris [1989] used a model partly based on expectancy theory to examine various research articles, including several discussed here, dealing with many aspects of individual behavior in professional accounting firms. This paper differs from Dillard and Ferris [1989] by focusing only on performance evaluation and doing so using a model of performance evaluation instead of an overall model of individual behavior.

Table 3

THE EFFECT OF CONTEXTUAL FACTORS ON FINAL RATINGS

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Ferris [1977]	Performance	Questionnaire	51 staff accountants from two Big 8 firms	Expectancy models were weak predictors of job performance
Jiambalvo [1979]	Performance	Questionnaire	33 seniors and 25 managers who had recently supervised them in one large firm	Performance affected by motivation; expectancy model performed better for manager than self-ratings
Blocher [1980]	Various environmental factors	Field study	58 seniors from two large firms	Neither assignment complexity nor whether it was repeat engagement for firms affected ratings
Ferris and Larcker [1983]	Performance, Reward structure	Field study	90 staff auditors from one firm	Current salary level independent of rated performance
Moizer and Pratt [1988]	Performance and rewards	Questionnaire	220 U.K. chartered accountants (all ranks)	Ratings and promotion likelihood perceived as being mainly affected by ability as opposed to effort or luck
Hassell et al. [1992]	Purpose of evaluation	Experiment	14 Big 8 managers, partners, and principals	Salary purpose much less important than promotion or career development decision
Hunt and Messier [1995]	Purpose of evaluation	Experiment	120 seniors, supervisors, and managers from nine large international firms	Purpose affected final ratings

Moizer and Pratt [1988], English chartered accountants perceived their performance evaluations and promotion probabilities as determined more by ability than effort. Luck was seen as having virtually no influence.

Ferris and Larcker [1983] used actual evaluations of staff auditors and found that rated performance was based on ratee motivation, supporting the expectancy model. On the other hand, current salary level was independent of rated performance. There was a significant relationship between ratee physical attractiveness and salary level. This may have been due to more attractive individuals having received higher starting salaries. On the other hand, end of year determination of rewards may have been based upon different criteria than evaluations made by other raters after each audit engagement. These results are disturbing for two reasons. First, Wright [1980] found that personnel administrators considered engagement reviews to be the greatest source of information for promotion and salary decisions of staff auditors. Personnel administrators' use of physical attractiveness to allocate salary increases may thus indicate a lack of self-insight. Second, if auditors become aware that good performance evaluations do not necessarily lead to rewards, disillusionment and reduced effort may follow. Although Wright [1985] found that only 22% of seniors perceived engagement evaluations to be the most important factor in salary and promotion decisions, a large majority believed it to be among the top three factors.

Ferris and Larcker [1983] found that rated performance was not significantly influenced by congruence of attitudes (either on professional issues or on social and political issues) between auditors and their subordinates. This is encouraging, because it suggests that ratees are not penalized for having different views from their supervisors.

Since different skills may be needed in auditing, tax, and MAS, cue weighting should differ across such firm subunits. Jiambalvo et al. [1983] found differences in both cue weighting and overall evaluations for members of different subunits of one office within a large CPA firm. There was a lack of interrater agreement between auditing and tax, but not between auditing and MAS. It was felt that this might have been due to the team approach generally taken in the latter two areas. Auditing and MAS individuals placed considerable focus on the "ability to work with people," while tax staff emphasized "creativity" more. Jiambalvo et al. [1983] did not find significant differences among subunits (audit, tax, management advisory services) in the level of self-insight.

Blocher [1980] found that neither assignment complexity, length, nor whether it was a repeat engagement for the firm significantly affected actual senior accountants' ratings. Blocher [1980] viewed this as positive, indicating a lack of "bias." On the other hand, these findings may be viewed as disturbing in that raters did not make allowances for the level of difficulty of the assignment.

The purpose of the performance evaluation to provide input into later decisions was examined in two studies. Hassell et al. [1992] found that the salary decision was viewed as much less important than the promotion/retention or career development/job assignment decision. Results were inconclusive in comparing the latter two purposes to one another. However, the instructions to subjects were unclear as to whether this was an evaluation on a particular engagement or an overall yearly evaluation. Subjects may have been more accustomed to doing the

latter, given their rank (manager and partner). Hunt and Messier [1995] found purpose of evaluation (to be used in later salary decisions or to schedule the ratee's future assignments) on a particular engagement affected final ratings.

Research on the use of performance evaluation ratings has value in determining the relationship between ratings and decisions ostensibly based on them. As noted earlier Ferris and Larcker [1983] found that relationship may be more tenuous than is often believed. Even if performance evaluation ratings are a major input for later decisions, the use of the same ratings for different decisions may result in inaccuracies.

No experimental research has examined the effect of time pressure on performance evaluation in CPA firms. Wright [1985] found that only 51% of seniors reported that they generally had enough time to properly evaluate performance. Seventy-seven percent noted that at times they were "somewhat hurried" in their approach. This corresponds to the 70% who admitted to performing or discussing an evaluation in a hurried, incomplete way. One aspect of performance evaluation that was frequently omitted was providing detailed feedback, including ways to improve performance, to subordinates.

3.1.10 *Rating Instruments*

This section deals with the effect of the rating instrument on final ratings.[11] If different raters use different criteria for what constitutes "good," "poor" or "average" performance on various performance dimensions (as opposed to differences in weighting dimensions), then it will be difficult for ratees to obtain useful feedback. In Wright [1980], auditors indicated that the most common difficulties with performance evaluation were vagueness in performance criteria and scales. To deal with this problem, Wright [1986] described the preparation and use of a BARS. BARS provide descriptions of behavior expected from subordinates at various levels of performance. Wright [1986] advocated the use of a diary of subordinate behavior as a memory aid with a BARS. Memory aids were deemed likely to be useful due to the considerable time lapse that may occur between observation of subordinate behavior and preparation of the formal performance evaluation. Memory aids have been supported as a means to enhance rating accuracy and defensibility and to provide feedback for staff development. Research on BARS and memory aids is summarized in Table 4.

Harrell and Wright [1990] provided support for the validity and reliability of BARS. They compressed Wright's [1986] four major dimensions of nine items into three major categories. In questionnaire responses, raters perceived elements of BARS to reflect proper elements and actual work performance better than conventional rating scales. In a longitudinal study, BARS ratings were associated with promotion to higher rank, annual salary increases, and ratings of retention

[11] Although the instruments described in this section are designed to be useful in organizing the auditor's attention to and categorization of subordinate behavior as well as subsequent retrieval from memory, Behaviorally Anchored Rating Scale (BARS) research is discussed under the category of final ratings because prior research has not examined specifically how BARS affects earlier parts of the model.

Table 4

THE EFFECT OF RATING FORMS AND MEMORY AIDS ON FINAL JUDGMENTS

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Wright [1986]	BARS, Behavior Diaries	Field Study	Seniors and supervisors from three firms	A BARS was developed with help from subjects. The use of a BARS with behavior diaries was advocated.
Harrell and Wright [1990]	BARS	Longitudinal field study; survey	152 raters evaluating 218 auditors from 1 large firm	Subjects indicated preference for BARS over conventional rating forms; BARS was better associated with promotions and salary increases

desirability. Harrell and Wright [1990] did not, however, examine how the conventional ratings used by the firm correlated to these outcomes.

BARS research appears to have value. Providing raters with exemplars of various levels of performance may help achieve greater rating consensus, as well as improvements in feedback and the ratee's subsequent skill development. Emphasizing the rater's need to take an active part in the rating process and stressing the observation of behaviors are also useful. On the other hand, if raters categorize subordinates (see following section) and state a category prototype as a recalled behavior, BARS may be of limited usefulness.

Behavior diaries appear useful with regard to dimensional ratings, but less so for overall evaluations. There is evidence [Hunt and Messier, 1995] that overall ratings do not change appreciably with a time delay, thereby reducing the potential need for behavior diaries. Also, the use of review notes may reduce the need for behavior diaries. Finally, there is evidence in the psychological literature [e.g., Lichtenstein and Srull, 1987] that raters immediately interpret what they observe. If recorded observations are in judgmental form (e.g., "Smith is slow," instead of "Smith exceeded the time budget by two hours") in a behavior diary, such diaries will be of limited value.

Also, behavior diaries may be influenced by preconceived notions about the ratee. A supervisor can elicit the expected behavior from a subordinate and then record it in the diary [Snyder and Swann, 1978]. If a supervisor considers a subordinate "lazy," instances of "lazy" behavior may be more salient and thus more likely to be noted than instances in which the subordinate was working hard [Ilgen and Feldman, 1983].

3.2 Categorization of Ratees

3.2.1 *Attribution*

The manner in which the auditor/rater categorizes a subordinate (e.g., "good auditor") appears to be an important part of the rating process. A first step in categorization involves determining how much of an observed level of ratee performance is due to the ratee and how much is due to the environment. Several accounting studies have looked at how raters attribute poor performance.

Determining an outcome (such as being over budget in a particular audit area) and then deciding to what extent to attribute that performance to internal causes (the ratee) or external causes (the environment) are important parts of performance evaluation. Ratees will likely resent being held responsible for factors outside their control. On the other hand, performance that can be attributed to a stable, internal cause would be considered representative of normal performance. Subjects in Kaplan and Reckers [1985] and Stolt [1985] attributed poor performance to a ratee with a poor work history (a stable, internal cause) rather than auditing a client exhibiting steady growth. A steadily growing client was viewed as offering fewer unexpected audit problems than a client with erratic growth. Research in attribution is summarized in Table 5.

The practical significance of attributing performance to the ratee or the environment lies in the subsequent action taken, such as preparing high or low perfor-

Table 5

CATEGORIZATION OF RATEES: ATTRIBUTION AND ACCOUNTABILITY

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Kaplan and Reckers [1985]	Attributions	Experiment	60 Big 8 audit managers and seniors	Poor performance attributed to subordinate when work history was poor and client history stable
Stolt [1985]	Attributions	Experiment (Mail)	89 advanced staff and 147 seniors in three offices of one large firm	Offering excuses increased net external attributions; internal attributions increased with rater experience
Kaplan and Reckers [1991]	Attributions Accountability	Experiment (Mail)	107 CPAs from all Big 8 firms	Attributions did not affect immediate action taken, but did affect later scheduling decision
Kaplan and Reckers [1993]	Attributions Accountability	Experiment	88 experienced auditors, mainly seniors	Subordinate's excuses for poor performance accepted despite evidence to contrary

mance ratings. Several papers extended attribution research into the area of final judgments. Kaplan and Reckers [1985] and Stolt [1985] found that subjects who made internal attributions indicated they would blame poor results more on the subordinate than the environment, but otherwise did not indicate what further action they would take. Kaplan and Reckers [1991] found attributions did not significantly affect the specific actions taken in response to poor subordinate performance. Most subjects stated that they would document on the evaluation form that the poor performance was due primarily to poor work by the subordinate, while noting the difficulty of the assignment. The lack of an attribution effect may have been due to the rather extreme performance level in the experiment reducing the subjects' range of choices. Attributions affected certain decisions, however. Auditors were more likely to seek out the subordinate for future work or support the subordinate's request to work with the superior in the future if the former's failure on the recent audit was attributed to task difficulty. Kaplan and Reckers [1993] found similar results regarding the desirability of working with the subordinate in a variety of settings. Causal attributions also affected end of job performance evaluations.

3.2.1.1 *Rater and Ratee Characteristics Affecting Attributions*

Individual differences have been examined in several attribution studies. Kaplan and Reckers [1993] found neither tolerance for ambiguity nor social deference (an indirect measure of need for approval) were associated with causal attributions. Sizeable individual differences were present, however, across auditors. Stolt [1985] found that when seniors perceived staff to be unlike themselves, they provided more internal attributions.

Results of the effect of experience on attributions for performance have been inconsistent, due to differences in how experience and performance were measured. Kaplan and Reckers [1985] found no significant differences between seniors' and managers' attributions. Stolt [1985] found greater internal attributions by seniors than advanced staff. Kaplan and Reckers [1991] used a finer measure of experience (years of work experience rather than rank) and found that internal attributions increased with experience.

3.2.1.2 *Context Variables in Attribution*

Since there is not a continuous superior/subordinate relationship in public accounting, work history information may come from others, such as supervisors who have previously worked with the subordinate. Research indicates that raters may give unwarranted attention to the opinions of others. Stolt [1985] found that ratees called "superstars" by others were seen as having better previous performance than those with similar performance but without such designation. Hunt and Messier [1995] found that knowing another senior's evaluation of the ratee on a previous engagement affected the current evaluation.

Attribution theory appears useful in determining how raters respond to poor performance. The manner in which raters attribute better than expected performance, however, has not been examined. Also, although attribution research has exam-

ined the relationship between attributions and final ratings [Kaplan and Reckers, 1991], research still has not investigated the intervening step of assigning the ratee to a category. A major contribution of attribution theory is its use as a basis for accountability research, as discussed in the following section of this paper.

3.2.2 *Accountability*

The ratee is not merely a passive person to be observed but can influence the way that he or she is perceived (and later evaluated) by the rater. This is called impression management in both the psychological literature and in Stolt [1985]. Kaplan and Reckers [1993] referred to this phenomenon as falling under the category of accountability (see Messier and Quilliam [1992]). Accountable individuals feel social pressure to justify their judgments to significant others [Tetlock, 1985]. An individual can use either accounts, explanations, or excuses on the one hand or apologies on the other in an attempt to mitigate poor performance. Since much of a subordinate's work in public accounting is unobserved, superiors may rely on the ratee's explanation for problems that occur in the audit.

Stolt [1985] found that net attribution responses were more internal when the ratee offered an internal excuse for poor performance. Seniors given an external excuse and apology had lower expectations of future subordinate failure than those not presented with such tactics.

Similarly, Kaplan and Reckers [1993] found that a subordinate offering an external account was seen as contributing less to exceeding a time budget and failing to meet a client's deadline than a similar performer who had no explanation. Contrary to expectations, this result occurred regardless of whether a ratee's work history and client financial condition were improving or declining. Thus, additional information did not cause raters to doubt the explanations of subordinates.

As previously noted, many respondents in Wright [1985] complained of time constraints and reported occasionally preparing evaluations in an incomplete, casual, or tardy way. This may indicate a lack of perceived accountability in performance evaluation vis-a-vis other auditing tasks, such as completing the audit on a timely basis.

Although accountability relates to ratee actions to mitigate the effect of poor performance, the focus of the above research has been on the rater's response to such behavior. No study has looked at the ratee's use of explanation strategies under varying individual and context conditions in an accounting performance evaluation context. Also, no study has examined the rater's accountability to both firm executives (to provide valid ratings for later decisions) and the subordinate. Accountability research is summarized along with attribution research in Table 5.

3.3 Information Search

The studies reviewed thus far do not explicitly consider the information search phase of the performance evaluation process. Most of the studies were done by mail questionnaire, so that the researcher did not observe the subjects performing the task. Many of the experimental studies which elicited performance ratings used repeated measures designs in which the subjects evaluated a large number of

hypothetical persons. Thus, little information could be provided about each one. Subjects were presumed to have used all of the limited information provided. The evaluation task was the only one they were expected to perform, and they made decisions immediately after viewing the stimulus. Thus, with the exception of Hunt and Messier [1995], the rater's information search and rating at a *later* time were not examined.[12] These factors limit the applicability of these findings to actual performance evaluation tasks. Information search is discussed in this section and delayed rating in section 3.4. Information search is important because prior work in psychology (e.g., Murphy et al. [1982]) has shown that recognizing and attending to relevant information greatly affects the subsequent accuracy of ratings. One needs to discover what information raters currently attend to in order to determine the best methods of encouraging them to obtain appropriate amounts of relevant information. Research on information search and final rating is summarized in Table 6.

Hunt and Messier [1995] examined information search and rating judgments by looking at two important context variables, purpose of evaluation and preconceived notions. Rater experience with performance evaluation was also examined. After obtaining as much behavioral information about a hypothetical subordinate as they wished, subjects reviewed audit workpapers and wrote review notes before rating the subordinate's performance. The review note preparation task added realism since raters rarely concentrate solely on obtaining information for rating purposes. Increased experience (as measured by number of evaluations prepared in the auditor's career) resulted in less time spent reviewing workpapers, but not in observing other subordinate behaviors. Hunt and Messier [1995] indicate that this finding might be due to the senior receiving less feedback from a manager on workpaper review than on evaluation of a staff accountant's behavior. Thus, experience does not necessarily lead to greater efficiency in the evaluation of a subordinate's behaviors.[13] There could instead be a motivational reason for this finding. Since the staff accountant's actions are not observed by the manager, the senior may feel little responsibility to learn how to observe and then accurately evaluate them.[14]

Hunt and Messier [1995] found that the amount of information searched for and time spent in information search were not significantly affected by different

---

[12] These studies treated performance evaluation as a stimulus-based task, in which the judgment is made as soon as information becomes available. The evaluation task might better be considered a memory-based task. The information the rater has attended to in the past and how it is organized and stored in memory affect recall and the subsequent judgment [Ilgen and Feldman, 1983].

[13] Such lack of feedback can lead to overconfidence. A number of studies described in Keasey and Watson [1989] have found that overconfidence seems to increase with the difficulty of the task. Receiving little feedback should increase task difficulty. Wright [1985] found that subjects expressed a high degree of confidence in their ability to evaluate well in all categories except assessing a subordinate's motivation level. The latter finding is disturbing, since motivation level tied for first place in subjective importance of rating categories. There was high consensus among auditors in this confidence. Wright's [1982] finding that auditors received little training in performance evaluation raises the question of whether this confidence is misplaced.

[14] Various studies in the accountability literature (reviewed in Tetlock [1985]) indicate that subjects who expect to be held accountable pay greater attention to information and process it more thoroughly.

Table 6

THE EFFECT OF CONTEXTUAL FACTORS AND RATER CHARACTERISTICS ON RATERS' INFORMATION SEARCH

STUDY	ISSUES EXAMINED	RESEARCH DESIGN	SUBJECTS	MAJOR RESULTS
Hunt and Messier [1995]	Information search; delayed rating	Experiment	120 seniors, supervisors, and managers from nine large international firms	Neither purpose of evaluation nor preconceived notions affected information search; experience level affected time spent in workpaper review, but not in examining other behaviors. Rating delay did not affect ratings.

purposes of evaluation.[15] Information search was also unaffected by the existence or nonexistence of a preconceived notion about the ratee. Both purpose and preconceived notions affected final ratings, however. In a second experiment Hunt and Messier [1995] report that subjects receiving a poor preconceived notion rated the hypothetical subordinate lower than did those receiving either a good preconceived notion or no preconceived notion.

Hunt and Messier [1995] also found many subjects provided ratings in categories of behaviors that they had not observed. Thus raters may not obtain enough information to evaluate performance in all relevant areas, instead simply inferring unobserved behavior from observed behavior.

3.4 Memory

No study has looked directly at memory issues in performance evaluation. Concern about the time delay between observing a subordinate's behaviors and evaluating that person has led to the suggested use of rating diaries [Wright, 1986]. However, no accounting study has examined how such aids improve memory and whether they affect final ratings. Hunt and Messier [1995] found that delays of one week had little effect on ratings. Memory, however, was not directly examined except to determine that raters had poor memory for specific behaviors when making delayed ratings. These results are consistent with the model outlined earlier which predicts that raters constantly revise their judgments about a subordinate during the audit and later have to remember only their last judgment not actual subordinate behaviors.

A categorization approach implies that the appropriate standard for ratings is classification accuracy, which involves identifying a ratee as a member of the correct category (such as "good auditor") [Lord, 1985], but not necessarily remembering specific behaviors. Behavioral accuracy (observing and recalling behaviors and then determining a category of behavior) is another matter. Both types of accuracy would seem to be important, but it appears to be difficult to achieve both high behavioral accuracy and high classification accuracy. For example, if information is processed according to a categorization model (e.g., Feldman, [1981]) and high classification accuracy is achieved, then behavioral accuracy is likely to be low because raters recall only category prototypes rather than actual behavioral observations. High classification accuracy may be useful in identifying good overall performers, but poor behavioral accuracy makes it difficult to provide necessary feedback to subordinates regarding their performance on a number of dimensions.

[15] This may be due to subjects' receiving instructions that they were to obtain information to complete the audit in an efficient and effective manner before being given one of two purposes of the evaluation. Completing the audit may have overridden the salary increase or scheduling purposes. Alternatively, purpose of the evaluation may not impact information search unless the rater is making the ultimate decision instead of providing one of many evaluations to be used by others in later decision-making.

4.0 FUTURE RESEARCH

4.1 General Comments

One motivation for research in performance evaluation of auditors is to gain knowledge useful in improving practice. Improvement of performance evaluation requires greater collaboration between practitioners and researchers (Banks and Murphy [1985]). Research should examine performance appraisals in situations more closely simulating the CPA firm environment.

Another way to make research more useful is for it to relate to possible improvements in rater training. Research looking at performance evaluation as a multi-step process should be useful in determining the most effective areas to emphasize in training.

Studies describing auditors' perceptions of the performance evaluation system (e.g., Wright [1985]) are at least a decade old. New research of this type is necessary to discover promising avenues for experimental research. Much has happened in the accounting profession since many of these studies were performed. Mergers of large firms, the slowed growth of the profession (which has caused many firms to drastically reduce hiring), an increasingly diverse workforce, and greater concern for "quality of life" issues are but a few of the changes that have occurred in public accounting in recent years. How these factors have changed firms' performance evaluation systems remains to be seen.

Following is a discussion of suggested research in the four major areas of the performance evaluation process. While research in individual areas of the model may still yield useful insights, integrating several parts of the process [cf., Hunt and Messier, 1995] should provide valuable information as to where various factors have an effect. Such research could also help determine the descriptive validity of the process model in an audit context.

4.2 Raters' Information Search

Information search is a relatively unexplored area that appears to offer many possibilities for research. Individual differences could be examined to determine if they affect information search. For example, those less tolerant of ambiguity might want to obtain more information to resolve discrepancies between inconsistent subordinate behaviors.

Inconsistent behaviors could also be used to determine if the order in which behavioral information is obtained affects the amount of further information search. Obtaining inconsistent information should increase processing time [Feldman, 1981].

Feedback, as a major output of the performance evaluation process, is a fertile area for research. The rater's further information search (as well as categorization and final rating) should be affected by the ratee's response to such feedback.

The effect of time pressure on information search could be examined. Such pressure could reduce information search, perhaps by increasing reliance on preconceived notions about the ratee.

Accountability could be examined to determine if a need to defend ratings increases information search. Research could attempt to separate accountability effects from those of lack of feedback in explaining why search for ratee behavior information did not vary with experience in Hunt and Messier [1995].

Different research methods should be considered. Within-subject designs could be used to determine if individuals adapt search strategies to the situation or use one strategy across situations. Hunt and Messier [1995] performed their experiment one-on-one with the subjects, providing them the information requested. Future research in rater information search could use computer information boards or process tracing.

4.3 Categorization

Attribution theory research should be extended. For example, investigating how good, rather than poor, performance is attributed to the ratee or the environment should prove useful.

Accountability appears to be a promising area for research. Kaplan and Reckers [1991] suggested collecting survey data to determine the extent to which subordinates attempt to mitigate poor performance with excuses or explanations, since their results indicate that subordinates can benefit from such behavior. Accountability could also serve as the basis for examining behavior such as underreporting of time or premature signoff of audit procedures, both of which may be done to improve one's performance evaluation by increasing perceived auditing efficiency.

Determining what categories auditors use in evaluating subordinates, how these categories differ with evaluation experience, and how they affect recall and rating judgment could help in designing rater training programs. Training to create a common categorization system might reduce variance among raters in overall evaluations. Categorization research could also determine the extent of classification versus behavioral accuracy.

The effect of contextual variables on categorization could be examined. For example, Hunt and Messier [1995] found that purpose of the evaluation and preconceived notions about the ratee affected final ratings but not information search. Categorization may have been the mechanism by which this occurred.

Research should also examine the effects of individual differences on categorization since Stolt [1985] found that raters are more likely to evaluate failure as being caused by the subordinate when the latter is not like them. In addition to personality variables, individual differences could include race, sex, and age, since the workforce is becoming increasingly diverse. If such factors proved to be correlated to perceived differences in performance, further investigation would be necessary to determine if the result represented bias or actual performance differences.

Experimental studies looking at the effect of a rater's like or dislike for the ratee on ratings would help determine the extent of impartiality in the performance evaluation system. Such effects could occur during categorization or final rating.

4.4 Memory

There is considerable opportunity for research in recall of performance information from memory. Research could determine what types of ratee behaviors are most likely to be remembered. Such research could help determine whether behaviors discrepant with a subordinate's overall performance are more or less easily remembered than other behaviors. This evidence would be helpful in determining whether auditors are able to identify and communicate areas of needed improvement to subordinates. Research could determine if more recent information is more likely to be recalled and affect rating judgments than earlier information.

4.5 Final Ratings

The relationship between engagement and year-end evaluations needs to be examined. The latter appears more likely to be performed in a group mode, while the former is an individual matter. A field study would be useful to provide information on how year-end decisions are made, including to what degree individual engagement evaluations are considered. Longitudinal studies aimed at determining the relationship between accuracy and congruence and voluntary turnover might also prove productive.

Previous accountability studies focused on accountability tactics of the ratee and their affect on raters' attributions and decisions. The senior rater is also accountable to the staff accountant. Raters may be reluctant to give less than good evaluations because they have to discuss the evaluation later with the ratee. Research could determine whether accountability caused the rating inflation noted in Wright [1980, 1985] and the perceived lack of honesty in discussions with subordinates following evaluations found by Wright [1985]. The rater is also accountable to top executives to provide accurate ratings for salary and promotion decisions. How raters balance their accountability to subordinates and superiors is an interesting research topic.

5.0 CONCLUDING REMARKS

This paper has reviewed existing research in performance evaluation in public accounting firms through the use of a cognitive information processing model. Much previous research looked at the final step in the performance evaluation process, the rating judgment, as the process itself. This research yielded considerable useful information about final ratings. However, two other areas of the process, information search and memory, leading up to final ratings received little attention. It is necessary to expand research into these aspects of the process and to examine several aspects of the process in the same study if results of research in performance evaluation in accounting firms are to prove useful in improving practice. Although such a cognitive processing approach appears useful, this research poses a significant challenge since it "deals with variables that are simply difficult to measure and for which research methodologies are primitive" [Banks and Murphy, 1985, p. 339].

ANNOTATED BIBLIOGRAPHY

1. Ferris, K., and D. Larcker. 1983. Explanatory variables of auditor performance in a large public accounting firm. *Accounting, Organizations and Society* 8 (1): 1–12.

This field study tested expectancy theory in an accounting environment and also looked at the relationship between performance evaluations and rewards. Ratee motivation affected rewarded performance as expected, but performance was not significantly related to rewards in the firm. This implies that different criteria may be used for year-end evaluations and end-of-job evaluations.

2. Harrell, A. and A. Wright. 1990. Empirical evidence on the validity and reliability of behaviorally anchored rating scales for auditors. *Auditing: A Journal of Practice and Theory* (Fall): 134–149.

This study consisted of two parts. In the first part, a survey found that auditors perceived a BARS to better represent performance than did conventional rating scales. In the second part, a longitudinal study indicated that BARS ratings corresponded well with salaries, promotions, and other rewards within CPA firms.

3. Hassell, J.M. and C.E. Arrington. 1989. A comparative analysis of the construct validity of coefficients in paramorphic models of accounting judgments: A replication and extension. *Accounting, Organizations and Society* 14 (5/6): 527–537.

The purpose of this study was to examine the construct validity of coefficients of cue importance. Personnel partners evaluated hypothetical staff accountants on an annual review. Each subject performed procedures used in Jiambalvo et al. [1983] and then the Analytic Hierarchy Process (AHP). Cue weighting was found to be dependent on the modeling technique used (regression or AHP).

4. Hunt, S.C. and W.F. Messier, Jr. 1995. Auditor performance evaluation: Factors affecting information search and the rating decision. Working paper.

This study examines auditors' search for information about a subordinate or a delay between observation of ratee behavior and ratings. Purpose of evaluation and preconceived notions did not affect information search. More experienced raters reviewed workpapers more quickly, but search for other ratee behaviors was unaffected by experience. This implies that training auditors in non-technical areas of performance evaluation may be indicated. Ratings made after a one-week delay and those made immediately did not significantly differ.

5. Jiambalvo, J., D.J.H. Watson, and J.V. Baumler. 1983. An examination of performance evaluation decisions in CPA firm subunits. *Accounting, Organizations and Society* 8 (1): 13–25.

This research looks at differences in performance evaluation between various subunits (tax, audit, management advisory services) of CPA firms. In a questionnaire study, raters evaluated hypothetical seniors. Significant differences existed

on cue weighting (but not self-insight) among departments, indicating that departments place different values on different categories of performance (for example, tax departments emphasized "creativity" more than audit). Subjects showed high consistency.

6. Kaplan, S., and P.M.J. Reckers. 1985. An examination of auditor performance evaluation. *The Accounting Review* 60 (July): 477–487.

This is the first published study dealing with attributions in accounting performance evaluation research. Manager and senior subjects attributed poor performance of a hypothetical subordinate to the ratee when the latter's performance had been declining and the client's financial position was stable.

7. Kaplan, S. and P.M.J. Reckers. 1993. An examination of the effects of accountability tactics on performance evaluation judgments in public accounting. *Behavioral Research in Accounting* 5: 101–123.

This laboratory experiment looked at ratees' use of accountability tactics (providing an explanation for poor performance). Results indicated that such tactics could be successful. Subjects did not appear to be influenced by factors such as previous ratings of subordinate or client's financial condition that were hypothesized to affect the likelihood of accepting the explanation. Causal attributions affected performance ratings and the likelihood of the ratee working under the auditor again on another engagement.

8. Luckett, P.F. and M.K. Hirst. 1989. The impact of feedback on inter-rater agreement and self-insight in performance evaluation decisions. *Accounting, Organizations and Society* 14 (5/6): 379–387.

In an experiment involving Australian supervisors and seniors, the authors examined the effect of training on accuracy, consensus, and self-insight. Feedback improved consensus and accuracy, but not self-insight, which was high at the start.

9. Wright, A. 1982. An investigation of the engagement evaluation process for staff auditors. *Journal of Accounting Research* 20 (Spring): 227–237.

This study focused on determining the objective criteria used to evaluate staff auditors. Seniors participated in an experiment in which they rated hypothetical staff auditors. Technical performance was by far the highest weighted dimension. Subjects had problems with self-insight; most thought they were using a number of cues instead of focusing on technical ability. Very high consensus was found.

REFERENCES

Albrecht, S., S.W. Brown, and D.R. Field. 1983. Toward increasing job satisfaction of practicing CPAs. *CPA Journal* 43 (October): 61–66.

Apostolou, B., and J.M. Hassell. 1993. An empirical examination of the sensitivity of the analytic hierarchy process to departures from recommended consistency ratios. *Mathematical and Computer Modelling* 17 (4/5): 163–170.

Banks, C.G., and K.R. Murphy. 1985. Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology* 38 (Summer): 335–345.

Blocher, E. 1979. Performance effects of different audit staff assignment strategies. *The Accounting Review* 54 (July): 563–573.

Blocher, E. 1980. CPA firms' staff evaluation process. *The CPA Journal* 50 (July): 41–47.

Bonner, S.E. 1990. Experience effects in auditing: The role of task-specific knowledge. *The Accounting Review* 65 (January): 72–92.

Cantor, J.H. 1976. Individual needs and salient constructs in interpersonal perception. *Journal of Personality and Social Psychology* 34 (September): 519–525.

DeNisi, A.S., T.P. Cafferty, and B.W. Meglino. 1984. A cognitive view of the performance appraisal process: A model and research proposition. *Organizational Behavior and Human Performance* 33 (June): 360–396.

Dillard, J.F., and K.R. Ferris. 1989. Individual behavior in professional accounting firms: A review and synthesis. *Journal of Accounting Literature* 8: 208–234.

Favero, J.L. 1983. The effects of rater prototypicality and time on rater's observation and evaluation accuracy. Unpublished Master's thesis, Purdue University, 1983.

Feldman, J.M. 1981. Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology* 66 (April): 127–148.

Feldman, J.M. 1986. A note on the statistical correction of a halo error. *Journal of Applied Psychology* 71 (February): 173–176.

Ferris, K.R. 1977. A test of the expectancy theory of motivation in an accounting environment. *The Accounting Review* 52 (July): 605–615.

Ferris, K., and D. Larcker. 1983. Explanatory variables of auditor performance in a large public accounting firm. *Accounting, Organizations and Society* 8 (1): 1–12.

Fischhoff, B., P. Slovic, and S. Lichtenstein. 1977. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance* 3 (November): 552–564.

Harrell, A., and A. Wright. 1990. Empirical evidence on the validity and reliability of behaviorally anchored rating scales for auditors. *Auditing: A journal of Practice and Theory* (Fall): 134–149.

Hassell, J.M., and C.E. Arrington. 1989. A comparative analysis of the construct validity of coefficients in paramorphic models of accounting judgments: A replication and extension. *Accounting, Organizations and Society* 14 (5/6): 527–537.

Hassell, J.M., H.W. Hennessey, Jr., and J.E. Rebele. 1992. A reexamination of the relative importance of CPA firms' performance evaluation criteria. *Advances in Accounting* 10:121–142.

Hellreigel, D., and G. White. 1973. Turnover of professionals in public accounting: A comparative analysis. *Personnel Psychology* 26 (Summer): 239–249.

Hirst, M.K., and P.F. Luckett. 1992. The relative effectiveness of different types of feedback in performance evaluation. *Behavioral Research in Accounting*. 4: 1–22.

Hunt, S.C., and W.F. Messier, Jr. 1995. Auditor performance evaluation: Factors affecting information search and the rating decision. Working paper.

Ilgen, D.L., and J.M. Feldman. 1983. Performance appraisal: A process focus. In *Research in Organizational Behavior*, vol. 5, edited by L.L. Cummings and B.M. Staw, pp. 141–197. Greenwich, CT: JAI Press.

Ilgen, D.R., J.L. Barnes-Farrell, and D.R. McKellin. 1993. Performance appraisal research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes* 54 (April): 321–368.

Jiambalvo, J. 1979. Performance evaluation and directed job effort: Model development in a CPA firm setting. *Journal of Accounting Research* 17 (Autumn): 436–455.

Jiambalvo, J. 1982. Measures of accuracy and congruence in the performance evaluation of CPA personnel: Replication and extension. *Journal of Accounting Research* 20 (Spring): 152–161.

Jiambalvo, J., D.J.H. Watson, and J.V. Baumler. 1983. An examination of performance evaluation decisions in CPA firm subunits. *Accounting, Organizations and Society* 8 (1): 13–25.

Kaplan, S., and P.M.J. Reckers. 1985. An examination of auditor performance evaluation. *The Accounting Review* 60 (July): 477–487.

Kaplan, S., and P.M.J. Reckers. 1991. An attributional analysis of the performance evaluation process. *Advances in Accounting* 9: 227–248.

Kaplan, S., and P.M.J. Reckers. 1993. An examination of the effects of accountability tactics on performance evaluation judgments in public accounting. *Behavioral Research in Accounting* 5: 101–123.

Keasey, K., and R. Watson. 1989. Consensus and accuracy in accounting studies of decision-making: A note on a new measure of consensus. *Accounting, Organizations and Society* (4): 337–345.

Kida, T. 1984. Performance evaluation and review meeting characteristics in CPA firms. *Accounting, Organizations and Society* 9 (2): 127–141.

Lance, C.E., D.J. Woehr, and S.A. Fisicaro. 1991. Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior* 12 (January): 1–20.

Landy, F.J., J. Barnes, and K. Murphy. 1978. Correlates of perceived fairness and accuracy in performance appraisal. *Journal of Applied Psychology* 63 (December): 751–754.

Landy, F.J. and J.L. Farr. 1980. Performance rating. *Psychological Bulletin* 87 (January): 72–107.

Libby, R., and J. Luft. 1993. Determinants of judgment performance in accounting settings: Ability, knowledge, motivation, and environment. *Accounting, Organizations and Society* 18 (5): 425–450.

Lichtenstein, M., and T.K. Srull. 1987. Processing objectives as a determinant of the relationship between recall and judgment. *Journal of Experimental Social Psychology* 23 (March): 93–118.

Lord, R.G. 1985. Accuracy in behavioral measurement: An alternative definition based on raters' cognitive schema and signal detection theory. *Journal of Applied Psychology* 70 (February): 66–71.

Luckett, P.F., and M.R. Hirst. 1989. The impact of feedback on inter-rater agreement and self insight in performance evaluation decisions. *Accounting, Organizations and Society* 14 (5/6): 379–387.

Maher, M.W., K.V. Ramanathan, and R.B. Peterson. 1979. Performance congruence, information accuracy, and employee performance: A field study. *Journal of Accounting Research* 17 (Autumn): 476–503.

McNair, C.J. 1991. Proper compromises: the management control dilemma in public accounting and its impact on auditor behavior. *Accounting, Organizations and Society* 16 (7): 635–653.

Messier, Jr., W.F. and W. Quilliam. 1992. The effect of accountability on judgment: Development of hypotheses for auditing. *Auditing: A Journal of Practice and Theory* (Supplement): 123–138.

Moizer, P., and J. Pratt. 1988. The evaluation of performance in firms of chartered accountants. *Accounting and Business Research* 18: 227–237.

Murphy, K.R., M. Garcia, S. Kerkar, C. Martin, and W.K. Balzer. 1982. Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology* 67 (June): 320–325.

Padgett, M.Y., and D.L. Ilgen. 1989. The impact of ratee performance characteristics on rater cognitive processes and alternative measures of rater accuracy. *Organizational Behavior and Human Decision Processes* 44 (October): 232–260.

Ramanathan, K., R. Peterson, and M. Maher. 1976. Strategic goals and performance criteria in CPA firms. *Journal of Accountancy* 141 (January): 56–64.

Regel, R.W., and D. Murray. 1989. Staff performance evaluation by auditors: The effect of training on accuracy and consensus. *Advances in Accounting* 7: 223–239.

Reinstein, A., and J.E. Smith. 1981. CPA firms' performance appraisal procedures. *Journal of Accountancy* 160 (August): 48–55.

Rhode, J., J. Sorensen, and E. Lawler. 1977. Sources of professional staff turnover in public accounting firms revealed by the exit interview. *Accounting, Organizations and Society* 2 (2): 165–175.

Rush, M.C., and J.E.A. Russell. 1985. Leader prototypes and prototype-contingent consensus in leader behavior descriptions. *Journal of Experimental Social Psychology* 24 (January): 88–104.

Snyder, M., and W.B. Swann, Jr. 1978. Behavioral confirmation in social interaction: From social perception to social reality. *Journal of Experimental Social Psychology* 14 (March): 148–162.

Stolt, S.M.J. 1985. An attributional analysis of performance evaluation in public accounting. Unpublished dissertation, Arizona State University.

Tetlock, P.E. 1985. Accountability: the neglected social context of judgment and choice. In L. Cummings and B.M. Staw, eds. *Research in Organizational Behavior* 7, Greenwich, CT: JAI Press: 297–332.

Williams, K.G., A.S. DeNisi, A.G. Blencoe, and T.P. Cafferty. 1985. The role of appraisal purpose: Effects of purpose on information acquisition and utilization. *Organizational Behavior and Human Decision Processes* 35 (June): 314–339.

Wright, A. 1980. Performance appraisal of staff auditors. *The CPA Journal* 50 (November): 37–43.

Wright, A. 1982. An investigation of the engagement evaluation process for staff auditors. *Journal of Accounting Research* 20 (Spring): 227–237.

Wright, A. 1985. Rating the raters: Indications of seniors' performance in evaluating staff auditors. *Advances in Accounting* 2: 185–198.

Wright, A. 1986. Performance evaluation of staff auditors: A behaviorally anchored rating scale. *Auditing: A Journal of Practice and Theory* 6 (Spring): 95–110.