Distributional Ratings of Performance Levels and Variability

AN EXAMINATION OF RATING VALIDITY IN A FIELD SETTING

DIANA L. DEADRICK
Old Dominion University
DONALD G. GARDNER
University of Colorado at Colorado Springs

The performance distribution assessment (PDA) method was purported to be a breakthrough in performance appraisal methodology; however, little research has been conducted to determine the usefulness of this method. This article describes some of the critical features of the PDA method and presents evidence supporting the validity of the PDA in an organizational setting. The performance and ability data of 397 sewing machine operators were analyzed to determine the validity of multiple performance measures derived from the PDA, the relative accuracy of the PDA compared with an evaluative rating method, and differential criterion-related validities for the multiple PDA performance measures. Results revealed significant correlations between the PDA-derived performance measures and objective measures of job performance, differential correlations between ability and the multiple PDA-derived performance measures, and equivalent levels of rating accuracy for the PDA and the evaluative measure of typical performance. Implications for research and practice are discussed.

Performance management systems are becoming increasingly popular as a means of developing a more strategic approach to managing employee and organizational performance (Lee, 1996; Masterson & Taylor, 1996). In contrast to traditional performance appraisal systems, a performance management system focuses attention on system as well as individual causes of

An earlier version of this article was presented at the Southern Management Association meeting in Orlando, FL, November 1995. The authors contributed equally to this article and wish to thank Dianna Stone and three anonymous reviewers for their helpful comments. *Address correspondence and requests for reprints to Diana L. Deadrick, Department of Management, College of Business & Public Administration, Old Dominion University, Norfolk, VA 23529-0218.*

Group & Organization Management, Vol. 22 No. 3 September 1997 317-342

performance and performance variation, thus providing a more integrated means of promoting continuous performance improvement. From an organizational perspective, this shift toward performance management is evidenced by the current emphasis on total quality management (TQM) practices. For example, advocates of the TQM systems approach espouse the importance of identifying and separating sources of performance variability that are attributable to the organization from sources attributable to the person (see Dobbins, Cardy, & Carson, 1991, for a review of the perspective). In terms of performance appraisal, a systems approach would incorporate measures of individual performance variation that reflect performance changes due to the person (e.g., ability, motivation, disposition) as opposed to changes in the job or job context (e.g., job requirements and/or work technologies).

From a research perspective, this shift toward performance management is evidenced by the current trend toward focusing attention on individual differences in performance variability and change patterns (Austin, Villanova, Kane, & Bernardin, 1991; Borman, 1991; Hofmann, Jacobs, & Baratta, 1993; Murphy, 1989). For example, Murphy (1989) developed a dynamic model of job performance that emphasizes the need for examining temporal changes in individual performance and identifying the extent to which those changes are caused by structural changes in the job or job content, or caused by changes in employee characteristics (e.g., motivation). The implication of his model is that performance does (and should!) change over time. The challenge, then, is to develop performance appraisal and management systems that capture this phenomenon of performance and provide a means of better understanding the nature and causes of performance and performance variability over time.

The performance distribution assessment (PDA) method of performance appraisal was proposed by Kane (1982, 1984, 1986) as somewhat of a breakthrough in performance appraisal technology. Whereas most performance appraisal methods focus on only an average level of performance, the PDA method focuses on an individual's distribution of feasible performance outcomes over a specified period of time. In this respect, the PDA method is highly compatible with the current trends toward better understanding individual differences in performance variability and change patterns. In addition, this method has the potential to provide more valid measures of performance than other appraisal methods. To the extent that rating errors and inaccuracies are attributable to the subjectivity in an appraisal system, the PDA method should reduce subjectivity by minimizing the judgmental role of raters in the measurement procedure (Kane, 1982, 1984, 1986).

To date, the usefulness of this appraisal method is unknown (Austin et al., 1991; Borman, 1991; Dobbins et al., 1991). Only two empirical evaluations

have been published (Jako & Murphy, 1990; Steiner, Rain, & Smalley, 1993), and those investigators came to different conclusions regarding the value of PDA-type (distributional) ratings. Jako and Murphy (1990) concluded that the benefits of distributional ratings may be "limited," whereas Steiner et al. (1993) concluded that distributional ratings "hold promise" as an appraisal format. In light of these mixed findings, the purposes of this article are to (a) describe some of the unique features of the PDA method of performance appraisal, (b) review the existing evidence pertaining to the validity of the PDA, and (c) present additional evidence pertaining to the validity of the PDA in a field setting.

THE PDA METHOD OF APPRAISAL

Kane (1982) proposed the PDA as a new methodology for performance appraisal. In addition to providing a new format for appraisal, the PDA also provides a means of developing a more complete performance management system. According to Kane, job performance is defined as the "record of outcomes achieved over the multiple instances of carrying out the function during a specified period of time" (p. 3). As a result, performance can best be represented as a distribution of outcome levels achieved, per job function, task, or duty, over a specified time period.

By conceptualizing performance as a distribution of outcomes achieved per job function, performance can be measured using both central tendency and variation-based parameters. Central tendency measures reflect the average, or typical, level of performance, whereas the variation-based measures reflect the degree of consistency or, conversely, the degree of variability in performance. This explicit focus on performance distributions and performance consistency has been largely ignored in the literature yet provides a means of better understanding (and managing) individual performance (Austin et al., 1991; Borman, 1991; Dobbins et al., 1991; Murphy, 1989).

Performance distributions can provide both researchers and managers with a richer description of performance and a better understanding of exactly what "typical" performance means (Borman, 1991). For example, two employees might have the same average level of performance, but this does not mean that they both consistently perform at that average level. One employee might be fairly consistent, plus or minus 10% about his or her average, whereas the other employee's performance might vary widely from very poor to very high. Assuming that raters can reliably estimate ratee performance distributions, it would then be possible to draw inferences about the relative impact of ability, motivation, and situational constraints on ratee performance (Austin et al., 1991; Borman, 1991).

Performance distributions can also be used to examine the prevalence of performance inconsistency at the individual level of analysis. The PDA method represents a means of identifying and examining the existence of individual performance variation that reflects performance changes due to the person (e.g., ability, motivation, disposition) as opposed to changes in the job or job context (e.g., job requirements and/or work technologies). Because the PDA method explicitly accounts for feasible performance outcomes and distributions, the focus is on performance parameters that are attributable to person influences rather than situational (system) influences (Austin et al., 1991; Borman, 1991; Dobbins et al., 1991; Kane, 1986; Murphy, 1989).

With the PDA method, the rating task is divided into three stages: preappraisal, performance appraisal, and performance measurement. In the preappraisal stage, job experts (e.g., incumbents, supervisors) describe the multiple outcome levels (e.g., high, average, and low levels) associated with each job function (e.g., word processing, filing, taking dictation), assign utility values to each of the outcome levels, and then determine the feasible distribution of performance for each job function for the appraisal period. After this preappraisal stage has been completed, raters then appraise ratee performance, which consists of estimating the percentage of time that the ratee actually achieved each of the outcome levels defined for each job function. Based on these frequency of occurrence estimates, it is then possible to derive a variety of scores that pertain to different performance parameters of interest to the organization. These derived scores are computed after the ratings have been conducted and do not require the raters themselves to calculate performance scores. Thus, after the rating distributions for all employees are obtained, the potential for rater errors is reduced because the raters are no longer involved in the process.

Three measures of particular importance for performance management are (a) mean performance, which refers to a ratee's average or typical level of performance; (b) consistency of performance, which refers to the amount of variability in ratee performance over a specified period of time; and (c) negative-range avoidance (NRA), which refers to how successful the ratee was in avoiding those outcome levels associated with negative utility values (see Kane, 1984, 1986, for suggested formulas). NRA is the percentage of time that an employee performs at levels that have positive marginal revenue product for a company, in labor economics terms. NRA represents a useful shortcut measure for assessing the costs/benefits of performance variability. For example, an employee who had a relatively low level of average performance yet a 100% success rate in avoiding the negative performance range might be considered a better performer than an employee with a higher level of average performance yet only a 30% NRA rate (depending on the nature

of the job; e.g., piloting an airline vs. sweeping floors). In an organization setting, a measure of NRA could be useful for monitoring the impact of performance variation at both the individual and the aggregate levels of measurement.

To the extent that the PDA is in fact a breakthrough in appraisal technology, this method should yield valid measures of the multiple parameters of performance. Although there is some debate about whether rating accuracy is the primary concern among practicing managers and executives, there is no question that the effectiveness of personnel decision making is based on the validity of the performance measurement system (Austin et al., 1991; Borman, 1991; Dobbins et al., 1991; Harris, 1994; Kane, 1994; Longenecker, Sims, & Gioia, 1987; Sulsky & Balzer, 1988). Proponents of the PDA method (Bernardin & Beatty, 1984; Kane, 1982, 1984, 1986) have identified several features of the PDA method that should produce more valid measures of performance.

First, raters may be able to accurately report relative occurrence ratings. Kane (1984) cited evidence that suggests that humans may store frequency information in terms of occurrence rates, which are likely to be recalled more accurately than raw frequencies. In the more general context of social information processing, Zuroff (1989) found that (a) specific judgments of frequency of occurrence were significantly influenced by actual variations in frequency of occurrence, in both immediate and delayed rating conditions; (b) specific judgments of frequency were unaffected by manipulations of schemata that were introduced prior to the observation task; and (c) global judgments of frequency were strongly affected by manipulations of schemata. These findings are supportive of Kane's contention that raters should more accurately report occurrence rates for specific job function outcome levels as opposed to omnibus ratings of performance. A determination of the extent to which raters can provide reliable and accurate judgments of occurrence rates is central to establishing the usefulness of the PDA method of appraisal (Borman, 1991). If raters are not sensitive to ratee performance variations, and, therefore, cannot provide reliable estimates of the occurrence ratings, then the value of the PDA methodology would be questionable.

A second feature of the PDA methodology that should improve rating validity is the separation of the performance appraisal task (i.e., producing performance ratings) from the performance measurement task (i.e., developing performance scores), which results in derived measures of performance effectiveness. It is well documented that raters will intentionally distort performance appraisal ratings for a variety of reasons (e.g., Harris, 1994; Kane, 1994; Longenecker et al., 1987). Using the PDA method, the relationship between the ratings and the resultant performance scores is essentially

concealed from the raters (Kane, 1986). Therefore, the distributional ratings elicited from the raters should be less susceptible to deliberate rater distortion (Kane, 1984, 1986, 1994). Without specific knowledge of the value placed on each rated level of performance, raters are not able to intentionally manipulate PDA ratings as easily as other rating methods (like graphic rating scales).[1]

A third feature of the PDA methodology that should improve validity is the descriptive nature of the performance ratings. Kane (1982, 1984, 1986) argued that the descriptive nature of the distributional ratings minimizes the cognitive demands place on raters—that is, raters are not required to cognitively calculate performance parameters. Because most rating systems do not specify which performance distribution parameters should be considered or excluded from consideration, raters are left to their own devices to come up with a rating (Kane, 1982). Rather than describing what they have observed and then evaluating it, raters often rate how good or bad performance was and then justify it with selective examples of ratee performance (the classic "tell-and-sell" approach).

PRIOR RESEARCH

The two studies that have been conducted provide some insight into different features of PDA-type ratings. Jako and Murphy (1990) published the first empirical study investigating PDA-type ratings. In their study, they compared the levels of rating accuracy and interrater agreement obtained from a PDA-type (distributional) rating format with those obtained from an evaluative (Likert-type) rating format. In addition, they examined the impact of different levels of judgment decomposition (i.e., overall, dimensional, and behavioral ratings) on rating accuracy and interrater agreement. Using undergraduate students and videotaped lectures, Jako and Murphy found that (a) the decomposed (behavioral) ratings resulted in more reliable and accurate ratings, regardless of rating format; and (b) the distributional format did not produce significantly different levels of reliability or accuracy when compared with the evaluative ratings. Based on these findings, Jako and Murphy concluded that "judgments collected in a distributional format are no more accurate than global evaluative judgments" (p. 504).

Steiner et al. (1993) argued that the study conducted by Jako and Murphy (1990) was an inadequate evaluation of the properties of distributional ratings. Because Jako and Murphy used a simpler rating format than that proposed by Kane (1986) and used stimuli that were relatively homogeneous in performance levels, Steiner et al. argued that the distributional ratings should not be expected to produce improvements in accuracy or reliability.

Based on these arguments, Steiner et al. conducted a study with the purpose of examining the construct validity of distributional ratings. They also compared the distributional format with a behavioral observation scale (BOS) format to examine the extent to which the different formats produced similar levels of interrater agreement. Using undergraduate students and videotaped lectures, Steiner et al. found that (a) the distributional ratings were sensitive to variability in performance (i.e., the standard deviation of the distribution increased as true performance variability increased), (b) the distributional format resulted in greater interrater agreement under conditions of greater performance variability, and (c) the distributional format did not produce significantly better levels of interrater agreement than the BOS format. Steiner et al. concluded that the distributional format can provide reliable estimates of variability in ratee performance and thus has the potential for providing richer information than traditional rating formats.

Although these two studies appear to yield conflicting findings, they are not directly comparable because the investigators based their conclusions on different aspects of PDA rating validity and accuracy. Whereas Jako and Murphy (1990) based their conclusions on findings pertaining to the relative accuracy of evaluative and distributional ratings, Steiner et al. (1993) based their conclusions on findings pertaining to the accuracy of the occurrence ratings. Furthermore, several limitations of these studies preclude a determination of the usefulness of the PDA method for research or practice.

First, the findings of the two studies reviewed here suggest that PDA-derived ratings are no better than evaluative ratings. However, in both of these studies, the outcome levels used for the distributional ratings were worded in evaluative rather than descriptive terms. As a result, neither study provides an adequate test of the relative accuracy of descriptive PDA ratings as opposed to evaluative ratings.

Second, neither study examined the validity of the multiple performance measures. Given that one of the most significant features of the PDA is the ability to develop measures of different parameters of performance, an important question to be examined is whether those estimates are equivalent to "true" performance parameters. An assessment of equivalence would involve a comparison of PDA performance measures with true-score measures of those same parameters. For example, if PDA measures of consistency of performance are construct valid, they should correlate with objective measures of performance consistency (e.g., the within-employee standard deviation of performance).

In addition, criterion-related validities for the PDA and true-score parameter estimates should be compared. Although there have been no empirical evaluations of the criterion-related validity of the different PDA performance

parameters, Kane (1982) provided some insight for making predictions. Kane argued that the central tendency (average) of a performance distribution reflects primarily the influence of ability on performance—that is, average outcomes reflect the performer's "fixed ability level in multiplicative combination with his or her average motivation level" (Kane, 1984, p. 241). In contrast, the variation-based parameters are idiosyncratic and reflect the influence of effort and/or extraneous constraints on performance—that is, variation around the average level of performance reflects fluctuations due to motivation and situational constraints. Therefore, ability should be differentially related to these two performance parameters such that ability is more strongly related to central tendency parameters than to variation-based parameters.

The third limitation is that the usefulness of the PDA in a real organizational setting is untested. An important question to be examined is whether raters in a real work organization are sensitive to performance variability and can provide reliable distributional ratings. Currently, there is skepticism about whether raters in real-world settings are capable of distributional rating accuracy (e.g., Murphy & Cleveland, 1995). Both Jako and Murphy (1990) and Steiner et al. (1993) argued that future research on the usefulness of the PDA needs to be conducted in field settings. The real value of the PDA depends on the ability of organizational raters to accurately estimate performance distributions, and some of the critical issues to be examined are the extent to which raters use actual samples of on-the-job performance when making occurrence rate estimates and whether they can provide accurate distributional ratings in light of the competing demands on their time.

PRESENT RESEARCH

The study reported here examines the validity of the PDA rating method in a field setting. The performance domain of interest here was quantity of production, and rating validity was evaluated using objective and subjective performance data. The focus on a single dimension of performance is important, inasmuch as past research suggests that objective and subjective measures of performance are more likely to converge when the performance dimension is held constant (e.g., Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Smith, 1976).

Performance ratings for a cohort of new hires were collected after approximately 6 months on the job, and the raters (supervisors) used both the PDA format and an evaluative rating format. Prehire ability data were used to

provide a more thorough analysis of PDA validity and to better understand the multiple parameters of performance derived from the PDA. Based on existing research, the following three hypotheses were examined:

Hypothesis 1: There is a strong convergence between PDA and analogous objective measures of the three performance parameters.

This hypothesis examines the extent to which the PDA estimates of mean performance, consistency of performance, and NRA will be significantly related to the objectively measured counterparts. We expect to find high convergence due to the matched specificity of the performance dimension and performance parameter (e.g., Binning & Barrett, 1989; James, 1973; Smith, 1976), the derived nature of the performance measures, and the salience of production for both ratees and raters. Whereas the convergence between the consistency of performance estimates reflects the reliability of the actual (occurrence) ratings, the convergence between the mean performance and NRA estimates reflects the accuracy of the derived performance scores.

Hypothesis 2: The convergence between PDA and objective measures of typical (average) performance is stronger than the convergence between evaluative and objective measures of typical performance.

This hypothesis contradicts the findings of Jako and Murphy (1990) and Steiner et al. (1993). We expect to find higher convergence using the PDA due to the descriptive nature of the PDA ratings. In this study, the PDA outcome levels were described using nonevaluative anchors to provide a more thorough comparison between the PDA format and a traditional evaluative format.

Hypothesis 3: The relationship between ability and mean performance measures is greater than the relationship between ability and consistency of performance measures, regardless of the method of performance measurement (i.e., PDA vs. objective measures).

This hypothesis examines the extent to which the different performance parameters are differentially related to ability.

METHOD

The data reported here were gathered as part of a larger study conducted

for the Virginia Employment Commission to evaluate the validity generalization scoring procedures for the General Aptitude Test Battery (GATB). The present study is not concerned with validity generalization procedures but, rather, the construct validity of the PDA methodology.

PARTICIPANTS

Performance and ability data were collected for sewing machine operators employed at five nonunionized garment manufacturing plants in the Southeast. All five plants were owned by the same company, produced the same kind of garments, used similar equipment and operating procedures, and operated under a uniform set of management practices, policies, and record keeping. The samples sizes varied across the plants (ns = 80 to 338), but the demographic and ability characteristics were quite similar (100% female, 73% White, 27% Black, 53% married, 65% no previous work experience in this industry, average GATB score of 100.29).

The analyses reported here were conducted using a combined sample of the five plants, which initially consisted of 932 operators hired during a 10-month period. Due to turnover and missing data, the data from only 397 operators are reported here. The characteristics of this final sample (67% White, 33% Black, 50% married, 67% no previous experience in this industry, average GATB score of 100.65) were similar to those of the initial cohort.

PROCEDURE

Performance data were collected using objective and subjective methods of measurement. The objective measures were reported on a weekly basis, and the subjective measures were collected after approximately 6 months on the job. Ability data were prehire measures of ability; however, these data were not used for selection purposes, nor were supervisors cognizant of employees' test scores.

Two performance appraisal instruments were developed for this study: an evaluative rating instrument and a PDA instrument. The evaluative instrument was based on a job analysis and included six dimensions of operator performance (quantity of work, quality of work, flexibility, receptivity, dependability, and work attitudes), plus a single-item overall rating. The dimensional rating scales were 5-point scales with graphic-type anchors that described each performance dimension in terms of company-specific operator goals. The rating instrument was developed with the input of various plant officials (i.e., management and industrial engineering department personnel). The PDA instrument was developed for only one performance dimension—

quantity of work. Based on discussions with plant officials, eight feasible outcome levels were identified and described in terms of company-specific production levels.

After the rating instruments were developed, feedback sessions were conducted with groups of supervisors (raters) to ensure that the rating instruments were relevant, unambiguous, and accurate. At a later date, rating sessions were conducted with groups of supervisors to train them and collect their performance ratings. Rater training was provided by two human resource management experts (a faculty member and a doctoral student), used a lecture-type format with question-and-answer discussion, and focused on the importance of rating accuracy. The training portion of these sessions lasted approximately 1 hour.

The actual rating session immediately followed rater training, and supervisors were told to focus on the performance of ratees during the past 6 months and to omit ratings for any employees they felt they had not adequately observed. Supervisors then rated their immediate subordinates using both the evaluative format and the PDA format. During both the training and rating sessions, supervisors were informed and assured that their ratings were to be used for research purposes only.

PERFORMANCE DATA

Production output (quantity) was the most critical dimension of sewing machine operator success, and individual operators had substantial control over their own production rates. In addition, output was diligently measured, and both operators and supervisors were aware of operators' production rates. Three types of performance data were collected: objective data, PDA ratings, and evaluative ratings.

Objective performance data. The objective performance data were a major source of interest because they served as the benchmark for evaluating rating validity. Objective output was measured using average hourly production (piece-rate) earnings per week—that is, total production earnings divided by the number of hours actually worked. Because the reported production earnings did not include any minimum wage guarantee, time not worked, or rework time, they represent a relatively pure measure of operator production output. Furthermore, the piece-rate standards were determined by industrial engineering studies typical of the industry. Within the limits of error in those engineering studies, jobs were equated along a common scale, and differences in production earnings reflected differences in production output performance.

Three objective performance variables were measured for each operator. Mean production performance (MPP-OBJ) was measured using the average hourly production earnings for Weeks 13 through 24. Consistency of production (COP-OBJ) was measured using the within-subject standard deviation of average hourly production earnings for Weeks 13 through 24. Negative-range avoidance (NRA-OBJ) referred to the proportion of time that production performance had positive utility and was calculated as the percentage of weeks during Weeks 13 through 24 that average hourly production earnings were greater than $3.45.

We chose Weeks 13 through 24 as the time frame for analysis because the company considered the first 12 weeks to be on-the-job training. In addition, we based our measure of NRA-OBJ on the company's explicit minimum standard of average hourly production earnings. According to company officials, if an operator was earning less than $3.35, the company had to pay "make-up pay," which they considered to have negative utility. Due to the nature of the PDA rating instrument and the defined outcome levels (discussed below), we rounded the minimum standard to $3.45 so that we could directly compare the objective and PDA-derived measures of NRA. Operators were allowed 12 weeks to meet the minimum standard, after which the company expected the operators to perform (earn) above the minimum standard 100% of the time.

PDA performance data. The focal job function for the PDA rating was also production output, which was defined in terms of average hourly production earnings. Eight outcome levels were described using average hourly production earnings as anchors. The anchor descriptions were production earnings rather than behaviors because (a) production earnings reflected outcomes, which were the focus of the study; and (b) operators, supervisors, and other plant officials normally described production performance in terms of production earnings. The outcome levels ranged from earnings that were *Less than $3.00* (Level 0) to *$6.50 and over* (Level 7). The PDA ratings consisted of supervisors estimating the percentage of time that employees performed at each of the eight outcome levels, with the percentages summing to 100%.

Three PDA-derived performance variables were measured for each operator. Mean production performance (MPP-PDA1) referred to the average level of production earnings and was calculated by dividing the weighted sum of outcome-level ratings by 100. The weights were the midpoint value (i.e., production earnings) of each outcome level. Because the ratings were coded as whole numbers, dividing by 100 yielded a value equivalent to the average hourly production earnings, but expressed as supervisor ratings. Consistency

of production (COP-PDA1) referred to the perceived variability in performance and was calculated using the formula for the standard deviation of grouped/frequency data (cf. McClave & Benson, 1988). Negative-range avoidance (NRA-PDA1) referred to the estimated proportion of time that production earnings had positive utility and was calculated by summing the outcome-level ratings for Level 2 through Level 7 (i.e., these outcome levels were greater than $3.45) and dividing by 100.

Kane (1984, 1986) provided different formulas for these three indices of performance. His formulas include estimates of utility and feasible performance distributions as the basis for developing effectiveness scores that are comparable across different jobs. Because our study focused on only one job and because the PDA ratings already included utility values (production earnings), we did not need to estimate those values. However, in the interest of obtaining a thorough examination of the PDA methodology, we computed mean production performance, consistency of production, and NRA using Kane's formulas and compared those measures to our simpler, more direct measures.

Development of the additional PDA measures of performance followed procedures detailed in Kane (1986). Effectiveness of mean performance (MPP-PDA2) was calculated by dividing the weighted sum of actual employee ratings by the weighted sum of maximum feasible performance ratings. The weights reflected how valuable the described level of performance was to the company, using the highest level as a reference point worth 100 points. We assigned –100 points to the lowest defined level of performance (less than $3.00) and then computed the intermediate utility weights using the formulas in Kane. This resulted in the following utility weights for the eight levels of performance: –100, –50, –25, 0, 50, 67, 83, and 100. The numerator for MPP-PDA2 is the sum of each employee's percentage rating (expressed as a proportion) multiplied by the corresponding utility rate. The denominator in Kane's methodology is the sum of proportion by utility weights for a hypothetical maximum feasible performance level, the latter determined by job experts.

We deviated slightly from Kane's (1986) methodology at this point because we had the actual distribution of earnings data and thus did not need to estimate maximum feasible performance. That is, we used the proportions of employees whose earnings fell at each of the eight levels to compute the maximum feasible performance. This assumes that the actual earnings really were the maximum feasible for this sample, which is probably an underestimate to some unknown degree. However, this would not affect our results because it would merely decrease MPP-PDA2 ratings by a constant and would have no effect on our analyses.

Consistency of performance (COP-PDA2) was calculated using the formula in Kane (1986):

$$\frac{S_L - S_{ACT}}{S_L - S_M}$$

where:

S_L = the standard deviation of the least consistent distribution,
S_{ACT} = the standard deviation of each employee's distribution, and
S_M = the standard deviation of the most consistent distribution.

In the present study, we used the actual lowest (0.11) and highest (1.04) standard deviations for the distribution ratings over the entire sample, rather than calculating them with the procedures outlined in Kane (1986). Lastly, negative-range avoidance (NRA-PDA2) was calculated by summing the proportions of the levels of performance that had positive utility weights (Levels 4 through 7).

Evaluative performance data. Although evaluative ratings were collected on six dimensions of operator performance, only one dimension was of interest here—quantity of work. This dimension was defined as the typical, or average, output, and supervisors evaluated operators on a 5-point rating scale. Each of the scale anchors described production output with respect to observable operator outcomes and agreed-upon production goals (e.g., 1 = *Production is at or below minimum standard* and 5 = *Production is among the best in the plant*).

ABILITY DATA

The ability measure used in our analyses was an equally weighted raw-score composite of cognitive, perceptual, and psychomotor abilities, as measured by the GATB (United States Employment Service, Department of Labor). We chose to use a composite measure of ability because there was no basis for expecting or predicting different ability factors to differentially relate to the different performance parameters. Our interest was not in performance prediction per se but, rather, the differential relationships between ability and the multiple performance parameters.

ANALYSES

Product-moment correlations were computed to determine the direction and strength of relationships between all of the study variables. Generally

speaking, the stronger the correlations between the objective and subjective measures of performance, the greater the convergent validity, and the stronger the evidence of rating validity.

RESULTS

Table 1 summarizes the descriptive statistics and intercorrelations of study variables. Of note in Table 1 is that there is a high degree of convergence between Kane's (1986) formulations of the PDA ratings (indicated by the PDA2 extensions) and those formulations based on the economic parameters supplied by the host organization (indicated by the PDA1 extensions). The average correlation between analogous PDA1 and PDA2 parameters was .76 ($p < .01$). Thus, Kane's formulations, designed for use when the dollar value of performance is ambiguous, show good convergence with formulations when the dollar value of performance is well known.

Hypothesis 1 stated that the PDA ratings of performance would converge with the objective measures of performance. The correlations between actual mean production earnings and the two MPP-PDA measures were .59 ($p < .01$). Similar results were found for NRA, in which actual NRA correlated .48 and .50 ($p < .01$) with the NRA-PDA formulations. These findings support the proposition that the PDA ratings would be accurate measures of typical and NRA parameters of production performance.

The consistency of performance indices (COP-PDA1 and COP-PDA2) were significantly correlated with the actual standard deviation of performance (.10 and .09, $p < .05$, respectively). However, these correlations were not substantial in magnitude as compared with the correlations for the other performance parameters. Because a standard deviation is a sum of squares about the mean (for each subject), we further examined the convergence between consistency measures in a way that would reflect whether raters observed how often each subject performed at each of the eight levels of performance outcomes, irrespective of the subject's average level of performance. Specifically, we correlated PDA ratings for each subject, for all eight levels of performance, with the actual proportion of weeks that each subject actually performed at each of the eight levels (out of the 12 total weeks). The results from this analysis are reported in Table 2.

As can be seen in Table 2, all eight of the correlations between rated and actual time spent performing at the eight levels of performance were statistically significant ($p < .001$). The average correlation was .29, which is

TABLE 1

Descriptive Statistics and Intercorrelations of Study Variables

Variable	*Mean*	SD	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1. MPP-OBJ	4.20	1.04	1.00									
2. COP-OBJ	1.44	0.76	.24	1.00								
3. NRA-OBJ	0.74	0.35	.84	.11	1.00							
4. MPP-PDA1	4.11	1.03	.59	.10	.51	1.00						
5. COP-PDA1	3.98	2.65	.03	.10	−.07	−.07	1.00					
6. NRA-PDA1	0.77	0.36	.47	.01	.48	.85	−.18	1.00				
7. MPP-PDA2	0.50	3.10	.59	.08	.54	.98	−.10	.86	1.00			
8. COP-PDA2	0.37	0.19	.12	.09	.12	.26	.63	.32	.22	1.00		
9. NRA-PDA2	0.45	0.41	.56	.07	.50	.83	−.06	.66	.92	.08	1.00	
10. GRS	2.58	1.05	.66	.13	.56	.58	.07	.52	.60	.24	.55	1.00
GATB	100.68	10.35	.24	.11	.21	.26	.09	.27	.27	.19	.22	.27

NOTE: $N = 397$. MPP-OBJ = actual quantity of production; COP-OBJ = actual consistency of performance; NRA-OBJ = actual negative-range avoidance; MPP-PDA1 = performance distribution assessment (PDA) mean production performance; COP-PDA1 = PDA consistency of performance; NRA-PDA1 = PDA negative-range avoidance; MPP-PDA2 = Kane mean production performance; COP-PDA1 = Kane consistency of performance; NRA-PDA1 = Kane negative-range avoidance; GRS = graphic rating of quantity; GATB = General Aptitude Test Battery composite of cognitive, perceptual, and psychomotor abilities.
$^{*}p < .05$. $^{**}p < .01$.

TABLE 2

Relationships Between Performance Distribution Assessment (PDA) Ratings and Actual Proportions of Time Operators Performed at Eight Levels of Performance

Level	*Range*	*Mean PDA Rating*	*Mean Actual*	*Correlation of Rating and Actual*
0	Less than $3.00	.09	.15	.41***
1	$3.00-3.45	.14	.12	.18***
2	$3.46-3.90	.15	.15	.24***
3	$3.91-4.30	.17	.13	.16***
4	$4.31-4.74	.20	.13	.25***
5	$4.75-5.60	.19	.22	.36***
6	$5.61-6.49	.04	.09	.38***
7	$6.50 and over	.02	.02	.30***

***$p < .001$.

substantially higher than the correlations between the actual and PDA-derived standard deviations of performance (.09 and .10, respectively). Interestingly, the convergence between rated and actual proportions of time tended to be higher at the extremes in performance outcomes (Levels 0, and Levels 4 through 7) than the middle levels. Apparently, raters were more accurate at estimating the performance levels of outliers than they were at estimating the midrange levels of performance (consistent with some recent laboratory and field research; Sanchez & De La Torre, 1996; Varma, DeNisi, & Peters, 1996). Overall, our first hypothesis received support; the PDA occurrence ratings were reliable estimates of performance variability.

Hypothesis 2 stated that the PDA ratings would be more accurate measures of typical performance than the evaluative rating. As shown in Table 1, the correlation between the evaluative and actual measures of typical performance was .66 ($p < .01$), whereas the analogous correlations were .59 for the two PDA formulations. While all three ratings showed substantial convergence with actual performance, neither the PDA ratings nor the GRS ratings appeared psychometrically superior to the other. Thus, our second hypothesis was not supported.

Hypothesis 3 stated that ability would be more strongly related to mean levels of performance than to consistency of performance. As shown in Table 1, the correlations between the GATB and mean performance (PDA and objective) ranged from .24 to .27 and were stronger than those between the GATB and the measures of consistency of performance (.09 to .19). This finding is

consistent with Kane's (1982) argument that typical performance levels, as opposed to performance variability, are primarily a function of employee ability. Thus, our third hypothesis was supported.

DISCUSSION

The findings reported here provide somewhat mixed support for the stated hypotheses. Although we found significant convergence between the objective and PDA measures of the multiple performance parameters and although we found similar correlations between ability and the matched objective and PDA performance measures, we did not find support for the hypothesis that the PDA measure of typical performance would be better than the evaluative measure.

Together, these findings represent good news and bad news. The good news is that performance can be accurately rated in a field setting; the bad news is that, on the surface, rating methods do not seem to make much of a difference when rating typical performance. In contrast to other format comparison studies, these findings are significant because (a) accuracy was benchmarked using actual production levels rather than expert ratings and (b) accuracy was examined in a setting in which rater-ratee relationships had been established and job environment demands and influences were present. Although this study appears to provide support for the generally accepted finding that appraisal methods do not matter when it comes to rating accuracy (e.g., Landy & Farr, 1980), there are several considerations that make such a conclusion unwarranted.

If the only performance parameter of interest is typical performance, our findings suggest that valid and accurate measures can be developed using either an evaluative or a distributional (descriptive) rating method provided that the frame of reference is explicitly and carefully defined. Previous studies have found that less than 25% of the variance in ratings can be accounted for by objective indices (see Cascio & Valenzi, 1978); however, stronger relationships should be evidenced when using rating formats with greater behavioral specificity and objective measures that are reliable and generally free from the effects of various situational constraints (Bernardin & Beatty, 1984). Our findings support this prediction. In this study, the focal criterion was quantity of production, and the rating anchors used in both rating instruments were outcome oriented. Interestingly, the average correlation between the actual quantity of production and the various ratings of quantity of performance in our study (.61) was similar to the average found by Bommer et al.

(1995) in their meta-analysis of matched objective and subjective measures of dimensional performance (.60; p. 596). It could be argued that there is no need for ratings when reliable objective performance data are readily available. However, this type of research could be used to support an argument that ratings on other nonverifiable dimensions of performance are also valid (Bernardin & Beatty, 1984), thus a means of establishing rater validity.

Although we expected the descriptive nature of the rating task to be a determinative factor in rating validity, the outcome-anchored rating scales used in both of our rating methods (PDA and GRS) probably obscured some of the expected differences in the rating methods. The evaluative rating instrument used here was carefully developed using managerial input, and the rating scale was defined using observable, outcome-oriented descriptors. Because it was neither a traditional graphic rating scale nor a behaviorally anchored rating scale, it is difficult to compare our results to prior research. In addition, the rating situation in this study was atypical of most laboratory or field-setting studies: The raters had both the opportunity and the responsibility to monitor production on an ongoing basis throughout the day. Production quantity was a highly salient performance factor for this company, and we expect that both supervisors and workers were quite cognizant of production levels of everyone in their work groups. This is in contrast to typical work situations in which standards for performance are ill defined and/or supervisors do not have the opportunity to observe employee performance on a daily basis. It could be argued that both types of ratings collected in this study were confounded, or biased, by the raters' heightened knowledge of production standards and actual production levels. However, this type of rating bias is exactly what rater training programs are intended to achieve.

Our inability to find differences in accuracy for the different rating methods actually provides support for the arguments made by James (1973), Smith (1976), and Binning and Barrett (1989) regarding performance construct validity: Equivalent performance measures can be developed if criterion specificity, time span of performance, and closeness to organizational goals are matched. In this study, our performance measures (objective, PDA, and evaluative) were matched in terms of a specific time period for evaluation (6 months), a specific job function (quantity of production), and specific job outcomes (production earnings). Furthermore, our results suggest that the outcome-anchored ratings represented a rating feature that provided an effective mechanism by which raters could accurately evaluate performance. Both rating formats defined production objectively, which permitted raters to rely on their natural cognitive processes as a means of processing information (Borman, 1991; Feldman, 1986).

Future research needs to be conducted to determine whether accurate measures of typical performance can be developed in more complex jobs for other dimensions of performance in which outcomes are not so readily observable (e.g., quality), perhaps using nonsupervisory raters. Although the simplistic nature of the rating task in our study (i.e., routine job, one dimension of performance, observable performance indicators, readily available production reports) allowed us the opportunity to examine the validity of the PDA in a controlled work setting, the usefulness of this method requires additional research in other job settings.

If performance parameters other than typical performance are of interest, then our findings suggest that the PDA represents the more useful method of appraisal. In this study, we found that raters were able to provide reliable information that yielded accurate measures of performance consistency, NRA, and typical performance. In terms of rating validity, we found evidence for both the internal and the external validity of the multiple performance parameters. These findings suggest that if the "conceptual criterion" (Astin, 1964) refers to the consistency and/or utility of performance over time, then the PDA represents a means of developing criterion measures that more faithfully depict the important parameters of interest.

Borman (1991) argued that performance ratings have several potential advantages, including (a) being sensitive to ratee performance over time and across different job situations, (b) being a flexible means for indexing performance on almost any dimension, and (c) using actual job performance as input into the evaluation of performance. However, he further argued that until rater accuracy can be improved, these potential advantages cannot be realized. In our study, we found that raters were capable of realizing the potential advantages described by Borman. In addition, the PDA system has the additional advantage of being a flexible means of indexing performance on multiple parameters.

Future research needs to examine the validity of the multiple performance parameters in other job situations in which performance variability is monitored and in which its minimization is of critical importance. We have demonstrated that raters are capable of rating performance variability; future research needs to address the question of "so what?" in a setting in which variability of performance is as important as quantity of production (typical performance) was in our research setting. Future research should also examine whether rater training can increase the correlations between distributional ratings and actual employee performance distributions. Follow-up questionnaires or interviews could be conducted to determine what types of information raters use when making distributional ratings and the perceived usefulness of the PDA from the rater's perspective. This type of qualitative research

could also be used to examine the user-friendliness of the PDA versus simpler rating methods. Bernardin and Beatty (1984) identified several potential problems with the PDA method, which include increased administrative involvement and questionable information processing advantages. Therefore, a comprehensive study of the utility of the PDA method should directly address these concerns in addition to the psychometric qualities examined here.

In light of the above considerations, we can conclude that the PDA method did yield valid and accurate measures of the multiple production performance parameters. From a research perspective, the PDA provides a means of mapping individual performance over time, thus a method for examining dynamic performance criteria at the individual level of analysis (e.g., Hofmann et al., 1993; Murphy, 1989). Feasible performance distributions represent the extent to which performance varies over time. When used in conjunction with information about individual differences in ability and motivational factors, researchers can begin to examine and better understand the causes and consequences of interindividual differences in performance variability over time.

IMPLICATIONS FOR MANAGERS AND PRACTITIONERS

From a practical perspective, the PDA fits well with the popular corporate philosophy of TQM. One of the fundamental premises of TQM is that performance appraisal systems in most organizations are dysfunctional because they fail to distinguish between performance variability caused by the system or organization (*common causes*) and performance variability caused by the employees (*local causes*). If, in fact, most performance variability is due to common causes, as suggested by Deming (1986), it is unjust to hold employees accountable for variations due to factors out of their control. The PDA methodology explicitly accounts for what is feasible in terms of employee performance; therefore, it can be adapted to factor out known common causes and consider primarily local causes that are under employees' control. In this respect, organizations can hold employees accountable for the effects of controllable variations in performance (Kane, 1982, 1986). The PDA method could also be used to develop performance control charts that would enable organizations to track individual performance variability and identify occurrences of negative-range performance (i.e., unacceptable or poor performance levels). Depending on the nature of the job and the degree to which performance is determined by group, as opposed to individual, effort, these performance control charts can be used to measure and monitor a range of group performance indices.

In this study, we found that raters in a field setting were sensitive to on-the-job performance variations and were able to provide valid information about employee performance. This finding has several important managerial implications for PDA as a performance management tool.

First, because the PDA explicitly measures consistency of performance, it allows managers to provide more meaningful performance-related feedback and coaching to subordinates in work environments in which consistency is paramount. Recent surveys regarding employee satisfaction with performance appraisals reveal that one of the most common sources of dissatisfaction is the lack of meaningful feedback (Lee, 1996), which is one of the reasons Deming (1986) condemned traditional appraisal methods. In the spirit of a true performance management system, the PDA method can be used to identify common and local variations in performance; follow-up meetings with employees can then be conducted to determine the causes of those variations. In this sense, the PDA method represents a dual feedback system in which employees receive feedback about their performance change patterns over time and the organization receives feedback from employees about structural as well as motivational constraints on performance.

Second, because NRA is explicitly measured, raters may be more motivated to accurately observe employee performance to ensure that employees are not costing the company real dollars. In addition, raters may be more inclined to provide accurate ratings and not use the performance appraisal system for such other purposes as politics and favoritism (Longenecker et al., 1987). In theory, performance management systems should provide a direct linkage between individual and organizational goals; in practice, this linkage is unclear (Antonioni, 1994; Lawler, 1994; Lee, 1996; Markowich, 1994). The NRA measure is perhaps the most direct means of linking the organization's appraisal system with the larger system of organizational goals and effectiveness.

In this study, company officials were surprised to learn that there were any employees performing in the negative range; company policy was that, after 12 weeks, all employees should be performing above $3.35 per hour 100% of the time. They were even more surprised to find out that this was occurring at most of the plants. This information led the company to begin a study to determine (a) why employees were not in the NRA levels 100% of the time (e.g., a thorough examination of on-the-job training procedures), (b) why negative-range performance occurred across the different plants (e.g., outdated and/or inefficient equipment), and (c) why plant managers had not rectified the problem (e.g., lack of information about the severity and pervasiveness of the problem). In the spirit of TQM, a simple measure of NRA could be the stimulus for investigating the causes of both system and person

problems regarding performance.

Third, to the extent that all of the PDA performance parameters are measured, performance can be more effectively tied to the organization's reward system. Those employees who actually contribute the most to achieving organizational goals (as opposed to getting the highest rating) could be commensurately rewarded. One of the major problems with pay-for-performance (PFP) plans lies in the nature and perceived accuracy of the performance measurement procedure. If the purpose of PFP is to recognize and reward those employees who contribute to the overall success of the organization, then success should be defined in terms of not only multiple dimensions of performance but also multiple parameters of performance.

For instance, TQM-oriented companies espouse the importance of continuous improvement, and performance management systems are supposed to be the means of monitoring and facilitating performance improvement efforts. Using the PDA method, performance improvement can be gauged using all of the distributional measures, and continuous improvement can be monitored in terms of individual, work-group, subunit, and organizational distributional measures over time. Depending on the goals of the organization, both individual and group-based rewards can be directly linked to those performance parameters that are directly linked to the organization's goals. If the PDA method is developed and administered using employee involvement programs and frequent work-performance-review sessions, it is likely to be perceived as an effective and fair procedure for implementing PFP programs.

Fourth, the PDA measures of performance provide a means of directly linking the performance management system with training needs assessment. To the extent that employees' mean performance levels are deficient and/or declining, job-related abilities might be the source of the problem. If so, improving the selection system and/or providing training would be viable solutions. On the other hand, if consistency of performance is the major performance-related problem, motivational and/or situational constraints might be the cause. In this case, organization design factors (i.e., organizational structure and culture, job design and opportunity biases, goal setting and reward systems) need to be examined. The different parameters of performance suggest different likely causes of performance problems. Just as training needs assessment is conducted at three levels of analysis (organization, job, person), the PDA method can be used to measure the different performance parameters at these same three levels of analysis. And just as training needs assessment is used to determine whether there is, in fact, a need for training, the PDA method can be used to provide the type of data necessary for such a determination. Thus, for organizations that are commit-

ted to performance management, the PDA methodology is an invaluable method for developing and administering a true performance management and development system.

CONCLUSION

In conclusion, we agree with Steiner et al. (1993) that distributional ratings hold promise for both research and practice. Future research needs to be conducted to determine the extent to which our findings about PDA validity generalize to other organizational settings. In addition, future research should continue to benchmark the relative accuracy of the PDA with more commonly used organizational rating methods, with one caveat: To the extent that typical performance is the conceptual criterion of interest, such comparisons are meaningful. However, if the conceptual criterion is a different performance parameter, such as variability of performance, then such method comparisons are meaningless to the extent that other ratings methods do not directly measure parameters other than typical performance.

NOTE

1. Of course, raters can always distort the ratings without this knowledge. For example, in the case of leniency errors, they could rate an employee as performing at the highest outcome level 100% of the time. However, if the rater is held accountable for the ratings produced (cf. Harris, 1994), then it would be more difficult to provide evidence that the ratee did, in fact, perform at the highest levels of performance all of the time. There would be others within the organization (e.g., coworkers) readily available to dispute the erroneous rating. This is in stark contrast to graphic rating scales, in which the rater is largely free to justify ratings with any sample of ratee performance that the rater chooses. We do not assert that the PDA methodology is error proof, just that it should be less so than most conventional methods of performance appraisal. In Kane's (1994) terminology, PDA lowers P_{xj} (probability of producing volitional errors) and increases P_A (probability of rating accurately).

REFERENCES

Antonioni, D. (1994, May-June). Improve the performance management process before discontinuing performance appraisals. *Compensation & Benefits Review, 26*, 29-37.

Astin, A. W. (1964). Criterion-centered research. *Educational and Psychological Measurement, 24*, 807-822.

Austin, J. T., Villanova, P., Kane, J. S., & Bernardin, H. J. (1991). Construct validation of performance measures: Definitional issues, development, and evaluation of indicators. In K. M. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 9, pp. 159-233). Greenwich, CT: JAI.

Bernardin, H. J., & Beatty, R. W. (1984). *Performance appraisal: Assessing human behavior at*

work. Boston: Kent.

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Bommer, W. H., Johnson, J. L., Rich, G. A., Podsakoff, P. M., & MacKenzie, S. B. (1995). On the interchangeability of objective and subjective measures of performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists.

Cascio, W. F., & Valenzi, E. R. (1978). Relations among criteria of police performance. *Journal of Applied Psychology, 63*, 22-28.

Deming, W. E. (1986). *Out of the crisis*. Cambridge, MA: MIT Initiative for Advanced Engineering Study.

Dobbins, G. H., Cardy, R. L., & Carson, K. P. (1991). Examining fundamental assumptions: A contrast of person and system approaches to human resource management. In K. M. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 9, pp. 1-38). Greenwich, CT: JAI.

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K. M. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 4, pp. 45-99). Greenwich, CT: JAI.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*, 737-756.

Hofmann, D. A., Jacobs, R., & Baratta, J. E. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology, 78*, 194-204.

Jako, R. A., & Murphy, K. R. (1990). Distributional ratings, judgment decomposition, and their impact on interrater agreement and rating accuracy. *Journal of Applied Psychology, 75*, 500-505.

James, L. R. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin, 80*, 75-83.

Kane, J. S. (1982, November). *Rethinking the problem of measuring performance: Some new conclusions and a new appraisal method to fit them*. Paper presented at the Fourth Johns Hopkins University National Symposium on Educational Research, Washington, DC.

Kane, J. S. (1984). Performance distribution assessment: A new breed of appraisal methodology. In H. J. Bernardin & R. W. Beatty, *Performance appraisal: Assessing human behavior at work* (pp. 325-341). Boston: Kent.

Kane, J. S. (1986). Performance distribution assessment. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 237-273). Baltimore: Johns Hopkins University Press.

Kane, J. S. (1994). A model of volitional rating behavior. *Human Resource Management Review, 4*, 283-310.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

Lawler, E. E. (1994, May-June). Performance management: The next generation. *Compensation & Benefits Review, 26*, 16-19.

Lee, C. (1996, May). Performance appraisal: Can we "manage" away the curse? *Training, 33*, 44-59.

Longenecker, C. O., Sims, H. P., & Gioia, D. A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1*, 183-193.

Markowich, M. M. (1994, May-June). We can make performance appraisals work. *Compensation & Benefits Review, 26*, 25-28.

Masterson, S. S., & Taylor, M. S. (1996). Total quality management and performance appraisal: An integrative perspective. *Journal of Quality Management, 1*, 67-89.

McClave, J. T., & Benson, P. G. (1988). *Statistics for business and economics* (4th ed.). San Francisco: Dellen.

Murphy, K. R. (1989). Is the relationship between cognitive ability and job performance stable over time? *Human Performance, 2*, 183-200.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.

Sanchez, J. I., & De La Torre, P. (1996). A second look at the relationship between rating and behavioral accuracy in performance appraisal. *Journal of Applied Psychology, 81*, 3-10.

Smith, P. (1976). Behaviors, results, and organization effectiveness: The problem of criteria. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 745-775). Chicago: Rand McNally.

Steiner, D. D., Rain, J. S., & Smalley, M. M. (1993). Distributional ratings of performance: Further examination of a new rating format. *Journal of Applied Psychology, 78*, 438-442.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73*, 497-506.

Varma, A., DeNisi, A. S., & Peters, L. H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology, 49*, 341-360.

Zuroff, D. C. (1989). Judgments of frequency of social stimuli: How schematic is person memory? *Journal of Personality and Social Psychology, 56*, 890-898.

Diana L. Deadrick is an associate professor of management at Old Dominion University. Her research interests include performance management and organization change and development.

Donald G. Gardner is a professor of management at the University of Colorado at Colorado Springs. His research interests include oganizational behavior and teams.