

Understanding the Latent Structure of Job Performance Ratings

Steven E. Scullen
North Carolina State University

Michael K. Mount
University of Iowa

Maynard Goff
Personnel Decisions International

This study quantified the effects of 5 factors postulated to influence performance ratings: the ratee's general level of performance, the ratee's performance on a specific dimension, the rater's idiosyncratic rating tendencies, the rater's organizational perspective, and random measurement error. Two large data sets, consisting of managers ($n = 2,350$ and $n = 2,142$) who received developmental ratings on 3 performance dimensions from 7 raters (2 bosses, 2 peers, 2 subordinates, and self) were used. Results indicated that idiosyncratic rater effects (62% and 53%) accounted for over half of the rating variance in both data sets. The combined effects of general and dimensional ratee performance (21% and 25%) were less than half the size of the idiosyncratic rater effects. Small perspective-related effects were found in boss and subordinate ratings but not in peer ratings. Average random error effects in the 2 data sets were 11% and 18%.

Job performance is a fundamentally important construct in organizational practice and research. From a practical perspective, it plays a central role in most personnel decisions, such as merit-based compensation, promotion, and retention. It is also used as an important source of developmental feedback. From a theoretical perspective, researchers have long been interested in understanding the causal mechanisms that lead to effective job performance.

Although job performance has been measured in many ways (e.g., volume of sales, quantity or quality of items produced, absences, number of promotions), the most frequently used measure is a supervisory performance rating. In recent years, multirater or 360° feedback programs have gained popularity. This means that peer, subordinate, and self-ratings also play an important role in the assessment of job performance. Given the significance of the job performance construct and the growing dependence on ratings from multiple sources as a means for measuring that construct, it is important to identify and quantify the factors that influence performance ratings.

The present study examines the latent structure of job performance ratings. We present a model that contains five factors that are postulated to affect ratings in a multirater feedback program. In

doing so, we address two major objectives. The first is to measure the effects associated with each of the five factors in the model. Other research has examined subsets of the five factors we investigate here, but to our knowledge, none has investigated the effects of all five of the factors simultaneously. This is important, because failure to investigate each of these effects provides an incomplete representation of the latent structure of performance ratings. The second major objective is to examine the influence of each of the five factors on ratings made by raters from four different perspectives (boss, peer, subordinate, and self) and on three dimensions of managerial performance. This allows us to determine whether the effects are consistent across different types of raters and different aspects of performance.

Factors That Influence Performance Ratings

Wherry's theory of rating (Wherry & Bartlett, 1982) indicates that there are three broad types of factors that influence performance ratings: the ratee's actual job performance, various rater biases in the perception and recall of that performance, and measurement error.¹ The model we developed for this study is based on those general distinctions. It was not our intention, however, to model Wherry's detailed mathematical relationships in their entirety, but rather to build a more general model that can be used to measure the ratings variance associated with ratee (performance) effects, rater (bias) effects, and random error. Our conceptualizations of those factors are described in the sections that follow.

Lance's (1994) study of the latent structure of performance ratings is one of the few that have been based on Wherry's

Steven E. Scullen, College of Management, North Carolina State University; Michael K. Mount, Henry B. Tippie College of Business, University of Iowa; Maynard Goff, Personnel Decisions International, Minneapolis, Minnesota.

A version of this article was presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, Georgia, April 1999. We thank Frank Schmidt for his helpful comments on this article.

Correspondence concerning this article should be addressed to Steven E. Scullen, College of Management, North Carolina State University, Raleigh, North Carolina 27695-7229. Electronic mail may be sent to steve_scullen@ncsu.edu.

¹ Wherry's theory explicitly excludes intentional rater distortions. His theory includes the effects of environmental or opportunity biases, but those are not of interest in the current study.

(Wherry & Bartlett, 1982) theory of rating and is therefore relevant to the current study. Lance modeled specific components of Wherry's theory to measure their effects on first- and second-level supervisors' ratings of performance in entry-level clerical and professional-technical personnel. Like Lance's, our study examines the latent structure of performance ratings, but our study differs from Lance's in three important ways. First, Lance's model was patterned specifically after Wherry's theory, whereas ours was not. Second, the four rater perspectives in our study (boss, peer, subordinate, self) represent more diverse viewpoints than do the two perspectives (first- and second-level supervisor) in Lance's study. And finally, the ratees in our study are managers, whereas the ratees in Lance's study were not. Nonetheless, the similarities in design and purpose between our study and Lance's invite a comparison of findings. As we indicate in the Discussion section, there are clear similarities between Lance's conclusions and ours.

Actual Job Performance

This broad category refers to the effects of the actual performance of the ratee on observed performance ratings and consists of two major components. One is the effect of the ratee's general level of job performance. For most ratees, however, the level of performance varies somewhat across job dimensions. The second component of performance accounts for those differences in performance across dimensions. We discuss each in more detail later in the article.

General performance is reflected in a general factor that underlies all judgments of a ratee's performance across all raters and performance dimensions (King, Hunter & Schmidt, 1980). This conceptualization of general performance is related to the concept of true halo (Cooper, 1981). Some degree of true halo is expected, because many of the antecedents of performance (e.g., mental ability and conscientiousness) are similar across the various dimensions of performance (Motowidlo, Borman, & Schmit, 1997). To the extent that ratings reflect actual performance, we expect to find evidence of a general factor in performance ratings.

The second component of actual performance relates performance on a particular dimension to the ratee's general level of performance. Similar to King et al. (1980), we conceptualize this in terms of a deviation from the general level of performance (i.e., as a residual). For those dimensions on which a ratee performs above (or below) his or her general level of performance, this component has a positive (or negative) value. If performance on a particular dimension is equal to the general level of performance, then a value of zero is attached to that dimension.

In our model, then, a ratee's actual level of performance on a given dimension is that ratee's general performance component plus or minus the appropriate dimensional component. Because the dimensional component is defined as a residual from the general performance component, the general and dimensional components are uncorrelated (King et al., 1980).

Ideally, most of the variance in observed performance ratings would be accounted for by the ratee's actual performance. Unfortunately, however, the general conclusion drawn from the literature is that actual job performance has a positive but less than optimal effect on ratings. Meta-analytic evidence suggests, for example, that corrected correlations between ratings and "objective" measures such as quality and quantity of work are only

moderate, ranging from approximately .10 to .40 (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Heneman, 1986; Viswesvaran, 1993). Similarly, two confirmatory factor analyses (CFAs) of objective and subjective (i.e., ratings) measures of jet engine mechanics' performance (Lance, Teachout, & Donnelly, 1992; Vance, MacCallum, Coover, & Hedge, 1988) revealed that ratings and work samples measure similar but not identical constructs.

Rater Biases

This broad category of effects refers to the systematic variance in performance ratings that is associated in some way with the rater and not with the actual performance of the ratee. We distinguish between two major types of rater bias effects. One is idiosyncratic tendencies exhibited by individual raters. These include several types of effects, of which halo and leniency are the most widely researched. *Halo error* refers to the tendency of raters to allow an overall impression of a ratee to influence judgments along several quasi-independent dimensions (King et al., 1980; Lance, LaPointe, & Stewart, 1994). *Leniency error* refers to a rater's tendency to assign ratings that are generally higher (or lower) than are warranted by the ratees' actual performance. Other types of effects (e.g., Rater \times Ratee interaction) also fall into this category. In this article, we refer to the aggregation of all of these as idiosyncratic rater effects. It is important to note that random measurement error is not a component of idiosyncratic rater bias in our model.

To understand the magnitude of the effects of leniency error on observed ratings, it is necessary to make an important but often neglected distinction between crossed and nested ratings. In crossed rating systems, each rater rates the performance of all ratees. In nested rating systems, each ratee's performance is rated by different raters. This distinction is important because the amount of idiosyncratic ratings variance is expected to be larger in a nested design than in a crossed design. The reason is that each rater in a nested design may exhibit a different degree of leniency. Thus, rater leniency differences introduce an element of variability into nested designs that is not present in crossed designs. To our knowledge, the magnitude of the difference in idiosyncratic variance between these designs is not yet known, largely because research has failed to acknowledge the distinctions between the two types of designs.

As we describe later, our data fall into the nested design category. Thus, in this study, both halo and leniency differences contribute to idiosyncratic rating variance. Because our intention in this study was to quantify the overall effect of this category of biases rather than to quantify the unique effects of each specific type of bias, we did not attempt to partition idiosyncratic variance into halo- and leniency-related components.

The defining feature of the effects in the idiosyncratic bias category is that they represent rating variance that is systematic within an individual rater but is not associated with the actual performance of the ratee or with ratings made by other raters. Two recent studies examined idiosyncratic rating variance and found its effects to be substantial. Conway (1996) found that 25% of the observed ratings variance was method related (idiosyncratic), and a meta-analysis by Viswesvaran, Ones, and Schmidt (1996) of intrarater and interrater reliabilities showed that 29% of the observed variance in ratings was idiosyncratic in origin. As we

discuss later, it is often very difficult to determine from a written report of multirater research whether the design was nested or crossed. However, if some or all of the studies included in the Conway and the Viswesvaran et al. data involved crossed data, then the amount of idiosyncratic rater bias in these studies should be less than we report in ours.

The second type of rater bias refers to effects associated with the rater's organizational perspective (self, subordinate, peer, or boss). Several researchers have argued that a rater's perspective may influence performance ratings independently of the idiosyncratic tendencies we have described (e.g., Borman, 1974; Murphy & Cleveland, 1995; Pulakos, Schmitt, & Chan, 1996; London & Smither, 1995; Tsui, 1984). Borman (1997) advanced three reasons why it is plausible to hypothesize that perspective-related biases affect performance ratings. First, raters from different organizational perspectives might focus their attention on different aspects of the ratee's performance. Second, raters from different perspectives might attend to the same aspects of performance but attach different weights to them. Third, raters from different perspectives often observe different samples of a ratee's behavior. Thus, ratings might differ across perspectives because of real differences in the behaviors that are observed. Borman's third point, and perhaps his first as well, suggests that differences across perspectives may not be biases at all. Instead, they may reflect portions of the true performance criterion space (Lance, Woehr, & Fisicaro, 1991) that are unique to ratings from each perspective. For now, we place perspective effects in the rater bias category, but later in the article we discuss in some detail the importance of recognizing and investigating the possibility that these effects represent true performance variance.

Evidence as to the existence of perspective-related effects is equivocal (Pulakos et al., 1996). Mount, Judge, Scullen, Sytsma, and Hezlett (1998) conducted the only published study we know of that simultaneously examined both idiosyncratic and perspective-related effects. They found that idiosyncratic effects are stronger than perspective-related effects, but they did not assess the magnitude of those effects in an absolute sense. The current article extends prior research by separating and quantifying the effects of both idiosyncratic and perspective-related effects.

Random Measurement Error

This factor refers to unsystematic variance in performance ratings. It is important to understand the magnitude of such errors, because random measurement error limits the extent to which measurements are reliable, and that, in turn, limits the validity of inferences made from those measurements. Meta-analytic evidence (Viswesvaran et al., 1996) indicates that the mean coefficient of stability for bosses' ratings of overall job performance (computed from ratings made by the same rater at two different times) is .81. This means that approximately 19% (i.e., $1.00 - .81$) of the variance in performance ratings made by bosses is due to random measurement error, transient error, and other unidentified factors. The present study extends the Viswesvaran et al. findings to three other types of raters—peers, subordinates, and self—and to three types of performance dimensions.

Summary

Wherry and Bartlett (1982) proposed three broad categories of factors that influence performance ratings: actual job performance of the ratee, rater biases, and random measurement error. In this study, we quantify the effects of each of these factors by using an expanded model that includes each of the following: (a) the effect of the ratee's general level of performance, (b) the effects of the ratee's performance on a particular performance dimension, (c) the effects of the rater's idiosyncratic rating tendencies, (d) the effects of the rater's organizational perspective, and (e) the effects of random measurement error. This is the first study to quantify all of these effects simultaneously and to compare them across different rater perspectives and different performance dimensions.

Method

The data for this study were taken from two large, independent data sets. One is new to the research literature, and one was used for related purposes in Mount et al. (1998). The research we report here focuses primarily on the new data, but results based on the Mount et al. data are presented as well. The Mount et al. data were used to examine the robustness of the major conclusions derived from the new data. Although the new data set for this study was developed using methods similar to those used in Mount et al., we emphasize that our rating instrument was different from the one used by Mount et al. and that our sample is completely independent from theirs. Except where otherwise noted, the following description pertains to the newly gathered data.

Participants and Instruments

Participants in this study were 2,142 managers, representing a wide variety of industries, functional areas, and levels of management. The majority of the ratees were White (88%), male (70%), and college graduates (82%). Seven ratings were available for each manager. In addition to providing a self-rating, each manager was rated by two bosses, two peers, and two subordinates. All ratings were made for developmental purposes. Participation was voluntary in most cases, and ratees were allowed to select their raters.

The Profilor, a multirater feedback instrument developed by Personnel Decisions International, Inc. (Hezlett, Ronnkvist, Holt, & Hazucha, 1997), was the instrument used to collect ratings data. The Profilor contains 135 items, grouped by the publisher into 24 scales. The Profilor is based on several decades of consulting experience and research on management. It was developed from a review of the management and psychology literatures, exhaustive analysis of the large Management Skills Profile (MSP) data base, and job analysis questionnaires and interviews of hundreds of managers representing many functional areas and most major industries. The Profilor is intended to represent behavioral performance competencies that are generally required of managers.

Like Mount et al. (1998), we used the conceptually similar frameworks proposed by Mann (1965) and Katz (1974) to identify three theoretically important dimensions of managerial performance—administrative (e.g., planning, organizing, assigning to tasks), human (working with and through people to accomplish objectives), and technical (knowledge of relevant methods and techniques in the functional area). Profilor scales were then associated with performance dimensions as follows.

Six doctoral students majoring in either human resources management or organizational behavior were given definitions for each of the three managerial performance dimensions and a brief description (from the publisher) of each of the 24 Profilor scales. The students were asked to identify the performance dimension that they believed each Profilor scale most clearly represents. Scales that were associated with the same performance

dimension by at least four of the six graduate students were assigned to that dimension. One scale (Leading Courageously) could not be assigned to a performance dimension and therefore was dropped. The remaining scales were assigned as follows: For the administrative dimension, the scales were Establish Plans, Manage Execution, Provide Direction, Coach and Develop, and Champion Change; for the human dimension, the scales were Foster Teamwork, Motivate Others, Build Relationships, Display Organizational Savvy, Manage Disagreements, Foster Open Communication, Listen to Others, Act With Integrity, Demonstrate Adaptability, and Develop Oneself; and for the technical dimension, the scales were Analyze Issues, Speak Effectively, Use Sound Judgment, Drive for Results, Show Work Commitment, Use Technical/Functional Expertise, Know the Business, and Influence Others.

The structure of the second data set (Mount et al., 1998) was similar to that of the Profilor data previously described. The Mount et al. data also included seven sets of developmental ratings (two bosses, two peers, two subordinates, and self) for each target manager ($n = 2,350$). The instrument Mount et al. used was the MSP (Sevy, Olson, McGuire, Frazier, & Paajanen, 1985), which was also developed by Personnel Decisions International, Inc. The two instruments differ in terms of the number (24 vs. 16, respectively) and the nature of the scales they contain, and there are no items common to both instruments. Mount et al. assigned the MSP scales to the same three managerial performance dimensions that we used for the Profilor scales.

Confirmatory Factor Analysis

A variety of CFA models have been applied to multitrait-multirater data. The correlated traits, correlated methods (CTCM; Widaman, 1985) model is often recommended (e.g., Kenny & Kashy, 1992), because it is most consistent with the Campbell and Fiske (1959) standards. However, the CTCM model was not a viable option in this study for both theoretical and practical reasons. Theoretically, the CTCM model is not completely consistent with our purpose. It partitions variance into three components (trait, method, and error), but the purpose of our study was to estimate five components of variance. Thus, modifications of some type would have been required. Because theoretically appropriate modifications seemed to be possible, we did attempt to fit the standard CTCM model to our data. Unfortunately, this resulted in several improper parameter estimates. This was not unexpected. Research (Becker & Cote, 1994; Conway, 1996; Kenny & Kashy, 1992; Marsh & Bailey, 1991) has shown that the CTCM model often presents serious problems with nonconvergence and improper parameter estimates. Because of the difficulties we encountered with the CTCM model, we do not discuss it here.

With the failure of the CTCM model, we turned to a form of the correlated uniquenesses (CU; Kenny, 1979) model, as recommended by several authors (Kenny & Kashy, 1992; Marsh, 1989; Marsh & Bailey, 1991). More specifically, we used a modified form of the CU-CFA method (Scullen, 1999) to estimate the magnitudes of the five components of observed ratings. An overview of the CU-CFA method is presented in the next section. A more complete rationale for the CU-CFA method and a detailed description of the technique are presented by Scullen (1999).

The CU-CFA method is a two-step process by which variance in a multitrait-multimethod (MTMM) matrix is partitioned into trait (performance, in this context), method (rater), and random measurement error components. In the first step, a correlated uniquenesses (CU) model is used to partition observed variance into performance-related and unique variance components. Then, in the second step, a CFA partitions the unique variance into method-related and measurement error components. At that point, variance has been partitioned into three components: performance-related, rater-related, and measurement error.

To achieve the goals of the current study, however, additional partitioning was required. That is, two types of performance-related variance had to be estimated, general and dimensional performance. In addition, the rater-

related variance had to be subdivided into perspective-related and idiosyncratic components. To do so, we modified the CU-CFA technique as follows.

The CU-CFA method begins by subjecting the MTMM matrix to a CU analysis. In the typical CU analysis, each variable is allowed to load on the appropriate trait factor, and the error terms for each variable measured by the same method are allowed to covary. Trait (performance) influences on the measures are modeled by loadings on the trait factors, and common method (rater) influences are represented by covariances among the error (uniqueness) terms.

Our model (Figure 1) includes factors representing the three performance dimensions in this study. The factors were allowed to correlate freely. To distinguish general performance variance from variance associated with particular dimensions of performance, we included a first-order general factor (Gustafsson & Balke, 1993; Mulaik & Quartetti, 1997) in the model. In our model, every observed rating loads on the general performance factor, and each also loads on one of the three trait (performance dimension) factors. As we discussed earlier, the general factor is orthogonal to each of the trait factors. Thus, the general factor represents variance that is common to all ratings, regardless of rater or performance dimension, and that is not associated with any of the trait factors. It would also have been possible to model the general factor as a second-order factor loading on each of the three dimensional factors. We chose not to do so because models with a second-order general factor place specific constraints on the relationships between the general factor and the observed variables, whereas the first-order general factor model does not (Gustafsson & Balke, 1993).

We recognize that if there are nonzero correlations among the method (rater) effects in our data, the general factor is likely to contain a certain amount of method or shared halo (Lance, 1994) variance in addition to performance-related variance. This could include the effects of what Lance et al. (1991) called nonperformance-based components (e.g., physical attractiveness). We believe, however, that it is appropriate to classify the common variance as general performance variance for two reasons. First, true ratee performance is likely to be the largest source of the variance that is common to all judgments made by all raters. Lance et al. found, for example, that performance-based general impression effects were clearly stronger than were nonperformance-based general impression effects. For shared halo error to be present in our model, it would be necessary for several raters to arrive at similar misjudgments about individual ratees. Therefore, any nonperformance-based effects are likely to be minimal.

A second and related reason for assigning common variance to general performance is that shared biases are most likely to occur between raters from the same perspective, and our model (as we discuss in the following) assigns variance shared by same-perspective raters to the perspective-related bias component. Although the correlations between method effects associated with different-perspective raters are constrained to zero in our model, we believe it is unlikely that these constraints have had any significant effect on our estimates of method variance. Our results show that the correlations between method effects associated with same-perspective raters are relatively small, and correlations of method effects across rater perspectives are likely to be smaller still. Marsh and Bailey (1991) have shown that modeling correlated method effects as if they were orthogonal generally results in only a trivial bias in estimates of trait and method variance.

Once the CU model has been estimated, the squares of the standardized loadings on the general factor and on the dimensional factors estimate the proportions of observed variance that are related to general and dimensional performance, respectively (Widaman, 1985). The error (uniqueness) terms in the CU model represent variance in the observed variables that is not associated with either general performance or dimensional performance. This includes measure-specific variance, the effects of measurement method (biases), and random measurement error.

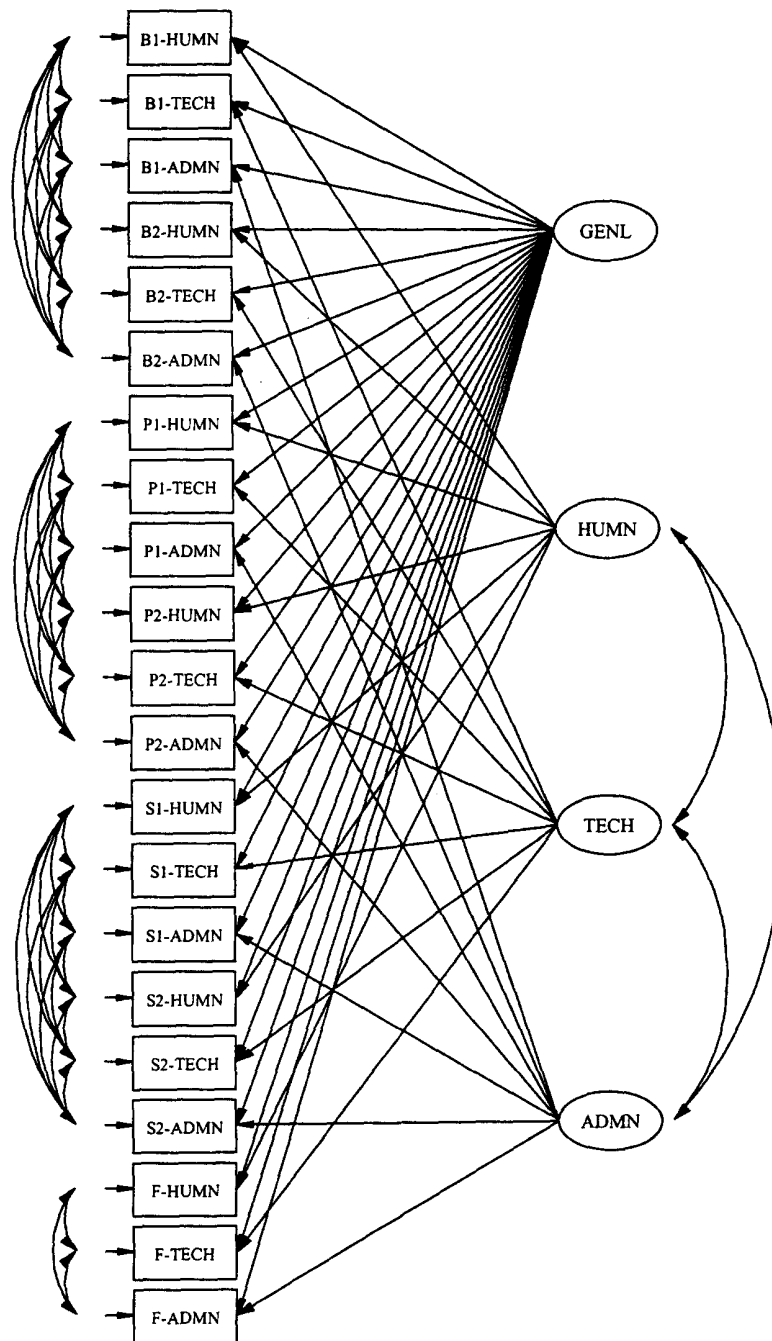


Figure 1. Original correlated uniquenesses model. B1 = Boss 1; B2 = Boss 2; P1 = Peer 1; P2 = Peer 2; S1 = Subordinate 1; S2 = Subordinate 2; F = self; HUMN = human; TECH = technical; ADMN = administrative; GENL = general performance.

The second phase of a standard CU-CFA analysis separates method variance from specific variance and error variance. This is done by conducting CFA on the matrix of correlated uniquenesses associated with each measurement method (i.e., there is a separate CFA for each measurement method). The current study required modifications in the CU-CFA technique, because the method variance had to be subdivided into perspective-

related and individual (idiosyncratic) components. Subdividing the method variance was accomplished as follows.

The first modification designed to separate perspective-related variance from idiosyncratic variance actually involved a change in the CU phase of our analysis. In a typical CU analysis, each rater is modeled as a separate measurement method. Thus, the error terms for each variable associated

with an individual rater are normally allowed to covary. We modified this by proceeding as if all of the variables associated with either one of the raters from a given perspective had been measured by a single method. For example, all of the ratings made by either Boss 1 or Boss 2 were treated as if they had been made by a single measurement method (i.e., boss ratings). Therefore, we allowed the error terms for all six of the boss ratings variables to covary (Figure 1). The same technique was used for the six peer ratings and for the six subordinate ratings. Because there was only one set of self-ratings for each ratee, there are only three correlated error terms for that perspective.

Continuing with the boss ratings example, the 6×6 (3 dimensions and 2 bosses) matrix of correlated uniquenesses of the error terms for bosses (computed in the CU phase) was fitted to a CFA model in which the three ratings made by Boss 1 loaded on a Boss 1 factor and the three ratings made by Boss 2 loaded on a Boss 2 factor (Figure 2). The Boss 1 and Boss 2 factors were allowed to correlate. If this CFA is conducted in the unstandardized (covariance) metric and if the factor variances are set to unity, then the squares of the factor loadings in the CFA will estimate the proportion of variance in each variable from the original CU analysis that is method-related (Scullen, 1999). Once the proportion of method variance in each observed variable is known, it can be partitioned into perspective-related and idiosyncratic components by considering the correlation between the factors.

Just as the correlation between ratings made by two raters (i.e., interrater reliability) measures the proportion of total variance that is systematic across raters, the correlation between the Boss 1 and Boss 2 factors represents the proportion of variance in either factor that is systematic across the factors. In effect, then, the factor intercorrelation indexes the proportion of the method variance in each observed variable that is shared with other raters from the same perspective.² The remainder of the method variance is unique to the individual rater.

After computing the amount of method variance in each variable (the squares of the loadings on the Boss 1 and Boss 2 factors), we computed the

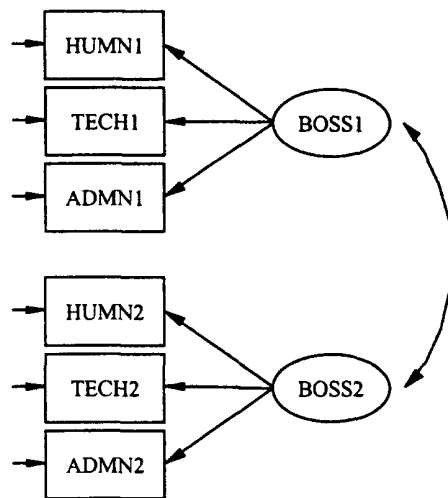


Figure 2. Confirmatory factor analysis (CFA) model for estimating perspective-related and idiosyncratic components of method effects. Variables in the boxes are the six correlated error terms for boss ratings as computed in the correlated uniquenesses model. HUMN1, TECH1, and ADMN1 refer to the error terms for human, technical, and administrative ratings made by Boss 1, respectively. HUMN2, TECH2, and ADMN2 are the corresponding error terms for ratings made by Boss 2. Input for this CFA is the 6×6 matrix of correlated uniquenesses for boss ratings. Similar models would be appropriate for the peer and subordinate perspectives.

perspective-related component for each variable by multiplying its total method variance by the factor intercorrelation. Idiosyncratic variance was computed as method variance multiplied by the difference between unity and the factor intercorrelation (or as total method variance minus perspective-related variance). Random variance for each variable is represented by the error variance for each variable in the CFA. Results of our analyses are presented in the next section.

Because our purpose was to compute estimates of variance rather than to compare the fits of different models, our main concern was that our model provided adequate fit to the data. Most researchers advise the use of multiple indicators for making that determination. Recent studies by Hu and Bentler (1998, 1999) indicate that a two-index strategy for judging model fit is most appropriate. For studies using the maximum likelihood method of estimation, Hu and Bentler advocated reporting the standardized root mean square residual (SRMSR; Bentler, 1995) as one index and suggested that the SRMSR be supplemented by at least one of several other indices. Among those recommended as a supplemental index by Hu and Bentler are the Tucker–Lewis Index (TLI; Tucker & Lewis, 1973), the comparative fit index (CFI; Bentler, 1990), and the root mean squared error of approximation (RMSEA; Browne & Cudeck, 1993). Those indices are among those suggested by other researchers (e.g., Byrne, 1994; Hoyle & Panter, 1995; Medsker, Williams, & Holahan, 1994) as well.

Following the suggestions we have just outlined, we report values for the SRMSR, TLI, RMSEA, and CFI (χ^2 is also reported). Hu and Bentler (1998, 1999) found that the commonly used rule of thumb that fit indices of .90 or higher are indicative of relatively good fit may be inappropriate. Their study suggested that the cut-off value should be approximately .95 for TLI and CFI. They also found that cut-off values of .08 for the SRMSR and .06 for the RMSEA (lower values of SRMSR and RMSEA indicate better fit) were most appropriate. We adopted the Hu and Bentler standards for the assessment of our model's fit.

Results

For each rater–ratee combination, we computed a mean for the items in each of the 21 scales used in this study. We computed ratings for each of the three performance dimensions from the scale means. That is, we computed each dimensional rating as the mean of the scale means for the scales we associated with that performance dimension. Each ratee thus received a total of 21 ratings—ratings on three dimensions from each of 7 raters. Mount et al. (1998) used similar procedures to generate the same type of matrix for the MSP data.

The correlation matrix and standard deviations for the Profilor ratings are presented in Table 1. The median different dimension–different rater correlation in the Profilor data was .16, the median same dimension–different rater correlation was .20, and the median different dimension–same rater correlation was .87. These results indicate modest convergence across dimensions and strong method effects.

We conducted our CU-CFA analyses using LISREL 8 (Jöreskog & Sörbom, 1996). Input for the CU phase was the 21×21

² This technique separates idiosyncratic from perspective-related variance in much the same way as hierarchical CFA (HCFA; Marsh & Hocevar, 1988) partitions variance in a lower order factor (LOF) into variance that is associated with a higher order factor (HOF) and variance that is not associated with the HOF. Because there would be only two LOFs in the models used in the current research (e.g., the Boss 1 and Boss 2 factors), an HCFA model would not be identified unless their two loadings on a HOF were constrained to be equal. Simply estimating the correlation between the factors accomplishes the same purpose.

Table 1
Standard Deviations and Intercorrelations of Performance Ratings

Perspective and skill	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Boss 1—human	0.551	—																				
2. Boss 1—technical	0.527	78	—																			
3. Boss 1—administrative	0.548	85	86	—																		
4. Boss 2—human	0.547	41	30	32	—																	
5. Boss 2—technical	0.521	30	41	34	80	—																
6. Boss 2—administrative	0.542	32	35	37	85	16	—															
7. Peer 1—human	0.606	25	18	18	26	16	18	—														
8. Peer 1—technical	0.563	21	30	23	21	28	23	83	—													
9. Peer 1—administrative	0.580	21	23	23	22	22	24	88	88	—												
10. Peer 2—human	0.609	30	19	21	31	20	21	25	19	20	—											
11. Peer 2—technical	0.555	24	30	25	26	29	25	21	28	22	84	—										
12. Peer 2—administrative	0.576	24	23	24	25	23	25	20	22	22	89	88	—									
13. Subordinate 1—human	0.704	19	10	14	22	12	16	20	15	16	20	14	16	—								
14. Subordinate 1—technical	0.609	16	18	17	19	20	19	17	23	19	16	21	18	87	—							
15. Subordinate 1—administrative	0.671	14	13	16	17	14	18	15	16	17	15	16	17	91	90	—						
16. Subordinate 2—human	0.700	21	12	15	18	10	12	24	19	20	20	13	16	33	28	28	—					
17. Subordinate 2—technical	0.612	17	20	18	17	19	16	21	26	22	16	20	17	27	33	28	87	—				
18. Subordinate 2—administrative	0.670	16	14	17	15	12	15	20	20	21	15	15	17	27	28	28	91	90	—			
19. Self—human	0.376	17	06	12	16	06	11	14	08	11	17	10	12	12	06	09	14	09	11	—		
20. Self—technical	0.390	07	18	15	08	19	15	04	16	10	06	16	11	02	11	07	03	12	07	73	—	
21. Self—administrative	0.416	07	09	16	07	10	15	03	08	10	06	09	11	07	09	11	08	11	13	79	78	—

Note. Decimals have been omitted from the correlations.

covariance matrix for each instrument. Completely standardized maximum likelihood estimates of the effects of interest were computed in the CU phase for both data sets. The fit statistics for the CU model indicated excellent fit, Profilor: $\chi^2(117, N = 2,142) = 127.50$, SRMSR = .015, TLI = 1.000, CFI = 1.000, RMSEA = .006; MSP: $\chi^2(117, N = 2,350) = 205.10$, SRMSR = .019, TLI = .996, CFI = .998, RMSEA = .018. However, one aspect of the solution created a minor problem. In both data sets, the CU analysis yielded very small and statistically nonsignificant negative correlations between the error terms for Peer 1 and the error terms for Peer 2. Median correlations were $-.06$ in the Profilor data and $-.02$ in the MSP data. Because these negative correlations would lead to small negative estimates of perspective-related variance for peer ratings, we re-estimated the model with the cross-rater correlated uniquenesses for peer raters constrained to zero. That is, the error terms for Peer 1 were not allowed to covary with the error terms for Peer 2 (Figure 3).

Parameter estimates under this model were virtually identical to the corresponding estimates under the original model. Fit statistics for the revised model were also very good, Profilor: $\chi^2(126, N = 2,142) = 133.18$, $p > .31$, SRMSR = .016, TLI = 1.000, CFI = 1.000, RMSEA = .005; MSP: $\chi^2(126, N = 2,350) = 221.71$, $p > .01$, SRMR = .020, TLI = .996, CFI = .997, RMSEA = .018. The chi-square value for the MSP analysis was large relative to the degrees of freedom, but even excellent models typically yield statistically significant chi-square values when the sample size is large (Hu & Bentler, 1995). With sample sizes of well over 2,000 in both data sets, large chi-square values were not unexpected in this study. The remaining indices indicated excellent fit for the model in both data sets. Chi-square difference tests support the hypothesis that there is no difference in model fit for either data set, Profilor: $\Delta\chi^2(9, N = 2,142) = 5.68$, $p > .75$; MSP: $\Delta\chi^2(9, N = 2,350) = 16.61$, $p > .05$. The other indices also indicated very little difference in fit across models. We therefore adopted the revised model.

All loadings on the general factor for boss, peer, and subordinate raters in both models were at least eight times their standard errors. For self-ratings, all but one of the loadings on the general factor were at least twice their standard errors. All of the performance factor loadings for all of the rater sources in both models were at least three times their standard errors. Therefore, the 95% confidence intervals for all of the dimensional loadings and all but one of the general factor loadings excluded zero. Because our purpose was to present the best estimates of variance associated with each of the dimensions and methods, we retained all of the factor loadings in our model, even the one that was not statistically different from zero. The 95% confidence interval for every method factor loading (CFA step) in both models also excluded zero.

The correlated uniquenesses for each method were subjected to CFA in the second phase of our analysis. Scullen (1999) discussed factors that are likely to artificially improve fit statistics and reduce standard error estimates in this type of analysis. He recommended that researchers not use those statistics to judge the fit of their models. Therefore, although those statistics appeared to be very good, we do not report them here. We used the procedures described previously to first estimate the total method variance in each variable and then partition the method variance into perspective-related and idiosyncratic components. As we indicated earlier, the proportion of the method variance that is perspective-

related was estimated by the correlation between factors in the CFA analysis. For bosses, those estimates were .18 (Profilor) and .15 (MSP). For subordinates, the Profilor and MSP estimates were .21 and .20, respectively. Estimated random error variances are represented in this model by error variances in the CFA analysis.

Table 2 presents the five estimated variance components for the three dimensions and the four perspectives in both data sets. To avoid unnecessary repetition and detail, the following comments are framed primarily around the Profilor results. References to MSP results are specifically identified as such. Readers will note in Table 2 that although the MSP and Profilor results are not identical, virtually all of the general conclusions that follow are supported by both sets of results.

Of the five effects investigated, it is clear that idiosyncratic rater effects are the major source of variance in observed ratings for every rater perspective and every performance dimension. On the basis of the averages across the three performance dimensions, the effects ranged from a low of 51% for boss ratings to a high of 71% for self-ratings. The average idiosyncratic rater effect was 62%.

Effects for the other source of rater bias, organizational perspective, were substantially smaller, ranging from a low of 0% for peers to a high of 17% for subordinates. There is no unique perspective-related effect for peer ratings³ (on either instrument), but there are small but meaningful effects for boss and subordinate ratings. Across dimensions and rater perspectives (excluding self), the mean perspective-related effect was 9%. Results in the Mount et al. (1998) study had indicated only that idiosyncratic effects are stronger than perspective-related effects are. The present results extend that finding by quantifying the two types of effects. Averaged across the boss, peer, and subordinate perspectives, idiosyncratic effects are about seven times greater than perspective-related effects (59% compared with 9% on the Profilor, and 49% compared with 7% on the MSP).

Table 2 also shows that, on average, general performance (13%), dimensional performance (8%), and random error (11%) contributed to observed variance at a much lower level than did the idiosyncratic effect. More specific results for each rater perspective are examined next.

Findings by Rater Perspective

Boss ratings. As Table 2 shows, there is a perspective effect associated with boss ratings. This effect represents rating variance that is systematic across ratings made by different bosses but is not shared in ratings made by raters from other perspectives. It accounts for approximately 11% of the observed variance in boss ratings, averaged across dimensions (8% in the MSP boss ratings). There is less idiosyncratic rating variance in boss ratings than in ratings from the other perspectives (51% vs. 62–71% in ratings from other perspectives in the Profilor data—43% vs. 52–64% in the MSP data). However, the effect of the idiosyncratic factor for boss ratings was still more than twice as large as any other factor

³ A reviewer suggested that it would be more accurate to state that perspective-related effects for peer ratings are not estimable in our model. It is our view, however, that the very small and statistically nonsignificant correlations between interpeer error terms in the original model effectively indicated that performance-related factors account for all of the relationship between ratings made by different peers.

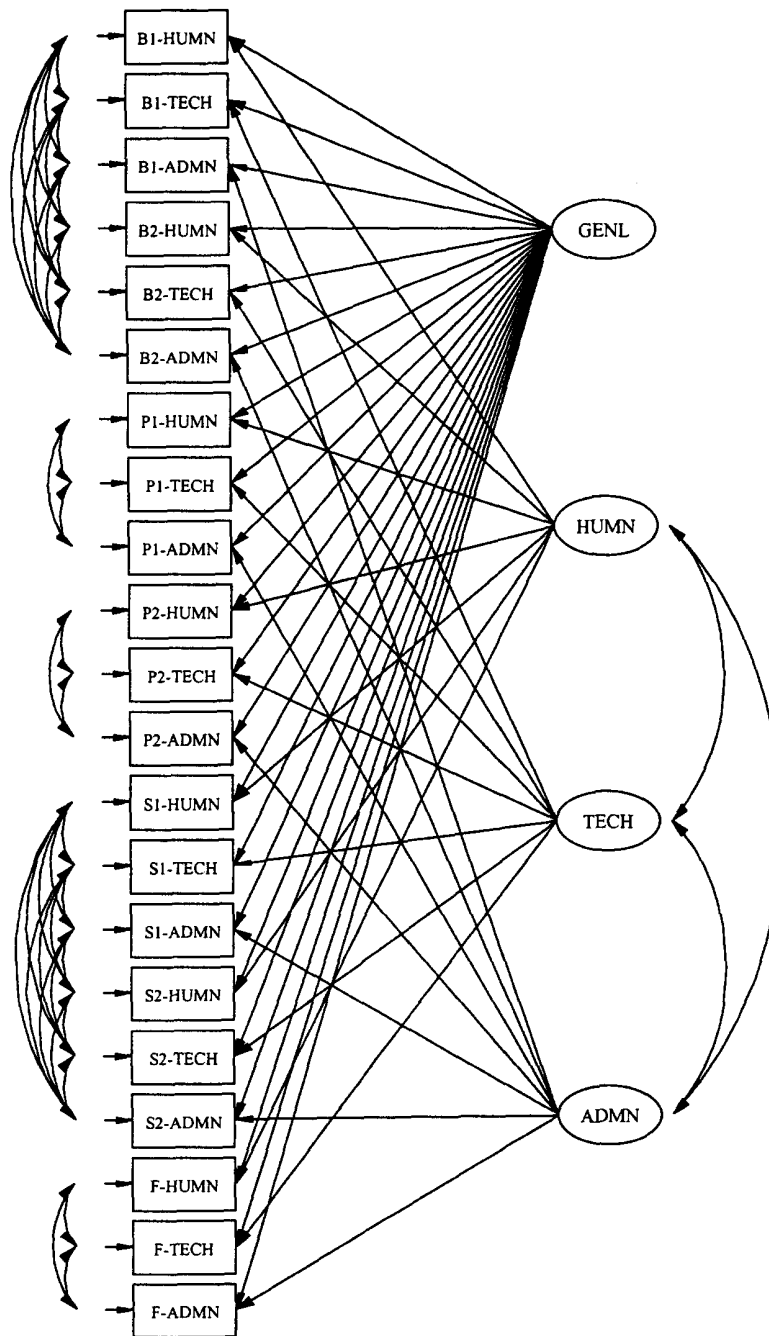


Figure 3. Revised correlated uniquenesses model with error terms not allowed to covary across peers. B1 = Boss 1; B2 = Boss 2; P1 = Peer 1; P2 = Peer 2; S1 = Subordinate 1; S2 = Subordinate 2; F = self; HUMN = human; TECH = technical; ADMN = administrative; GENL = general performance.

for bosses. The effect of actual ratee performance was larger in boss ratings than in ratings made by others. Averaged across the three traits, dimensional and general performance together accounted for 27% (9% and 18%, respectively) of the rating variance for bosses. The corresponding figure for the MSP ratings was 32%. This is a meaningful amount of variance, but it is substantially

smaller than the rater bias effects, which accounted for 62% of the variance in boss ratings (51% in the MSP data).

Peer ratings. Both data sets indicate that there was no unique perspective effect associated with peers. In other words, there was no systematic tendency for peer raters to agree more with each other than with raters from other perspectives. Well over half of

Table 2
Percentage of Variance in Observed Ratings Accounted for by the Five Factors

Perspective	Ratee performance								Rater bias											
	General				Dimensional				Perspective				Idiosyncratic				Measurement error			
	D1	D2	D3	M	D1	D2	D3	M	D1	D2	D3	M	D1	D2	D3	M	D1	D2	D3	M
Profilor																				
Boss	25	17	14	18	5	11	12	9	11	11	12	11	49	50	56	51	11	12	7	10
Peer	23	22	17	21	4	8	7	6	0	0	0	0	64	59	69	64	9	10	7	9
Subordinate	11	12	8	10	3	4	4	4	17	16	17	17	62	58	65	62	7	10	6	8
Self	5	2	0	2	7	14	15	12	—	—	—	—	75	67	71	71	13	17	14	15
M				13				8				9				62				11
MSP																				
Boss	16	20	24	20	18	13	5	12	6	7	9	8	36	39	52	43	23	20	9	17
Peer	16	23	22	20	14	8	5	9	0	0	0	0	48	45	63	52	22	24	11	19
Subordinate	8	13	10	11	9	3	7	7	12	13	14	13	49	52	57	53	22	19	11	17
Self	2	6	3	3	15	11	16	14	—	—	—	—	62	63	67	64	21	21	14	19
M				14				11				7				53				18

Note. Dashes represent quantities that could not be estimated. D1 = human dimension; D2 = technical dimension; D3 = administrative dimension; MSP = Management Skills Profile.

the observed variance in peer ratings (64% on average in the Profilor data, and 52% in the MSP data) was idiosyncratic, indicating considerable variation across individual peer ratings. The amount of peer rating variance that was accounted for by ratee performance was meaningful (Profilor 27%, MSP 29%) and was similar to the amount accounted for by boss ratings, in terms of both general performance and dimensional performance. Peer ratings were also similar to boss ratings in the proportion of variance accounted for by rater biases (62% for bosses and 64% for peers in the Profilor data—51% for bosses and 52% for peers in the MSP data), although the composition of the bias component was different. In peer ratings, the bias component was entirely idiosyncratic, whereas in boss ratings, a meaningful portion of the bias component was perspective related.

Subordinate ratings. There was a unique effect associated with the subordinate perspective, as was the case for boss ratings. On average, the subordinate perspective accounted for 17% of the observed ratings variance in the Profilor data and 13% in the MSP data. In both data sets, this was the largest of the perspective-related effects. Unlike either the boss or the peer ratings, the perspective-related effect for subordinates was as large as or larger than either of the performance-related effects. Thus, organizational perspective was a more important source of variance for subordinate ratings than for ratings from the other perspectives. This result supports the conclusion drawn by Conway and Huffcutt (1997) in their recent meta-analysis of multisource performance ratings that subordinates have "a relatively unique perspective" (p. 349).

Subordinate ratings were also quite idiosyncratic. Like the peer ratings, over half of the observed variance (62% averaged across traits) was associated with the idiosyncratic component. In contrast, the two components of ratee performance accounted for only 14% of the subordinate rating variance. This effect is substantially smaller than for bosses or peers. Rater biases accounted for a total of 79% of the variance in subordinate ratings (66% in the MSP). This was larger than for boss or peer ratings but was about the same as for self-ratings.

Self-ratings. It was not possible to distinguish between perspective-related and idiosyncratic effects for self-ratings, be-

cause there can be only one self-rater for each ratee. We chose to categorize this effect as idiosyncratic, but a plausible argument could be made that this is also a perspective-related effect. One notable characteristic of self-ratings is that dimensional variance was relatively high (Profilor 12%, MSP 14%) compared with other perspectives across the three performance dimensions. At the same time, effects associated with general performance were low (Profilor 2%, MSP 3%) compared with the other perspectives. Because of the possible confounding of perspective-related and idiosyncratic variance in self-ratings, it is not meaningful to compare the self-ratings idiosyncratic component with the corresponding components from other perspectives.

Comparison of Profilor and MSP Results

Table 2 shows that the same general pattern of results held true for both the MSP and the Profilor data. Idiosyncratic rater effects in the MSP ratings were the largest by far, accounting for about half of the observed variance on average. Of the two performance effects, those associated with general performance were larger than those for dimensional performance. Perspective effects were evident for boss and subordinate ratings but not for peer ratings. Thus, the MSP results lead to the same conclusions as do the Profilor results regarding the relative effects of the five factors that we have postulated influence performance ratings.

Discussion

Main Findings

The main purpose of this study was to examine the latent structure of job performance ratings by quantifying the effects of five major factors that influence observed performance ratings and by comparing those effects across raters from four perspectives and across three performance dimensions. Although details of the results varied somewhat between the two data sets we used, both support the following conclusions.

Our most important and robust finding is that idiosyncratic variance was the largest component of variance for all combinations of rater perspective and performance dimension. For peer, subordinate, and self-ratings, the idiosyncratic component was larger than all of the other effects combined. Idiosyncratic variance was also the dominant component in boss ratings, but to a lesser extent.

Ideally, the rating variance associated with the performance of the ratee would be large relative to the variance associated with biases of the rater. In other words, what is being rated should account for more variance than does who is doing the rating. Our results show that this is not generally true. For boss and peer ratings, actual ratee performance (i.e., the sum of general and dimensional variance) accounted for only about 30% of total variance in both data sets. For subordinate and self-ratings, the effects of actual performance were smaller still, ranging from about 15–20%. Thus, actual performance accounted for approximately 20–25% of the variability in performance ratings when averaged across dimensions, perspectives, and instruments.

The general performance components in our results are typically larger than the corresponding dimensional components for all perspectives except self. This means that the dimensional factors contributed fairly little unique information beyond what was associated with the general ratings factor. In other words, aspects of ratee performance that are specific to a particular dimension had a relatively minor influence on ratings. This finding is consistent with those of Viswesvaran (1993), who found support for a strong general performance factor in 25 distinct measures of job performance. It is not yet clear, however, why the general factor exerts a more powerful influence on ratings than do the dimensional factors. Later in the article, we offer suggestions for research regarding the roles of the general and dimensional factors in performance ratings.

The other primary finding in this study concerns the existence of perspective-related effects. Results from both of our data sets suggest that perspective has some effect in boss ratings and that it is a fairly important component of subordinate ratings variance, where it accounts for about half of the variance that is systematic across raters. That is, interrater reliabilities for subordinate ratings (the sum of general, dimensional, and perspective-related variance) were .31 in both of our data sets, and the perspective-related components (Profilor 17%, MSP 13%) accounted in each case for about half of the 31% of variance that was systematic across raters. Our results also support the conclusion that there is no perspective-related effect for peer ratings. We conclude, therefore, that both boss and subordinate ratings capture something that is unique to that perspective, but peer ratings do not.

Up to this point, we have viewed actual performance as comprising general performance and dimensional performance. It is possible, however, that perspective-related effects should be included in the actual performance category as well. As we discussed earlier, Borman (1997) posited that raters from different perspectives might rate differently because they observe different aspects of ratee performance. This suggests that perspective-related rating differences could be more a function of true differences in the performance observed by each type of rater (Borman, 1974; Kavanagh, Borman, Hedge, & Gould, 1987) than of differences in the observers themselves. If so, then the perspective-related variance components do not represent rater biases; instead, they represent

specific aspects of the criterion space that are not represented in ratings from other perspectives (Lance et al., 1992; Tornow, 1993). Under this view, perspective-related variance should be added to the general and dimensional performance variance to fully account for performance-related variation in ratings.

If that is done with boss ratings, the percentage of variance associated with actual ratee performance would be the sum of the dimensional (9%), general (18%), and perspective-related (11%) components—a total of 38%. Performance-related variance sums to only 27% if perspective-related variance is not included. Similarly, the percentage of performance-related variance in subordinate ratings would increase from 14% to 31% with the addition of perspective-related variance. There is no corresponding gain for peer ratings, because they exhibited no perspective-related variance.

The fact remains, however, that even if perspective-related effects are considered to be a part of true performance, the rater (i.e., idiosyncratic) effects still overshadow the performance effects. Our findings parallel those of Lance (1994), who concluded that “ratings were stronger reflections of raters’ overall biases than of true performance factors” (p. 768), and are surprisingly consistent with a generalizability theory analysis of ratings by Greguras and Robie (1998). Greguras and Robie estimated several sources of ratings variance and then compared the total variance associated with raters (i.e., rater main effects and Ratee \times Rater interaction effects) with the variance associated with ratees (i.e., true score effects). For boss ratings, Greguras and Robie found that the rater effects were 1.17 times as large as the ratee effects. The corresponding figures for peer and subordinate ratings in their study were 2.11 and 2.22, respectively. Analogous computations in the current study involve dividing the idiosyncratic rater variance by the sum of the general performance, dimensional performance, and perspective-related effects for each rater perspective. Results from our study, averaged across instruments, yield values of 1.21 for boss ratings, 2.08 for peer ratings, and 1.86 for subordinate ratings. Thus, our study supports and extends the findings in the Lance (1994) and Greguras and Robie (1998) studies.

Our study also supports several conclusions about the validities of ratings from the various perspectives. Regardless of whether perspective-related effects are classified as actual performance or bias, our results indicate that boss ratings capture more of the ratee’s actual job performance than do ratings from any other perspective. This suggests that the validity of boss ratings is higher than the validity of ratings from the other perspectives. Because true performance levels are unknown, none of the validities can be determined with certainty. But especially if perspective-related variance is included in actual performance, boss ratings capture the most performance-related variance and, at the same time, are the least idiosyncratic. This is strong evidence that boss ratings are the most valid. This is a reasonable conclusion, given that bosses are the most likely to have had training and experience in rating performance.

Peer ratings may be less valid than are boss ratings, although there is little difference if perspective-related effects are discounted. That is, if perspective effects are treated as bias, the effects of actual performance are about the same for peer ratings as for boss ratings. If perspective effects are viewed as actual performance, however, then peer ratings reflect considerably less performance than do boss ratings. Although there appears to be no

perspective-related effect in peer ratings, we believe that peers are an important source of ratings information. First, peer ratings contain considerable amounts of performance variance. This in itself makes peer ratings valuable. Second, peers are a good source of ratings because of their numbers. As we discuss in the following, it is beneficial to have a large number of raters, because this allows for the aggregation of ratings.

The validity of subordinate ratings depends to a relatively large extent on how perspective-related variance is viewed. If the unique subordinate perspective is seen as a form of undesirable bias, then only 14% of the rating variance is associated with actual ratee performance. Thus, the validity of subordinate ratings would be considerably lower than it is for either boss or peer ratings. But if the unique characteristics of subordinate ratings are seen as valuable information that is not available from the other perspectives, then their validity is much higher. Including perspective effects approximately doubles (from 14% to 31%) the amount of true performance variance in subordinate ratings. It is important for researchers to determine the nature of this perspective effect. We point out that because participants were allowed to select their own raters, this large perspective-related effect could reflect a tendency for managers to solicit ratings primarily from in-group (Dansereau, Graen, & Haga, 1975) rather than out-group subordinates. It is very possible that the views of in-group subordinates are more homogeneous than are the views of subordinates in general. Researchers should examine the extent to which the perspective-related effects found here are representative of all subordinates.

Implications

Taken as a whole, our findings suggest some potential problems and pitfalls in the use of performance ratings for research and administrative purposes. We preface our remarks with a reminder that the ratings we analyzed in this study were made for developmental purposes only. The administrative implications we discuss are based on the premise that the relative magnitudes of the variance components in administrative ratings are similar to those in developmental ratings.

Performance ratings are used in practice to make decisions concerning pay raises, promotions, and terminations. Our results show that a greater proportion of variance in ratings is associated with biases of the rater than with the performance of the ratee. Although this may have already been known in a general sense, previous research had not quantified these effects for four rater perspectives and three dimensions. In scientific research, corrections for unreliability can account for the effects of measurement error; however, there is no analogous correction factor that can be used in organizations to eliminate the effects of idiosyncratic rater bias. The obvious implication of our finding is that decision makers should be aware of the impact of idiosyncratic bias and attempt to control its effects. This could be done by seeking a variety of types of performance information, possibly including objective measures or ratings made by multiple individuals.

Our results also illustrate the significant benefit organizations can gain from using multirater systems. Generalizability theorists (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) have shown that if ratings are averaged across n raters, each of the error components is divided by n , whereas the true variance components

remain unchanged. This effectively increases the proportion of true variance.

Of course, the larger the error components are, the greater is the advantage of using multiple raters. Because the error components, especially idiosyncratic variance, are large, averaging across several raters can significantly reduce the effects of bias and random error.

Recommendations for Future Research

Our results have important implications for theories of rating. Any causal model seeking to explain performance ratings must account for rater effects, as these are the largest source of rating variance. Our study does not investigate the causes or the nature of individual (i.e., idiosyncratic) and perspective-related effects, and research of this type is clearly needed. We make suggestions in another section for research aimed at understanding the nature of these job performance components and their effects on ratings. We also discuss the implications of our research for those interested in the random measurement error component of ratings.

Rater effects. One implication of our results is that models seeking to explain performance ratings should include factors associated with the perspective of the rater. Boss and subordinate perspectives accounted for approximately 10–15% of the observed variance in our data. Some researchers (e.g., Murphy & Cleveland, 1995) have argued that bosses tend to emphasize those aspects of performance that can be objectively measured, such as reaching production goals or remaining within budget. Empirical evidence supports that notion. Oppler, Campbell, Pulakos, and Borman (1992) found that correlations between ratings and nonratings measures were consistently higher for ratings made by bosses than for ratings provided by peers. There is also evidence (Fox & Bizman, 1988) that subordinates are more attuned to other aspects of the target manager's performance, such as interpersonal skills.

Bollen and Paxton (1998) suggested that researchers investigate the nature of method effects by including hypothesized determinants of those effects in their structural models. Lance et al. (1992) is an example of that type of research. Further studies of that nature should continue to provide valuable insights into the validity of ratings made by different types of raters and might shed important light on how ratings can be improved.

Researchers should also continue to study the nature of idiosyncratic effects. Past research has investigated influences such as racial and gender biases (Mount, Sytsma, Hazucha, & Holt, 1997; Pulakos et al., 1996), interpersonal affect (Varma, DeNisi, & Peters, 1996), and implicit theories of performance (Borman, 1987), to name just a few. Each of those research streams contributes to our understanding of the largest component of ratings variance.

Our study was concerned more with the overall magnitude than with the composition of idiosyncratic rater effects. Our findings regarding the effect of idiosyncratic variance in boss ratings (51% in the Profilor data and 43% in the MSP data) are somewhat at odds with past research. Viswesvaran et al. (1996) reported that idiosyncratic tendencies account for only 29% of the variance in boss ratings. We believe that differences in rating designs are largely responsible for the discrepancy. As we argued earlier, interrater differences in leniency cause total variance in nested designs to be higher than in crossed designs. Our data were

gathered using a nested design, so to the extent that the studies represented in the Viswesvaran et al. meta-analysis were of the crossed design type, their estimate of idiosyncratic variance should be smaller than ours.

We examined a representative sample of the studies included in the Viswesvaran et al. (1996) meta-analysis to determine which type of design had been used. In many cases, the rating procedures were not described precisely enough by the authors of the primary studies for us to determine whether the reported indices referred to (a) agreement between two raters, each of whom rated the entire set of ratees, or (b) agreement between two raters, for which the pair of raters is different for each ratee. Therefore, we could not determine the exact nature of the data and, consequently, we do not know how much of the difference between the Viswesvaran et al. estimate of rater bias and our estimate could be explained by differences in rating design.

Nonetheless, it is interesting to speculate about whether the difference between our estimates of idiosyncratic variance in boss ratings and the Viswesvaran et al. (1996) estimate represents a valid estimate of the variance introduced by interrater differences in leniency. If all or most of the studies in the Viswesvaran et al. meta-analysis involved crossed rating designs, then somewhere between 14% (43% – 29%, using our MSP estimate of idiosyncratic variance) and 22% (51% – 29%, using our Profilor estimate) of the observed ratings variance in nested systems may be due to leniency differences between raters. Future research should examine that possibility.

We advise all researchers in this area to consider carefully the implications of rating system design in planning their own studies and urge them to explicitly state the nature of their rating designs when reporting those studies.

Performance-related factors. Our findings show that the general factor tends to have somewhat greater influences on performance ratings than do the dimensional factors, although the exact proportions vary by rater perspective and dimension. The possibility that the general factor and the dimensional factors have different determinants suggests one important line of research. One might consider, for example, the personality dimension of conscientiousness, which has been shown to have the most generalizable validity of the five factors of personality (Barrick & Mount, 1991). It is possible that its validity generalizes across jobs because it predicts underlying general performance, which has commonalities across jobs. In addition, it also predicts performance on specific dimensions that may be relevant to a particular job. On the other hand, one possible reason why the validity of a different personality trait, such as openness to experience, does not generalize across jobs is because it predicts only specific aspects of performance, which we have shown to be quite small, but does not tap into the general performance component. This is only one example of the type of research that is needed to understand the determinants and consequences of these two distinct types of rater performance components. Further research is needed to understand the causal antecedents of these two components of performance.

Random measurement error. Our MSP estimate of random error variance (18%) is similar to the 19% figure reported by Viswesvaran et al. (1996) for boss ratings of overall performance (random measurement error). The estimate of random error variance we derived from the Profilor data (11%) was somewhat smaller than either the MSP or the Viswesvaran et al. estimates. A

comparison of the variance components for the MSP and the Profilor shows only minor differences between instruments in the general, dimensional, and perspective-related components. Differences in the idiosyncratic and random error components are somewhat larger. Our study contributes to the literature by showing that the amounts of systematic variance in performance ratings are essentially the same for peer, subordinate and self-ratings as for boss ratings. Although we found some variability across instruments, there was relatively little variation within instruments. Stated differently, the effects of random measurement error are very similar across rating perspectives. Future research should examine the factors that contribute to differences between groups or between instruments in the distributions of variance across factors.

Limitations

Two aspects of this research could limit its generalizability to other contexts. One concerns the developmental nature of these ratings. It is unclear how our results would generalize to ratings made for administrative purposes. Moreover, ratees in this study were allowed to choose their raters. It is not possible to determine how this may have affected our results. It is possible that this resulted in higher or less variable ratings than would have been observed if ratees had not been allowed these choices.

The other possible limitation concerns the fact that two scales (Leadership and Coaching) were omitted from the MSP analysis and one (Leading Courageously) was omitted from the Profilor analysis. As a reviewer pointed out, each of these scales relates to several aspects of managerial performance, and this probably explains why judges were not able to agree on a single dimension that any one of them best represents. It is possible that the omission of these scales affected our results, particularly the proportion of variance associated with performance. This is an interesting question but one that would require a different theoretical perspective on managerial performance than the one examined here. We encourage researchers to develop and test models that incorporate a leadership dimension. These should lead to valuable insights as to how managers view the role of leadership in management.

We add one final caveat to aid the interpretation of our results. It is important to make a distinction between nomothetic (between-persons) and idiographic (within-person) uses of multirater performance ratings (Allport, 1937; Pelham, 1993; Zevon & Tellegen, 1982). This distinction is important because multirater data are used in both ways, yet the types of psychometric evidence required to support each are different. The purpose and use of a measure, and the consequences of the decisions it is expected to support, should determine the appropriate evidentiary basis on which scores are interpreted (Messick, 1995).

Idiographic use of multirater feedback can be viewed as a clinical appraisal of an individual in which the user is the ratee, not the organization. It is usually anonymous, confidential, within-person, developmental feedback and, as a result, there is no need to generalize findings to others. Thus, the psychometric characteristics of between-persons uses of ratings may have only limited implications for idiographic uses of such measures (Runyan, 1983). Rather, the meaning of idiographic ratings for the individual may be more dependent on such issues as the context of the measurement, individual and organizational item relevance, idio-

syncratic job characteristics, career aspirations, interrater consensus, and within-person contrasts. With appropriate attention to these issues, confidential multirater development feedback can be a systematic, quantitative, and dependable (as measured by consensus) source of strictly idiographic information, yielding indirect economic value to the organization.

In contrast, between-persons ratings have important implications for administrative and performance decisions. As such, they carry a different and greater psychometric burden that is commensurate with the greater organizational risk from decision errors. The meaning of between-persons use of multirater data is critically dependent on such factors as understanding the magnitude of different sources of rating variance, maximizing valid ratee performance variance (both specific and general), and generalizing relationships to other performance and organizational variables. This is especially important because, as we have noted, no corrections are available for organizations to use to ameliorate measurement error and bias in ratings, nor are any available to correct for any resulting economic effects of decision errors. Organizations would be well advised to demand and assure that the psychometric rigor used is commensurate with the intended use of the ratings.

Summary

The main contribution of this study is to enhance our understanding of the latent structure of performance ratings. Our results quantify the magnitudes of five major effects on performance ratings and show how they compare for bosses, peers, subordinates, and self-ratings on three performance dimensions using two large, independent data sets. Using an entirely nested design, we find that that, on average, idiosyncratic rater effects account for over half of the variance in performance ratings. Given that performance ratings are the most frequently used measure of performance, this presents a major challenge to the field of industrial-organizational psychology. In light of these findings, we renew the call for research investigating ways to decrease idiosyncratic rater biases while increasing the amount of actual ratee performance in performance ratings. We also call for additional research that investigates the antecedents and consequences of the two distinct components of ratee performance as well as the perspective-related effects present in boss and subordinate ratings. Solving these vexing problems will enable industrial-organizational psychology research and practice to have greater impact.

References

- Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-25.
- Becker, T. E., & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management, 20*, 625-641.
- Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bollen, K. A., & Paxton, P. (1998). Detection and determinants of bias in subjective measures. *American Sociological Review, 63*, 465-478.
- 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 employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.
- Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance, 12*, 105-124.
- Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307-322.
- Borman, W. C. (1997). 360° ratings: An analysis of assumptions and a research agenda for evaluating their validity. *Human Resource Management Review, 7*, 299-315.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows: Basic concepts, applications, and programming*. Newbury Park, CA: Sage.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Conway, J. M. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management, 22*, 139-162.
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331-360.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*, 218-24.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- Dansereau, F., Graen, G., & Haga, W. J. (1975). A vertical dyad linkage approach to leadership within formal organizations. *Organizational Behavior and Human Performance, 13*, 46-78.
- Fox, S., & Bizman, A. (1988). Differential dimensions employed in rating subordinates, peers, and superiors. *Journal of Psychology, 122*, 373-382.
- Greguras, G. J., & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83*, 960-968.
- Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407-434.
- Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39*, 811-826.
- Hezlett, S. H., Ronnkvist, A. M., Holt, K. E., & Hazucha, J. F. (1997). *The PROFILOR(R) technical summary*. Minneapolis, MN: Personnel Decisions International.
- Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.
- Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure modeling: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software.
- Katz, R. L. (1974). Skills of an effective administrator. *Harvard Business Review*, 52, 90-102.
- Kavanagh, M. J., Borman, W. C., Hedge, J. W., & Gould, R. B. (1987). *Job performance measurement in the military: A classification scheme, literature review, and directions for research* (AFHRL-TR-87-15). Brooks Air Force Base, TX: Air Force Human Resources Laboratory, Training Systems Division.
- Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165-172.
- King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65, 507-516.
- Lance, C. E. (1994). Test of a latent structure of performance ratings derived from Wherry's (1952) theory of ratings. *Journal of Management*, 20, 757-771.
- Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology*, 79, 332-340.
- Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, 77, 437-452.
- Lance, C. E., Woehr, D. J., & Fisicaro, S. A. (1991). Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior*, 12, 1-20.
- London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes: Theory-based applications and directions for research. *Personnel Psychology*, 48, 803-839.
- Mann, F. C. (1965). Toward an understanding of the leadership role in formal organizations. In R. Dubin, G. C. Homans, F. C. Mann, & D. C. Miller (Eds.), *Leadership and productivity* (pp. 68-77). San Francisco: Chandler.
- Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335-361.
- Marsh, H. W., & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Applied Psychological Measurement*, 15, 47-70.
- Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107-117.
- Medsker, G. J., Williams, L. J., & Holohan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. *Journal of Management*, 20, 439-464.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.
- Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and conceptual performance. *Human Performance*, 10, 71-83.
- Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology*, 51, 557-576.
- Mount, M. K., Sytsma, M. R., Hazucha, J. F., & Holt, K. E. (1997). Rater-ratee race effects in developmental performance ratings of managers. *Personnel Psychology*, 50, 51-69.
- Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling*, 4, 193-211.
- Murphy, K. R., & Cleveland, J. C. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Oppler, S. H., Campbell, J. P., Pulakos, E. D., & Borman, W. C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. *Journal of Applied Psychology*, 77, 201-217.
- Pelham, B. (1993). The idiographic nature of human personality: Examples of the idiographic self-concept. *Journal of Personality and Social Psychology*, 64, 665-677.
- Pulakos, E. D., Schmitt, N., & Chan, D. (1996). Models of job performance rating: An examination of ratee race, ratee gender, and rater level effects. *Human Performance*, 9, 103-119.
- Runyan, W. M. (1983). Idiographic goals and methods in the study of lives. *Journal of Personality*, 51, 413-437.
- Scullen, S. E. (1999). Using confirmatory factor analysis of correlated uniquenesses to estimate method variance in multitrait-multimethod matrices. *Organizational Research Methods*, 2, 275-292.
- Sevy, B. A., Olson, R. D., McGuire, D. P., Frazier, M. E., & Paajanen, G. (1985). *Managerial skills profile technical manual*. Minneapolis, MN: Personnel Decisions.
- Tornow, W. W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resources Management*, 32, 221-230.
- Tsui, A. S. (1984). A multiple-constituency framework of managerial effectiveness. In J. G. Hunt, D. Hosking, C. A. Schriesheim, & R. Stewart (Eds.), *Leaders and managers: International perspectives on managerial behavior and leadership* (pp. 28-44). New York: Pergamon Press.
- Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.
- Vance, R. J., MacCallum, R. C., Coover, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, 73, 74-80.
- Varma, A., DeNisi, A. S., & Peters, L. H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology*, 49, 341-360.
- Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor?* Unpublished doctoral dissertation, University of Iowa, Iowa City.
- Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81, 557-574.
- Wherry, R. J., Sr., & Bartlett, C. J. (1982). The control of bias in ratings: a theory of rating. *Personnel Psychology*, 35, 521-551.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.
- Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, 43, 111-122.

Received February 10, 1999

Revision received January 10, 2000

Accepted January 13, 2000 ■