Journal of Applied Psychology
2003, Vol. 88, No. 1, 50–66

0021-9010/03/$12.00 DOI: 10.1037/0021-9010.88.1.50

Evidence of the Construct Validity of Developmental Ratings of Managerial Performance

Steven E. Scullen
North Carolina State University

Michael K. Mount
University of Iowa

Timothy A. Judge
University of Florida

The construct validity of developmental ratings of managerial performance was assessed by using 2 data sets, each based on a different 360° rating instrument. Specifically, the authors investigated the nature of the constructs measured by developmental ratings, the structural relationships among those constructs, and the generalizability of results across 4 rater perspectives (boss, peer, subordinate, and self). A structure with 4 lower order factors (Technical Skills, Administrative Skills, Human Skills, and Citizenship Behaviors) and 2 higher order factors (Task Performance and Contextual Performance) was tested against competing models. Results consistently supported the lower order constructs, but the higher order structure was problematic, indicating that the structure of ratings is not yet well understood. Multisample analyses indicated few practically significant differences in factor structures across perspectives.

Ratings are used for a number of purposes (e.g., administrative, research, and developmental) in organizations. Administrative ratings are focused on assessing performance for the purpose of making personnel (e.g., compensation) decisions. Research-based ratings serve to provide the performance information needed for purposes such as evaluating training programs or testing hypothesized relationships among organizational constructs. The basic premise underlying the use of developmental job performance ratings is that they provide ratees with information they can use to improve their job performance. To be effective for any of those purposes, ratings must reflect true job performance; that is, they must be construct valid. Unfortunately, despite repeated calls for construct validation of performance measures (Austin & Villanova, 1992), it remains true that relatively little is known about the construct validity of ratings (Lance, 1994).

The present study examined the construct validity of developmental ratings of job performance. It had three main purposes: The first was to determine whether developmental ratings measure four aspects of managerial performance derived from the research literature. The second was to examine the interrelationships among ratings of performance for those four components of managerial work. The third was to assess the generalizability of those findings across rater perspectives (boss, peer, subordinate, and self) and across two philosophically distinct rating instruments.

Before proceeding, we emphasize that ours was a study of developmental ratings only. Prior research has shown that some of the psychometric properties of performance ratings vary according to the purpose for which they were made. For example, administrative ratings tend to be more lenient and less accurate than developmental ratings or research ratings (Bernardin & Orban, 1990; Dobbins, Cardy, & Truxillo, 1988). Less is known about the factor structures of ratings and especially about how they might be affected by rating purpose. We speculate later in the article about how those factor structures might compare, but we urge caution in generalizing our results to ratings made for purposes other than development.

After a brief discussion of the construct validation of performance measures, we develop a hypothesized factor model for managerial performance ratings. We then use confirmatory factor analyses (CFAs) to test the hypothesized model and to compare it with other plausible models. We report several multisample analyses that tested the generalizability of the factor structures across rater perspectives. Finally, we discuss the theoretical and practical significance of our findings.

Construct Validity of Performance Ratings

Construct validity is "a term used to indicate that the test scores are to be interpreted as indicating standing on the psychological

Steven E. Scullen, College of Management, North Carolina State University; Michael K. Mount, Henry B. Tippie College of Business, University of Iowa; Timothy A. Judge, Warrington College of Business, University of Florida.

This article is based on Steven E. Scullen's doctoral dissertation, completed at the University of Iowa under the direction of Michael K. Mount and Timothy A. Judge. An earlier version of this article was presented at the meeting of the Academy of Management in Chicago, Illinois, August 1999.

We thank Sara Rynes, Frank Schmidt, and Steve Dunbar for their contributions to this research as members of the dissertation committee. We also thank John Fleenor and Jean Leslie of the Center for Creative Leadership, Greensboro, North Carolina, and Maynard Goff of Personnel Decisions International, Minneapolis, Minnesota, for their cooperation.

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

construct measured by the test" (American Psychological Association, 1999, p. 174). Ratings are construct valid to the extent that there is a high degree of correspondence between the ratings and the true levels of performance. Because true performance levels are usually unknown, and perhaps unknowable, construct validity must be assessed indirectly. One way that scholars have done this is by studying relationships between ratings and other measures of performance, such as job samples (Lance, Teachout, & Donnelly, 1992; Vance, MacCallum, Coovert, & Hedge, 1988) or other objective measures of performance (Bommer, Johnson, Rich, Podsakoff, & MacKenzie, 1995; Heneman, 1986). These studies support the idea that ratings reflect, at least to some extent, actual performance on the job.

However, much of that research was based on entry- to mid-level enlisted military jobs, and it is not clear that the constructs underlying (ratings of) performance in those jobs would generalize to management jobs. Moreover, the measurement of performance in managers is less amenable to the methodologies used in those studies (e.g., work samples, which are a more appropriate method for measuring performance for enlisted military personnel than for managers).

A more common method for assessing the construct validity of performance ratings for managers has been to examine multitrait–multirater (MTMR) matrices for evidence of convergent and discriminant validity (e.g., Lawler, 1967; Mount, 1984). In recent years, confirmatory factor-analysis techniques (Kenny & Kashy, 1992; Widaman, 1985) have largely supplanted earlier methods based on D. T. Campbell and Fiske's (1959) standards or on the analysis of variance. Findings have consistently shown some degree of convergence across rating sources but relatively little discriminant validity across dimensions of performance.

Brannick (as cited in Hurley et al., 1997) suggested a way to improve on construct validation research using MTMR methods, noting that

> A stronger cross validation in the sense of understanding what the factors really mean in some population would be achieved by using new variables (items or measures) to measure the same constructs in a new sample [different samples]. Finding a similar structure under such a circumstance would be strong evidence in support of the model.

This quotation (pp. 675–676) exemplifies the general method used in this study. We tested our hypothesized factor structure with data gathered from independent samples and from two philosophically diverse rating instruments. Strong similarities in the factor structures across rater perspectives, samples, and rating instruments is evidence of substantial construct validity in performance ratings.

Following Murphy's (1989) view that construct validation should begin with an explication of the constructs of interest (as opposed to an inductive approach in which the researcher attempts to discover the nature of the performance dimensions through factor analysis or some similar statistical technique), the first step in this study was to identify and define the hypothesized dimensions of performance ratings. We defined dimensions in accordance with an overall conception of what job performance is. According to Murphy, this type of construct-oriented approach is superior because (a) it places no arbitrary statistical limitations on the nature of performance dimensions, as factor analysis might, and (b) it makes explicit the assumption that performance is a construct and that the ultimate definition of performance dimensions depends entirely on one's conceptual definition of performance.

We begin with an examination of the theoretical dimensions of performance but point out that other factors that might also affect the way managers perceive performance must be considered as well. In the following sections we examine theoretical perspectives on the components of work and then turn to other factors, such as potential effects of differences in rater perspective (i.e., boss, peer, subordinate, and self), that might impact the factor structures of performance ratings.

Theoretical Perspectives on Job Performance

A number of theoretical perspectives have been put forth, some representing performance on jobs in general (e.g., Borman & Motowidlo, 1993; J. P. Campbell, 1990; D. Katz, 1964; Murphy, 1989; Viswesvaran, 1993) and some specific to managerial jobs (e.g., Borman & Brush, 1993; R. L. Katz, 1974; Mann, 1965; Mintzberg, 1975; Tornow & Pinto, 1976; Yukl, 1989). In one sense, any of these perspectives could have been used to generate performance dimensions for our study. But it was critical to this research that the selected dimensions represent, as closely as we could anticipate them to, the constructs that are reflected in ratings that real-world managers actually give.

A number of scholars (e.g., Borman, 1983, 1987; Feldman, 1981; Ilgen & Feldman, 1983; Krzystofiak, Cardy, & Newman, 1988; Murphy & Cleveland, 1995) have argued that the dimensions derived from a scientific theory of job performance or from a formal job analysis are not necessarily the constructs that real-world raters use when they evaluate performance on that job. Borman (1983, 1987) maintains that through experience and observation, raters develop implicit theories about how the job should be performed and that they use these theories as a framework for making sense of and evaluating observed behavior. Thus, theoretical perspectives on performance may be of value for hypothesizing dimensions of ratings but only to the extent that the theoretical performance dimensions coincide with raters' implicit theories.

Managers typically spend much of their time facing significant time pressures and many competing demands on their attention, most of them not appraisal related. Raters in this type of environment are likely to focus their attention on relatively few performance cues. Therefore, we expect those theories with fairly simple and straightforward constructs to more accurately represent the perspectives of time-pressured managers than do more complex theories of performance.

For conceptual and practical reasons, we chose to base our hypothesized dimensions on the D. Katz (1964), Mann (1965), R. L. Katz (1974), and Borman and Motowidlo (1993) typologies. One reason was that these typologies lead to a straightforward, commonsense, and empirically supportable set of hypothesized performance dimensions. Another was that a relatively simple set of dimensions would facilitate our testing the generalizability of a factor structure across philosophically different rating instruments. Especially with the very large and diverse samples we used in this research, we saw no reason to believe that the implicit theories (i.e., the actual rater constructs) of raters who completed one instrument would differ in any systematic way from the implicit theories of raters who completed the other instrument. We did suspect, however, that the specific factor structures of the two instruments would be somewhat different. Hence, responses to any instrument could be expected to exhibit a factor structure in which there would be evidence of both the raters' implicit theoretical constructs and influences specific to the philosophical foundation of the instrument. If that is correct, the unique aspects of the factor structure of either instrument would be of less interest than the commonalities across instruments. We therefore hypothesized a relatively simple structure that would represent the set of factors that are common across instruments.

Mann (1965) and R. L. Katz (1974) proposed similar three-skill (technical, human, and administrative or conceptual) approaches to managerial effectiveness. A synthesis of their approaches was adopted as the starting point for the model in this study. Brief discussions of each skill are presented here; more explicit definitions follow in conjunction with the hypothesized model.

Technical skills refers to the manager's proficiency in specific methods, processes, and techniques within the manager's special function. *Human skills* include the ability to work effectively as both a team member and a team leader. *Administrative skills* involve understanding the organization as a whole and how the various parts of the organization are interdependent. Examples of the three skills include planning, organizing, delegating, inspecting, and coordinating.

There is empirical support for this three-factor structure. A cluster analysis of managerial effectiveness criteria (Brush & Licata, 1982) resulted in three clusters, which were labeled employee-centered activities, technical competence, and functional managerial skills (planning, coordinating, and decision making). In another study (Lau, Newman, & Broedling, 1980), a factor analysis of 50 items based on Mintzberg's (1975) 10 management roles yielded three factors, which were identified as Supervision, Planning, and Technical Problem Solving. Parallels with the R. L. Katz (1974) and Mann (1965) three-factor typologies are clear.

There is evidence that a fourth aspect of job performance, which we call *citizenship behaviors* (Coleman & Borman, 2000; Organ, 1997), also influences performance ratings. Several constructs are relevant to this component of performance. Three of them are organizational citizenship behaviors (OCBs; Organ, 1988; Smith, Organ, & Near, 1983), prosocial organizational behaviors (Brief & Motowidlo, 1986), and organizational spontaneity (George & Brief, 1992). Each differs a bit from the others, and there has been a good deal of discussion (e.g., Coleman & Borman, 2000; Organ, 1997; Van Dyne, Graham, & Dienisch, 1994) as to exactly how this dimension of performance should be characterized. In general, these constructs refer to "work behavior that contributes, at least in the long run, to organizational effectiveness, but which is sometimes overlooked by the traditional definitions and measures researchers use to assess job performance" (Van Dyne et al., 1994, p. 766). Examples include being cooperative, loyal, or persistent beyond expectations. Our definition is presented in detail in conjunction with our hypothesized model.

Studies of ratings made by supervisors of military mechanics (Motowidlo & Van Scotter, 1994), life insurance agents (MacKenzie, Podsakoff, & Fetter, 1991), and secretaries (Werner, 1994) have all supported the idea that citizenship behaviors account for a considerable amount of variance in ratings of overall performance, beyond what is predictable from objective measures of performance or from ratings of in-role behaviors.

It is important to note, however, that none of that research involved ratings of managers. Organ (1988) raised the interesting point that as one rises in organizational rank, role specifications tend to become more diffuse, and the distinction blurs between in-role and extrarole behaviors. Organ as well as Borman and Motowidlo (1993) argued, however, that even at the highest levels, there are identifiable differences between managers in these types of behaviors. A study involving mostly management-level ratees (Williams & Anderson, 1991) yielded evidence that ratings of managerial performance are influenced by two of these types of behaviors—those intended to benefit specific individuals and those intended to benefit the organization. Behaviors of both types explained variance in ratings over and above the effects of in-role performance.

In summary, empirical evidence suggests that ratings of managerial performance are influenced by at least four types of factors: the three hypothesized by Mann (1965) and R. L. Katz (1974) plus citizenship behaviors. These factors form the basis of the model we hypothesize here. Our hypothesized model also incorporates Borman and Motowidlo's (1993) factors of task performance and contextual performance. We present our use of task and contextual performance in our discussion of the hypothesized model. However, before presenting the hypothesized model, we examine the possibility that the constructs measured by ratings may differ across rater perspectives.

Differences by Rater Perspective

Borman (1997) discussed several possible reasons why raters from different perspectives might rate differently. One is that bosses, peers, and subordinates attend to different dimensions of performance. This might stem from differences in either the self-interests (Tsui, 1984) or the implicit theories of performance (Borman, 1974; Murphy & Cleveland, 1995) of these raters.

Some empirical evidence (e.g., Lance, 1994; Pulakos, Schmitt, & Chan, 1996) supports the notion of differences in the factor structures of ratings across rater perspectives. It is reasonable to postulate that bosses would emphasize objective measures of performance, such as reaching production goals or staying within budget constraints. Research (Oppler, Campbell, Pulakos, & Borman, 1992) suggests that correlations between ratings and nonratings measures are higher for boss ratings than for peer ratings. Similarly, Conway (1999) found that bosses give greater weight to these types of factors in judging overall job performance than do peers.

Peers have a different type of relationship with the ratee than do bosses. Whereas bosses can rely on formal authority if necessary to secure needed resources from a subordinate, a peer is more likely to depend on interpersonal relationships. Thus, teamwork and cooperation could receive greater attention from peers than from bosses. Empirical evidence supports the notion that raters are more likely to consider interpersonal relationships when rating their peers than when rating their subordinates (Conway, 1999; Fox & Bizman, 1988).

Subordinates have yet a different type of relationship with the ratee, and could be particularly interested in leadership and fairness issues as well as their bosses' ability and willingness to help them develop their technical and administrative competence. Again, the research of Fox and Bizman (1988) supports this position. They found that raters were particularly likely to emphasize managerial abilities, interpersonal relations, and professional knowledge when evaluating their bosses' performance.

Another body of research suggests, however, that there are few, if any, differences across rater perspectives in terms of their conceptualizations of performance dimensions (Facteau & Craig, 2001) or the calibration of their ratings (Maurer, Raju, & Collins, 1998). Similarly, Tsui and Ohlott (1988) concluded that there were no material differences in what they called *managerial effectiveness models* for supervisors, peers, and subordinates. Thus, as Borman (1997) argued, any evidence of differences in rating constructs across rater perspectives is weak at best.

The empirical evidence provided no compelling reason for us to hypothesize that factor structures differ across the boss, peer, subordinate, and self-ratings perspectives. We therefore began with the premise that the same factor structure is appropriate for all four rater perspectives. Specifically, we hypothesized a four-factor structure that generalizes across the four rating sources. We then tested the validity of the assumption of invariance across rater perspectives.

Hypothesized Factor Structure

Our hypothesized factor structure includes four lower order performance factors (Technical Skills, Administrative Skills, Human Skills, and Citizenship Behaviors) and two higher order factors (Task Performance and Contextual Performance). The conceptual model is depicted in Figure 1. Each of the performance factors is defined below.

Lower Order Factors

Technical Skills refers to two types of proficiency. The first is the manager's ability to perform the core substantive and technical tasks that pertain directly to the organizational function (e.g., accounting or production) in which the manager works. These concern the degree to which the manager has the specialized knowledge, skills, and analytical abilities that are normally associated with professional roles and affiliations in his or her specific discipline. The second type of proficiency is more general and refers to proficiency in the financial, quantitative, and other types of data analysis that are common to managers in all organizational functions.

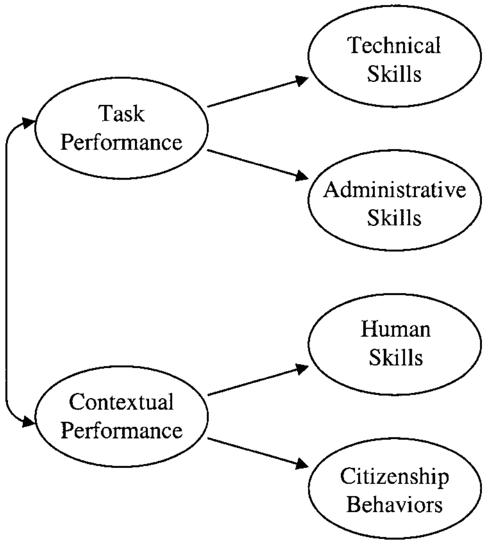

Figure 1. Hypothesized factor structure for performance dimensions. Measured variables, lower order factors, and disturbance terms are omitted for simplicity.

Administrative Skills refers to the manager's ability to think and act in terms of the particular organizational system in which the manager operates. This requires an understanding of how the people, structures, procedures, and policies operate in his or her organization to attain certain objectives. Administrative skills include planning, programming, and organizing work; setting goals for the work unit; working through nontechnical problems or crises that threaten goal attainment; delegating tasks and authority; inspecting work; and coordinating the efforts and activities of different organizational members, levels, and departments.

Human Skills refers to a manager's ability to work with and through people to accomplish goals. This encompasses both the ability to work effectively as a group member and the ability to elicit effort within the team the manager leads. Performance on this dimension is facilitated by an understanding of the general principles of human behavior, particularly those involving motivation and maintaining interpersonal relationships, and by the skillful use of these principles while interacting with others in the work situation. This factor also includes the manager's ability to anticipate possible reactions to various actions she or he might take, his or her skill in communicating with others, and her or his ability to clearly represent the needs and goals of members at different levels in the organization to each other.

Our definition of Citizenship Behaviors is consistent with the recent construct clarification by Coleman and Borman (2000). It refers to three types of acts that managers may perform beyond what is expected of them: interpersonal (assisting, supporting, developing, and cooperating), organizational (demonstrating commitment, loyalty, allegiance, and compliance), and job task conscientiousness (persistence, dedication to one's job, and desire to perform well). The interpersonal aspect of citizenship behaviors is similar to what Van Scotter and Motowidlo (1996) called *interpersonal facilitation*, and the combination of the organizational and job task conscientiousness aspects is similar to their job dedication construct.

Higher Order Factors

The two hypothesized higher order factors in our model generally parallel Borman and Motowidlo's (1993) factors of Task Performance and Contextual Performance. Borman and Motowidlo outlined two types of task performance. One concerns the transformation of raw materials into the goods and services produced by the organization. The other includes activities required to support and maintain the technical core, including planning, coordinating, purchasing, distributing, and so on. These two types of task performance are analogous to the technical and administrative skills dimensions that we use in our study, except that our dimensions are framed around the manager's specific department or functional area rather than the overall organization.

Borman and Motowidlo (1993) defined contextual performance as activities that “support the broader organizational, social, and psychological environment in which the technical core must function” (p. 73). We hypothesized that there are two distinct aspects of contextual performance. On the one hand, managers are expected to exert interpersonal influence to accomplish work through other people. This could include a willingness to set and adhere to high standards for work, to hold people accountable, and to confront problem employees. On the other hand, managers may interact with others in a supportive and positive way to help them and the organization function more effectively. We see the human skills factor in our study as including the ability to lead and challenge others to perform, whereas our citizenship behaviors factor includes actions that stem from a desire to establish and maintain positive working relationships with others. Although these are different types of performance, we believe that both types directly affect the social and psychological environment in which employees work and that both are aspects of contextual performance.

Although it could be argued that for managers, human skills are an aspect of task performance, we had two reasons for categorizing human skills as a component of contextual performance. First, as we argued above, human skills are consistent with the the definition of contextual performance. And second, Motowidlo, Borman, and Schmit (1997) maintain that task performance is primarily a function of cognitive ability, whereas the main antecedent of contextual performance is personality. In our view, performance on the human skills component of our model is determined more by personality characteristics than by cognitive ability. Coleman and Borman (2000) acknowledged that the specific contents of contextual performance, citizenship behaviors, and similar performance dimensions are “somewhat arbitrary, with these decisions depending largely on the construct definition intended by the author” (p. 41). We believe that human skills and citizenship behaviors, as we have defined them, are distinct factors, both of which fall under the heading of contextual performance.

Figure 1 illustrates the hypothesized relationships among our performance constructs. Past research (e.g., Conway, 1999) suggests that the correlation between Task Performance and Contextual Performance may be substantial, perhaps .60 or larger. This raises the possibility that a higher order general factor could be included in the model. For technical reasons concerning model identifiability, we did not do so. That decision was not based on an assumption that no such factor exists but instead on the fact that we would not be able to estimate its effects.

Method

Instruments

Data sets based on two multirater feedback instruments were analyzed in this research. Each has been used extensively in practice to provide managers with performance feedback from their bosses, peers, and subordinates as well as to allow them to compare that feedback to their perceptions of their own performance. The two instruments are published by different organizations, and each is based on a different theoretical foundation. Descriptions of the instruments and their theoretical foundations are presented below.

The Management Skills Profile (MSP) was developed by Personnel Decisions International, Inc. (see Sevy, Olson, McGuire, Frazier, & Paajanen, 1985). It consists of 116 items, which are grouped by the publisher into 18 scales of 4–10 items. The 18 scales are further grouped into an eight-component model of management competency. The model was developed through applied research and consulting experience. Components of the management competency model and the scales included in each are Administrative (Planning, Organizing, and Personal Organization and Time Management), Leadership (Leadership Style & Influence, Motivating Others, Delegating & Controlling, and Coaching & Development), Interpersonal Skills (Human Relations and Conflict Management), Communication (Informing, Listening, Oral Communications, Written Communications), Personal Adaptation (Personal Adaptability), Motivation and Commitment (Personal Motivation), Occupational/Technical Knowledge (Occupational/Technical Knowledge), and Cognitive Skills (Problem Analysis & Decision Making, Financial & Quantitative). Raters provide anonymous ratings indicating how well each item describes observed behaviors of the ratee using a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*to a very great extent*). There is also a *does not apply* option.

The second instrument, Benchmarks (Lombardo & McCauley, 1994), was developed at the Center for Creative Leadership, Greensboro, North Carolina. Benchmarks is based on interview and survey research in which executives described key experiences in their careers and the lessons they learned from them. It contains 106 items, which the publisher groups into 16 scales (Resourcefulness, Doing Whatever It Takes, Quick Study, Decisiveness, Leading Employees, Setting a Developmental Climate, Confronting Problem Employees, Work Team Orientation, Hiring Talented Staff, Building and Mending Relationships, Compassion & Sensitivity, Straightforwardness & Composure, Balance Between Personal Life and Work Life, Self-Awareness, Putting People at Ease, and Acting with Flexibility). Raters indicate the extent to which the ratee displays the characteristic described in each item. Responses are made on a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*to a very great extent*). This section also includes a *does not apply* option. A second section of Benchmarks, Potential Flaws, was not used in this research.

The psychometric properties of the instruments may be less relevant in this study than in most because of the manner in which we used the instruments. Nonetheless, internal consistency reliabilities for the 18 scales of the MSP (see Sevy et al., 1985) range from .70 (Conflict Management) to .91 (Human Relations). For the 16 Benchmarks scales (see Lindsey, Homes, & McCall, 1987), internal consistency reliabilities range from .75 (Decisiveness) to .97 (Hiring Talented Staff). Thus, reliabilities reach acceptable levels for all scales on both instruments. We also emphasize that in a review of a number of instruments designed to measure managerial skills, both of the instruments used in this study were included among 16 deemed to reflect accepted standards of instrument development (Van Velsor & Leslie, 1991).

Participants

Most of the participants were working managers enrolled in self-development programs. Ratings were made for developmental purposes only. Managers were allowed to choose their raters when that was possible. For reasons that are described later, most of the samples used in this research were limited to ratees for whom ratings were available from more than one rater per perspective (except self). For ratees who had been rated by more than two raters from a given perspective, two of those raters were selected randomly for inclusion in this study.

Both the MSP and Benchmarks data sets consisted of data from managers representing a crosssection of industries, functions, and levels within their organizations. The number of MSP participants varied (N = 3,424–14,388), depending on rater perspective. Most of the MSP participants were White (87%), male (74%), and college graduates (76%). The mean age was 42 years. The number of managers in the Benchmarks data set also varied across rater perspectives (N = 1,546–1,722). Most of the Benchmarks participants were also White (90%), male (68%), and college graduates (88%). The mean age was 42 years.

Procedures

Phase 1: Preparation for confirmatory factor analysis. Six advanced doctoral students served as subject matter experts (SMEs). SMEs were given definitions of the four performance dimensions hypothesized in this research (technical skills, administrative skills, human skills, and citizenship behaviors) and a randomly ordered list of the individual items (not the scales) in the instruments. The SMEs independently assigned each item to the most appropriate of the four hypothesized performance dimensions. Items that were assigned to the same factor by at least four SMEs were retained. The remaining items were dropped.

For both conceptual and practical reasons, the analyses in this study were based on item parcels (means of multiple items) rather than on individual items. Conceptually, it is important to match the depth and breadth of the observed measures to the depth and breadth of the constructs of interest (Bagozzi & Edwards, 1998). Bagozzi and Edwards (1998) argued that there are two basic ways in which constructs can be modeled: disaggregation and aggregation. Disaggregation models, based on individual items or aggregations of items representing a particular component, are "suited to fine-grained analyses from which one desires to examine components . . . and obtain detailed information" (Bagozzi & Edwards, 1998, p. 55). In aggregation models, items are aggregated into more "abridged or condensed representations of a construct" (Bagozzi & Edwards, 1998, p. 57). It was not our intention in the current study to fit the specific nuances of each instrument's factor structure. Instead, we wanted to test the ability of a single set of more broadly defined (i.e., abridged or condensed) factors to fit both of the instruments and all four rater perspectives in our study. Our parceling strategy of random assignment of items to parcels representing broadly defined constructs was representative of the aggregation type of model and is an appropriate way to represent the molar nature of our hypothesized constructs (Kishton & Widaman, 1994).

Parceling was also attractive from a practical standpoint. Parcels greatly reduce the number of required parameter estimates and therefore increase the likelihood of convergence and proper solutions (West, Finch, & Curran, 1995). They also increase the reliabilities of the indicator variables, they reduce the likelihood that parameters will be affected by item-specific variance (Lance, Woehr, & Fisicaro, 1991), their distributions are more likely to approximate a normal distribution than are the distributions of individual items, and the results based on parcels are more likely to be stable (generalizable) across samples (West et al., 1995).

We formed 16 parcels per instrument: four parcels for each of our four lower order performance dimensions. This was done by randomly assigning the items that the SMEs had associated with each factor into four groups of as nearly equal size as possible. Parcel scores were then computed for each ratee as that ratee's mean rating (from that rater) across the items assigned to that parcel. Missing data were replaced with the mean rating (across all ratees) awarded on that item by all raters from that perspective.[1]

Each ratee in the boss, peer, and subordinate samples was rated by two raters. Therefore, each ratee received a total of 32 parcel scores (16 from each rater) from each of those rater perspectives. Covariance matrices (32 × 32) were computed for each rater perspective except self. Because there was only one set of self-ratings for each ratee, the correlation and covariance matrices for self-ratings were 16 × 16.

Phase 2: CFA. The objective of this phase was to use CFA to test the hypothesized model and to compare its fit to the fits of three other factor structures. The first competing factor structure was a unidimensional model in which a single factor loads on all parcels. This model was tested to ensure that ratings measure more than a single construct. The second competing model was the four-correlated-dimensions model. This model tested the viability of the four hypothesized lower order performance dimensions. The third model was a one higher order general factor model that examines the possibility that all of the four lower order performance dimensions are facets of a single higher order factor (i.e., that there is no empirical distinction between task and contextual performance). Each of those models is described in more detail in the Results section.

Our analyses involved an adaptation of two CFA models, the hierarchical confirmatory factor analysis model (HCFA; Marsh & Hocevar, 1988) and the correlated uniquenesses model (CU; Kenny, 1979). The HCFA feature of the model was used to control for two types of idiosyncratic rater variance: halo error and differences in leniency across raters. So that differences across individual raters could be controlled, it was necessary to have more than 1 rater per perspective (except self) for each ratee. This was the purpose for limiting our samples to those ratees who had been rated by at least two raters from a given perspective; with two raters per ratee, rater-specific variance can be modeled separately from variance that is common to both raters.

HCFA models use multiple indicators for each combination of dimension and rater, allowing each combination to be represented by a latent variable. We refer to each of these latent variables as a dimension–rater factor. For example, Technical Skills as rated by Peer 1 is a dimension–rater factor, and the four Technical Skills parcels as rated by Peer 1 serve as indicators for that factor. Most of our models have eight dimension–rater factors (four dimensions rated by two raters). Trait (i.e., performance dimension) effects are modeled by allowing the dimension–rater factors to load on higher order performance dimension factors (see Figure 2).

Method (rater) effects were modeled by using an adaptation of the CU model. In the typical CU analysis, method effects are modeled by allowing error terms for observed variables measured by the same method to covary. In this research, it was disturbances for the factors representing ratings made by the same rater that were allowed to covary. The disturbances represent variance in one rater's ratings that is systematic across rating dimensions but which is not associated with any performance factor or with any of the other raters. This is assumed to be method variance, stemming from rater-specific leniency and halo. The CU feature of the model carries the additional advantage of allowing for multidimensionality in the method effects, which the traditional CFA model does not (Marsh & Bailey, 1991).

CFAs were performed using LISREL 8 (Jöreskog & Sörbom, 1996) with maximum-likelihood estimation. Each of the models was fitted to each of seven data matrices (described below). Inputs for the CFAs were the covariance matrices described previously.

Hu and Bentler (1998, 1999) suggested that when maximum-likelihood estimation is used, the standardized root-mean-square residual (SRMSR; Bentler, 1995) and at least one of several other indexes should be used to judge model fit. Following Hu and Bentler (1998, 1999), we report values for the SRMSR, the root-mean-square error of approximation (RMSEA; Browne & Cudeck, 1993), the nonnormed fit index (NNFI; Tucker & Lewis, 1973), and the comparative fit index (CFI; Bentler, 1990). Chi-square values are also reported. Hu and Bentler (1998, 1999) also found that cutoff values, indicating relatively good fit, should be approximately .95 for the NNFI and CFI, .08 for the SRMSR, and .06 for the RMSEA. Those standards were adopted in this study.

Phase 3: Generalizability of factor structures across perspectives. In the final phase of the data analysis, we used the multisample analysis feature of LISREL 8 (Jöreskog & Sörbom, 1996) to determine whether there were meaningful differences in the factor structures across rater perspectives on each rating instrument.

Results

Of the 116 MSP items, 98 (84%) were assigned to a hypothesized factor. Eighty-two (77%) of 106 Benchmarks items were

[1] In response to a reviewer's concern that our results could be peculiar to the specific combinations of items that were randomly assigned to each parcel, we twice randomly reassigned MSP items to parcels and repeated our analyses. Results in both cases were similar to those that are reported here.

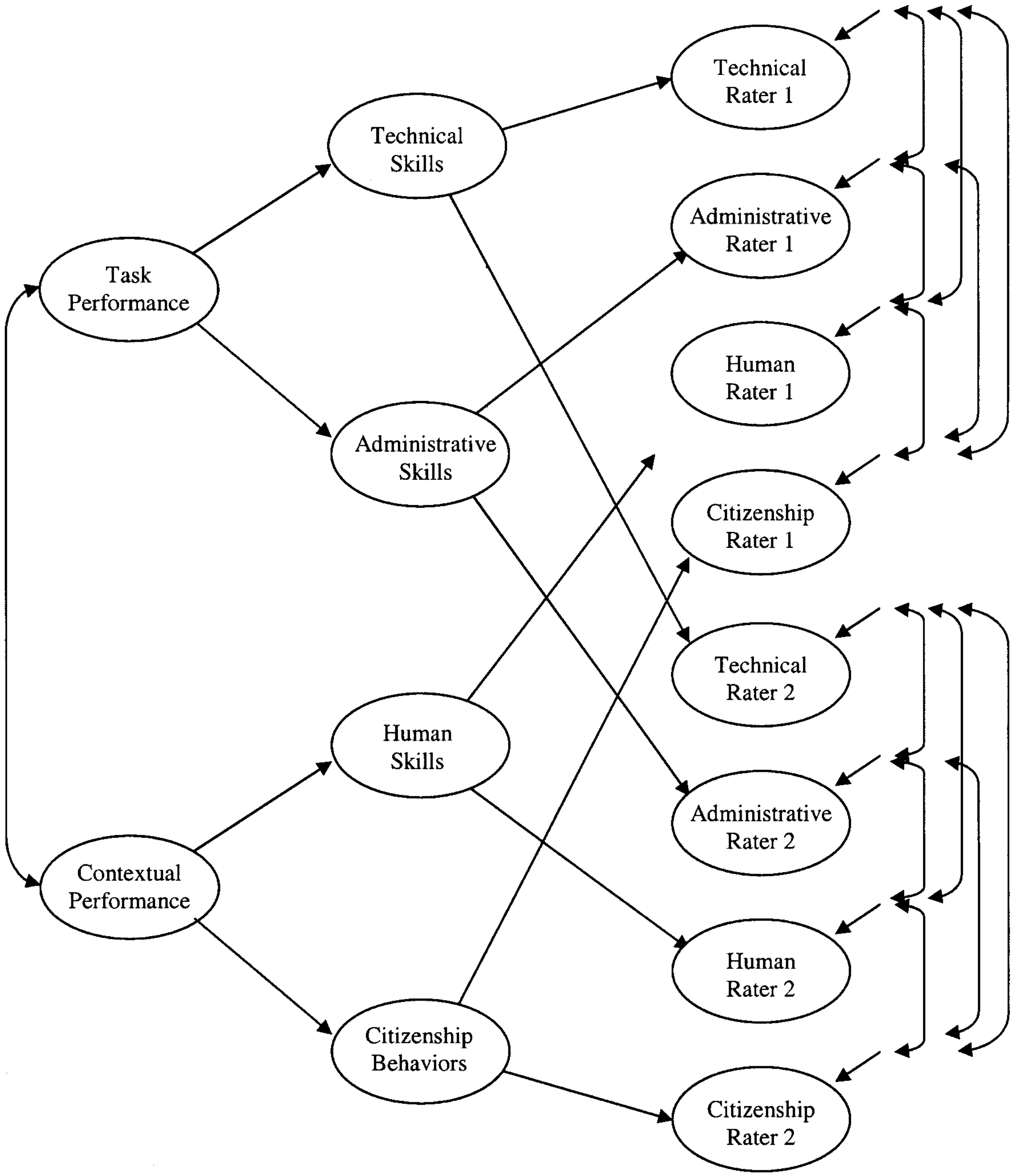

Figure 2. Hierarchical confirmatory factor analysis model with correlated disturbance terms. Observed variables (parcels) and their error terms are omitted for simplicity.

assigned. Because the categories to which SMEs assigned items constitute a nominal measurement system, the reliability of their assignments was assessed by computing Cohen's (1960) kappa for each possible pair of SMEs. Kappa indicates the percentage of items on which a pair of raters agree, beyond what would be expected by chance. Landis and Koch (1979) offered a rule of thumb suggesting that kappa values of .41–.60 indicate moderate agreement and that values from .61–.80 indicate substantial agreement. Umesh, Peterson, and Sauber (1989) also suggested that kappa values be interpreted in comparison with the maximum value possible for the observed proportion of agreement between raters.

The median kappa values in our study were .57 and .50 for the MSP and Benchmarks data, respectively. The corresponding maximum possible values for the median kappa values (Umesh et al., 1989) were .64 and .62. As mentioned, kappa is a measure of pairwise agreement. But because the assignments of items to performance dimensions were made on the collective assessments of six SMEs, the reliability of the item assignments was considered acceptable.

Phase 1: Preparation for CFA

In the MSP data set, boss analyses were based on 3,424 managers for whom ratings from two bosses were available. Peer analyses were based on 10,625 pairs of ratings, and subordinate analyses were based on 12,671 pairs of ratings. Self-ratings data included 14,328 sets of ratings. The Benchmarks data set did not contain a sufficient number of ratees with multiple boss ratings to do a meaningful analysis of those ratings. Therefore, only the peer, subordinate, and self-ratings were analyzed. (Note, however, that Benchmarks boss ratings were included in the multisample analysis described later.) Peer analyses were based on 1,698 pairs of ratings. Subordinate analyses were based on 1,546 pairs of ratings. There were 1,722 sets of self-ratings.

Because most of the correlation matrices in this study are large (five of them are 32 × 32), they are not presented here. All of the input matrices are available from Steven E. Scullen. As is common in MTMR research, strong rater (method) effects were apparent in our data. In the MSP ratings, the average heterotrait–monomethod correlations were .57, .60, and .63 for bosses, peers, and subordinates, respectively. The mean monotrait–heteromethod correlations were only .37, .29, and .29, respectively. In the Benchmarks ratings, mean heterotrait–monomethod correlations (.51 for peers and .50 for subordinates) were again considerably higher than the mean monotrait— heteromethod correlations (.20 for peers and .25 for subordinates). Thus, same-rater correlations of different traits were higher than different-rater correlations of the same trait for all rater perspectives and for both instruments.

Phase 2: CFA

A total of 28 CFAs were performed initially, four models for each of seven combinations of rater perspective and instrument. For brevity, some results are presented in summary form only. Table 1 contains fit statistics for each of those models.

Unidimensional model. In this model, a single factor loaded on all measured variables (i.e., parcels). As expected, the fit statistics for this model were consistently poor, indicating that it is not reasonable to conclude that all item parcels are measures of the same performance construct for any rater perspective or rating instrument.

Four-correlated-dimensions model. This model included the four hypothesized lower order performance dimensions (Technical Skills, Administrative Skills, Human Skills, and Citizenship Behaviors). Thus, it is similar to the hypothesized model (see Figure 2), except that the hypothesized higher order performance dimensions (Task Performance and Contextual Performance) were not included. Instead, the lower order performance factors were allowed to intercorrelate freely. This model was tested to determine whether the four hypothesized lower order performance dimensions could be supported.

For all seven data sets, the four-correlated-dimensions model fit the data well. All of the fit indices reached the standards suggested by Hu and Bentler (1998, 1999) as signifying good fit. All of the chi-square values were large relative to degrees of freedom, but even excellent models generally yield large chi-square values when sample sizes are large, as they were in this research.

Many of the interfactor correlations in this model were quite high. Tables 2 and 3 show that across the seven factor correlation matrices, over half (25 of 42) of the correlations were .70 or higher, and one was greater than .90. In nearly every instance, the largest correlation was between the Human Skills and Citizenship Behaviors factors. Therefore, we tested a series of three-factor solutions, with the Human Skills and Citizenship Behaviors factors combined, to examine the possibility that those two factors are redundant. In every case, however, the decrement in fit was substantial. Chi-square values increased by 18%–57%, and CFI values fell by .01–.02. Cheung and Rensvold (1999) found that changes of that magnitude in the CFI are likely to indicate real differences in models. For those reasons and because Human Skills and Citizenship Behaviors had been hypothesized as separate factors, all four performance dimension factors were retained.

We also examined the factor loadings for the parcels in each data set to ensure that the parcels had been properly assigned to performance dimensions. Results confirmed that parcels were assigned appropriately. The loadings of all parcels in all of the analyses were at least 13 times their standard errors. Most were at least 40 times their standard errors. There was no evidence of significant cross loading. Given the viability of the four lower order performance dimensions, we proceeded to test the simplest possible higher order factor structure. This involves one higher order general factor that loads on all four of the performance dimensions.

One higher order general factor model. This model has a single higher order performance factor (i.e., Task Performance is not distinguished from Contextual Performance). Such a model cannot fit the data more closely than does the four-correlated-dimensions model, because the one higher order general factor model attempts to represent all of the relationships among the four dimensional factors in terms of their relationships to the general factor (Rindskopf & Rose, 1988). However, this model is more parsimonious than is the four-correlated-dimensions model and thus might be preferred if its fit is comparable with the fits of the other models.

Table 1 shows that the fit of this model was similar to the fit of the four-correlated-dimensions model. In six of the seven data sets, the fit indexes for the one higher order general factor model were identical to the corresponding indexes for the four-correlated-dimensions model. The only data set in which there were differences was the Benchmarks self-ratings, for which the RMSEA, the NNFI, and the CFI were slightly poorer for the one higher order general factor model.

We note that in most cases, the Administrative Skills, Human Skills, and Citizenship Behaviors factors in this model loaded much more heavily on the general factor than did the Technical Skills factor (see Table 4). Median loadings were .69 for the Technical Skills factor and .92, .89, and .92 for the Administrative Skills, Human Skills, and Citizenship Behavior factors, respectively. Hence, the general factor tends to share considerably less of its variance with Technical Skills ratings than with ratings on the other aspects of performance. This point has implications for understanding the nature of ratings data, and we return to it later.

Hypothesized model. This model posited that the Technical Skills and Administrative Skills factors would be more highly

Table 1
Fit Statistics for the Four Models for Each Instrument and Each Perspective

Perspective and model	χ^2	*df*	SRMSR	RMSEA	NNFI	CFI
Management Skills Profile[a]						
Bosses ($N = 3{,}424$)						
Unidimensional	69,070	464	.22	.34	.40	.44
Four correlated dimensions	5,679	442	.04	.06	.95	.96
Hypothesized[b]	5,757	443	.04	.06	.95	.96
One higher order general factor	5,975	444	.04	.06	.95	.96
Peers ($N = 10{,}625$)						
Unidimensional	214,913	464	.26	.34	.40	.44
Four correlated dimensions	14,288	442	.03	.05	.96	.96
Hypothesized[b]	14,473	443	.03	.05	.96	.96
One higher order general factor	14,734	444	.03	.05	.96	.96
Subordinates ($N = 12{,}671$)						
Unidimensional	257,447	464	.27	.35	.42	.46
Four correlated dimensions	16,475	442	.03	.06	.96	.97
Hypothesized	16,563	443	.03	.06	.96	.97
One higher order general factor	16,765	444	.03	.06	.96	.97
Self ($N = 14{,}388$)						
Unidimensional	33,397	104	.09	.18	.78	.81
Four Correlated dimensions	5,261	98	.04	.06	.96	.97
Hypothesized	5,303	99	.04	.06	.96	.97
One higher order general factor	6,136	100	.04	.06	.96	.97
Benchmarks[c]						
Peers ($N = 1{,}698$)						
Unidimensional	25,631	464	.23	.29	.41	.44
Four correlated dimensions	2,345	442	.04	.05	.95	.96
Hypothesized	2,350	443	.04	.05	.95	.96
One higher order general factor	2,379	444	.04	.05	.95	.96
Subsordinates ($N = 1{,}546$)						
Unidimensional	22,622	464	.22	.29	.44	.47
Four correlated dimensions	2,122	442	.04	.05	.96	.96
Hypothesized	2,124	443	.04	.05	.96	.96
One higher order general factor	2,151	444	.04	.05	.96	.96
Self ($N = 1{,}722$)						
Unidimensional	3,042	104	.07	.15	.77	.80
Four correlated dimensions	753	98	.05	.06	.95	.96
Hypothesized	780	99	.05	.06	.95	.95
One higher order general factor	823	100	.05	.07	.94	.95

Note. SRMSR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; NNFI = nonnormed fit index; CFI = comparative fit index.
[a] Sevy et al. (1985). [b] This model produced one or more improper parameter estimates. [c] Lombardo and McCauley (1994).

correlated with each other (as facets of Task Performance) than with the other factors, and that the same would be true of the Human Skills and Citizenship Behaviors factors (facets of Contextual Performance). However, as indicated in Tables 2 and 3, the Administrative Skills factor was generally more highly correlated with the two hypothesized Contextual Performance factors than with the Technical Skills factor.

CFAs of the hypothesized structure reflected the same problem. For all rater perspectives on both instruments, Table 5 suggests that Task Performance is much more highly correlated with Contextual Performance, $\varphi(2, 1)$, than with Technical Skills, $\gamma(9, 1)$. This is inconsistent with the hypothesized relationships among those factors. A comparison of the Administrative Skills and Technical Skills factors' loadings on the Task Performance factor also suggests problems. The Administrative Skills factor's standardized loadings were all .97 or higher, with two of them resulting in improper (i.e., >1) estimates. The Technical Skills factor loadings (*Mdn* = .69) were much lower. Thus, whereas Administrative Skills shared essentially all of its variance with the Task Performance factor, Technical Skills shared less than half its variance with Task Performance.

For those reasons, it was clear that the Administrative Skills factor was misplaced in the hypothesized model. This is especially evident in the MSP data. It could be argued that the hypothesized model is appropriate for the Benchmarks data, as there were no improper solutions and the overall fit statistics for the hypothesized model were generally just as good as for the other models. We argue, however, that the hypothesized model is not the model of choice in those data, because the factor loadings are not consistent with the notion that Administrative Skills are best grouped with Technical Skills.

Table 2
Performance Dimension Factor Intercorrelations for the MSP

Performance dimension	1	2	3	4
Bosses and peers				
1. Technical Skills	—	.63	.38	.51
2. Administrative Skills	.64	—	.79	.81
3. Human Skills	.45	.82	—	.84
4. Citizenship Behaviors	.57	.82	.83	—
Subordinates and self				
1. Technical Skills	—	.68	.56	.63
2. Administrative Skills	.67	—	.86	.86
3. Human Skills	.54	.82	—	.88
4. Citizenship Behaviors	.62	.87	.91	—

Note. In the top portion of the table, entries above the diagonal are for boss ratings and entries below the diagonal are for peer ratings. In the lower portion of the table, entries above the diagonal are for subordinate ratings and entries below the diagonal are for self-ratings. MSP = Management Skills Profile (see Sevy et al., 1985).

We believe that the choice for the most appropriate model must be made between the four-correlated-dimensions model and the one higher order general factor model. If chi-square difference tests were applied, the four-correlated-dimensions model would be the clear choice in every data set. But it is well known that chi-square statistics are directly related to sample size, and with samples as large as the ones used in this study, even minor differences between models are likely to be statistically significant. Monte Carlo research suggests that a combination of chi-square testing and comparisons of fit indexes is the most effective method for detecting those differences (Cheung & Rensvold, 1999).

In six of our seven data sets, there are no differences between the four-correlated-dimensions and the one higher order general factor models in terms of SRMSR, RMSEA, NNFI, or CFI. In the remaining data set (Benchmarks self), there are differences of .01 in three of those indexes. We selected the one higher order general factor model over the four-correlated-dimensions model because it offers the better combination of model fit, parsimony, and conceptual clarity. The one higher order general factor model provides more insight into the higher order nature of performance ratings by drawing more attention to the Technical Skills factor's somewhat unique relationships with the other lower and higher order performance factors than does the four-correlated-dimensions model.

Table 3
Performance Dimension Factor Intercorrelations for Benchmarks[a]

Performance dimension	1	2	3	4
Peers				
1. Technical Skills	—			
2. Administrative Skills	.72	—		
3. Human Skills	.51	.70	—	
4. Citizenship Behaviors	.66	.78	.82	—
Subordinates and self				
1. Technical Skills	—	.83	.65	.75
2. Administrative Skills	.70	—	.76	.82
3. Human Skills	.54	.78	—	.81
4. Citizenship Behaviors	.69	.81	.84	—

Note. In the top portion of the table, entries below the diagonal are for peer ratings. In the lower portion of the table, entries above the diagonal are for subordinate ratings and entries below the diagonal are for self-ratings.
[a] Lombardo and McCauley (1994).

Table 4
Standardized Loadings on the General Factor in the One Higher Order General Factor Model

	Performance dimension			
Perspective	Technical Skills	Administrative Skills	Human Skills	Citizenship Behaviors
Management Skills Profile[a]				
Bosses	.57	.92	.89	.92
Peers	.63	.93	.90	.91
Subordinates	.69	.93	.93	.94
Self	.65	.90	.91	.98
Benchmarks[b]				
Peers	.71	.86	.84	.95
Subordinates	.86	.93	.83	.92
Self	.72	.92	.84	.92

[a] Sevy et al. (1985). [b] Lombardo and McCauley (1994).

Phase 3: Generalizability of Factor Structures Across Rater Perspectives

The first component of this phase was to conduct tests of our conclusion that a single model can adequately represent the data

Table 5
Standardized Parameter Estimates in the Hypothesized Model

	Parameter				
Perspective	$\gamma(9, 1)$[a]	$\gamma(10, 1)$[a]	$\gamma(11, 2)$[b]	$\gamma(12, 2)$[b]	$\varphi(2, 1)$[c]
Management Skills Profile[d]					
Bosses	.59	1.08[e]	.91	.92	.81
Peers	.64	1.03[e]	.92	.91	.87
Subordinates	.69	.99	.94	.94	.93
Self	.68	.99	.92	.99	.89
Benchmarks[f]					
Peers	.75	.97	.86	.97	.84
Subordinates	.85	.97	.87	.94	.90
Self	.72	.97	.89	.95	.90

[a] $\gamma(9, 1)$ and $\gamma(10, 1)$ are the standardized loadings of the Task Performance factor on Technical Skills and Administrative Skills, respectively. [b] $\gamma(11, 2)$ and $\gamma(12, 2)$ are the standardized loadings of the Contextual Performance factor on Human Skills and Citizenship Behaviors, respectively. [c] $\varphi(2, 1)$ is the estimated correlation between the Task Performance and Contextual Performance factors. [d] Sevy et al. (1985). Improper parameter estimates. [e] Values are improper. [f] Lombardo and McCauley, (1994).

from all rater perspectives. The second was to determine whether parameter estimates are invariant across perspectives. Tests were conducted using the multisample feature of LISREL 8 (Jöreskog & Sörbom, 1996). Multisample analyses accept data from multiple samples simultaneously but, unfortunately, require the same number of variables to be measured in each sample. Because there were differences in the total number of parcels across rater perspectives (two sets of parcels for boss, peer, and subordinate ratings, but only one set for self-ratings), we decided to estimate two types of multisample models. In the first type, we included only those perspectives for which there were two raters per ratee. For the MSP data, this included ratings from the two bosses, the two peers, and the two subordinates (but not the self-ratings). For the Benchmarks data, we conducted a similar analysis using the ratings from the two peers and the two subordinates. This type of model allowed us to retain the multiple-raters feature of the boss, peer, and subordinate samples, but it did not allow us to include self-ratings in the analyses.

To include self-ratings in this testing, we conducted a second set of multisample analyses using only one rater per perspective per ratee. In the MSP data, we used the ratings from Boss 1, Peer 1, Subordinate 1, and self. Because there were many Benchmarks ratees ($N = 1{,}582$) whose performance had been rated by one boss (although only a handful had been rated by two bosses), we were also able to include the boss ratings in this type of multisample analysis with the Benchmarks data. Therefore, all four of the rater perspectives were included in both the MSP and the Benchmarks analyses.

We followed a sequence of measurement equivalence tests recommended by Vandenberg and Lance (2000). The first test determines whether there are differences in the covariance matrices across the multiple samples. Equivalence of the covariance matrices indicates that the same models will fit each data set well and that the parameter estimates will be comparable across all samples. Thus, no further testing is required. If the covariance matrices are not equivalent, subsequent tests can determine the source of the differences across samples.

Our test with the MSP multiple-rater samples (i.e., with boss, peer, and subordinate ratings) suggested there are few, if any, differences across perspectives, $\chi^2(1{,}056, N = 26{,}720) = 9{,}521$; SRMSR = .09; RMSEA = .03; NNFI = .99; and CFI = .99. Similarly, the Benchmarks (peers and subordinates) data indicated no differences in covariance matrices across perspectives, $\chi^2(528, N = 3{,}244) = 846$; SRMSR = .04; RMSEA = .02; NNFI = .99; and CFI = 1.00. Chi-square values were large. But because the samples were large and the other fit indexes were very good, we accepted the equivalence across perspectives of the covariance matrices for each instrument. The equivalence of the covariance matrices indicates that there are no differences in factor structures across boss, peer, and subordinate groups for the MSP or Benchmarks. We emphasize that our results signify equivalence across perspectives within these instruments, but they do not imply equivalence across instruments.

Results were somewhat different when we examined the equivalence of the covariance matrices based on one rater per perspective, including self: MSP, $\chi^2(408, N = 41{,}058) = 10{,}130$; SRMSR = .26; RMSEA = .05; NNFI = .98; and CFI = .99; Benchmarks, $\chi^2(408, N = 4{,}966) = 2{,}334$; SRMSR = .26; RMSEA = .05; NNFI = .97; and CFI = .98. The SRMSR values suggest that there are important differences across perspectives on each instrument. Hu and Bentler (1998, 1999) have shown that the SRMSR is more sensitive to misspecifications of factor covariances than are the other fit indexes. Factor covariances are an important aspect of the current study because of their implications for understanding the higher order relationships among rating constructs.

To determine the nature of those differences, we began with a multisample analysis of a model with four first-order performance factors (Technical Skills, Administrative Skills, Human Skills, and Citizenship Behaviors) and a second-order general factor. This model is the one-rater analogue of the one higher order general factor we selected above. We first tested configural invariance by fitting this model with no equality constraints across samples, which tests the hypothesis that each group is using the same conceptual frame of reference in responding to the items (Vandenberg & Lance, 2000). Results of these and the subsequent analyses are presented in Table 6. For each instrument, results generally indicated that this model fit the data well. RMSEA values in all of these analyses were admittedly slightly higher than the .06 standard specified by Hu and Bentler (1998, 1999), but as we indicate below, it is the SRMSR values that we considered to be most crucial to these analyses. Thus, we concluded that discrepancies in the covariance matrices did not stem from differences in the factor

Table 6
Fit Statistics for Multisample Analyses With One Rater Per Perspective

Equality constraints	χ^2	*df*	SRMSR	RMSEA	NNFI	CFI
Management Skills Profile[a]						
None	23,325	400	.04	.07	.96	.96
$\lambda\text{-}_{y}$[b]	24,722	436	.05	.07	.96	.96
θ_{ϵ}[c]	28,510	484	.05	.08	.96	.96
γ[c]	28,884	493	.07	.08	.96	.96
ψ[d]	29,987	505	.08	.08	.96	.95
ϕ[e]	32,761	508	.26	.08	.95	.95
Benchmarks[d]						
None	3,826	400	.05	.08	.95	.96
$\lambda\text{-}_{y}$[b]	4,483	436	.06	.08	.95	.95
θ_{ϵ}[c]	4,998	484	.07	.08	.95	.94
γ[e]	5,072	493	.07	.08	.95	.94
ψ[f]	5,292	505	.09	.08	.94	.94
ϕ[g]	5,578	508	.26	.08	.94	.94

Note. Equality constraints force corresponding parameter estimates to be equal across all rater perspectives for a given instrument. The equality constraints in each of the models also include all prior constraints. SRMSR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; NNFI = nonnormed fit index; CFI = comparative fit index.
[a] Sevy et al. (1985). [b] Corresponding loadings of the four performance factors on the measured variables (parcels) were constrained to be equal. [c] Corresponding error variances on the measured variables were constrained to be equal. [d] Lombardo and McCauley (1994). [e] Corresponding loadings of the general factor on the performance dimension factors were constrained to be equal. [f] Corresponding disturbance terms for the performance dimension factors were constrained to be equal. [g] Variance of the general factor was constrained to be equal.

patterns. We then proceeded to introduce a series of increasingly restrictive equality constraints.

The first constraint forced the first-order factor loadings (i.e., λ–y, the loadings on the performance dimension) for like parcels to be equal across all perspectives for a given instrument, which tests the equality of scaling units across groups (see Vandenberg & Lance, 2000). Results again indicated good fit in both instruments (see Table 6). We then constrained the corresponding unique variances in the measured variables (θ_ϵ) and again found that the overall model fit was good in both instruments. This indicated that it was factor variances and/or covariances that were responsible for the differences in the observed covariance matrices.

Vandenberg and Lance (2000) recommend that researchers test the equivalence of factor variances and covariances before testing the equality of structural parameters (general factor loadings in this case) across samples. Results of those tests revealed that in each instrument, the factor variances and covariances were not equivalent across all rater perspectives. It is interesting that for both instruments, there was little, if any, difference across rater perspectives in the variances of the Technical Skills factor. For the remaining factors (i.e., Administrative Skills, Human Skills, and Citizenship Behaviors), however, variances for self-ratings were consistently lower than those for ratings from bosses, peers, and subordinates. This supports the conclusion that self-raters, as compared with other raters, tend to use smaller ranges of the construct continua (Vandenberg & Lance, 2000) for all dimensions other than Technical Skills. This is additional evidence suggesting that ratings of technical skills are somewhat unique.

Given that there were differences in factor variances and covariances across perspectives for both instruments, we estimated models in which the corresponding loadings on the general factor (i.e., γ) were constrained to be equal across perspectives. Disturbance terms for the four performance factors were not constrained nor was the variance of the general factor. Results indicated that this model fit well for all data sets (see Table 6). We then also constrained the disturbances of the performance factors (ψ) to be equal across perspectives for each instrument and again found that the model fit reasonably well for each instrument (see Table 6). In the next step, we constrained the general factor variances (ϕ) to also be equal across perspectives for each instrument. SRMSR values (.26 for both instruments) in Table 6 show that this model fit poorly for each instrument. We then relaxed the requirement that the variance of the self-ratings general factor be equal to the variances of the general factors for the other perspectives and found that this model again fit reasonably well. That is, for each instrument the model fit well when the variance of the general factor for self-ratings was allowed to differ from the variances of the general factor for the other perspectives. This supports our earlier conclusion that self-ratings are generally confined to a smaller portion of the construct continuum than are ratings from other perspectives, at least on the administrative, human, and citizenship aspects of performance.

Discussion

Despite the widespread use of developmental ratings, little is known about their construct validity. In response to this, our study was designed to (a) determine whether developmental ratings measure four conceptually meaningful aspects of managerial performance, (b) examine structural relationships among ratings of those aspects of performance, and (c) assess the generalizability of our findings across four rater perspectives and two rating instruments.

With respect to the lower order dimensions of performance, every analysis reported in this study supported the existence of the four hypothesized performance factors—Technical Skills, Administrative Skills, Human Skills, and Citizenship Behaviors—as conceptually and empirically distinguishable ratings factors. This suggests that ratings reflect not only the three skill types associated conceptually with successful managerial performance by R. L. Katz (1974) and Mann (1965), but also citizenship behaviors that have been linked conceptually (D. Katz, 1964) and empirically (Podsakoff, Ahearne, & MacKenzie, 1997) to organizational effectiveness. Our results clearly show that our hypothesized lower order factors generalized across rater perspectives and rating instruments. Overall, these results provide strong support for the construct validity for the shared (across raters) variance in developmental boss, peer, subordinate, and self-ratings.

It is significant that the nature of the performance constructs and the interrelationships among them generalized across all of the rater perspectives and all of the rating instruments included in this study. Borman's (1997) review of the current state of knowledge regarding 360° feedback systems indicated that one possible reason for the low levels of interrater agreement across perspectives is that different-perspective raters either use different performance dimensions to evaluate performance or define the dimensions differently. The generalizability of the factor structures across rater perspectives in the current study argues against that possibility, suggesting instead that raters from all perspectives attend to a similar set of core performance factors. Our study extends Facteau and Craig's (2001) finding that raters across various perspectives (bosses, peers, subordinates, and self) share a common conceptualization of managerial performance dimensions. Because their research focused on a single feedback system and a single organization, they were unable to test the generalizability of their results to other systems and organizations. Our results suggest that their conclusions do generalize and are clearly on the side of few, if any, differences across perspectives in terms of the nature of the constructs commonly used by raters from that perspective.

The generalizability of our results was also demonstrated by the consistency with which they indicated that the Administrative Skills factor in our hypothesized model is more closely related to the Human Skills and Citizenship Behaviors factors than to the Technical Skills factor. We are not the first to have wrestled with how to categorize administrative skills. Conway (1999) suggested that researchers examine whether administrative performance is distinct from technical performance. Our study suggests that the two are distinct and that they have very different relationships with the human and citizenship components of performance.

The misplacement of the Administrative Skills factor in turn indicated that the higher order structure of the ratings is somewhat different from what we had hypothesized. Although our results did not support the Borman and Motowidlo (1993) higher order structure in exactly the ways that we had hypothesized, this does not imply that Borman and Motowidlo's conceptual distinction between task and contextual performance is inappropriate. In fact, we believe our results lend indirect support to the Borman and Motowidlo model. Our Technical Skills factor is similar to Borman and Motowidlo's Task Performance in that it includes the functional activities associated with the manager's functional area. The same is true of the relationship between their Contextual Performance and our Administrative Skills, Human Skills, and Citizenship Behavior factors. Behaviors associated with the Human Skills and Citizenship Behaviors factors undoubtedly contribute to the organization's social and psychological climate in the ways outlined by Borman and Motowidlo. Elements of administrative performance also impact the social psychological climate. For example, a manager's ability to set appropriate goals, to plan effectively for their accomplishment, and to delegate tasks and authority to the right people surely affects the quality of the social environment in which people work.

That, coupled with our general factor's relationship with the Technical Skills factor being quite different from its relationship with the other three performance factors, suggests the possibility that the highest order distinction that raters make is between technical performance and performance on other (i.e., nontechnical) tasks. We see that distinction as being somewhat similar to the one between task and contextual performance. It is also generally consistent with Kavanagh, Borman, Hedge, and Gould's (1987) position that there are two types of job performance dimensions, technical competency skills and job-relevant interpersonal skills, and with Murphy's (1989) identification of task accomplishment and interpersonal relations as performance dimensions that are universal to all jobs. Ideally, we would have been able to test a model with Technical Skills and Nontechnical Skills as higher order factors against the higher order general factor model. No statistical comparison can be made, however, because the two models have the same degrees of freedom and fit indexes.

Despite the similarities across rater perspectives, our findings did indicate two differences. One was a tendency for subordinates' ratings to be more highly correlated across dimensions than were ratings from other perspectives (see Tables 2 and 3). This is consistent with Borman's (1987) suggestion that rater constructs develop and sharpen over time. To the extent that our subordinate ratings were made by less experienced raters than those who provided the boss ratings, the subordinates' performance constructs could be less well-developed and less sharply focused than the constructs used by bosses. Other empirical evidence also supports the idea that experience results in more highly differentiated systems (Ilgen & Feldman, 1983) and that more knowledgeable raters are less prone to rely on general impressions (Kozlowski & Kirsch, 1987) than are less knowledgeable raters.

Our results also showed that self-ratings of most aspects of performance tend to be less variable than boss, peer, or subordinate ratings. This conclusion is similar to one reached by Facteau and Craig (2001). However, our results suggest that technical skills may be an exception to this rule. The variance of self-ratings of technical skills in our study differed little, if at all, from the variances of ratings made by raters from other perspectives. We note that in the Facteau and Craig study, self-ratings variance was most similar to the variances of ratings from other sources in business knowledge, a dimension that resembles aspects of our Technical Skills factor. Thus, both our study and the one conducted by Facteau and Craig suggest that although self-ratings are generally less variable (across ratees) than ratings made by raters from other perspectives, this may not be the case in areas involving function-specific aspects of the job.

When interpreting all of the results of our study, it is important to be aware of two issues. First, the effects of the performance factors on observed ratings must be considered in the context of other influences, particularly idiosyncratic effects. Most of this research examined factors that are present across multiple raters per perspective (for bosses, peers, and subordinates). Hence, effects that are idiosyncratic to individual raters (except self) were, by definition, outside our area of interest. But idiosyncratic factors have strong effects on ratings (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998), and so far, very little research has examined the possible differences across rater perspectives in terms of the nature of idiosyncratic rater effects that are present or the types of influences that might generate those effects. Second, researchers must also be aware that even if raters from different perspectives attend to analogous constructs, this does not imply agreement across perspectives in terms of the constructs they consider to be most important or in the standards they use to judge performance.

Suggestions for Future Research

Conway (1999) observed that research has been moving from very simple toward somewhat more complex performance models. The four-factor models in this study fit better than the three-factor models, but other models with more or different factors might have fit even better. Performance models advocated by J. P. Campbell (1990), Borman (1987), Borman and Brush (1993), Tornow and Pinto (1976), and Yukl (1989) suggest many potential factors. Future researchers could use those models to test the possibility that managers use different or finer-grained distinctions than those studied in this research.

The high correlations between the Human Skills and Citizenship Behaviors factors in our research suggest that none of the rater perspectives in our research discriminated very strongly between those factors. This could be due to the lack of precision with which these types of constructs have been defined in the literature. Organ (1997) reviewed the various conceptualizations of OCB and concluded that "[it's] construct clean-up time" (p. 85). Researchers (e.g., Coleman & Borman, 2000) have been working on doing so. It is possible that the interfactor correlations in this study were inflated because of that lack of clarity and that future research will reveal how the items composing those factors could be divided into more homogeneous and conceptually distinct groupings that would be less highly correlated.[2]

Also, we join with other researchers (e.g., Conway, 1999; Van Scotter & Motowidlo, 1996) in their call for research that continues to examine the higher order structure of performance. At the same time, however, we reiterate our view that the structure of job performance is not necessarily the same as the structure of performance ratings. If the higher order structure of performance actually does differ from the structure of ratings, it could have important

[2] We thank two reviewers for pointing out that another possible reason for the lack of discrimination between human and citizenship behaviors was that our instruments were not specifically designed to make that distinction.

ramifications for the validity of performance ratings. It would lead to significant questions such as the following: Do the differences in factor structures have a practically significant impact on ratings? If so, is it possible through, for example, rater training to modify the factor structure of ratings? If not, which model should serve as the conceptual basis for designing instruments used to measure performance? For these and many other reasons, researchers should continue to investigate the higher order structure of job performance itself, the higher order structure of ratings of performance, and the correspondence between the two.

Another important area for future research to examine is the generalizability of our factor structure to ratings made for purposes other than development. So far, very little research has explored the factor structure of administrative ratings. We encourage researchers to engage in that type of research. We speculate briefly here about how their results might compare with ours. Research has shown that, compared with developmental ratings, ratings made for administrative purposes tend to be more lenient and more prone to exhibit halo error (Farh, Cannella, & Bedeian, 1991; Jawahar & Williams, 1997). Disregarding potential complicating factors (e.g., ceiling effects), we argue that a rater's consistent tendency to be lenient or harsh would have no effect on the factor structure, because adding the same constant to all observations has no effect on that variable's correlations with other variables.

The tendency for administrative ratings to exhibit a greater degree of halo may seem at first to suggest that CFAs like those done for this study, but based on administrative ratings, would result in higher interdimensional correlations or higher loadings on the general factor for the performance dimensions. However, we do not believe that this would be the case. Instead, we hypothesize that (a) theoretically, the only relevant source of variance and covariance for the latent performance factors is performance itself; (b) this renders the expected values of the variances and covariances of the performance factors in administrative ratings equal to those for developmental ratings; and, therefore, (c) expected values of the performance factor intercorrelations, or their loadings on the general factor, are identical for the two types of ratings.

In terms of our Figure 2, what we would expect to differ across administrative and developmental factor structures is the relative magnitudes of the loadings of the dimension–rater factors (e.g., Technical Skills–Rater 1) on the performance dimensions (e.g., Technical Skills). Those loadings should be smaller (and the disturbance terms for the dimension–rater factors should be larger) in the case of administrative ratings, because rater effects are expected to be stronger in administrative ratings than in developmental ratings. The essence of our prediction is that the factor structures of developmental and administrative ratings would be similar in their nature to, but different in their loading patterns from, developmental ratings exhibiting larger performance dimension loadings and smaller rater loadings than administrative loadings.

Implications for Practice

The findings of this study are relevant to several areas of practice. First, our results, especially in combination with those of Facteau and Craig (2001), support the common practice of comparing 360° feedback results across rater sources. Facteau and Craig found, using one instrument in one organization, that conceptual frames of reference were similar across rater perspectives. Our findings extend theirs by confirming that results generalize across rating instruments and across managers in multiple industries, functions, and organizational levels.

Second, our higher order general factor model suggests a means through which ratees might be better able to make sense of their ratings. Ratees in developmental feedback systems with many (often 15–25) scales and several rater sources receive a great deal of performance information. It might be helpful for ratees if their results were framed, at least initially, in terms of their general performance and then on the technical, administrative, human, and citizenship aspects of the job. This might aid ratees in their assimilation of feedback information.

Finally, if we are correct in hypothesizing that the factor structure of administrative ratings is similar to the factor structure of developmental ratings, then another area of application is in the content and design of performance evaluation systems and the associated rater-training systems. DeNisi (1996) has argued that evaluation systems organized around important rater constructs may be more meaningful and less cognitively demanding for raters than are systems built around other constructs. This could mean that if raters used systems designed to measure what they consider to be the right constructs, they would be less hesitant to rate performance and would be better prepared to explain or defend their evaluations if necessary. Research also shows that ratings exhibit more variance (i.e., raters make greater distinctions) when the system is based on the rater's own personal constructs than when based on other dimensions (Borman, 1983). Similarly, ratees may be more likely to accept ratings if the constructs match their own (DeNisi, 1996). All of this suggests that both raters and ratees prefer performance evaluation systems that are aligned with their own views about the important components of performance.

The similarity of the factor structures across rater perspectives in our research and the fact that this research was based on samples of raters and ratees from a variety of industries, functions, and organizational levels suggest that development of an evaluation instrument based on common rater and ratee constructs could be less difficult than might be expected. It is possible, of course, that subgroups of raters (e.g., managers in a specific organization, industry, or function) might have performance models with features that are unique to that group, but the results of this study suggest that even if this is true, the factors included in this research are still likely to be prominent.

Concerning rater training, DeNisi (1996) argued that much of the training designed to reduce rater errors (e.g., leniency and halo) is relatively short-lived and that refocusing the goals of training might produce more widespread and longer lasting effects. He suggested that organizations study the performance theories that guide their raters' evaluation processes, determine which theories best serve the organization's interest, and then design training toward helping raters understand and adopt those theories. The performance factors hypothesized in the current research could serve as a starting point for those studies. It would also be valuable for them to identify individual raters whose performance models are not consistent with organizational goals. Training could be designed to help those raters adopt a more appropriate theory of performance. The net result for the organization could include greater consensus on how good performance is defined in the organization and how it can be recognized in practice.

Limitations

Several factors could limit the generalizability of our results. These include the purpose of the ratings, characteristics of the samples, and features of the design. Concerning the purpose for the ratings, and consistent with most 360° feedback research, all of the data in this study were taken from developmental multirater feedback programs. Research has shown that some of the psychometric characteristics of ratings (e.g., means) may vary somewhat, depending on whether ratings have pay, promotion, or retention consequences for ratees. But little research has compared correlational relationships between performance dimensions in developmental ratings with those in administrative ratings. Although we have speculated about the factor structures of administrative versus developmental ratings, we acknowledge that the generalizability of our results to ratings made for administrative purposes is unknown.

Characteristics of the samples used here should also be considered. The participants in these samples represented a wide range of industries, functions, and organizational levels. The possibility that results could vary across subgroups within any of those larger classifications was not examined in this study. Another potentially important issue, again consistent with other 360° feedback research, is that many of the ratees in this study chose their raters. It is possible that some chose raters who they thought would be the most candid and that others chose raters who would give the most favorable ratings. The effects of this freedom of choice on the generalizability of these results to a situation in which ratees are not allowed to choose their own raters is unknown. However, there are two reasons why this might not represent a serious limitation. First, for most of the ratees in our study, the two peer raters and the two subordinate raters were randomly selected from groups of several (often five or more) people who had provided ratings. If the ratees' freedom to choose raters affected our results, we believe the magnitude of that effect is smaller than it would have been if ratees had been allowed to handpick just two raters from each perspective. Another reason why allowing ratees to select their raters may not have been problematic is that this is not an uncommon practice in developmental feedback programs. Therefore, this freedom of choice is a realistic condition for making and studying developmental ratings.

Concerning features of the design, the major drawback was that the difficulty in obtaining ratees with two or more bosses prevented us from including boss ratings in some of our Benchmarks analyses. We point out, however, that our Benchmarks data did allow us to include boss ratings in our one-rater type of multisample analyses and that there were no indications that any of our conclusions would have been materially different if we had been able to incorporate boss ratings in the remaining Benchmarks analyses.

Conclusion

Despite the research community's longstanding concern with "the criterion problem" (Austin & Villanova, 1992), relatively little past research has attempted to validate ratings, developmental or administrative. One of the potentially major problems is that implicit or folk theories may exert powerful and systematic influences on performance ratings, but "[at] present we know very little about which dimensions of performance [raters] typically emphasize, about variation from supervisor [rater] to supervisor [rater] in the dimensions that are attended to, or about the circumstances that will lead to either widespread consensus or widespread disagreement" (Murphy & Cleveland, 1995, p. 120).

Our study provides insights into some of those questions. It suggests that developmental ratings measure constructs that are relevant to individual and organizational success and that the performance constructs reflected in the ratings variance that is shared across raters vary little from one perspective to another. A significant strength of this study is that its results withstood cross-validation in very large samples taken from conceptually diverse measures. This is strong evidence that our results are not peculiar to the samples or to the instruments used in this research.

References

American Psychological Association. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology, 77,* 836–874.

Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods, 1,* 45–87.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Bernardin, H. J., & Orban, J. E. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology, 5,* 197–211.

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. (1983). Implications of personality theory and research for the rating of work performance in organizations. In F. J. Landy, S. Zedeck, & J. C. Cleveland (Eds.), *Performance measurement and theory* (pp. 127–165). Hillsdale, NJ: Erlbaum.

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.

Borman, W. C., & Brush, D. H. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance, 6,* 1–21.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 71–98). San Francisco: Jossey-Bass.

Brief, A. P., & Motowidlo, S. J. (1986). Prosocial organizational behavior. *Academy of Management Review, 11,* 710–725.

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.

Brush, D. H., & Licata, B. J. (1982, August). *Managerial performance: Toward a definition of its construct domain.* Paper presented at the 90th Annual Convention of the American Psychological Association, Washington, DC.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 1, 2nd ed., pp. 687–732). Palo Alto, CA: Consulting Psychologists Press.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25,* 1–27.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Coleman, V. I., & Borman, W. C. (2000). Investigating the underlying structure of the citizenship performance domain. *Human Resource Management Review, 10,* 25–44.

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84,* 3–13.

DeNisi, A. S. (1996). *Cognitive approach to performance appraisal: A program of research.* New York: Routledge.

Dobbins, G. H., Cardy, R. L., & Truxillo, D. M. (1988). The effects of purpose of appraisal and individual differences in stereotypes of women on sex differences in performance ratings: A laboratory and field study. *Journal of Applied Psychology, 73,* 551–558.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86,* 215–227.

Farh, J., Cannella, A. A., & Bedeian, A. G. (1991). Peer ratings: The impact of purpose on rating quality and user acceptance. *Group & Organization Studies, 16,* 367–386.

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

Fox, S., & Bizman, A. (1988). Differential dimensions employed in rating subordinates, peers, and superiors. *Journal of Psychology, 122,* 373–382.

George, J. M., & Brief, A. P. (1992). Feeling good–doing good: A conceptual analysis of the mood at work–organizational spontaneity relationship. *Psychological Bulletin, 112,* 310–329.

Heneman, R. L. (1986). The relationship between supervisory ratings and results-oriented measures of performance: A meta-analysis. *Personnel Psychology, 39,* 811–826.

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.

Hurley, A. E., Scandura, T. A., Scriesheim, C. E., Brannick, M. T., Seers, A., Vandenberg, R. J., & Williams, L. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior, 18,* 667–683.

Ilgen, D. R., & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (Vol. 5, pp. 141–197). Greenwich, CT: JAI Press.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal-purpose effect. *Personnel Psychology, 50,* 905–925.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Katz, D. (1964). The motivational basis of organizational behavior. *Behavioral Science, 9,* 131–146.

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* (Rep. No. 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.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757–765.

Kozlowski, S. W. J., & Kirsch, M. P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology, 72,* 252–261.

Krzystofiak, F., Cardy, R., & Newman, J. (1988). Implicit personality and performance appraisal: The influence of trait inferences on evaluations of behavior. *Journal of Applied Psychology, 73,* 515–521.

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., 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.

Landis, J. R., & Koch, G. C. (1979). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lau, A. W., Newman, A., & Broedling, L. A. (1980). The nature of the managerial work in the public sector. *Public Administration Review, 40,* 513–520.

Lawler, E. E., III. (1967). The multitrait–multirater approach to measuring managerial job performance. *Journal of Applied Psychology, 51,* 369–381.

Lindsey, E., Homes, V., & McCall, M. W., Jr. (1987). *Key events in executives' lives* (Tech. Rep. No. 32). Greensboro, NC: Center for Creative Leadership.

Lombardo, M. M., & McCauley, C. D. (1994). *Benchmarks: A manual and trainer's guide.* Greensboro, NC: Center for Creative Leadership.

MacKenzie, S. B., Podsakoff, P. M., & Fetter, R. (1991). Organizational citizenship behavior and objective productivity as determinants of managerial evaluation of salespersons' performance. *Organizational Behavior and Human Decision Processes, 50,* 123–150.

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., & Bailey, M. (1991). Confirmatory factor analysis 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.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83,* 693–702.

Mintzberg, H. (1975). The manager's job: Folklore and fact. *Harvard Business Review, 53,* 49–61.

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.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79,* 475–480.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37,* 687–702.

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.

Murphy, K. R. (1989). Dimensions of job performance. In R. F. Dillon & J. W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives* (pp. 218–247). New York: Praeger.

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.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome.* Lexington, MA: Lexington Books.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10,* 85–97.

Podsakoff, P. M., Ahearne, M., & MacKenzie, S. B. (1997). Organizational citizenship behavior and the quantity and quality of work group performance. *Journal of Applied Psychology, 82,* 262–270.

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.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67.

Sevy, B. A., Olson, R. D., McGuire, D. P., Frazier, M. E., & Paajanen, G. (1985). *Management skills profile technical manual.* Minneapolis, MN: Personnel Decisions.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68,* 653–663.

Tornow, W. W., & Pinto, P. R. (1976). The development of a managerial job taxonomy: A system for describing, classifying, and evaluating executive positions. *Journal of Applied Psychology, 61,* 410–418.

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.

Tsui, A. S., & Ohlott, P. (1988). Multiple assessment of managerial effectiveness: Interrater agreement and consensus in effectiveness models. *Personnel Psychology, 41,* 779–803.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Umesh, U. N., Peterson, R. A., & Sauber, M. H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement, 49,* 835–850.

Vance, R. J., MacCallum, R. C., Coovert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology, 73,* 74–80.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.

Van Dyne, L., Graham, J. W., & Dienisch, R. M. (1994). Organizational citizenship behavior: Construct redefinition, measurement, and validation. *Academy of Management Journal, 37,* 765–802.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81,* 525–531.

Van Velsor, E., & Leslie, J. B. (1991). *Feedback to managers: Volume II. A review and comparison of sixteen multi-rater feedback instruments.* Greensboro, NC: Center for Creative Leadership.

Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor?* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Werner, J. M. (1994). Dimensions that make a difference: Examining the impact of in-role and extra-role behaviors on supervisory ratings. *Journal of Applied Psychology, 79,* 98–107.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Williams, L. J., & Anderson, S. A. (1991). Job satisfation and organizational commitment as predictors of organizational citizenship behaviors. *Journal of Management, 17,* 601–617.

Yukl, G. (1989). Managerial leadership: A review of theory and research. *Journal of Management, 15,* 251–289.

Received June 19, 2000
Revision received April 1, 2002
Accepted April 9, 2002 ■