

## **RATER SOURCE EFFECTS ARE ALIVE AND WELL AFTER ALL**

BRIAN HOFFMAN

University of Georgia

CHARLES E. LANCE AND BETHANY BYNUM

University of Georgia

WILLIAM A. GENTRY

Center for Creative Leadership

Recent research has questioned the importance of rater perspective effects on multisource performance ratings (MSPRs). Although making a valuable contribution, we hypothesize that this research has obscured evidence for systematic rater source effects as a result of misspecified models of the structure of multisource performance ratings and inappropriate analytic methods. Accordingly, this study provides a reexamination of the impact of rater source on multisource performance ratings by presenting a set of confirmatory factor analyses of two large samples of multisource performance rating data in which source effects are modeled in the form of second-order factors. Hierarchical confirmatory factor analysis of both samples revealed that the structure of multisource performance ratings can be characterized by general performance, dimensional performance, idiosyncratic rater, and source factors, and that source factors explain (much) more variance in multisource performance ratings whereas general performance explains (much) less variance than was previously believed. These results reinforce the value of collecting performance data from raters occupying different organizational levels and have important implications for research and practice.

Traditionally, organizations have relied primarily on employees' immediate supervisors to provide job performance ratings (Murphy, 2008). In recent years however, organizations have begun to evaluate work performance from multiple rater sources. To this end, multisource performance ratings (MSPRs, also often referred to as 360° ratings) have enjoyed increased popularity as performance evaluation tools (Church & Allen, 1997). Briefly, MSPR systems require the collection of ratings of job-related competencies from raters occupying multiple sources. Then, organizational constituents' ratings are presented to the target, separated by skill dimension and rater source (typically, supervisors, direct reports, and peers). Although MSPR systems are occasionally used for administrative

---

Correspondence and requests for reprints should be addressed to Brian Hoffman, University of Georgia, Psychology Department, 228 Psychology Building, The University of Georgia, Athens, GA 30602; [hoffmanb@uga.edu](mailto:hoffmanb@uga.edu).

© 2010 Wiley Periodicals, Inc.

purposes (e.g., promotion, raises), the preponderance of MSPR tools are used for employee development (Timmreck & Bracken, 1997).

One key assumption underlying the use of MSPRs is that raters from different sources provide unique performance-relevant information to the ratee that would not be captured by traditional supervisory ratings alone. In essence, due to the complex nature of the job performance construct and interpersonal relations at work, it is believed that a single supervisor's ratings are not sufficient to provide a full picture of a target's work performance. In fact, the implementation of MSPR systems is predicated on the assumption that raters from different sources will *disagree* with respect to their perceptions of target performance (Borman, 1974). Indeed, if all raters provided the same information regarding a ratee's performance, there would be little need to collect performance ratings from multiple raters.

Consistent with the assumptions underlying the use of MSPRs, much of the existing research typically indicates that both rating source and performance dimension factors account for significant proportions of variance in MSPRs (Lance, Teachout, & Donnelly, 1992; Woehr, Sheehan, & Bennett, 2005). Despite this consistent pattern of results, some recent research has questioned the meaningfulness of source effects in multi-source ratings and, by extension, MSPR programs themselves. In particular, Mount, Judge, Scullen, Sytsma, and Hezlett (1998) argued that the latent structure of MSPRs is best characterized by performance dimension factors plus idiosyncratic rater factors, and *not* broader more traditional rater source factors. As a result, Mount et al. suggested that "ratings made by raters within the same source (e.g., two peers or two subordinates) are no more similar to each other than ratings made by raters from different sources (e.g., a boss and a peer or a peer and subordinate)" (p. 572). Scullen, Mount, and Goff (2000) extended the work of Mount and his colleagues by examining a wider range of models of the latent structure of MSPRs but also concluded that source effects accounted for substantially less variance than was accounted for by dimension, rater, or error factors, and thus, these authors also questioned the importance of source effects. Viswesvaran, Schmidt, and Ones (2002) arrived at a similar conclusion in their meta-analysis of supervisor and peer interrater correlations and more recently noted that the source effect now "has been disconfirmed by empirical research" (Viswesvaran, Schmidt, & Ones, 2005, p. 110). Finally, based on their analyses of MSPRs using interrater correlations, intraclass correlations, and within-group interrater agreement indices, LeBreton, Burgess, Kaiser, Atchley, and James (2003) also concluded that ratings were no more similar for different raters within versus between sources. Collectively, these findings appear to undermine the value of MSPR programs that routinely aggregate ratings within source and then interpret

the aggregated performance ratings on the assumption of the existence of systematic source effects.

Although valuable, we believe that (a) this recent research oversimplifies the structure of MSPR data, and (b) the conclusion that MSPR source effects are not meaningful (or nonexistent) is premature at best. In its most extreme form, the argument that is suggested by these findings is that all rater effects on MSPR data are entirely idiosyncratic, and existing research that has supported the presence of source related variance (Scullen et al., 2000) has attributed at most modest variance to the source providing ratings. Although we acknowledge the importance of idiosyncratic rater effects, we also argue that raters from the same source also share perspectives on ratee performance that can be captured at the level of a higher-order, second-order factor (SOF). Therefore, the purpose of this study was to reexamine the structure of MSPRs using the data reported by Mount et al. (1998) and an additional large independent data set to determine the plausibility of rater source factors at the SOF level. Hence, this study contributes to the literature by attempting to clarify the structure of MSPRs, the existence and nature of MSPR source effects, as well as the relative proportion of variance accounted for by the various factors that influence performance ratings.

### *Structure of MSPRs*

Beginning with Wherry's (1952; as cited in Wherry & Bartlett, 1982) seminal work on the subject, the structure and components of performance ratings have been the subject of considerable theoretical and empirical attention for over half a century. Using a variety of methodologies (e.g., analysis of variance, interrater agreement, confirmatory factor analysis, and generalizability theory), prior research supports a variety of components that explain variance in performance ratings including performance dimension effects, general performance effects, idiosyncratic rater effects, rater source effects, and measurement error. Although the primary focus of this study is a reassessment of the role that rater source effects play in MSPRs, source effects can only be understood when viewed in the context of other rating components. Indeed, the failure to model each variance source can result in biased estimates of portion of variance accounted for by the remaining structured components. In the following sections we review these rating components.

### *Dimensional Performance*

Historically, the preponderance of performance taxonomies (e.g., Borman & Brush, 1993; Mann, 1965; Mintzberg, 1975; Smith, Organ, &

Near, 1983), psychometric models of ratings (Kenny & Berman, 1980; King, Hunter, & Schmidt, 1980; Wherry, 1952), and performance evaluation instruments (Austin & Villanova, 1992; DeVries, Morrison, Shullman, Gerlach, 1986; McCauley & Lombardo, 1990) conceptualize work performance as consisting of multiple related, yet distinct, dimensions. For instance, one of the more popular performance structures supports a distinction between task performance and organizational citizenship behaviors (Hoffman, Blair, Meriac, & Woehr, 2007; Smith et al., 1983). Performance taxonomies specific to managerial work also hypothesize multiple dimensions of performance ranging from planning and organizing the work of others to being the public face of the organization (e.g., Borman & Brush, 1993; Mintzberg, 1975). Although three major psychometric models of ratings (Kenny & Berman, 1980; King et al., 1980; Wherry, 1952) vary somewhat in their operationalization of dimensional performance, functionally, performance dimensions effects are represented by the variance that is common across all raters' ratings of a given performance dimension. On the basis of classical test theory, this dimension-based common variance is seen as arising as a function of variation in ratees' actual job performance levels on each dimension. In that the majority of theoretical and operational work is based around the notion of performance dimensions, relatively strong performance dimensions effects would be expected in MSPRs. Despite this expectation, prior research rarely evidences strong support for the impact of dimensional performance on ratings (Mount et al., 1998; Scullen et al., 2000; Viswesvaran et al., 2002). Instead, performance dimensions often explain less than 10% of the variance in ratings (Conway, 1996; Mount et al., 1998; Scullen et al., 2000). Based on these findings, it is not surprising that although performance dimension effects are present in MSPRs, a variety of other factors have explained the majority of the variance in performance ratings.

### *General Performance*

Typified by Guilford's (1954) general performance factor, Kenny and Berman's (1980) true correlation, Cooper's (1981) true halo, and Viswesvaran et al.'s (2005) actual correlation, the idea of a *valid* general factor in performance ratings has persisted almost as long as ratings have been used to assess human performance. In fact, of the three major psychometric models of ratings, Wherry's (1952) is the only one that does not specify a valid general performance factor. Although there are some differences in the specific conceptualizations and operationalizations of this valid general performance component, in general it reflects variance common to all raters' ratings of all performance dimensions (King et al.,

1980; Scullen et al., 2000; Viswesvaran et al., 2005). Conceptually, this general performance factor reflects the degree to which there exists a “true positive manifold among job performance dimensions” (Viswesvaran et al., 2005, p. 108). Indeed, given common predictors relevant across distinct performance domains (e.g., cognitive ability, conscientiousness, etc.), it should not be surprising that a factor representing the degree to which a ratee is generally effective or ineffective emerges in performance ratings (Cooper, 1981; Feldman, 1981; Scullen et al., 2000; Viswesvaran et al., 2005). Despite the near universality of a valid general performance factor, existing research is unclear as to the magnitude of its effects on performance ratings. For instance, Viswesvaran and his colleagues’ (2005) meta-analysis of peer and supervisor ratings concluded that at the observed level (e.g., uncorrected), the general performance factor explains 27.4% of the variance in performance ratings. Scullen et al. (2000) also supported a general performance factor in two large samples of MSPRs; however, the proportion of variance explained by general performance in Scullen et al. (13.5%) was half the size of Viswesvaran et al.’s uncorrected estimate. Despite these differences in prior estimates of the proportion of variance in ratings attributable to a general factor, existing evidence and theory point toward the presence of a general factor in performance evaluations. Accordingly, we also expected a general performance factor to explain significant variance in MSPRs.

### *Rater Effects*

Rater effects refer to two distinct sources of variance in performance ratings: variance attributable to the individual rater (idiosyncratic rater effects) and variance attributable to rater source. Existing research has adopted two primary approaches to assess the presence and pervasiveness of source effects. First, research has compared the correspondence of ratings from same-source raters (e.g., two supervisors) to that of different-source raters (e.g., a peer and a supervisor). Based on this stream of research, conventional wisdom is that there exists a systematic effect of rater source that transcends individual rater differences within different sources (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Viswesvaran, Ones, & Schmidt, 1996). For example, the average within source correlation reported in a meta-analysis by Conway and Huffcutt (.40) was greater than the average relationship of ratings across sources (.22).

The second major line of psychometric research on MSPRs has used confirmatory factor analysis of multitrait-multirater matrices in a quasi-multitrait-multimethod framework to assess the latent factor structure of performance ratings (e.g., Lance et al., 1992). The original multitrait-multimethod framework was designed to facilitate inferences regarding

the construct validity of measures by examining the degree to which the same trait measured by different methods was related (convergent validity) and different traits were distinct from one another (discriminant validity; Campbell & Fiske, 1959). In essence, multitrait-multirater investigations of MSPRs have been assumed to represent a special case of the more general multitrait-multimethod methodology, where the performance dimensions represent the focal traits and the sources providing the rating are assumed to represent variations in the measurement method (hence multitrait-multirater; MTMR).

Lawler's (1967) investigation of the relative impact of performance dimension and rating source on performance ratings represented one of the first attempts to study MSPRs using this quasi-multitrait-multimethod methodology and found modest convergent validity evidence among supervisor, peer, and self-ratings and strong rater source effects. Since Lawler's initial evaluation, substantial research has investigated the relative impact of source and dimension factors on ratings provided by raters from different sources using a variety of samples, rating sources, rating instruments, and performance dimension structures. Results consistently indicate that performance ratings made by raters from different sources are characterized by both source and dimension effects and that source effects are substantial relative to dimension effects (Campbell, McHenry, & Wise, 1990; Coover, Craiger, & Teachout, 1997; Holzbach, 1978; King et al., 1980; Klimoski & London, 1974; Lance et al., 1992; Lawler, 1967; Vance, MacCallum, Coover, & Hedge, 1988; Woehr et al., 2005; Zedeck & Baker, 1972). This line of research is consistent with the theoretical underpinnings of MSPRs, which suggest that raters from different sources observe and emphasize different aspects of ratee behavior (Borman, 1974, 1997).

Despite the consistent findings that support dimension and (large) source factor effects in MSPR research, recent research has questioned the presence and importance of rater source effects. First, Mount et al. (1998) suggested that this stream of research suffers from an important limitation, namely that this research has typically used single raters from each source or has aggregated ratings within sources prior to examining the structure of MSPRs. As a result, Mount et al. suggested that the often-found source effects actually represent variance attributable to the individual rater as opposed to source effects per se. As an empirical test of this possibility, Mount et al. used data from a sample of 2,350 managers evaluated with the Management Skills Profile (Sevy, Olson, McGuire, Frazier, & Paajanen, 1985) to compare a number of different models of the structure of MSPRs. Importantly, this study differed from previous research in that Mount et al. included multiple raters from each source (self-ratings accompanied by ratings from two peers, two supervisors, and

two subordinates) as opposed to aggregating ratings within sources prior to examining the structure. Results indicated that a model that specified seven idiosyncratic rater factors plus three performance dimension factors fit the data substantially better than did a model that specified three performance dimension factors plus four source factors alone, supporting their contention that commonly found rater source factors should more properly be interpreted as representing idiosyncratic rater factors. Scullen et al. (2000) substantially replicated these findings in two additional large samples using a correlated uniqueness parameterization. Others have also questioned the presence of source effects on the basis of meta-analytic results (Viswesvaran et al., 2002) and by using alternate approaches to assess interrater agreement (LeBreton et al., 2003).

The implications of this stream of research are clear: (a) previous MSPR research has confounded idiosyncratic rater and source effects, (b) evidence does not support widely held assumptions that rater source effects are important components of MSPRs (rather, these effects have been misattributed from what are actually idiosyncratic rater effects), and (c) researchers and practitioners should avoid the standard practice of aggregating ratings taken from the same source prior to conducting research on MSPRs and interpreting/presenting MSPR-based feedback (Mount et al., 1998; Viswesvaran et al., 2002, 2005). Clearly, these conclusions are at odds with the majority of MSPR research and especially practice. For example, from a practical standpoint, imagine presenting a feedback recipient with seven or eight different ratings taken from raters in the same source (e.g., multiple subordinates or peers) on multiple different performance dimensions, resulting in upwards of 20 or more distinct sets of ratings. A practice resulting in the presentation of this many distinct sets of ratings would be quite confusing for the feedback recipient. In addition, from a research standpoint, these findings call into question previous research that has examined MSPRs by looking only at differences due to source. Finally, the impact of Mount et al.'s conclusions on the performance rating field is evidenced by a recent large-scale meta-analysis of the structure of performance that, partly on the basis of Mount et al.'s conclusions, *assumed* the absence of source factors when developing and empirically testing a model of the structure of performance ratings (Viswesvaran et al., 2005).

By providing an initial explication of the distinction between idiosyncratic rater and source effects, Mount et al.'s (1998) and Scullen et al.'s (2000) studies provided important contributions to the literature. Nevertheless, we also believe that the case against the existence of systematic source effects has been overstated and that there are reasons to believe that systematic rater source effects exist in MSPRs at a broader conceptual level. A brief historical review of the

conceptualizations and operationalizations of idiosyncratic rater and rater source effects sheds light on the appropriate interpretation of these effects.

The pervasiveness of idiosyncratic rater effects has been a focus of psychologists for over a century (Wells, 1907). Alternately referred to as halo error (Thorndike, 1920), logical error (Newcomb, 1931), overall rater biases (Wherry, 1952), correlational bias (Kenny & Berman, 1980), illusory halo (Cooper, 1981), rater leniency/elevation (Saal, Downey, & Lahey, 1980), and idiosyncratic rater effects (Mount et al., 1998; Scullen et al., 2000), a variety of terms and explanations have been used to describe the variance associated with individual raters. Although the idiosyncratic rater effect has been attributed to many different factors and has been interpreted differently by prior researchers, operationally, idiosyncratic rater effects are present to the extent that all ratings from an individual rater covary with one another but not with the ratings provided by other raters. Idiosyncratic rater factors are distinguished from the previously discussed general performance factor in that the general performance factor represents variance common across all raters and is typically assumed to represent true score variance, whereas the idiosyncratic effect is a systematic effect that is common only to an individual rater and is often assumed to represent rater *bias* (Scullen et al., 2000). Previous rating research has consistently supported substantial idiosyncratic rater effects. In the context of MSPRs, Mount et al. estimated that idiosyncratic rater effects accounted for 72% of the variance in ratings, and Scullen et al. found that idiosyncratic rater effects accounted for an average of 58% of the variance in ratings across two large samples. Therefore, recent MSPR research is consistent with over a century of previous rating research findings (Cooper, 1981) that show large idiosyncratic rater effects that appear to be robust to a variety of interventions designed to reduce them (Kingstrom & Bass, 1981; Woehr & Huffcutt, 1994). Accordingly, we expected to support a MSPR structure consisting of substantial idiosyncratic rater factors.

In contrast to idiosyncratic rater effects, rater source effects reflect variance that is shared by raters from the same source. Recall that Mount et al.'s (1998) and Scullen et al.'s (2000) findings suggested that (a) idiosyncratic and source effects had been confounded in previous studies of the latent structure of MSPRs, and (b) idiosyncratic effects on ratings were far stronger than were source effects. Mount et al.'s critical test of the viability of idiosyncratic rater factors versus rater source factors was in reference to an alternative model that specified only more general source factors. In effect, these models compared one model of individual-level rater general impression effects (the idiosyncratic factors model) to an alternative model that requires that raters from the same



source share *identical* general impressions of ratee performance because the latter source-only model effectively restricts the correlation between the respective individual-level rater factors in the idiosyncratic factors model equal to 1.00. Although these general impressions may indeed be shared (in fact this is our argument for shared perspectives within source at the SOF level), it is unreasonable to expect them to overlap entirely. Also, research reviewed earlier on rater agreement suggests that ratees are rated more similarly by members of the same rater group (e.g., multiple subordinates or multiple peers) than members of different rater groups (bosses vs. subordinates). Only one study has modeled and estimated the effects of both idiosyncratic and source effects on MSPRs. Specifically, Scullen et al. (2000) supported small source effects (an average of 8% of the variance attributable to source across two samples) relative to idiosyncratic effects (an average of 58% of the variance) using a correlated uniqueness approach to model parameterization. Nevertheless, due to well-documented biases inherent to the correlated uniqueness parameterization under certain conditions (Conway, Lievens, Scullen, & Lance, 2004; Lance, Noble, & Scullen, 2002; Lance, Woehr, & Meade, 2007) there is reason to question the magnitude of source and idiosyncratic effects estimated by Scullen and his colleagues (we address the conditions that lead to bias in the correlated uniqueness model in more depth below). Given these limitations of previous studies, the main purpose of this study was to extend prior research distinguishing idiosyncratic rater from source effects (Mount et al., 1998; Scullen et al., 2000) by testing an alternative and more general model that specifies both idiosyncratic rater effects *and* systematic rater source factors via SOF analysis.

#### *Measurement Error*

Finally, consistent with classical test theory, a portion of the variance in MSPRs is assumed to be attributable to nonsystematic measurement error. Scullen et al. (2000) estimated that measurement error accounted for an average of 14.5% of the variance in MSPRs.

#### *Models Tested*

Based on our review of the MSPR literature, we tested six factor structures of MSPRs that model different ways in which raters and performance dimensions might contribute to the latent structure of MSPRs. Consistent with Mount et al.'s (1998) and Scullen et al.'s (2000) studies, we included seven raters for each ratee, including self-ratings, two peers, two supervisors, and two subordinates. The primary difference between each of the models we tested involves alternative specifications of the

rater factors (specification of the performance dimension factors will be discussed later in the method section).

The first model proposed that only performance dimension factors characterize MSPR data. This is similar to the trait-only model in multitrait-multimethod nomenclature and suggests that performance dimensions (and not sources/raters) characterize covariances among MSPRs. To our knowledge such a “trait-only” model has never provided an acceptable fit to MSPR data—we fit it here for comparison purposes only. The second model hypothesized that seven (idiosyncratic) rater factors characterized the data (one self and two supervisor, two peer, and two subordinate factors). Support for this model would suggest that only rater factors and not dimension factors characterize covariances among MSPRs. Based on other findings that have supported the presence of trait *and* method factors, we did not expect that this model would fit the data well either (e.g., Lance et al., 1992; Woehr et al., 2005).

The third model, a four source-three dimension model, specified that covariances among MSPRs are attributable to both rater source *and* performance dimension factors. This model is consistent with the model that is often supported in the MSPR literature (e.g., Woehr et al., 2005), with one important difference. As previously mentioned, in past MSPR research a single rater (or aggregate of raters) from each source has been used as input into subsequent analyses. In this study, and consistent with Mount et al. (1998) and Scullen et al. (2000), multiple raters from each source provided ratings, and thus the ratings obtained from different raters within the same source were loaded onto the same rater source factor, allowing us to unconfound idiosyncratic rater and source variance. As we mentioned earlier, the assumption reflected by this model is that the general impressions that are shared by raters within the same source are identical to one another because this source-only model effectively restricts the correlations between the respective individual-level rater factors equal to 1.00.

Fourth, a 10-factor model comprising three performance dimensions and seven idiosyncratic rater factors, one for each individual rater, was specified. This is the same 10-factor model on which Mount et al. (1998) based their conclusions that performance dimension and idiosyncratic rater effects, and not source effects, characterize MSPRs.

Next, we tested a fifth model, a 13-factor model that parameterized three rater source SOFs in addition to idiosyncratic first-order rater factors from the 10-factor model. In this model, the two idiosyncratic rater first-order factors (FOFs) corresponding to raters from the same source were parameterized as loading on a SOF representing the systematic variance shared by raters from the same source, resulting in three source-based SOFs corresponding to supervisor, peer, and subordinate raters.

Note that the assumption reflected by this model is that raters within the same source may share general impressions of the ratee but that these general impressions are not so similar as to be identical to one another.

Finally, in recognition of Scullen et al.'s (2000) important contributions to understanding the latent structure of MSPRs and the prominent role that a general performance factor has been accorded historically, we also tested a 14-factor model that included a general first-order performance factor in addition to the factors specified by 13-factor model. Consistent with prior research, the general factor is constrained to be orthogonal to the other latent factors in the model (King et al., 1980; Scullen et al. 2000). This model is presented in Figure 1. This model is consistent with Scullen et al.'s (2000) in proposing the same performance true score structure (general and dimensional performance) and the same rater factors (idiosyncratic and source). However, it differs importantly from Scullen et al. in that it parameterizes (a) rater effects explicitly as FOFs rather than as ad hoc correlations among uniquenesses as in a correlated uniqueness parameterization of rater effects, and (b) SOF source effects as representing shared, source-contingent perspectives. Based on prior research, we expected the 14-factor model to provide the best fit to the data.

### *Method*

#### *Participants*

Data reported here are from two samples: (a) MSPR data for 2,350 managers who were rated on the Management Skills Profile (Sevy et al., 1985) as reported in the correlation matrix presented by Mount et al. (1998) and (b) a sample of 22,420 managers who were rated on the Center for Creative Leadership's BENCHMARKS<sup>R 1</sup> MSPR instrument. Self-ratings and ratings provided by two supervisors, peers, and subordinates each were available from both samples (resulting in 16,450 and 156,940 total respondents for the Mount et al. study and BENCHMARKS<sup>R</sup> data, respectively). Specific sample and instrumentation information from the first sample completing the Management Skills Profile can be found in the Mount et al. study.

For the BENCHMARKS<sup>R</sup> sample, the target manager represented a wide variety of organizations, industries, and hierarchical levels. Each manager participated voluntarily for the purpose of professional development. The sample consisted primarily of White (76%) male (64%) college

---

<sup>1</sup>BENCHMARKS<sup>R</sup> is a registered trademark of the Center for Creative Leadership.

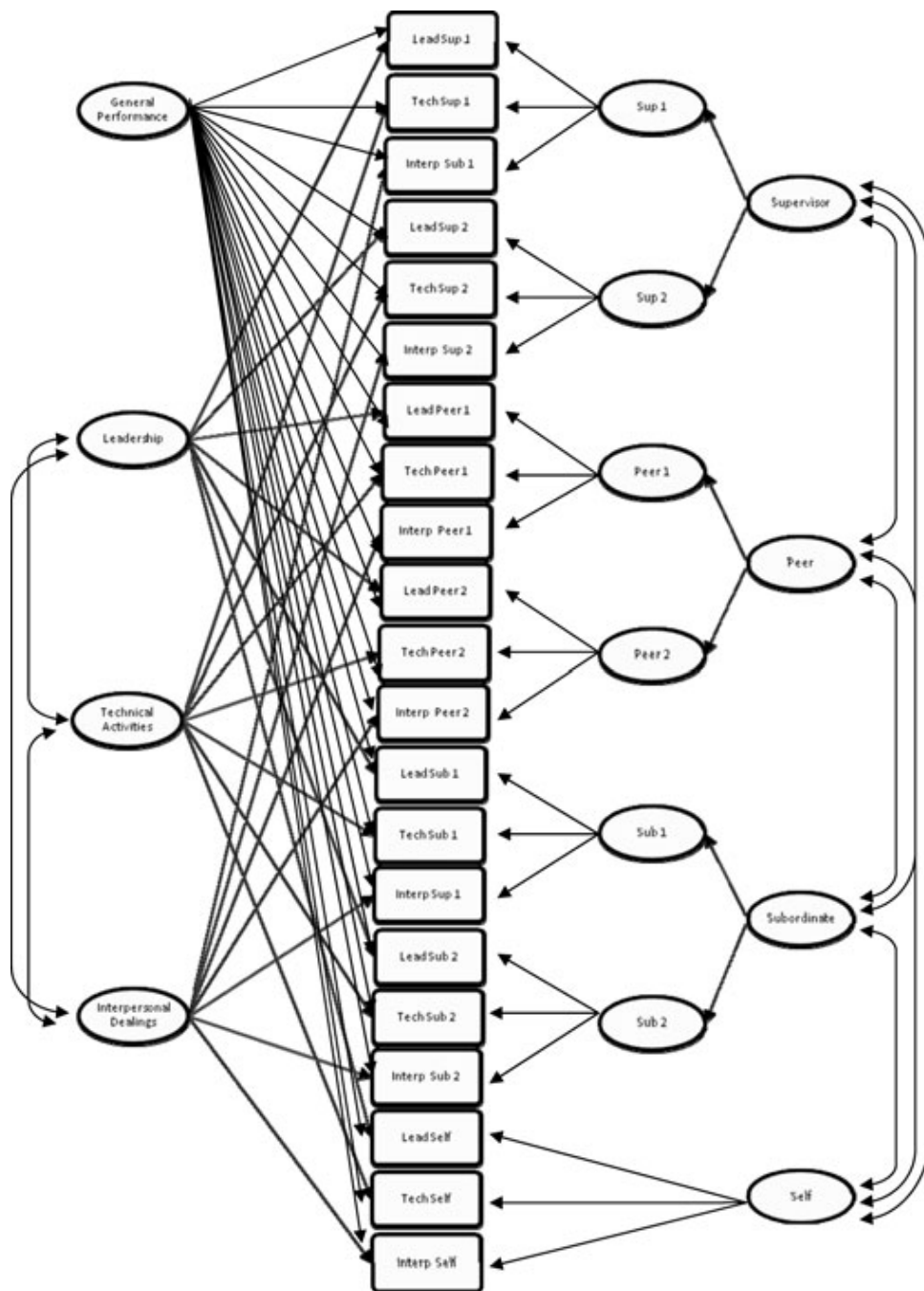


Figure 1: MSPR Second-Order Factor Model.

graduates (88%). The mean ratee age was 42 years. BENCHMARKS<sup>R</sup> is a well validated and reviewed multisource instrument used for leadership development purposes (Carty, 2003; Lombardo & McCauley, 1994; Lombardo, McCauley, McDonald-Mann, & Leslie, 1999; McCauley & Lombardo, 1990; McCauley, Lombardo, & Usher, 1989; Spangler, 2003;

Zedeck, 1995). BENCHMARKS<sup>R</sup> includes 115 items designed to assess 16 distinct dimensions of managerial behavior. Participants are rated on a scale of 1 = *not at all* to 5 = *to a very large extent*.

### *Dimensional Structure*

The dimensionality and structure of managerial performance has been the subject of a variety of reviews and taxonomies. Mount et al. (1998) used a three-dimensional taxonomy of managerial performance consisting of administrative, human relations, and technical skills/motivation suggested by Mann (1965). Although a useful framework, we chose to use an alternative framework in this study for a number of reasons. First, we were unable to reliably classify individual BENCHMARKS<sup>R</sup> dimensions into the administrative and technical skills factors. Other researchers have also had difficulty classifying performance constructs as representing either administrative or technical skills and have questioned the usefulness of this distinction (e.g., Conway, 1999). Also, previous research has conceptualized the BENCHMARKS<sup>R</sup> instrument around three broad dimensions of managerial performance that include leading people, meeting job challenges, and respecting self and others (e.g., Fleenor, McCauley, & Brutus, 1996). This taxonomy maps very closely onto a three-dimension taxonomy of managerial performance developed by Borman and Brush (1993), including leadership and supervision, technical activities and mechanics of management, and interpersonal dealings and communication. Based on the design of the BENCHMARKS<sup>R</sup> instrument and the accumulated empirical support for Borman and Brush's taxonomy (e.g., Conway, 1999), we used (a) leadership and supervision (b) meeting technical activities/mechanics of management, and (c) interpersonal dealings and communication as the three general performance dimensions that we modeled here.

To ensure that each of the dimensions proposed by Borman and Brush (1993) fit with the proposed taxonomy, the first and third authors classified each of the 16 dimensions assessed by BENCHMARKS<sup>R</sup> into one of the aforementioned three categories of managerial performance. Of the 16 dimensions, 13 were classified into one of the broad performance dimensions by both of the raters (100% agreement). The classification of these dimensions was as follows: (a) *leadership and supervision*: leading employees, confronting problem employees, participative management, and change management. (b) *technical activities/mechanics of management*: resourcefulness, being a quick study, and decisiveness; and (c) *interpersonal dealings and communication*: compassion and sensitivity, building and mending relationships, straightforwardness and composure, self-awareness, putting people at ease, and differences matter.

The remaining three dimensions could not be reliably classified (balance between personal life and work, doing whatever it takes, and career management) and were consequently dropped from analyses. To ensure that the BENCHMARKS<sup>R</sup> instrument conformed to this three-dimension structure, we used each of the 13 scales as a manifest indicator in a set of CFAs. On the basis of our a priori classification, we specified a 12-factor model that included self-ratings and ratings for one randomly selected supervisor, peer, and subordinate for each of the three broad dimensions. This 12-factor model fit the data well [ $\chi^2$  (1208) = 116,187.279, root mean squared error of approximation (RMSEA) = .076, Tucker–Lewis index (TLI) = .954, comparative fit index (CFI) = .958] and provided a better fit to the data than did a model that specified only a single general factor for each rating source ( $\chi^2$  (60) = 43,111.55,  $p < .001$ ; CFI = .016), supporting our classification of the 13 dimensions onto each of their respective three broad factors. Accordingly, based on our a priori classification, the design and typical uses of the instrument, and the results of the CFAs, individual ratings on BENCHMARKS<sup>R</sup> scales were aggregated up to the level of the three broader dimensions for subsequent analyses.

#### *Confirmatory Factor Analysis*

We tested the six models discussed earlier using confirmatory factor analysis using LISREL 8.7 (Jöreskog & Sörbom, 2004). Consistent with Hu and Bentler's (1998, 1999) recommendations, we used the  $\chi^2$  test, the standardized root mean squared residual (SRMSR), Steiger's (1990) RMSEA, the TLI (Tucker & Lewis, 1973), and Bentler's (1990) CFI to evaluate model fit. SRMSR is a summary index of the percentage of variance unaccounted for by the fitted model, whereas RMSEA represents a measure of lack of fit per degree of freedom (Browne & Cudek, 1993). TLI and CFI are relative fit indices that (a) evaluate model fit relative to a null model and (b) take into account the overall number of model parameters estimated. The criteria that Hu and Bentler's (1998, 1999) proposed for good model fit that we used here were SRMSR  $\leq$  .08, RMSEA  $\leq$  .06, CFI and TLI  $\geq$  .95. We also use three indexes to compare relative fit between nested models: the difference  $\chi^2$  ( $\chi^2$ ) test, CFI (Cheung & Rensvold, 2002), and a relative fit index (RFI) described by Lakey, Goodie, Lance, Stinchfield, and Winters (2007) and Lance et al. (1992). The  $\chi^2$  test provides a statistical test of whether some less restricted model ( $M_U$ ) provides a closer fit to the data than some more restricted model ( $M_R$ ) that is a special case of, or nested within,  $M_U$ . Nevertheless, with large sample sizes the  $\chi^2$  test is very powerful so that most models are rejected statistically (Bentler & Bonett, 1980). For this reason we used the CFI and RFI as supplementary relative fit indexes.

Based on their extensive Monte Carlo simulations, Cheung and Rensvold (2002) recommended a cutoff of  $\text{CFI} = .01$  as indicating a significant difference in fit between some  $M_R$  and an alternative  $M_U$ . Also, the RFI:

$$RFI = 1 - \frac{\chi^2_{M_R} - \chi^2_{M_U}}{\chi^2_{\text{Null}} - \chi^2_{M_U}} \quad (1)$$

indexes the goodness-of-fit of  $M_R$  relative to  $M_U$  as compared to the null or independence model. RFI ranges between 0 and 1.00 with values closer to 1.00 indicating increasingly good model fit.

### *Results*

Correlations among the three BENCHMARKS<sup>R</sup> performance dimensions used in this study as rated by the seven different raters are shown in Table 1 (see Table 1 of the Mount et al. [1998] study for their correlation matrix). For the BENCHMARKS<sup>R</sup> data, the mean different dimension-different rater correlation was .15, the mean same dimension-different rater correlation was .21, and the mean different dimension-same rater correlation was .81. As is typical in prior MSPR research, these findings indicate relatively weak convergent validity and strong method effects in the traditional multitrait-multimethod sense. Note that the mean same dimension-different rater correlation is almost exactly the same as the correlation between different source ratings (.22) reported in the meta-analysis by Conway and Huffcutt (1997). Also, these findings are consistent with those reported by Mount and his colleagues (mean different dimension-different rater  $r = .18$ ; mean same dimension-different rater  $r = .28$ ; mean different dimension-same rater  $r = .75$ ).

Correlation matrices were input into LISREL 8.7 for both samples to examine the structure of MSPRs. For the first two models (trait and rater only models), we estimated the correlations among all factors. For those models specifying both dimension and rater or source factors (the remaining models), we followed the correlated trait-correlated method approach to model estimation in which correlations between rater/source and dimension factors were fixed equal to zero, whereas rater/source and dimension factors are allowed to be correlated among themselves (Lance et al., 2002; Wildaman, 1985).

As is shown in the top panel of Table 2, we reproduced Mount et al.'s (1998) finding that the 10 idiosyncratic rater-plus-dimension factor model (Model 4) provided a better fit to the Management Skills Profile data as compared to the first three models. Results in the bottom panel of Table 2 show that we replicated these findings using the BENCHMARKS<sup>R</sup> as well. In both data sets, both the overall model goodness-of-fit indices

TABLE 1  
Correlations Among BENCHMARKS<sup>R</sup> Ratings

Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Boss #1—Technical	1.0																				
2. Boss #1—Interpersonal	.74	1.0																			
3. Boss #1—Leadership	.79	.87	1.0																		
4. Boss #2—Technical	.36	.23	.26	1.0																	
5. Boss #2—Interpersonal	.23	.32	.26	.75	1.0																
6. Boss #2—Leadership	.26	.26	.28	.79	.87	1.0															
7. Peer #1—Technical	.26	.20	.21	.26	.19	.21	1.0														
8. Peer #1—Interpersonal	.16	.26	.20	.16	.26	.21	.78	1.0													
9. Peer #1—Leadership	.18	.21	.22	.21	.21	.22	.82	.89	1.0												
10. Peer #2—Technical	.26	.19	.21	.20	.20	.21	.25	.18	.19	1.0											
11. Peer #2—Interpersonal	.14	.25	.20	.26	.26	.21	.18	.27	.21	.78	1.0										
12. Peer #2—Leadership	.17	.20	.21	.21	.21	.22	.19	.21	.22	.82	.89	1.0									
13. Subordinate #1—Technical	.19	.13	.15	.13	.13	.16	.18	.12	.14	.19	.13	.16	1.0								
14. Subordinate #1—Interpersonal	.12	.19	.16	.20	.20	.16	.14	.20	.17	.14	.21	.17	.82	1.0							
15. Subordinate #1—Leadership	.14	.16	.17	.16	.16	.17	.15	.16	.17	.15	.16	.18	.86	.91	1.0						
16. Subordinate #2—Technical	.20	.14	.16	.14	.14	.16	.19	.14	.16	.19	.13	.16	.26	.21	.23	1.0					
17. Subordinate #2—Interpersonal	.13	.22	.18	.22	.22	.18	.15	.22	.19	.14	.22	.18	.26	.36	.31	.82	1.0				
18. Subordinate #2—Leadership	.15	.16	.18	.16	.16	.17	.16	.17	.18	.15	.16	.18	.23	.24	.26	.86	.90	1.0			
19. Self-Technical	.18	.03	.09	.04	.04	.10	.15	.02	.07	.14	.02	.07	.16	.06	.10	.16	.06	.10	1.0		
20. Self-Interpersonal	.08	.17	.12	.17	.17	.12	.09	.16	.12	.08	.16	.11	.09	.17	.13	.10	.18	.13	.68	1.0	
21. Self-Leadership	.11	.10	.15	.10	.10	.15	.10	.09	.13	.09	.09	.13	.13	.13	.17	.13	.13	.16	.76	.80	1.0

Note. Leadership = Leadership and supervision, Technical = Technical activities/mechanics of management, Interpersonal = interpersonal dealings and communication. For  $r \geq .02$ ,  $p < .001$ .



TABLE 2  
Overall Model Goodness of Fit

Mount et al. (1998) sample	$\chi^2$	df	SRMSR	RMSEA	TLI	CFI	$\chi^2$	df	CFI	RFI
1. 3-factor Dimensions only model	27,618.093*	186	.163	.237	.276	.358				
1 vs. 4							27,287.885*	42	.638	.277
2. 7-factor Seven individual rater factors and no trait factors	6,930.811*	168	.052	.171	.802	.842				
2 vs. 4							6,600.603*	24	.154	.807
3. 7-factor Four rater source factors—one for each level: boss, peer, subordinate, and self; and three dimension factors	13,642.423*	159	.140	.182	.515	.633				
3 vs. 4							13,312.215*	15	.363	.587
4. 10-factor Seven individual rater factors and three dimension factors	330.208*	144	.020	.024	.994	.996				
4 vs. 5							8.155	9	.001	.999+
5. 13-factor Same as 10 factor but with 3 source SOFs	338.363*	153	.021	.023	.993	.995				
5 vs. 6							120.444*	21	.004	.997
6. 14-factor Same as 13 factor but with a general factor	218.186*	132	.013	.017	.997	.998				

TABLE 2 (continued)

BENCHMARKS <sup>R</sup>	$\chi^2$	df	SRMSR	RMSEA	TLI	CFI	$\chi^2$	df	CFI	RFI
1. 3-factor Dimensions only model	381,425.006*	186	.189	.265	.054	.162				
1 vs. 4							377,832.431*	42	.830	.055
2. 7-factor Seven individual rater factors and no trait factors	51,998.973*	168	.032	.151	.858	.886				
2 vs. 4							48,406.398*	24	.106	.867
3. 7-factor Four rater source factors— one for each level: boss, peer, subordinate, and self; and three dimension factors	152,351.049*	159	.147	.147	.579	.681				
3 vs. 4							148,758.474*	15	.311	.565
4. 10-factor Seven individual rater factors and three dimension factors	3,592.575*	144	.020	.033	.989	.992				
4 vs. 5							3.856	9	.001	.999*
5. 13-factor Same as 10 factor but with three source SOFs	3,596.431*	153	.019	.032	.990	.993				
5 vs. 6							1,755.508	21	.003	.996
6. 14-factor Same as 13 factor but with a general performance factor	1,840.923*	132	.011	.024	.994	.996				

(i.e.,  $\chi^2$ , CFI, TLI, SRMSR, and RMSEA) and the incremental fit indices (i.e.,  $\chi^2$ , CFI and RFI) showed that Model 4 provided a substantially better fit than did Models 1 through 3. These results support Mount et al.'s and others' (e.g., Scullen et al., 2000; Viswesvaran et al., 2002) contentions regarding the importance of idiosyncratic rater factors above and beyond more general source effects in MSPRs.

We also tested a fifth 13-factor model that parameterized source SOFs in addition to idiosyncratic FOFs. In this model, the two idiosyncratic rater factors from each rating source from the 10-factor model were parameterized as loading on a SOF representing the systematic variance shared by raters from the same source, and it too provided a very good overall fit to the data in both samples (see Table 2). In fact, in neither sample did the 10-factor model (Model 4) provide a better fit to the data than did the more parsimonious 13-factor model (Model 5), even according to the  $\chi^2$  test that, given the sample sizes reported here, has extraordinary power to detect differences in nested models' fit to the data. When evaluated according to the CFI and RFI, two indexes that are unaffected by sample size, the fit of Model 5 for the Management Skills Profile (CFI = .001, RFI = .999+) and the BENCHMARKS<sup>R</sup> ratings (CFI = .001, RFI = .999+)<sup>2</sup> was essentially identical to that of Model 4, also supporting the idea that the interrelationships among rater factors within source can be accounted for parsimoniously by more general second order source factors. Although it may not be obvious, the 13-factor model (Model 5) is indeed more parsimonious than the 10-factor model (Model 4) because the 13-factor model structures the 21 correlations among the seven idiosyncratic rater FOFs in terms of a smaller number of six SOF loadings plus six correlations among the four source factors (generating the difference  $21 - 12 = 9$  df reported in Table 2). The presence of source factors is evident in the correlations among first order rater factors. Specifically, for both the BENCHMARKS and the MSP, the mean same source correlation (.27 and .33, respectively) among rater factors was greater than the mean different source correlation among rater factors (.16 and .19, respectively).

Finally, we tested a 14-factor model that added a first-order general performance factor to Model 5. As Table 2 shows, the addition of the general performance factor resulted in significantly improved model fit in both the Management Skills Profile and BENCHMARKS<sup>R</sup> samples ( $\chi^2(21) = 120.44$ ,  $p < .01$  and  $\chi^2(21) = 1755.508$ , respectively), but CFI (.004 and .003, respectively) and RFI values (.997 and .996, respectively) indicated that the fit of Model 5 (the 13-factor model) was essentially equivalent to that of Model 6 (the 14-factor model) from a practical standpoint.

<sup>2</sup>In both data sets the RFI rounded to 1.00 at the fifth decimal place.

TABLE 3  
*First-Order Factor Loadings on Second-Order Factors (SOFs)*

	<i>MSP SOFs</i>			<i>BENCHMARKS<sup>R</sup> SOFs</i>		
	Boss	Peer	Subordinate	Boss	Peer	Subordinate
Boss 1	.633*	—	—	.515*	—	—
Boss 2	.649*	—	—	.527*	—	—
Peer 1	—	.546*	—	—	.456*	—
Peer 2	—	.570*	—	—	.447*	—
Subordinate 1	—	—	.533*	—	—	.438*
Subordinate 2	—	—	.537*	—	—	.561*

*Note.* MSP = Management Skills Profile; \* $p < .01$ .

TABLE 4  
*Performance Dimension and Rater Source Factor Correlations*

<i>MSP Dimensions:</i>	HumRel	Tech1	Admin	<i>Sources</i>	Boss	Peer	Subordinate	Self
HumRel	1.000			Boss	1.000			
Tech1	-.174*	1.000		Peer	.884*	1.000		
Admin	.353*	-.327*	1.000	Subordinate	.591*	.751*	1.000	
				Self	.220*	.263*	.202*	1.000
<i>BENCHMARKS<sup>R</sup> :</i>	<i>Lead</i>	<i>Tech2</i>	<i>Interp</i>	<i>Sources:</i>	Boss	Peer	Subordinate	Self
Lead	1.000			Boss	1.000			
Tech2	.447*	1.000		Peer	.844*	1.000		
Interp	.456*	-.065	1.000	Subordinate	.521*	.623*	1.000	
				Self	.136*	.140*	.194*	1.000

*Note.* MSP = Management Skills Profile. HumRel = human relations, Tech1 = technical skills/motivation, Admin = administrative, Lead = leadership and supervision, Tech2 = meeting technical activities/mechanics of management, Interp = interpersonal dealings and communication. \* $p < .01$ .

Therefore, although the addition of the general performance factor is theoretically important and did improve overall model fit, the improvement was very small in any practical sense.

SOF loadings from Model 6 are shown in Table 3 and these are all large, reasonably homogeneous, and very similar across samples (mean loadings = .58 and .49 for the Management Skills Profile and the BENCHMARKS<sup>R</sup> ratings, respectively). The large SOF loadings provide insight as to why Model 5 provided a complete and parsimonious account of the interrelationships among the FOFs estimated in Model 4.

Table 4 shows correlations among the dimension and source factors from Model 6. The low to moderate correlations among the dimension factors support their discriminability and the usefulness of the broad

taxonomies that we used to classify the individual dimension ratings. Consistent with previous literature (Mabe & West, 1982; Thornton, 1980), correlations between the self source factor and the other three source factors were the lowest among the source factor intercorrelations (mean  $r = .19$ ). For the remaining source factors, the supervisor and peer factors were the most strongly related of any of the source factors (mean  $r = .86$ ), followed by the peer-subordinate correlation (mean  $r = .69$ ), and the supervisor-subordinate source factors exhibiting the weakest of the across source correlations (mean  $r = .55$ ).

Finally, we decomposed rating variance according to the various factors estimated by Model 6. First, we calculated the percent of variance attributable to the general performance and dimension factors as the mean squared standardized dimension factor loading.<sup>3</sup> Next, we decomposed variance attributable to raters into two different components: idiosyncratic rater variance and variance attributable to rater source effects. To estimate the variance attributable to idiosyncratic raters, we first calculated the total percentage of variance attributable to the idiosyncratic rater factor as each rating's squared standardized FOF rater factor loading. Idiosyncratic rater variance was subsequently calculated as the product of each rating's squared standardized FOF rater factor loading times the respective idiosyncratic FOF's uniqueness because the FOF uniqueness represents that portion of the individual rater FOF variance that is *not* accounted for by the respective source SOF. Second, we calculated source variance as each rating's squared standardized FOF rater factor loading minus idiosyncratic variance. We took estimates in LISREL's  $\epsilon$  matrix as each rating's uniqueness (error and specific variance). Results are shown in Table 5.

Note first that the general performance (g) factor accounted for only 3.5% of the variance in MSPRs on the average. These estimates are far lower than those that have been previously reported (e.g., Scullen et al., 2000; Viswesvaran et al., 2005; we will have more to say about this later) and explain why Model 6 (which included the g factor) failed to provide a practical improvement in model fit as compared to Model 5 (which did not include the g factor). Consistent with prior research investigating the structure of performance ratings (e.g., Lance et al., 1992; Mount et al., 1998; Scullen et al., 2000; Woehr et al., 2005), performance dimension factors accounted for a modest amount of variance (an average of 7% of the variance), relative to other rating components. Next, idiosyncratic rater effects accounted for the largest proportions of variance in both the Management Skills Profile and the BENCHMARKS<sup>R</sup> ratings (an average

---

<sup>3</sup>In LISREL nomenclature, the mean squared "completely standardized" factor loading.

TABLE 5  
*Mean Percentages of Variance Accounted for in Ratings*

	g	Dimensions	Idiosyncratic	Source	Uniqueness
MSP ratings:					
Boss	.06	.06	.43	.30	.16
Peer	.03	.05	.53	.20	.19
Subordinate	.02	.05	.52	.24	.16
Self	.04	.08	.69	—	.19
Grand mean	.04	.06	.49	.25	.17
BENCHMARKS <sup>R</sup> ratings:					
Boss	.01	.09	.56	.21	.13
Peer	.02	.06	.64	.17	.12
Subordinate	.09	.04	.58	.20	.09
Self	.01	.14	.70	—	.15
Grand mean	.03	.08	.62	.19	.12

*Note.* MSP = Management Skills Profile; g = the general performance factor.

of 55% of the variance), and this too is consistent with prior research investigating the role of idiosyncratic rater factors in MSPRs (Mount et al., 1998; Scullen et al., 2000). Nevertheless, our results differ from these earlier studies' findings with respect to the relative importance of source effects. Mount et al. did not consider the possibility that *both* idiosyncratic rater *and* rater source factors could affect ratings simultaneously and concluded that *all* variance in ratings accounted for by rater effects was idiosyncratic. Scullen et al. did model idiosyncratic and source effects simultaneously and concluded that source effects accounted for about 8% of the variance in ratings. Our findings suggest that the impact of rater source effects is nearly three times as large (an average of 22% of the variance due to source effects in this study) as the value presented by Scullen et al., supporting the practical and theoretical relevance of rater source effects in MSPRs. Finally, uniqueness accounted for an average of 14.5% of the variance across the two samples.

### *Discussion*

The primary purpose of this study was to reexamine the latent structure of MPSRs, with a specific focus on the role that rater source plays in performance ratings. Across two large, independent samples of MPSRs, a set of confirmatory factor analyses supported a MSPR structure consisting of four systematic sources of variance, including performance dimension, general performance, idiosyncratic rater, and rater source effects. Our results differ from those of prior research in two important respects: the magnitude of rater source and general performance factors. Specifically,

a SOF model in which individual raters from the same source loaded on a second order source factor provided almost identical fit as a model that proposed only idiosyncratic rater effects in both samples. This SOF model is more parsimonious than the idiosyncratic factors only model and is consistent with previous theory on the presence of source effects. Our findings further suggest that source effects account for a substantially larger proportion of variance and that general performance accounts for a much smaller portion of variance than has been reported previously.

### *Components of MSPRs*

It is important to note that although we provided evidence for the presence of source effects, idiosyncratic rater effects still explained more variance in MSPRs than did any of the remaining modeled effects. Idiosyncratic rater effects accounted for an average of 55% of the variance in MSPRs in this study, 58% of the variance in Scullen et al. (2000), and 71% of the variance in Mount et al. (1998). Despite the apparent importance of idiosyncratic rater effects in MSPRs, relatively little research has systematically examined the meaning of these effects. Traditional psychometric theories of performance ratings (e.g., Thorndike, 1920) and recent MSPR research (Scullen et al., 2000) consider them as representing *biases* that should be minimized by well-developed performance appraisal systems. Others have suggested that there may be both valid and invalid aspects of raters' general impressions (Lance & Woehr, 1986; Lance, Woehr, & Fisicaro, 1991) and that different raters' general impressions represent (perhaps) equally valid but only moderately overlapping perspectives on ratee performance (e.g., Hoffman & Woehr, 2009; Borman, 1997; Lance, Baxter, & Mahan, 2006; Lance, Hoffman, Gentry, & Baranik, 2008). Additional research is needed that tests these competing hypotheses.

On the other hand, performance dimension effects accounted for less than 10% of the variance in MSPRs in this analysis as well as in Mount et al.'s (1998) and Scullen et al.'s (2000) previous studies. This estimate of the relative magnitude of performance dimension effects is consistent with other MSPR research (Campbell et al., 1990; Coover et al., 1997; Holzbach, 1978; King et al., 1980) and research in similar domains that use multitrait-multimethod-related approaches (e.g., assessment centers; Bowler & Woehr, 2006). Previously, the relatively "small" impact of rater source compared to idiosyncratic raters in MSPRs has led some to question the usefulness of collecting performance data from multiple sources and the corresponding practice of separating performance feedback on the basis of the source providing the feedback. Nevertheless, our results suggest that the impact of rater source is roughly three times larger than the impact of performance dimensions. Paradoxically, the

results presented here cast more serious questions on the use of performance *dimensions* in MSPRs than the separation of those performance dimensions based on the source providing the ratings. As has been suggested previously, further examination of the validity of MSPR *dimensions* is sorely needed (Arthur & Villado, 2008; Borman, 1997; Hoffman & Woehr, 2009; Scullen et al., 2000).

One of the primary distinctions between the findings of this study and that of past research is the magnitude of the source effects. Our findings indicated that source effects accounted for nearly three times more variance in MSPRs compared to the estimates provided by Scullen et al. (2000). Why the large discrepancy in results? The main difference between our two studies was that Scullen et al. used a correlated uniqueness parameterization of rater and source effects whereas we used a hierarchical confirmatory factor analysis adaptation of the general correlated trait-correlated method model. It is now well known that the correlated uniqueness model yields upwardly biased estimates of trait factor loadings and intercorrelations in the analysis of multitrait-multimethod data when method effects are strong and reasonably highly intercorrelated in the population, and that the correlated trait-correlated method model avoids these biases (Conway et al., 2004; Lance et al., 2002, 2007). In this case, rater source effects were indeed strong (see Table 3) and reasonably highly correlated (see Table 4), which are precisely the conditions that have been shown analytically (Lance et al., 2002) and empirically (Conway et al., 2004; Lance et al., 2007) to result in upwardly biased trait factor loadings and correlations in correlated uniqueness parameterizations. This raises the real possibility that Scullen et al.'s results reflected a misattribution of stronger dimension variance components and weaker rater source variance components than was actually warranted as a result of bias that is inherent to the correlated uniqueness model.

The second primary distinction between this study's findings and those of past research is the magnitude of the general performance factor's effect on ratings. In particular, the impact of general performance was quite small in this study (an average of approximately 4% of the variance across the two samples) relative to the recent work of Scullen et al. (an average of approximately 14% of the variance) and especially compared to the results of Viswesvaran et al.'s meta-analysis (approximately 27% of the variance). Given that Scullen et al. (2000) used a parameterization (the correlated uniqueness model) known to result in the misattribution of source-specific variance to trait factors (general performance and performance dimensions), it should not be at all surprising that the magnitude of the general performance factor in Scullen et al. was substantially larger than the portion of variance attributable to general performance accounted for in our results.



We believe that the discrepancy between the magnitude of the general performance factor in our study and Viswesvaran et al.'s (2005) is a function of differences in the underlying performance model specified. Specifically, based on some of the research reviewed here (e.g., Mount et al., 1998), Viswesvaran et al. (2005) developed a performance model based on the assumption that rater source effects are not present in performance ratings. In contrast, our findings reveal that source variance is indeed an important component of performance ratings. It is possible that Viswesvaran et al.'s failure to model rater source effects resulted in the misattribution of source variance to the general performance factor and ultimately, to the large discrepancy in findings across our studies.

To understand the broad implications of our results, it is instructive to compare the relative proportion of variance attributable to each component in our study to the relative importance of these constructs in similar research. Overall, the proportion of variance explained by systematic, cross-rater effects (e.g., average variance attributable to general performance, performance dimensions, and sources, depending on the performance model specified) is strikingly similar across Viswesvaran et al., Scullen et al., and this study (27%, 31%, and 32%, respectively). The primary distinction in these results is the magnitude of each of the components of the systematic, cross-rater variance. Specifically, Scullen et al. found idiosyncratic effects to be the largest, followed by measurement error, general performance, dimension performance, and finally, source effects. Viswesvaran et al. did not model each of these systematic effects, but their results clearly point to the substantial importance of a general performance factor. Although our results are consistent with Scullen et al. in pointing to idiosyncratic rater effects as the largest source of variance in MSPRs, the relative importance ascribed to the remaining variance sources differs widely. For instance, source effects were the second largest variance source in this study but the smallest in Scullen et al., whereas general performance was the second largest variance component in Scullen et al., but the smallest here. Clearly, such a disparity in results leads to widely different conclusions with respect to the importance of various rating components. Together, due to documented biases associated with Scullen et al.'s (2000) correlated uniqueness parameterization and Viswesvaran et al.'s specification of an incomplete performance model, we believe that our results provide the most accurate estimate of the portion of variance accounted for by general performance, performance dimension, idiosyncratic rater, and source factors to date.

An inspection of the pattern of cross-source latent factor correlations reveals several interesting patterns. The supervisor-peer (mean  $r = .86$ ) and subordinate-peer (mean  $r = .69$ ) latent factors share more common variance than either supervisor-subordinate (mean  $r = .55$ ) or self-other

ratings (mean  $r = .19$ ). Although the magnitude of the cross-source correlations may seem large given substantial evidence pointing to typically weak cross-source convergence, it is important to note that these correlations are at the level of latent variables and, as such, are disattenuated for measurement error. Further, these results are remarkably consistent in terms of both magnitude and pattern across the two large samples here as well as with the disattenuated cross-source correlations presented in Conway and Huffcutt's (1997) meta-analysis of the psychometric properties of MSPRs (supervisor-peer  $\rho = .79$ ; peer-subordinate  $\rho = .66$ ; supervisor-subordinate  $\rho = .57$ ; self-other  $\rho = .29$ ). These results provide indirect support for the impact of rater source on performance ratings by demonstrating the impact of rater-ratee proximity in the organizational hierarchy on cross-source agreement. In particular, supervisor-peer rating pairs and peer-subordinate rating pairs differ by a single organizational level and as a result would be expected to have a more similar perspective on the target's performance compared to supervisor-subordinate pairs, which are separated by multiple hierarchical levels. Although this pattern of results is suggestive, further research is needed examining the construct validity of MSPR source effects.

In addition, this study helps to clarify the nature and importance of various factors proposed to influence performance ratings by the major psychometric models of ratings (Kenny & Berman, 1980; King et al., 1980; Wherry, 1952). Although the proportion of variance attributable to performance dimensions was relatively small, our results are consistent with the three primary models of performance ratings, each specifying the impact of empirically distinguishable dimensional performance on ratings (Kenny & Berman, 1980; King et al., 1980; Wherry, 1952). Next, two of the three primary psychometric models of ratings hypothesized a true general performance factor (Kenny & Berman, 1980; King et al., 1980). Similar to our findings with respect to dimensional performance, although a general performance factor was present in performance ratings, the magnitude of this factor was smaller than would be expected based on the prominence of this component in prior research. Next, idiosyncratic rater factors, a persistent factor in ratings for over a century (Wells, 1907) and a component of the three primary psychometric models of ratings (Kenny & Berman, 1980; King et al., 1980; Wherry, 1952) was also evident as the largest effect in MSPRs. Therefore, these results are consistent with the psychometric rating models' emphasis on idiosyncratic rater effects. Finally, past models of ratings did not explicitly recognize the impact of rater source effects on performance ratings. Nevertheless, our results point to the importance of rater source effects in MSPRs, underscoring the need to explicitly model the impact of rater source when investigating the psychometric properties of performance ratings.

*Implications*

Our results have a number of implications for research and practice. First, Mount et al.'s (1998) and Scullen et al.'s (2000) results led them to question the typical practice of aggregating raters within a given source prior to presenting MSPR feedback, and Mount et al. went so far as to suggest that feedback be presented only on the basis of individual raters. Clearly, such recommendations are troubling for practitioners who typically separate developmental feedback on the basis of the source providing the ratings and for researchers who frequently examine source-based differences in MSPRs. On the other hand, our findings reaffirm the existence and importance of source effects in MSPRs, lending support to continuing the common practice of separating ratings by source when presenting feedback.

Although this study supports systematic source effects, their etiology is still not clear. Rater source effects have been interpreted as arising from (a) differences in rater opportunity to observe ratee behavior (Borman, 1974), (b) different conceptualizations of performance dimensions by different rater groups (Woehr et al., 2005), and (c) raters from different sources having different interaction goals with target managers (Lance et al., 2006; Lance et al., 2008). As such, a critical area for future research is to determine *how* rater general impressions are formed and the variety of factors that differentially impact the content of different raters' general impressions. Toward this end, recent work suggests that source factors have different nomological networks, providing empirical evidence for the theoretical distinctness of source effects (Hoffman & Woehr, 2009).

As pointed out by an anonymous reviewer, our study also illustrates an often overlooked issue with respect to source effects. In particular, prior research has at least implicitly conceptualized source effects as indicative of variance that is *unique* to raters from a given source. In contrast, our results clearly point to a substantial overlap among (disattenuated) latent source factors (mean  $r = .70$ ). Still, on average, peer, supervisor, and subordinate source factors share approximately 49% common variance, supporting the discriminability of these factors. Together, source effects are most appropriately viewed as partly overlapping, as opposed to wholly unique, variance attributable to rater source.

Despite the support for general performance, performance dimension, and source effects in MSPRs, idiosyncratic rater effects still explained substantially more variance in MSPRs than other systematic effects. In a traditional psychometric sense, idiosyncratic rater effects are viewed as a form of bias (Schmidt, Viswesvaran, & Ones, 2000), and despite decades of research revealing no (Landy & Farr, 1980; Kingstrom & Bass, 1981) or modest effects (Woehr & Huffcutt, 1994), many suggest a

continued need “for research investigating ways to decrease idiosyncratic rater biases” (Scullen et al., 2000, p. 969). If idiosyncratic rater effects represent bias, such a large portion of variance attributable to bias is quite problematic for the use of MSPRs in research, developmental, and particularly administrative settings. Nevertheless, this may not be the case. Leader-member exchange theory suggests that leaders interact differently with different subordinates, depending on the quality of the exchange relationship (Dansereau, Graen, & Haga, 1975). Assuming that managers behave differently when interacting with coworkers occupying the same source (Sparrowe & Liden, 1997; Yukl, 2006), a lack of convergence in the perceptions of raters from the same source (e.g., two peers) should not be surprising and certainly should not be rejected out of hand as indicative of biased or otherwise “incorrect” ratings. At least a portion of the variance unique to individual raters from the same source may in fact reflect meaningful, performance-relevant variance (Lance & Woehr, 1986; Lance et al., 1991; Murphy & DeShon, 2000). Future research investigating the extent to which idiosyncratic rater variance represents rater bias or substantively meaningful, performance relevant-variance is needed.

Although our results support the aggregation of ratings obtained from raters from the same source, idiosyncratic rater effects are too large to ignore. Practitioners interested in using MSPRs for developmental purposes are faced with a quandary. Aggregating same source information will result in the loss of substantial, potentially important information provided by individual raters. On the other hand, presenting divergent feedback from each individual rater would be cumbersome and likely quite confusing for feedback recipients. One alternative is to provide an indication of the variance in ratings, or “crystallization” in same-source raters’ perceptions, in addition to the typical mean-level feedback for each source. Such a practice would provide feedback recipients important developmental information regarding the differing perceptions of same source raters while maintaining a focus on meaningful cross-source differences and a relatively high degree of parsimony.

Finally, we note that uniquenesses ( $u^2$ s, including measurement error and specific variance) accounted for an average of 14.5% of the variance in the MSPRs reported here (see Table 5). Recognizing that  $1 - u^2$  is a lower-bound estimate of a measure’s reliability,<sup>4</sup> it is apparent that the average reliability of the MSPRs reported here was  $r_{xx} = .855$ , which

<sup>4</sup>More generally  $\hat{r}_{xx} = 1 - \frac{tr(\Sigma_{xx})}{\mathbf{1}'\Sigma_{xx}\mathbf{1}}$  is a lower-bound estimator of test reliability for a  $k$ -component test, where  $tr(\cdot)$  indicates the trace,  $\Sigma_{xx}$  is the  $k \times k$  diagonal matrix of component uniquenesses,  $\mathbf{1}$  is a  $k \times 1$  unit vector and  $\Sigma_{xx}$  is the  $k \times k$  component covariance matrix (Bentler & Woodward, 1983).

is nearly identical to the average intrarater reliabilities and substantially higher than the average interrater reliabilities for supervisor (.52) and peer ratings (.42) reported by Viswesvaran et al. (1996). Interestingly, it is these latter interrater reliability estimates that have been used for attenuation correction in a number of recent meta-analyses<sup>5</sup> in lieu of other, greater lower-bound reliability estimators (Bentler & Woodward, 1983; Li, Rosenthal, & Rubin, 1996). As a result, we suspect that these meta-analyses have overstated meta-analytically estimated relationships between job performance and a number of other variables and recommend that future meta-analyses invoke more reasonable attenuation corrections based on greater lower bound estimates of reliability such as the .855 reported here.

### Conclusion

Given the popularity of MSPRs, a clear understanding of their structure is critical to the appropriate use of these tools in research, administrative, and developmental settings. This study contributes to the literature by providing a re-assessment of the structure of MSPR and the magnitude of systematic rating factors, with a particular emphasis placed on understanding the role that rater source plays in MSPRs. In contrast to recent research, our results provide evidence that (a) source effects are present in MSPRs after all, (b) source effects account for a relatively large proportion of the variance in MSPRs but less variance than idiosyncratic rater effects, and (c) the effect of a general performance factor is substantially smaller than has been suggested previously.

### REFERENCES

- Arthur W, Villado AJ. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, 93, 435–442.
- Austin JT, Villanova P. (1992). The criterion problem: 1917–1992. *Journal of Applied Psychology*, 77, 836–874.
- Bentler PM. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler PM, Bonett DG. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Bentler PM, Woodward JA. (1983). The greatest lower bound to reliability. In Wainer H, Messick S (Eds.), *Principals of modern psychological measurement: A Festschrift to Frederic M. Lord* (pp. 237–253). Hillsdale, NJ: Erlbaum.

---

<sup>5</sup>As of 29 DEC 08 22 studies cited in the PsychInfo data base used Viswesvaran et al.'s (1996) interrater reliability estimates to correct for attenuation in meta-analyses involving performance ratings.

- Borman WC. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, 12, 105–124.
- Borman WC. (1997). 360 ratings: An analysis of assumptions and research agenda for evaluating their validity. *Human Resource Management Review*, 7, 299–315.
- Borman WC, Brush DH. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance*, 6, 1–21.
- Bowler MC, Woehr DJ. (2006). A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings. *Journal of Applied Psychology*, 91, 1114–1124.
- Browne MW, Cudeck R. (1993). Alternative ways of assessing model fit. In Bollen K, Long JS (Eds.), *Testing structural models* (pp. 136–162). Newbury Park, CA: Sage.
- Campbell DT, Fiske DW. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell JP, McHenry JJ, Wise LL. (1990). Modeling job performance in a population of jobs. *PERSONNEL PSYCHOLOGY*, 43, 313–333.
- Carty HM. (2003). Review of BENCHMARKS<sup>R</sup> [revised]. In Plake BS, Impara J, Spies RA (Eds.), *The fifteenth mental measurements yearbook* (pp. 123–124). Lincoln, NE: Buros Institute of Mental Measurements.
- Cheung GW, Rensvold RB. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255.
- Church AH, Allen DW. (1997). Advancing the state of the art of 360-degree feedback. *Group and Organization Management*, 22, 149–161.
- Conway JM. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management*, 22, 139–162.
- Conway JM. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, 84, 3–13.
- Conway JM, Huffcutt AI. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10, 331–360.
- Conway JM, Lievens F, Scullen SE, Lance CE. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, 11, 535–559.
- Cooper WH. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218–244.
- Coover MD, Craiger JP, Teachout MS. (1997). Effectiveness of the direct product versus confirmatory factor model for reflecting the structure of multimethod-multirater job performance data. *Journal of Applied Psychology*, 82, 271–280.
- Dansereau F, Graen G, Haga W. (1975). A vertical dyad linkage approach to leadership within formal organizations: A longitudinal investigation of the role making process. *Organizational Behavior and Human Performance*, 13, 46–78.
- DeVries, DL, Morrison AM, Shullman SL, Gerlach ML. (1986). *Performance appraisal on the line*. Greensboro, NC: Center for Creative Leadership.
- Feldman JM. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, 66, 127–148.
- Fleenor JW, McCauley CD, Brutus S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly*, 7, 487–506.
- Guilford JP. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.
- Harris MM, Schaubroeck J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *PERSONNEL PSYCHOLOGY*, 41, 43–62.
- Hoffman BJ, Blair C, Meriac J, Woehr DJ. (2007). Expanding the criterion domain? A meta-analysis of the OCB literature. *Journal of Applied Psychology*, 92, 555–566.
- Hoffman BJ, Woehr DJ. (2009). Disentangling the meaning of multisource feedback source and dimension factors. *Personnel Psychology*, 62, 735–765.

- Holzbach RL. (1978). Rater bias in performance ratings: Superior, self, and peer ratings. *Journal of Applied Psychology*, 63, 579–588.
- Hu LT, Bentler PM. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterization model misspecification. *Psychological Methods*, 3, 424–453.
- Hu LT, Bentler PM. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling*, 6, 1–55.
- Jöreskog K, Sörbom D. (2004) LISREL 8.70. Chicago: Scientific Software International Inc.
- Kenny DA, Berman JS. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin*, 88, 288–295.
- King LM, Hunter JE, Schmidt FL. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65, 507–516.
- Kingstrom PO, Bass AR. (1981). A critical analysis of studies comparing behaviorally anchored rating scale (BARS) and other rating formats. *PERSONNEL PSYCHOLOGY*, 34, 263–289.
- Klimoski RJ, London M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59, 445–451.
- Lakey CE, Goodie AS, Lance CE, Stinchfield S, Winters KC. (2007). Examining DSM-IV criteria for pathological gambling: Psychometric properties and evidence from cognitive biases. *Journal of Gambling Studies*, 23, 479–498.
- Lance CE, Baxter D, Mahan RP. (2006). Multi-source performance measurement: A reconceptualization. In Bennett W, Lance CE, Woehr DJ (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 49–76). Mahwah, NJ: Erlbaum.
- Lance CE, Hoffman BJ, Gentry WA, Baranik LE. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review*, 18, 223–232.
- Lance CE, Noble CL, Scullen SE. (2002). A critique of the correlated trait–correlated method (CTCM) and correlated uniqueness (CU) models for multitrait-multimethod (MTMM) data. *Psychological Methods*, 7, 228–244.
- Lance CE, Teachout MS, Donnelly TM. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. *Journal of Applied Psychology*, 77, 437–452.
- Lance CE, Woehr DJ. (1986). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal of Applied Psychology*, 71, 679–687.
- Lance CE, Woehr DJ, Fisicaro SA. (1991). Cognitive categorization processes in performance evaluation: Confirmatory tests of two models. *Journal of Organizational Behavior*, 12, 1–20.
- Lance CE, Woehr DJ, Meade AW. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, 10, 430–448.
- Landy FJ, Farr JL. (1980). Performance rating. *Psychological Bulletin*, 87, 72–107.
- Lawler EE, III. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, 51, 369–381.
- LeBreton, JM, Burgess JD, Kaiser RB, Atchley EK, James LR. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6, 80–128.
- Li H, Rosenthal R, Rubin DB. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, 1, 98–107.

- Lombardo MM, McCauley CD. (1994). *BENCHMARKS<sup>R</sup> : A manual and trainer's guide*. Greensboro, NC: Center for Creative Leadership.
- Lombardo MM, McCauley CD, McDonald-Mann D, Leslie JB. (1999). *BENCHMARKS<sup>R</sup> developmental reference points*. Greensboro, NC: Center for Creative Leadership.
- Mabe PA, West SG. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67, 280–296.
- Mann FC. (1965). Toward an understanding of the leadership role in formal organizations. In Dubin R, Homans GC, Mann FC, Miller DC (Eds.), *Leadership and productivity* (pp. 68–77). San Francisco: Chandler.
- McCauley C, Lombardo M. (1990). BENCHMARKS<sup>R</sup> : An instrument for diagnosing managerial strengths and weaknesses. In Clark KE, Clark MB (Eds.), *Measures of leadership* (pp. 535–545). West Orange, NJ: Leadership Library of America.
- McCauley C, Lombardo M, Usher C. (1989). Diagnosing management development needs: An instrument based on how managers develop. *Journal of Management*, 15, 389–403.
- Mintzberg H. (1975). The manager's job: Folklore and fact. *Harvard Business Review*, 53, 49–61.
- Mount MK, Judge TA, Scullen SE, Sytsma MR, Hezlett SA. (1998). Trait, rater, and level effects in 360-degree performance ratings. *PERSONNEL PSYCHOLOGY*, 51, 557–576.
- Murphy KR. (2008). Three models of the performance appraisal process. *Industrial and Organizational Psychology Perspectives on Research and Practice*.
- Murphy KR, DeShon R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *PERSONNEL PSYCHOLOGY*, 53, 873–900.
- Newcomb T. (1931). An experiment designed to test the validity of a rating technique. *Journal of Educational Psychology*, 22, 279–289.
- Saal FE, Downey RG, Lahey MA. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413–428.
- Schmidt FL, Viswesvaran C, Ones DS. (2000). Reliability is not validity and validity is not reliability. *PERSONNEL PSYCHOLOGY*, 53, 901–912.
- Scullen SE, Mount MK, Goff M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85, 956–970.
- Sevy BA, Olson RD, McGuire DP, Frazier ME, Paajanen G. (1985). *Managerial skills profile technical manual*. Minneapolis, MN: Personnel Decisions, Inc.
- Smith CA, Organ DW, Near JP. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology*, 68, 653–663.
- Spangler M. (2003). Review of BENCHMARKS<sup>R</sup> [revised]. In Plake BS, Impara J, Spies RA (Eds.), *The fifteenth mental measurements yearbook* (pp. 124–126). Lincoln, NE: Buros Institute of Mental Measurements.
- Sparrowe RT, Liden RC. (1997). Process and structure in leader-member exchange. *Academy of Management Review*, 22, 522–552.
- Steiger JH. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, 25, 173–180.
- Thorndike EL. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25–29.
- Thornton G. (1980). The relationship between supervisory and self appraisals of executive performance. *PERSONNEL PSYCHOLOGY*, 21, 441–455.
- Timmreck CW, Bracken DW. (1997). Multisource feedback: A study of its use in decision-making. *Employment Relations Today*, 24, 21–27.
- Tucker LR, Lewis C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1–10.



- Vance RJ, MacCallum RC, Coover MD, Hedge JW. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, 73, 74–80.
- Viswesvaran C, Ones DS, Schmidt FL. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81, 557–574.
- Viswesvaran C, Schmidt FL, Ones DS. (2002). The moderating influence of job performance dimension on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, 87, 245–354.
- Viswesvaran C, Schmidt FL, Ones DS. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, 90, 108–131.
- Wells FL. (1907). A statistical study of literary merit. (Columbia Univ. Cont. to Phil. & Psych., 16, 3.). *Archives of Psychology*, 7, 5–25.
- Wherry RJ. (1952). *The control of bias in ratings: VIII. A theory of rating* (PRB report No. 922, Contract No. DA-49-083 OSA69, Department of the Army). Columbus, OH: Ohio State University Research Foundation.
- Wherry RJ, Bartlett CJ. (1982). The control of bias in ratings: A theory of rating. *PERSONNEL PSYCHOLOGY*, 35, 521–551.
- Wildaman KF. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1–26.
- Woehr DJ, Huffcutt AI. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189–206.
- Woehr DJ, Sheehan MK, Bennett W. (2005). Assessing measurement equivalence across ratings sources: A multitrait-multirater approach. *Journal of Applied Psychology*, 90, 592–600.
- Yukl GA. (2006). *Leadership in organizations* (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Zedeck S. (1995). Review of BENCHMARKS<sup>®</sup>. In Conoley J, Impara J (Eds.), *The twelfth mental measurements yearbook* (Vol. 1, pp. 128–129). Lincoln, NE: Buros Institute of Mental Measurements.
- Zedeck S, Baker HT. (1972). Nursing performance as measured by behavioral expectation sales: A multitrait-multirater analysis. *Organizational Behavior and Human Performance*, 7, 457–466.

Copyright of Personnel Psychology is the property of Blackwell Publishing Limited and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.