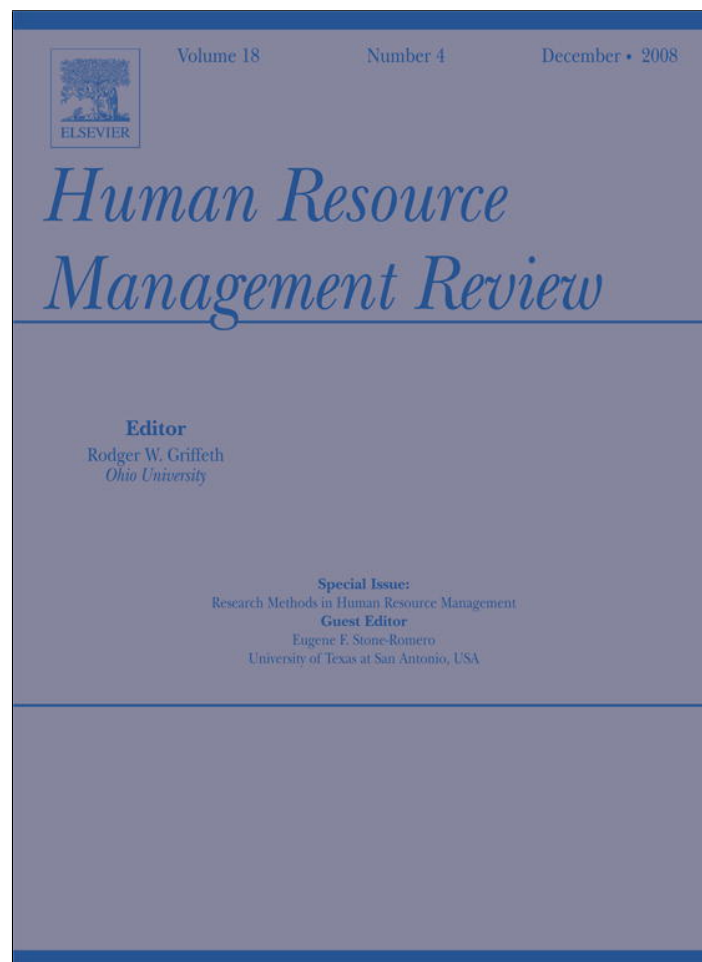


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Contents lists available at [ScienceDirect](#)

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Rater source factors represent important subcomponents of the criterion construct space, not rater bias

Charles E. Lance^{a,*}, Brian J. Hoffman^a, William A. Gentry^b, Lisa E. Baranik^a

^a The University of Georgia, United States

^b Center for Creative Leadership, United States

ARTICLE INFO

Keywords:

Multisource performance ratings
Multitrait-multimethod
MTMM
Confirmatory factor analysis
Psychometrics

ABSTRACT

We contrast normative accuracy and ecological perspectives on applications of the multitrait–multimethod methodology to multisource performance ratings and review existing research that provides critical tests of these perspectives. Existing research supports the ecological perspective which proposes that the rater source effects that are typically found in analysis of multisource performance ratings do not represent mere halo biases but alternative, perhaps equally valid perspectives on ratee performance. We suggest that future research view multifaceted research designs in the broader context of a prototype multidimensional data relational system such as that proposed by Lance, Baranik, Lau, and Scharlau (Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (in press). If it's not trait it must be method: (Mis) application of the multitrait–multimethod design in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences*. Mahwah, NJ: Erlbaum).

© 2008 Elsevier Inc. All rights reserved.

Productivity and job performance are cornerstone constructs in human resource management (Borman, 1991; Pritchard, 1992). For example, job performance measurement is an integral component in test validation research (Campbell, 1990; McDonald, 1999), training needs assessment and training program evaluation (Arvey & Cole, 1989; Campbell, 1988; Ostroff & Ford, 1989), promotion and succession planning (Burack & Mathys, 1987), salary administration (Cascio, 1989; Hammer, 1988), recruitment, selection and placement (Burke & Pearlman, 1988; Schneider & Schmitt, 1992), and developmental feedback toward performance maintenance and improvement (Smither, London, & Reilly, 2005).

The last two decades have seen a dramatic increase in popularity of a particular performance measurement and feedback approach that is known as multisource performance rating (MSPR; Conway & Huffcutt, 1997) or 360-degree assessment and feedback (London & Tornow, 1998). In a typical application of a MSPR program managers are rated by their supervisors, peers, and subordinates (and perhaps also by themselves and their clients), ratings are aggregated within each source where there are multiple raters, and then developmental feedback with respect to the aggregated ratings on relevant performance dimensions is given to ratees for performance review and planning purposes (Church & Bracken, 1997; London & Smither, 1995; Smither et al., 2005). From an applied perspective, MSPRs are thought to be valuable in part because ratings from different sources provide complementary views of the ratee's performance from different organizational perspectives (Borman, 1997). However, from a traditional psychometric perspective, research on MSPRs has consistently produced what has been interpreted as representing a pattern of troubling findings: despite the fact that ratings *within* sources have some demonstrated convergence, almost invariably there is low to moderate convergence (at best) in ratings *across* sources (Conway, 1996; Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; LeBreton, Burgess, Kaiser, & James, 2003). This relative lack of convergence across sources in MSPRs has been viewed as reflecting (a)

* Corresponding address. Department of Psychology, The University of Georgia, Athens, GA 0602-3013, United States. Tel.: +1 706 542 3053; fax: +1 706 542 3275.
E-mail address: clance@uga.edu (C.E. Lance).

valued complementary perspectives on rater performance by 360-degree feedback practitioners and, alternately, (b) the influences of pervasive unwanted, contaminating, rater source *bias* by those who have researched MSPRs from a traditional psychometric perspective. The purpose of this paper is to attempt a resolution of these conflicting interpretations of the same evidential data base.

We argue in this article that MSPR researchers may well have been misled into thinking that rater source effects represent rater *biases* through their cavalier adaptation of the multitrait–multimethod (MTMM) methodology to study the latent structure of MSPRs. Applications of the MTMM methodology to the study of MSPRs have routinely assumed that (a) the rating dimensions under study represent the *Traits* in the MTMM design, and (b) since rater sources do not represent the Traits of interest, they must therefore represent the *Methods* in the MTMM design. Combined with these assumptions and typical empirical findings that (a) correlations between the same performance dimensions as rated by different sources (corresponding to monotrait–heteromethod correlations in an MTMM matrix) are relatively low, and (b) correlations between different performance dimensions as rated by the same source (corresponding to heterotrait–monomethod correlations in an MTMM matrix) are relatively high, MSPRs have been assumed to demonstrate weak convergent validity (at best), and strong and pervasive measurement method effects, from a traditional psychometric perspective. In the remainder of this article we present background on the MTMM methodology, discuss two competing theories on the nature of rater source effects in MSPRs, review literature that has pitted these competing theories against one another empirically, and discuss implications of these findings both from scientific and applied perspectives.

1. Background

The MTMM matrix was introduced by D. T. Campbell and Fiske (1959) as an innovative approach to the study of convergent and discriminant validity of psychological measures. This article is now one of the most often cited in psychology (Fiske & Campbell, 1992) — as of 31 October 2007 it had been cited 4450 times in the Web of Science alone in a wide variety of disciplines including the social and physical sciences, law, medicine and education (see Lance et al., *in press*). One of the reasons for the widespread adoption of the MTMM methodology is that the provision of convergent and discriminant validity evidence is widely regarded as a cornerstone for the establishment of measures' construct validity (Benson, 1998; Messick, 1995).

D. T. Campbell and Fiske's (1959) criteria for convergent and discriminant validity and the presence of method effects are now widely recognized as being rather subjective (Widaman, 1985). As a result, a number of more objective, quantitative approaches to the analysis of MTMM matrices have developed over the years, including analysis of variance (e.g., Boruch, Larkin, Wolins, & McKinney, 1970; Kavanagh, MacKinney & Wolins, 1971), path analysis (Avison, 1978; Schmitt, 1978) multiple regression (e.g., Lehmann, 1988), and exploratory factor analysis (Golding & Seidman, 1974; Wothke, 1995). Today however, the most popular analytic model of choice is some form of a confirmatory factor analysis (CFA) model. A number of such models have been proposed including a family of additive models that specify latent Trait and Method factors' effects on observed measures (e.g., Widaman, 1985), models that include latent Trait factors but which model method effects as covariances among uniquenesses of measures of traits using the same measurement method (e.g., Marsh, 1989), hierarchical CFA models (e.g., Marsh & Hocevar, 1988), models that specify interactive Trait x Method effects (Browne, 1984; Campbell & O'Connell, 1967), in addition to others (e.g., Eid, 2000; Kenny & Kashy, 1992; Lance, Woehr, & Meade, 2007). The first two of these, sometimes referred to as the correlated trait–correlated method (CTCM) and the correlated trait–correlated uniqueness (CTCU) models, are the most widely accepted and implemented models. The CTCM model can suffer convergence and admissibility problems (Brannick & Spector, 1990; Kenny & Kashy, 1992) that the CTCU model often avoids, but Conway, Lievens, Scullen, and Lance (2004) and Lance, Noble and Scullen (2002) showed that the CTCU model suffers from a number of other serious conceptual and analytic problems. Consequently we only consider the general CTCM model (and several of its special cases) here.

2. An Illustration of the CTCM Model for MSPRs

As an illustration of applying the MTMM framework to MSPRs, we obtained data from a sample of 22,420 managers who were rated on the Center for Creative Leadership's Benchmarks®¹ (Center for Creative Leadership, 2004; Lombardo & McCauley, 1994; McCauley, Lombardo, & Usher, 1989) multisource feedback instrument by their supervisor, peers, subordinates and themselves. Benchmarks® consists of 16 separate measures, but in the interest of parsimony we combined these into three broad performance dimensions (Meeting Job Responsibilities, Respecting Self and Others, and Leading People) that map onto a generalized taxonomy of managerial performance developed by Borman and Brush (1993) and that has previously been used with the Benchmarks® instrument (see Fleenor, McCauley, & Brutus, 1996). Also, although common practice is to aggregate ratings within sources when presenting feedback, for illustrative purposes we randomly selected (a) one peer and one subordinate, and (b) a subset of 520 rateres for analysis. Thus the subsample reported here consists of 520 managers who provided self-ratings and who were rated by their supervisor and one peer and subordinate each.

Correlations among Benchmarks® ratings are shown in Table 1. A preliminary (and subjective) analysis of these ratings indicates some support for convergent validity: the average different source–same dimension correlation (mean $r = .18$) is somewhat larger than the mean different source–different dimension correlation (mean $r = .13$, $t(52) = 4.52$, $p < .01$), although this difference is not large in an absolute sense. On the other hand, there is evidence of strong rating source effects: the mean different dimension–same source correlation (mean $r = .81$) is substantially higher than either the average same dimension–different source correlation (mean $r = .18$, $t(28) = 33.96$, $p < .01$) or the average different dimension–different source correlation (mean $r = .13$, $t(46) = 39.74$, $p < .01$).

¹ Benchmarks® is a registered trademark of the Center for Creative Leadership.

Table 1
Correlations among Benchmarks® multisource ratings

Ratings	1	2	3	4	5	6	7	8	9	10	11
1. SUP-MJR	1.00										
2. SUP-RSO	.74	1.00									
3. SUP-LP	.79	.87	1.00								
4. PEER-MJR	.26	.20	.21	1.00							
5. PEER-RSO	.16	.26	.20	.78	1.00						
6. PEER-LP	.18	.21	.22	.82	.89	1.00					
7. SUB-MJR	.19	.13	.15	.18	.12	.14	1.00				
8. SUB-RSO	.12	.19	.16	.14	.20	.17	.82	1.00			
9. SUB-LP	.14	.16	.17	.15	.16	.17	.86	.91	1.00		
10. SELF-MJR	.18	.03	.09	.15	.02	.07	.16	.06	.10	1.00	
11. SELF-RSO	.08	.17	.12	.09	.16	.12	.09	.17	.13	.68	1.00
12. SELF-LP	.11	.10	.15	.10	.09	.13	.13	.13	.17	.76	.80

Note: SUP = Supervisor rating, PEER = Peer rating, SUB = Subordinate rating, SELF = Self rating, MJR = Meeting Job Requirements, RSO = Respecting Self and Others, LP = Leading People. For $r > .09$ $p < .05$, for $r > .12$, $p < .01$. Same source-different dimension correlations are shown in *italics*, different source-same dimension correlations are shown in **boldface**.

Finally, the generally low level of the different dimension-different source correlations provides some evidence of discriminant validity for the ratings. These results are typical in indicating strong rater source effects on, and weak convergent validity between dimensional performance ratings.

Results from CFA of the Benchmarks® ratings are shown in Table 2. The complete CTCM model (Model 1–4 Rating Sources × 3 Performance Dimensions) fit the data well according to conventional goodness-of-fit criteria (i.e., a statistically non-significant χ^2 statistic, SRMSR < .08, RMSEA < .06, TLI > .95, CFI > .95, see Hu & Bentler, 1998, 1999). Table 3 shows results for the complete model. The majority of the loadings on the Dimension factors (three right-most columns) were statistically significant, providing support for the ratings' convergent validity in assessing the latent Dimension factors (average Dimension factor loading = .265, SD = .09). Convergent validity is also supported by the $\Delta\chi^2$ test shown in Table 2 indicating that the full CTCM model (Model 1) fit the data significantly better than did the reduced 0 trait-correlated method (OTCM) model (Model 2). Discriminant validity is also supported by (a) the $\Delta\chi^2$ test comparing the full CTCM model to a reduced 1 trait-correlated method (1TCM) model (Model 3, see Table 2), and (b) relatively low and statistically non-significant correlations among the Dimension factors as is shown in the lower portion of Table 3 (mean factor correlation = .23). Finally, substantial rater source effects are evidenced (a) in Table 2 by the large and statistically significant $\Delta\chi^2$ test comparing the full CTCM model to a reduced correlated trait-0 method (CTOM) model (Model 4), and (b) by the large and statistically significant loadings of the measures on the rating source factors (first four columns in Table 3; mean Rater Source factor loading = .89, SD = .05). Additional evidence of the importance of the rating source factors is the fact that the CTOM model failed to converge in an admissible solution. This is one indication that the CTOM model was inconsistent with the data and is a common finding for MSPR data (Lance et al., 2002). Together, this example provides support for the construct (i.e., convergent and discriminant) validity of Benchmarks® multisource ratings but also indicates the presence of substantial rater source effects. In fact, the mean Rater Source factor loading (= .89; primary evidence for the presence of Source effects) was 236% larger than the mean Dimension factor loading (.265, primary evidence for convergent validity). This is not an uncommon finding.

3. CFA of multisource performance ratings

There now exists a large number of studies that have used the MTMM methodology to investigate the structure of multisource performance ratings in the manner illustrated in the previous example. In these applications a multitrait-multisource (MTMS) matrix is generated and in almost all cases the dimensions that are being rated are assumed to represent the trait facet of measurement and the rating sources are assumed to represent the method facet. Examples of the attributions that dimen-

Table 2
Model goodness-of-fit

Model	χ^2	df	SRMSR	RMSEA	TLI	CFI
1. 7-factor CTCM model 4 source factors plus 3 dimension factors	13.66	33	.012	.01	1.01	1.00
1 vs. 2: Convergent validity	415.01*	15	–	–	–	–
2. 4-factor OTCM model: 4 source factors only	428.67*	48	.029	.14	.91	.93
1 vs. 3: Discriminant validity	92.39*	3	–	–	–	–
3. 5-factor 1TCM model: 4 source factors plus 1 global "Trait" factor	106.05*	36	.011	.07	.98	.99
1 vs. 4: source effects	2969.97*	18	–	–	–	–
4. 3-factor CTOM model: 3 dimension factors only ^a	4036.07*	51	.300	.35	.10	.31

Note. df = model degrees of freedom, SRMSR = standardized root mean squared error, RMSEA = root mean squared error of approximation, TLI = Tucker–Lewis index, CFI = comparative fit index, CTCM = correlated trait-correlated method, OTCM = 0 trait-correlated method, 1TCM = 1 trait-correlated method, CTOM = correlated trait-0 method. * $p < .01$.

^a Model 4 failed to converge in an admissible solution.

Table 3
CTCM model results

Variables:	SUP	SUB	PEER	SELF	MJR	RSO	LP
<i>Factor loadings</i>							
Supervisor							
MJR	.82**	–	–	–	.32**	–	–
RSO	.90**	–	–	–	–	.28**	–
LP	.94**	–	–	–	–	–	.18
Subordinate							
MJR	–	.86**	–	–	.27**	–	–
RSO	–	.91**	–	–	–	.25**	–
LP	–	.95**	–	–	–	–	.15
Peer							
MJR	–	–	.88**	–	.23**	–	–
RSO	–	–	.93**	–	–	.19**	–
LP	–	–	.96**	–	–	–	.15*
Self							
MJR	–	–	–	.81**	.39**	–	–
RSO	–	–	–	.86**	–	.36**	–
LP	–	–	–	.86**	–	–	.41*
<i>Factor correlations</i>							
SUP	1.00						
SUB	.22**	1.00					
PEER	.16**	.17**	1.00				
SELF	.09	.08	.12*	1.00			
MJR	–	–	–	–	1.00		
RSO	–	–	–	–	–.07	1.00	
LP	–	–	–	–	.34	.42	1.00

Note. CTCM = correlated trait-correlated method, SUP = Supervisor rating, PEER = Peer rating, SUB = Subordinate rating, SELF = Self rating, MJR = Meeting Job Requirements, RSO = Respecting Self and Others, LP = Leading People. * $p < .05$, ** $p < .01$.

sions=traits and sources=methods abound. For example Conway wrote that “different trait-same rater correlations share a common method (i.e., the same rater)” (Conway, 1996, p. 143), and “[r]esearchers have often defined method variance in the Multitrait–Multirater (MTMR) sense...In the MTMR framework, method variance is the systematic dimension-rating variance specific to a particular source” (Conway, 1998, p. 29). Also, Mount, Judge, Scullen, Systma, and Hezlett (1998) summarized, “[s]tudies that have examined performance rating data using multitrait–multimethod matrices (MTMM) or multitrait–mutirater (MTMR) matrices usually focus on the proportion of variance in performance ratings that is attributable to traits and that which is attributable to the methods or raters” (p. 559) and that these studies have “documented the ubiquitous phenomenon of method effects in performance ratings” (p. 568; see also, Becker & Cote, 1994; Conway & Huffcut, 1997; Doty & Glick, 1998; Podsakoff, MacKenzie, Podsakoff, & Lee, 2003 for similar attributions).

Results of several CFA studies of MTMS data indicate that both dimension factors and source factors are important in explaining covariances among multisource ratings (e.g., Campbell, McHenry, & Wise, 1990; Covert, Craiger, & Teachout, 1997; Holzbach, 1978; King, Hunter, & Schmidt, 1980; Klimoski & London, 1974; Lance, Teachout, & Donnelly, 1992; Mount et al., 1998; Scullen, Mount, & Goff, 2000; Vance, MacCallum, Covert, and Hedge, 1988; Woehr, Sheehan, & Bennett, 2005; Zedeck & Baker, 1972). That is, results of these studies indicate that multisource ratings reflect both the dimensions they were designed to represent and the sources of measurement. This is somewhat disturbing from a traditional psychometric perspective because method (in this case rating source) effects are often thought of as sources of unwanted contaminating variance (Burns, Walsh, & Gomez, 2003), and one comprehensive review of this literature estimated that Source factors accounted for 56% more variance in ratings than did Dimension factors (Conway, 1996). Perhaps even more troubling is the fact that many other studies of this type fail to find support for any discriminable dimension (i.e., trait) factors at all. In these cases all covariance between ratings is due to (correlated) source effects or source effects plus a single undifferentiated general performance factor (see Lance et al., 2002 for a review). Thus from a traditional psychometric perspective multisource ratings often (a) fail to exhibit convergent validity in representing the performance dimensions they were designed to reflect, and (b) reflect substantial proportions of undesirable method variance.

A related body of literature has assessed levels of convergence among raters within and between sources (Borman, 1997; Murphy, Cleveland, & Mohler, 2001). A comprehensive meta-analysis by Conway and Huffcutt (1997) indicated that the average relationship between ratings provided by raters from different organizational levels is typically somewhat weak (average $r = .22$), and this finding is consistent with other meta-analyses (Harris & Schaubroeck, 1988; Viswesvaran, Ones, & Schmidt, 1996) and our findings reported in Table 3 (mean source factor correlation = .14). Together, findings from studies of CFA of MTMS data and other studies of interrater agreement indicate that (a) multisource ratings do not exhibit strong levels of convergent validity across sources in representing the dimensions they were intended to assess, and (b) rater source effects interject substantial proportions of contaminating method variance into correlations among ratings. This seems to be not so good news for multisource feedback programs. Or is it?

4. A contrast of two paradigms

4.1. Normative accuracy model

Recently, Lance, Baxter, and Mahan (2006) proposed two alternative interpretations for these sets of findings. One of these, a *normative accuracy model* is based in traditional psychometric theory and mathematical performance rating models such as Guilford's (1954), Kenny and Berman's (1980), King et al.'s (1980), Viswesvaran, Schmidt and Ones' (2000) and Wherry and Bartlett's (1982), all of which specify rater (source) bias factors as part of their theories. Stated generically, a *normative accuracy model* can be written as:

$$X = T + SB + E \quad (1)$$

where X is some (dimensional) performance rating, T is the ratee's corresponding (dimensional) true score, SB represents systematic rater bias and E refers to nonsystematic measurement error. Rater source effects are thought to interject the systematic bias specified by normative accuracy models and are modeled in CFA of MTMS matrices as putative method factors. Examples of SBs that have received attention in the performance rating literature include halo error, deliberate rating distortions, and "cognitive heuristics in observing, storing, retrieving ratee performance information" (Lance et al., 2006, p. 51). The key point here is that rater source effects commonly found in CFA of MTMS data are interpreted under a normative accuracy framework as representing unwanted, systematic *bias* effects on multisource ratings that in the past researchers have sought to minimize by developing rating technologies such as improved rating formats (e.g., Kingstrom & Bass, 1980; Landy & Farr, 1980) and various rater training programs (Woehr & Huffcutt, 1994).

4.2. Ecological perspective

Lance et al. (2006) also provided a second, competing perspective on findings from research on multisource ratings that they called an *ecological perspective*. Rooted in Gibson's (1950, 1979) early work on visual perception and adaptations of his work by others in the area of person perception (e.g., Funder, 1987; McArthur & Baron, 1983; Swann, 1984), an ecological perspective on multisource ratings "emphasizes the essential accuracy of perception-based knowledge" (McArthur & Baron, p. 230). Fundamentally, an ecological perspective views perception as serving an adaptive function, providing perceptual information that is useful in directing the organism toward goal attainment. Central to this notion is the idea that organisms are attuned to affordances offered by aspects of their environment, including aspects of their social environment. Affordances represent opportunities for the perceiver to act upon the environment or to be acted upon (Beauvois & Dubois, 2000). In the context of multisource ratings, it is expected that ratees will be attuned to different affordances and therefore have different interaction goals with different constituencies, and vice versa. For example, graduate student ratees may specify very different affordances in the roles of (a) instructor of an introductory research methods class, (b) a collegial member of their graduate student peer group, and (c) a student in a doctoral-level performance appraisal seminar. As a result, raters who occupy different organizational roles relative to the ratee will have different interaction goals with the ratee and, consequently, may be privy to very different sets of performance related behavior on the part of the ratee (Borman, 1974, 1997; Burns et al., 2003; London & Smither, 1995; Zedeck, Imparto, Krausz, & Oleno, 1974). Thus an ecological perspective views rating source effects as representing "overall assessments of different sets of performance-related behaviors (Lance et al., 1992, p.448), as "distinct views of a common individual's job performance [that] may be equally valid" (Landy & Farr, 1980, p. 76), or "meaningful differences in... behavior across sources, especially when each source rates...behavior in different situations" (Burns et al., 2003, p. 242). These ideas are not new (Murphy & DeShon, 2000a,b) but they have yet to be acknowledged in contemporary psychometric models of performance ratings (Le, Oh, Shaffer, & Schmidt, 2007; Schmidt, Viswesvaran & Ones, 2000; Viswesvaran et al., 1996). In the following section we review research that contrasts the normative accuracy and ecological perspectives.

5. Relevant research

To our knowledge there have been three direct competitive tests between the normative accuracy and ecological perspective interpretations of rater source effects on MSPRs. In the first of these, Lance et al. (1992) conducted hierarchical CFA (HCFA) of self-, supervisor and peer ratings obtained on and from 261 United States Air Force (USAF) Ground Equipment Mechanics and supported a second-order factor (SOF) structure that contained both the hypothesized performance dimensions and rater source factors. Next, they augmented the "core" HCFA with measures of additional "external" performance related constructs including mechanical aptitude, technical school grade, job knowledge, and job experience as a critical test of the normative accuracy model versus ecological perspective predictions. According to the normative accuracy perspective, rater source effects represent contaminating, performance-irrelevant rater biases (Viswesvaran et al., 2000) and therefore ought *not* to correlate with these external performance-related variables. Alternately, the ecological perspective considers rater source effects as representing different but complementary perspectives on ratee performance such that rater source factors *should* correlate with performance-related external variables. Results indicated that 8 out of 16 correlations between rater source SOFs and performance-related external variables were statistically significant and in the predicted direction, providing support for the ecological perspective. In a second study, Lance et al. (2006) substantially replicated these findings in a sample of 1017 incumbents in six additional USAF Specialties.

Finally, Hoffman and Woehr (submitted for publication) extended Lance et al.'s (1992, 2006) findings by proposing that different rater source factors should relate differentially to variables measured outside the core CFA of MSPRs in a broader nomological

network (Cronbach & Meehl, 1955) using MSPRs obtained from 440 participants in an executive MBA program. As predicted, Hoffman and Woehr found that the Subordinate source factor correlated most strongly with measures of the ratee's leadership skills, whereas the Peer source factor correlated most strongly with measures of the ratee's interpersonal skills. Together, Lance et al.'s and Hoffman and Woehr's findings answer calls for an examination of the nomological network surrounding the source factors characteristic of MSPRs (Borman, 1997; Conway, 2000; Woehr et al., 2005) and support contentions from an ecological perspective that rater source effects do not represent (mere) rater *biases*, but rather represent alternative but complementary *valid* perspectives on ratee performance (Borman, 1974, 1997; London & Smither, 1995; Tornow, 1993; Zedeck et al., 1974). Thus these findings support the idea that the strong and pervasive rater source factors that are routinely supported in analyses of MSPRs are more properly interpreted as representing what has been called "valid halo" (Bingham, 1939) or "true halo" (Cooper, 1981), or perhaps more appropriately "valid general impression" (Lance & Woehr, 1986) or "performance-based general impression" (Lance, Woehr, & Fiscaro, 1991) and *not* performance-irrelevant rater *error* bias (Viswesvaran et al., 2000).

But normative accuracy models remain dominant, at least in psychometric circles. For example, in one large-scale meta-analysis, Viswesvaran et al. (2000) concluded that there exists a general factor in peer and supervisory ratings that accounts for somewhere around "60% of total variance" at the construct level, but that within-source correlations are "substantially inflated by halo for both supervisory (33%) and peer (63%)" ratings (p. 108). Of course, the findings summarized here indicate that these within-source "halo" effects are at least partly performance-based. In fact there is other, indirect evidence that support the ecological perspective interpretation that rater general impressions that drive rater source effects are largely performance based. For example, Lance et al. (1991) distinguished empirically between performance-based (PBGI) and nonperformance-based aspects of raters' general impressions (nPBGI) and found that raters' overall performance ratings correlated much more strongly with PBGI ($r=.571, p<.01$) than with nPBGI ($r=.186, p<.05$). As a second example, Nathan and Tippins (1990) examined the moderating effects of halo on test validation results and found that selection tests were *more* valid when halo was greater and that validity diminished at lower levels of halo. As a third example, performance rating research has demonstrated that rater general impression is positively related to rating accuracy (Fiscaro, 1988). Finally, Burns et al. (2003) investigated the meaning of method effects in Parent and Teacher responses to questions about attention deficit hyperactivity disorder symptoms (ADHD) of target children. As is often the case, CFA of these multisource ratings supported the presence of both source (teacher and parent) and trait factors (ADHD symptoms). Based on the convergence of method effects over a three month time interval, the authors reasoned that source effects "represent the situational specificity of the child's behavior rather than a form of bias associated with characteristics of the rater" (Burns et al., p. 539). Thus, and consistent with the ecological perspective, raters' general impressions (the alleged "method" effects in many MTMR studies) have been shown to be largely performance-based and not highly error prone as was assumed under normative accuracy perspectives.

6. Whence the continued prominence of normative accuracy models?

So how did it come to pass that normative accuracy models became so prominent in rating research? We speculate three possibilities. First, it was the case that much of the earlier social judgment research focused on errors and mistakes that raters made in rating others, especially in circumscribed, primarily laboratory settings (Funder, 1987) and the job performance rating literature followed suit (Landy & Farr, 1980; Saal, Downey & Lahey, 1980). As such, the focus of much of the early performance rating literature was on what was wrong with ratings (including leniency and halo *errors*) and how to fix them through interventions such as improved rating formats and rater training programs.

The second possibility stems from the (implicit or explicit) assumption underlying most psychometric rating models (e.g., Guilford, 1954; King et al., 1980; Viswesvaran et al., 2000; Wherry & Bartlett, 1982), that lack of interrater agreement indicates that one or both raters is wrong, or biased in their ratings. Excellent examples of this attribution are (a) Viswesvaran et al.'s (2000) claim that the "part of the overall impression that is in common with other raters... is considered true variance.... The part that is unique to that rater – the idiosyncratic part – is halo error" (p. 109) and (b) Wherry and Bartlett's (1982) specification of overall rater bias ("BRO") components in their psychometric theory of rating. In both of these cases the assumption was that idiosyncratic rater general impressions, the source of rater (source) effects in MSPRs, represented errors in the rating process that were to be minimized.

A third possibility stems from what might be regarded as the rather cavalier adaptation of the MTMM methodology to the study of MSPRs. As we noted above, it has been a common practice to analyze MSPRs using a quasi-MTMM analytic approach under the default assumption that performance rating dimensions=traits and raters/sources=methods. Combined, traditional psychometric assumptions that (a) rater sources represent (mere) measurement methods, (b) measurement method effects represent undesirable sources of contaminating variance, along with the pervasive findings that (c) rater/source effects are found in MSPR data, points to the conclusion that MSPRs are substantially contaminated with measurement method bias in the form of halo error. However, the evidence reviewed here indicates otherwise, namely that MSPRs capture, in part, common parts of the criterion construct space that represents interrater (inter-source) convergence on ratee performance effectiveness and, in part, unique aspects of the criterion construct space that are captured by the different raters/sources.

7. Implications

One of the key assumptions that justifies operational 360° feedback programs is that the various rater constituencies provide nonredundant, perhaps minimally overlapping perspectives on ratee performance (e.g., London & Smither, 1995; Tornow, 1993). Otherwise, why go to the trouble of obtaining ratings from so many different rater sources? The first implication from the findings

reviewed here is that this assumption seems to be justified. That is, the findings reviewed here support the idea that “supervisory and peer ratings may represent two distinct views of a common individual's job performance and may be equally valid, even though they are not highly correlated” (Landy & Farr, 1980, p. 76). In other words, “supervisors likely evaluate an individual's job performance quite differently than his or her subordinates would, in that supervisors rate the focal individual in his or her role as a subordinate, and subordinates rate the focal individual in his or her role as a supervisor...[so that]... interrater ratings from different rating sources should not necessarily be in agreement, in that they are not assessing the same, but different, aspects of job performance” (Bozeman, 1997, p. 314).

Second, the veracity of performance rating psychometric model assumptions needs to be evaluated in the light of research that indicates that rater (source) effects do not represent mere halo error biases. Although useful in highlighting the multifaceted and multidimensional nature of ratings and criterion measures in general, we feel that some of their (implicit or explicit) assumptions have been too restrictive and have misled psychometric ratings researchers into painting the picture that performance ratings are routinely flawed, highly biased and in urgent need of repair. Rather, we suggest that performance ratings may not be as broken as they seem to be on the psychometric surface.

Third, findings from the research reviewed here suggest that corrections to validity coefficients for attenuation due to unreliability that are based on Viswesvaran et al.'s (1996) estimates of the reliability of job performance ratings, and which are now widely applied (e.g., Judge, Thoresen, Bono, & Patton, 2001; Le et al., 2007) are inappropriate. Viswesvaran et al.'s interrater estimate of the reliability of performance ratings is grounded in a normative accuracy paradigm and makes the implicit assumption that different raters are at least as interchangeable as congeneric tests in which their true score components are, minimally, unidimensional and linearly related (Lord & Novick, 1968). The research reviewed here suggests that these assumptions are unwarranted, at least as far as the interchangeability of raters from *different* levels is concerned as raters from different organizational perspectives appear to capture both common and unique aspects of the performance construct space (i.e., ratings contain both common and unique true score components). However other, related research bears directly on the issue. Specifically, Mount et al. (1998) and Scullen et al. (2000) found support for systematic and pervasive idiosyncratic rater effects above and beyond the rater source effects discussed here, suggesting that even raters *within the same level* (multiple peers or supervisors) are also not interchangeable (i.e., congeneric). The implications of these findings for corrections for attenuation can be seen by contrasting the normative accuracy and ecological perspectives in reference to the Venn diagram in Fig. 1 (note that the relative sizes of the areas in this Figure of are not intended to represent relative proportions of variance accounted for in any ecological sense). Most psychometric models (such as the one that forms the basis for Viswesvaran's interrater reliability estimate) consider only the overlap between raters' ratings as representing “true-score” or reliable variance in estimating interrater reliabilities. That is, only the shared, or common aspects of multiple raters' ratings, which may include relevant or valid variance – area A in Fig. 1 – or irrelevant variance (shared bias) – area D in Fig. 1, is considered to be reliable variance and is accounted for in the calculation of interrater reliabilities. But this ignores reliable and valid aspects that are *unique* to each rater's ratings (areas B in Fig. 1) that should rightfully be included in the reliability estimate. As is well known, test (or rater qua “test”) reliability is a theoretical quality that cannot be calculated directly but must somehow be estimated, and usually many different reliability estimates are available (Nunnally & Bernstein, 1994). For the reasons stated above, interrater reliabilities likely yield very lower-bound estimates of ratings' actual reliabilities and, when used in correction for attenuation formulae, yield inflated and non-credible estimates of disattenuated validities (James, 1996; Murphy & DeShon, 2000a). In fact, alternate greater lower-bound (GLB) reliability estimates that would more closely approximate ratings' actual reliabilities, and that would be more appropriate for attenuation corrections are routinely available (Bentler & Woodward, 1980; Drewes, 2000). For example, communality estimates from MTMM-related designs discussed here yield one such GLB reliability estimate. For example, squared multiple correlation communality estimates for the ratings reported in Table 2 ranged between .78 and .95 (mean $h^2 = .89$), which of course (a) are much higher than Viswesvaran's interrater reliability estimate of .52, (b) are much more in line with other ratings reliability estimates (e.g., internal consistency), and (c) when used in correction for attenuation formulae would yield less inflated and more credible estimates of disattenuated validities. As such,

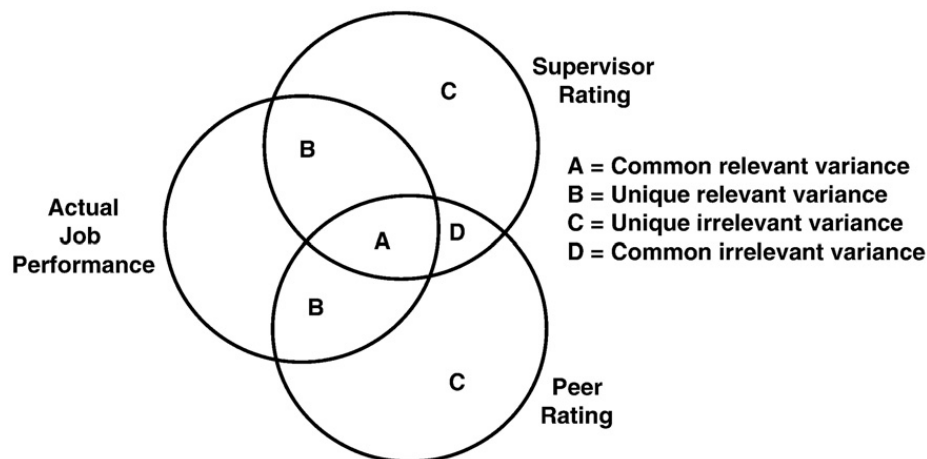


Fig. 1. Venn diagram of overlap between ratings and actual job performance.

interrater reliability estimates should be viewed as unrealistically lower-bound estimates of ratings' reliabilities, should not be used to disattenuate observed test validities, and should be replaced with more appropriate GLB estimates of ratings' reliabilities.

Finally, we urge MSPR researchers and researchers in general to take a broader perspective on multifaceted measurement designs. For example, in the context of MTMM designs, Doty and Glick (1998) suggested that at least two different forms of "measurement method" can be distinguished: (a) various measurement techniques (e.g., item formats, item wording, and data collection processes such as paper and pencil survey vs. online administration), and (b) data sources (e.g., multiple raters, multiple informants). Also, Lance et al. (in press) suggested that the default assumption that "if it ain't trait it must be method" (p. 1) has become institutionalized as a methodological urban legend, meaning that in a three faceted design where one facet represents research participants, of the remaining two facets it is common practice to assume that the one that is not the trait facet of interest is (necessarily) a measurement method facet. They urged researchers to resist this default mode of thinking and instead locate their particular research design within a prototype multidimensional data relational system consisting of six dimensions including:

- (a) persons (or groups of persons, or collectivities of groups of persons who may be the object of study), (b) focal constructs that constitute the relevant characteristics of the entities studied; (c) occasions, or temporal replications of measurement; (d) different situations in which measurement may occur; (e) observers or recorders of entities' behavior, and (f) response modalities/formats (p. 19).

Within this system one can see how a typical MSPR study involves the persons (ratees), focal constructs (performance dimensions), and observers or recorders (raters/sources) dimensions. It is now becoming clear that this latter dimension does not merely represent alternative methods of collecting performance rating data, but effects that are far more interesting.

8. Summary and conclusion

MSPRs continue as a popular method of diagnosing ratees' skills and informing subsequent activities directed toward performance improvement. Still, important questions remain as to the psychometric properties of these popular tools, not the least of which is, to what extent should we be concerned about systematic differences observed in different raters perspectives of target performance? Of course, traditional psychometric theory and normative accuracy models would suggest that we should be very concerned. In contrast, the ecological perspective argues that systematic source effects in MSPRs should be expected and are even desirable. In an attempt to reconcile these two views, we reviewed studies that provide direct comparisons of the normative accuracy and ecological perspectives. In demonstrating relationships between source effects and other constructs, this research provided support for the ecological perspective on ratings' validity. Thus our review indicates that source effects in MSPRs do not represent mere method bias, but instead, represent important and differentially valid performance relevant information. It is our hope that the present review will stimulate awareness of the important variance captured by MSPR source effects and stimulate further research examining the meaning of other non-trait components of measurement.

References

- Arvey, R. D., & Cole, D. A. (1989). Evaluating change due to training. In I. L. Goldstein (Ed.), *Training and development in organizations* (pp. 89–118). San Francisco: Jossey-Bass.
- Avison, W. R. (1978). Auxiliary theory and multitrait–multimethod validation: A review of two approaches. *Applied Psychological Measurement*, 2, 431–447.
- Beauvois, J., & Dubois, N. (2000). Affordances in social judgment: Experimental proof of why it is a mistake to ignore how others behave towards a target and look solely at how the target behaves. *Swiss Journal of Psychology*, 59, 16–33.
- Becker, T. E., & Cote, J. A. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management*, 20, 625–641.
- Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, 17, 10–22.
- Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, 45, 249–267.
- Bingham, W. (1939). Halo, valid and invalid. *Journal of Applied Psychology*, 23, 221–228.
- Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance*, 12, 105–124.
- Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), 2nd ed. *Handbook of industrial and organizational psychology*, Vol. 2. (pp. 271–326). Palo Alto, CA: Consulting Psychologists.
- Borman, W. C. (1997). 360 ratings: An analysis of assumptions and 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.
- Boruch, R. F., Larkin, J. D., Wolins, L., & McKinney, A. C. (1970). Alternative method of analysis: Multitrait–multimethod data. *Educational and Psychological Measurement*, 30, 833–854.
- Bozeman, D. (1997). Interrater agreement in multi-source performance appraisal: A commentary. *Journal of Organizational Behavior*, 18, 313–316.
- Brannick, M. T., & Spector, P. E. (1990). Estimation problems in the block-diagonal model of the multitrait–multimethod matrix. *Applied Psychological Measurement*, 14, 325–339.
- Browne, M. W. (1984). The decomposition of multitrait–multimethod matrices. *British Journal of Mathematical and Statistical Psychology*, 37, 1–21.
- Burack, E. H., & Mathys, N. J. (1987). *Human resource planning: A pragmatic approach to manpower staffing and development*. Lake Forest, IL: Brace-Park.
- Burke, M. J., & Pearlman, K. (1988). Recruiting, selecting, and matching people with jobs. In J. P. Campbell & R. J. Campbell (Eds.), *Productivity in organizations* (pp. 97–142). San Francisco: Jossey-Bass.
- Burns, G. L., Walsh, J. A., & Gomez, R. (2003). Convergent and discriminant validity of trait and source effects in ADHD–inattention and hyperactivity/impulsivity measures across a 3-month interval. *Journal of Abnormal Child Psychology*, 31, 529–541.
- Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Campbell, D. T., & O'Connell, E. J. (1967). Method factors in multitrait–multimethod matrices: Multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409–426.
- Campbell, J. P. (1988). Training design for performance improvement. In J. P. Campbell & R. J. Campbell (Eds.), *Productivity in organizations* (pp. 177–216). San Francisco: Jossey-Bass.

- Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L.M. Hough (Eds.), 2nd ed. *Handbook of industrial and organizational psychology, Vol. 1*. (pp. 687–732). Palo Alto, CA: Consulting Psychologists.
- Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology*, 43, 313–333.
- Cascio, W. F. (1989). *Managing human resources: Productivity, quality of work life, profits*, (2nd ed.). New York: McGraw-Hill.
- Center for Creative Leadership (2004). *BENCHMARKS® facilitator's manual*. Greensboro, NC: Center for Creative Leadership.
- Church, A. H., & Bracken, D. W. (1997). Advancing the state of the art of 360-degree feedback. *Group & Organization Management*, 22, 149–161.
- Conway, J. M. (1996). Analysis and design of multitrait–multirater performance appraisal studies. *Journal of Management*, 22, 139–162.
- Conway, J. M. (1998). Understanding method variance in multitrait–multirater performance appraisal matrices: Examples using general impressions and interpersonal affect as method factors. *Human Performance*, 11, 29–55.
- Conway, J. M. (2000). Managerial performance development constructs and personality correlates. *Human Performance*, 13, 23–46.
- Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, 10, 331–360.
- Conway, J. M., Lievens, F., Scullen, S. E., & Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, 11, 535–559.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218–244.
- Coover, M. D., Craiger, J. P., & Teachout, M. S. (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.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Doty, D. H., & Glick, W. H. (1998). Common method bias: Does common methods variance really bias results? *Organizational Research Methods*, 1, 374–406.
- Drewes, D. W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods*, 5, 214–227.
- Eid, M. (2000). A multitrait–multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.
- Fisicaro, S. A. (1988). A re-examination of the relation between halo error and accuracy. *Journal of Applied Psychology*, 73, 239–244.
- Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin*, 112, 393–395.
- Fleener, J. W., McCauley, C. D., & Brutus, S. (1996). Self-other rating agreement and leader effectiveness. *Leadership Quarterly*, 7, 487–506.
- Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin*, 101, 75–90.
- Gibson, J. J. (1950). *The perception of the visual world*. Boston: Houghton-Mifflin.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.
- Golding, S. L., & Seidman, E. (1974). Analysis of multitrait–multimethod matrices: A two step principal components procedure. *Multivariate Behavioral Research*, 9, 479–496.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- Hammer, T. H. (1988). New developments in profit sharing, gainsharing, and employee ownership. In J. P. Campbell & R. J. Campbell (Eds.), *Productivity in organizations* (pp. 328–366). San Francisco: Jossey-Bass.
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, 41, 43–62.
- Hoffman, B. J., & Woehr, D. J. (submitted for publication). Disentangling the meaning of multisource feedback: Expanding the nomological network surrounding source and dimension factors. Manuscript.
- Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self, and peer ratings. *Journal of Applied Psychology*, 63, 579–588.
- Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterization model misspecification. *Psychological Methods*, 3, 424–453.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling*, 6, 1–55.
- James, L. R. (1996, April). Quantitative issues in personnel selection research and practice. *Symposium presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego, CA*.
- Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, 127, 376–407.
- Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait–multimethod analyses of ratings. *Psychological Bulletin*, 75, 34–49.
- Kenny, D. A., & Berman, J. S. (1980). Statistical approaches to the correction of correlational bias. *Psychological Bulletin*, 88, 288–295.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165–172.
- King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional forced-choice performance evaluation scale. *Journal of Applied Psychology*, 65, 507–516.
- Kingstrom, P. O., & Bass, A. R. (1980). A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. *Personnel Psychology*, 65, 263–289.
- Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59, 445–451.
- Lance, C. E., Baranik, L. E., Lau, A. R., & Scharlau, E. A. (in press). If it's not trait it must be method: (Mis)application of the multitrait–multimethod design in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organizational and social sciences*. Mahwah, NJ: Erlbaum.
- Lance, C. E., Baxter, D., & Mahan, R. P. (2006). Multi-source performance measurement: A reconceptualization. In W. Bennett, C. E. Lance, & D. J. Woehr (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 49–76). Mahwah, NJ: Erlbaum.
- Lance, C. E., Noble, C. L., & Scullen, S. E. (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, 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. (1986). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal of Applied Psychology*, 71, 679–685.
- 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.
- Lance, C. E., Woehr, D. J., & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, 10, 449–462.
- Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, 87, 72–107.
- Le, H., Oh, I., Shaffer, J., & Schmidt, F. (2007). Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives*, 21(3), 6–15.
- LeBreton, J. M., Burgess, J. R. D., Kaiser, R. B., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, 6, 80–128.
- Lehmann, D. R. (1988). An alternative procedure for assessing convergent and discriminant validity. *Applied Psychological Measurement*, 12, 411–423.
- Lombardo, M. M., & McCauley, C. D. (1994). *BENCHMARKS®: A manual and trainer's guide*. Greensboro, NC: Center for Creative Leadership.
- London, M., & Smither, J. W. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for future research. *Personnel Psychology*, 48, 803–839.
- London, M., & Tornow, W. W. (1998). *Maximizing the value of 360-degree feedback*. Greensboro, NC: Center for Creative Leadership.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.
- Marsh, H. W. (1989). Confirmatory factor analysis of multitrait–multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335–361.
- Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod matrices: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107–117.

- McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, *90*, 215–238.
- McCauley, C., Lombardo, M., & Usher, C. (1989). Diagnosing management development needs: An instrument based on how managers develop. *Journal of Management*, *15*, 389–403.
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.
- 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., Cleveland, J. N., & Mohler, C. J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In D. W. Bracken, C. W. Timmreck, & A. H. Church (Eds.), *The handbook of multisource feedback* (pp. 130–148). San Francisco: Jossey-Bass.
- Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.
- Murphy, K. R., & DeShon, R. (2000). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913–924.
- Nathan, B. R., & Tippins, N. (1990). The consequences of halo 'error' in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology*, *75*, 290–296.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*, (3rd ed.) New York: McGraw-Hill.
- Ostroff, C., & Ford, J. K. (1989). Assessing training needs: Critical levels of analysis. In I. L. Goldstein (Ed.), *Training and development in organizations* (pp. 25–62). San Francisco: Jossey-Bass.
- Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., & Lee, J. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903.
- Pritchard, R. D. (1992). Organizational productivity. In M. D. Dunnette & L. M. Hough (Eds.), 2nd ed. *Handbook of industrial and organizational psychology*, Vol. 3. (pp. 443–472). Palo Alto, CA: Consulting Psychologists.
- Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.
- Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901–912.
- Schmitt, N. (1978). Path analysis of multitrait–multimethod matrices. *Applied Psychological Measurement*, *2*, 157–173.
- Schneider, B., & Schmitt, N. (1992). *Staffing organizations*, (2nd ed.). Prospect Heights, IL: Waveland.
- Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*, 956–970.
- Smither, J. W., London, M., & Reilly, R. R. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. *Personnel Psychology*, *58*, 33–66.
- Swann, W. B. (1984). Quest for accuracy in person perception: A matter of pragmatics. *Psychological Review*, *91*, 457–477.
- Tornow, W. W. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management*, *32*, 221–229.
- Vance, R. J., MacCallum, R. C., Covert, M. D., & Hedge, J. W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. *Journal of Applied Psychology*, *73*, 74–80.
- Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.
- Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2000). 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.
- Wherry, R. J., & Bartlett, C. J. (1982). The control of bias in ratings: A theory of rating. *Personnel Psychology*, *35*, 521–551.
- Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement*, *9*, 1–26.
- Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189–205.
- Woehr, D. J., Sheehan, M. K., & Bennett, W. (2005). Assessing measurement equivalence across ratings sources: A multitrait–multirater approach. *Journal of Applied Psychology*, *90*, 592–600.
- Wothke, W. (1995). Covariance components analysis of the multitrait–multimethod matrix. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp. 125–144). Hillsdale, NJ: Erlbaum.
- Zedeck, S., & Baker, H. T. (1972). Nursing performance as measured by behavioral expectation sales: A multitrait–multirater analysis. *Organizational Behavior and Human Performance*, *7*, 457–466.
- Zedeck, S., Imparto, N., Krausz, M., & Oleno, T. (1974). Development of behaviorally anchored rating scales as a function of organizational level. *Journal of Applied Psychology*, *59*, 249–252.