

TRAIT, RATER AND LEVEL EFFECTS IN 360-DEGREE PERFORMANCE RATINGS

MICHAEL K. MOUNT, TIMOTHY A. JUDGE, STEVEN E. SCULLEN
University of Iowa

MARCIA R. SYTSMA, SARAH A. HEZLETT
Personnel Decisions International, Inc.

Method and trait effects in multitrait-multirater (MTMR) data were examined in a sample of 2,350 managers who participated in a developmental feedback program. Managers rated their own performance and were also rated by two subordinates, two peers, and two bosses. The primary purpose of the study was to determine whether method effects are associated with the level of the rater (boss, peer, subordinate, self) or with each individual rater, or both. Previous research which has tacitly assumed that method effects are associated with the level of the rater has included only one rater from each level; consequently, method effects due to the rater's level may have been confounded with those due to the individual rater. Based on confirmatory factor analysis, the present results revealed that of the five models tested, the best fit was the 10-factor model which hypothesized 7 method factors (one for each individual rater) and 3 trait factors. These results suggest that method variance in MTMR data is more strongly associated with individual raters than with the rater's level. Implications for research and practice pertaining to multirater feedback programs are discussed.

Multirater or 360-degree feedback systems are characterized by the evaluation of an individual's performance by multiple raters from multiple levels. Although procedures vary, typically the individual is rated by others who interact frequently with the individual, who are knowledgeable about the individual's performance, and whose opinions are valued by the individual. The most common procedure is to include peers, subordinates, and bosses (in addition to self-ratings), but raters outside the organization, such as customers or suppliers, may also be included.

Multirater feedback programs differ from traditional appraisal programs in several ways. Aside from the use of multiple raters, multirater systems are used most frequently to enhance personal development and growth, rather than to help with salary administration, promotions, or

Portions of this paper were presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, Inc., St. Louis, April, 1997.

Correspondence and requests for reprints should be addressed to Michael K. Mount, Department of Management & Organizations, College of Business Administration, University of Iowa, Iowa City, IA, 52242-1000.

other administrative decisions. Further, ratings provided in multirater systems are made anonymously (with the exception of the immediate supervisor) and are not accompanied by face to face discussion. Such conditions are believed to increase the likelihood that raters will provide ratings that will be more honest and, therefore, more beneficial to the ratee.

Multirater feedback systems are believed to have a number of advantages over traditional appraisal systems (Hazucha, Hezlett, & Schneider, 1993; London & Beatty, 1993; London & Smither, 1995; Tornow, 1993). One is that because job performance is multidimensional, raters other than the immediate supervisor may be better suited to evaluating certain aspects of performance. Another is that even if raters have the same opportunity to observe performance, they may perceive and evaluate it differently. Generally speaking, multirater feedback systems are assumed to provide job relevant information to ratees that would otherwise not be available.

Most previous research in this area has examined the psychometric characteristics of ratings provided by raters from different levels (e.g., self, peer, subordinate, boss). Comparisons have focused on the degree of halo (e.g., Cooper, 1981; Lance & Woehr, 1986; Murphy & Anhalt, 1992) and interrater agreement between and—to a lesser extent—within levels (e.g., Harris & Schaubroeck, 1988; Viswesvaran, Ones, & Schmidt, 1996). One popular way to study the influence of traits and methods on performance ratings is the multitrait-multimethod (MTMM) or multitrait-multirater matrix (MTMR). The typical MTMR study examines ratees who have been rated by a single rater from each of several levels (bosses, peers, subordinates, self). Method effects reported in those studies have been tacitly assumed to emanate from differences in the raters' levels (i.e., bosses rate differently than peers, who rate differently than subordinates, etc.; e.g., Conway, 1996). However, another explanation for these method effects, as yet unexplored, also exists. They may simply reflect the fact that each rater is a different individual; that is, it is the idiosyncratic rating tendencies of individual raters rather than differences in the raters' levels that produce the observed method effects. When there is only one rater per level, as has been the case in most previous studies, the method effects (if any) that are associated with the rater's level are confounded with those due to the rater. Although this problem has been recognized by other researchers (e.g., Klimoski & London, 1974), to our knowledge it has not been empirically investigated. Therefore, the major purpose of this study is to examine the extent to which method effects in multirater data are associated with the level of the rater (self, peer, self, boss) or with each individual rater, or both.

Previous Research Findings

Correlations between ratings from different levels show that peers and bosses generally exhibit greater agreement with each other than with self-ratings. Harris and Schaubroeck (1988) found corrected correlations of .64 between peers and bosses, correlations of .27 for self- and boss ratings and .31 for self- and peer ratings. However, little research has examined correlations between subordinates' and others' ratings. The evidence available indicates that subordinates' ratings correlate more highly with bosses' and peers' than they do with self-ratings (e.g., Atwater & Yammarino, 1992; Furnham & Stringfield, 1994; McEvoy & Beatty, 1989; Mount, 1984; Schmitt, Noe, & Gottschalk, 1986; Wohlers, Hall, & London, 1993).

Correlations between ratings from the same level indicate that interrater agreement between bosses is somewhat higher than between peers. Viswesvaran, Ones, and Schmidt (1996) conducted a meta-analysis of published studies and reported interrater reliabilities for bosses of .52 for overall performance and .45 to .63 for dimensions of performance. These are very similar to values reported by Rothstein (1990) of .48 for duty ratings and .52 for ability ratings based on bosses' ratings of approximately 10,000 first-line supervisors. For peer ratings Viswesvaran et al. reported interrater reliabilities of .42 for overall performance and .34-.71 for dimensions of performance. Evidence is sparse regarding interrater reliabilities for subordinate ratings. Tsui and Ohlott (1988) found interrater reliability to be .26 and Tsui (1983) found the median reliability across dimensions to be .16. Mount (1984) reported intraclass correlations for subordinate ratings of different dimensions of managers' performance ranging from .15 to .28. The finding that the magnitude of within-source agreement does not differ much. The magnitude of between-source agreement suggests that individual raters are an important factor in explaining rating variance.

Studies that have examined performance rating data using multi-trait-multimethod matrices (MTMM) or multitrait-multirater (MTMR) matrices usually focus on the proportion of variance in performance ratings that is attributable to traits and that which is attributable to methods or raters. Trait variance is evidenced when the correlation between different methods (raters) assessing the same trait is high (convergent validity). Evidence of method variance or halo is present when the correlation between ratings of different traits made by the same method or rater is high. In most MTMM contexts, it is desirable to have a high proportion of trait variance and a low proportion of method variance. The predominance of trait variance over method variance suggests that

the different methods are measuring the same construct. However, results of recent analysis of 20 MTMR matrices from both published and unpublished studies revealed that the opposite relationship is generally observed. That is, the proportion of method variance is high relative to the proportion of trait variance (Conway, 1996).

High method variance indicates that ratings are strongly associated with the method (i.e., rater or rating level) from which the ratings were generated. There are at least two potential causes of high method variance in MTMR studies. One is halo, which occurs when the rater's ratings are heavily influenced by an overall evaluation of the ratee. Over 50 years of research in industrial and organizational psychology indicates that halo is one of the largest sources of measurement error in performance ratings. Another potential cause of strong method effects is that the construct(s) measured by boss, peer, subordinate, and self-ratings are not identical. That is, raters from different levels observe different aspects of performance and may also use different standards when judging performance.

Models to be Tested

Based on the literature reviewed above we tested five models that hypothesize different factor structures of rating methods (raters) and traits (managerial skills) that account for variance in performance ratings. Our data consisted of a set of seven ratings completed for each manager ratee: self-ratings and ratings made by two peers, two subordinates, and two bosses. Each model hypothesizes differing configurations of trait (managerial skills) and method factors. With respect to method factors, we distinguish between those associated with the level of the rater and those associated with the individual rater. The first model, the 3-factor model, is a trait-only model and hypothesizes that covariation in performance ratings is associated only with traits of the manager ratee and not with the rating method. The next two models hypothesize seven factors. The first hypothesizes that there are seven method factors (two bosses, two peers, two subordinates, and self) and no trait factors. This model posits that covariation in ratings is associated with individual raters and is not associated with the traits of the ratees. The second 7-factor model hypothesizes that there are four method factors (one for each level) and three trait factors. This model hypothesizes that covariation in ratings can be explained by the traits of the ratees and the levels of the raters. The 9-factor model hypothesizes that there are six method factors (one for each rater, except that bosses are combined into a single factor) and three trait factors. Hypothesizing a single factor for bosses is plausible for several reasons. Bosses receive training on and

have experience with observing, documenting, and rating performance, whereas other raters may not. They also may have held the subordinate manager's job in the past and, therefore, share a common frame of reference regarding the responsibilities of the job. Further, managers are held accountable for appraising the performance of others and, therefore, are more uniformly motivated in the rating process. The results of the Viswesvaran et al. (1996) meta-analysis, which showed higher interrater reliabilities for bosses than peers, suggests that this model is plausible. The 10-factor model hypothesizes that there are seven method factors, one for each individual rating perspective—two peers, two subordinates, two bosses, and self—and three trait factors. With respect to method effects, this model posits that covariation in ratings is associated with individual raters rather than with the four rating levels. This model differs from the first 7-factor model in that it hypothesizes three trait factors in addition to the seven method factors. The nature of the three hypothesized trait factors is discussed below.

Method

Participants

The population consisted of 2,350 managers who completed the Management Skills Profile (MSP) developed by Personnel Decisions International, Inc. (Sevy, Olson, McGuire, Frazier, & Paajanen, 1985). Managers' participation in the program was voluntary and results were used for personal and professional development. A set of seven ratings was available for each manager ratee: self-ratings and ratings made by two bosses, two peers, and two subordinates.

Managers represented essentially all functional areas and levels of management in several industry groups (manufacturing, banking, government, sales, health care, and non-profit). Most were White (87%), male (74%), and college graduates (76%). The mean age was 42 years. The two bosses, two peers, and two subordinates completed the same instrument as the one used for self-evaluations (with only minor variations in the demographic characteristics section).

Instrument

The MSP consists of 116 items depicting 18 dimensions of managerial behavior. Raters indicate how well the item describes observed behaviors of the ratee using a 5-point Likert scale ranging from 1 (*not at all*) to 5 (*to a very great extent*). There is also a "Does not apply" option.

Items are grouped into 18 performance dimensions, with 4 to 10 items per dimension.

Previous research investigating ratings of managers' performance has been hindered because researchers have not used a classification scheme to organize the numerous dimensions of managers' performance. Consequently, we used the framework of management performance dimension proposed by Mann (1965). This framework has been recommended by other researchers (e.g., Yukl & Van Fleet, 1990) as a useful scheme for classifying dimensions of managers' performance. It consists of three competence categories—administrative (e.g., planning, organizing, assigning to tasks), human relations (working with and through people to accomplish objectives), and technical competence (knowledge of relevant methods and techniques). In the present study, five trained raters assigned the 18 performance dimensions from the MSP to one of the three categories in Mann's taxonomy. A dimension was retained if at least four of the five (80%) raters agreed on the category assignment. Results indicated that for 12 of the 18 dimensions there was 100% agreement on the category assignment; for four of the dimensions there was 80% agreement; and, for two of the dimensions, Coaching and Leadership, there was relatively little agreement and, consequently, both were excluded from further analysis. The three categories and corresponding MSP skill areas are: *Administrative*: Planning, Organizing, Personal Organization and Time Management, Informing, Delegation; *Human Relations*: Conflict Management, Listening, Motivating Others, Human Relations, Personal Adaptability; *Technical Skills and Motivation*: Problem Analysis, Oral Communication, Written Communication, Personal Motivation, Financial and Quantitative, Occupational and Technical Knowledge. The dependent measures in the study were scores obtained for each of the seven raters on each of the three skill areas (Administrative, Human Relations, and Technical Skills) by averaging the scores on the relevant MSP scales.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA), conducted in the present study using LISREL 8 (Jöreskog & Sörbom, 1993), presents two distinct advantages. First, in principle CFA is ideally suited for investigating multi-trait-multimethod matrices (Lance, Teachout, & Donnelly, 1992; Widaman, 1985). The CFA model assumes that each variable contains method variance, trait variance, and unique variance (Conway, 1996). This allows determination of the degree to which methods (raters or levels in this study) and traits (skills in this study) account for covariation among the measures. Second, CFA allows comparison of alternative factor

structures that might explain the data (Jöreskog, 1993). Traditionally, decisions among alternative models are guided by nested comparisons based on the chi-square (χ^2) statistic or a χ^2 that takes parsimony as well as fit into account, such as Akaike's Information Criterion (AIC; Hu & Bentler, 1995). However, due to concerns over significance testing in general (Schmidt, 1996), and the χ^2 statistic in particular (Bollen, 1989), some researchers are using other methods to choose among alternative models. One rule is to reject a model if the standardized fit statistic (e.g., NFI, CFI, RFI—see below) is .01 or more lower than another model (Widaman, 1985).

An important consideration in confirmatory factor analysis is the sample size, because the number of estimated parameters relative to sample size is an important determinant of convergence, standard errors, and model fit (Idaszak, Bottom, & Drasgow, 1988). Although strict guidelines for minimum sample sizes do not exist, Bentler (1985) suggested that a sample size to parameter ratio of 5 or more is sufficient to achieve reliable estimates in maximum likelihood estimation. Because the most complex CFA model in the present study produced a sample size to estimated parameter ratio of 27:1, the sample size was considered adequate for the analyses.

Fit statistics are the central means through which alternative factor structures are compared. There are numerous statistics that can be used to describe a model's fit to the data. The most widely used measure of fit is χ^2 . Statistically significant χ^2 statistics suggest that the model does not adequately fit the data. However, it is widely recognized that χ^2 depends on the sample size and therefore even excellent fitting models will produce a significant χ^2 when the sample size is large (Hu & Bentler, 1995). Other popular fit statistics traditionally reported in the LISREL program include the root-mean-square residual (RMR) and the goodness-of-fit index (GFI). Although such rules are inherently subjective, values of at most .10 for RMR and at least .90 for GFI are thought to indicate acceptable fits (Medsker, Williams, & Holahan, 1994).

A problem with RMR and GFI is that they, like χ^2 , may depend on the sample size. Accordingly, researchers have suggested alternative fit statistics that depend less on the sample size and are thought to provide better information on model fit (Marsh, Balla, & McDonald, 1988). Four of these fit statistics were used in this study. These are the normed fit index and non-normed fit indexes (NFI and NNFI; Bentler & Bonnett, 1980), the comparative fit index (CFI; Bentler, 1990), and the relative fit index (RFI; Bollen, 1989). As with GFI, levels above .90 for these statistics imply adequate fit. The fit statistics reported include the four basic types of statistics recently reviewed by Hu and Bentler (1995).

Results

A complete data record for a manager consisted of 812 ratings (7 raters \times 116 MSP items). The length of these records increased the likelihood that there would be missing values for managers. If listwise deletion had been used, a manager would have been eliminated if even one of the 812 ratings was missing. Although fewer than 5% of the data points were missing for any single item, the cumulative effect over the 116 items was that a large number of managers would be excluded from the study. Therefore, we deemed it necessary to replace missing values using a mean substitution procedure. Missing values were replaced with the appropriate mean computed for each of the 18 MSP dimensions for each of the seven rater groups (Boss #1, Boss #2, Peer #1, and so on) across the 2,350 managers. In order to assess the effect of the substitution procedure, we compared a 21×21 correlation matrix (seven raters times three MSP factors, described earlier) with a similar matrix obtained using only those managers with no missing data ($n = 429$). The mean differences between corresponding correlations in the two matrices was .029. In light of such small differences, we believed that the use of the mean substitution procedure did not alter any substantive conclusions in the study.

Table 1 provides a correlation matrix of the three performance rating dimensions, measured from each of the seven raters. The average different skill–different rater correlation is .18. The average same skill–different rater correlation is .28. The average different–skill same-rater correlation is .75. These results suggest modest trait or skill convergence, but strong method effects.

Sample covariances served as input into the LISREL program. Maximum likelihood was chosen as the method of estimation. The CFA models with level or rater factors were specified such that the levels or raters were allowed to be correlated among themselves. Similarly, models estimated with trait factors were specified in such a way that the trait factor intercorrelations were freely estimated. For the models that contained both level/rater and trait factors, the intercorrelations between level/rater factors and trait factors were constrained to zero. In order to test the assumption that there were no rater by trait interactions (no rater–trait intercorrelations), models were estimated that allowed rater and trait factors to be intercorrelated. Although these models did yield lower χ^2 statistics than models constraining these rater–trait intercorrelations to zero, their standardized fit statistics (e.g., GFI, NFI, NNFI, CFI, RFI) were not superior to the statistics for models that constrained

TABLE 1
Intercorrelations of Performance Ratings

Perspective and skill	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1. Boss #1-human relations	-																				
2. Boss #1-technical	60	-																			
3. Boss #1-administrative	73	76	-																		
4. Boss #2-human relations	43	22	30	-																	
5. Boss #2-technical	22	42	30	60	-																
6. Boss #2-administrative	28	30	39	73	75	-															
7. Subordinate #1-human relations	26	09	13	23	08	12	-														
8. Subordinate #1-technical	17	20	17	15	21	17	73	-													
9. Subordinate #1-administrative	20	13	21	17	12	20	78	80	-												
10. Subordinate #2-human relations	26	10	13	26	08	14	34	24	25	-											
11. Subordinate #2-technical	16	22	17	17	23	19	22	30	24	73	-										
12. Subordinate #2-administrative	16	11	18	18	12	21	25	24	31	77	79	-									
13. Peer #1-human relations	32	15	21	30	15	22	27	19	20	26	17	21	-								
14. Peer #1-technical	19	32	24	16	31	24	12	22	15	12	23	17	64	-							
15. Peer #1-administrative	20	19	27	20	19	29	16	18	23	15	17	23	76	75	-						
16. Peer #2-human relations	33	15	20	32	14	20	26	16	17	25	15	16	31	16	19	-					
17. Peer #2-technical	21	33	25	19	34	26	12	20	14	15	25	17	16	30	21	63	-				
18. Peer #2-administrative	22	20	27	24	21	29	17	15	20	17	18	21	22	21	26	75	75	-			
19. Self-human relations	19	05	10	15	01	04	17	06	09	19	07	12	19	06	09	17	04	09	-		
20. Self-technical	04	23	14	02	25	12	02	11	04	04	16	08	06	20	10	03	20	10	62	-	
21. Self-administrative	06	07	19	04	06	16	06	05	15	06	07	18	11	09	18	07	08	15	70	68	-

Note: Decimals are omitted; $N = 782$.

TABLE 2
Fit Statistics for Alternative Models

Model	χ^2	df	RMR	GFI	NFI	NNFI	CFI	RFI
3-factor Trait-only model	27,660.04	186	.16	.50	.25	.16	.25	.16
7-factor Seven methods—one for each individual rater and no trait factors	6,223.08	168	.05	.68	.81	.77	.82	.76
7-factor Four methods—one for each level: boss, peer, subordinate, and self; and three skills factors	13,712.25	159	.14	.67	.63	.51	.63	.51
7-factor Six methods—one for each rater, except bosses combined into one; and three skills factors	1,113.09	150	.13	.96	.97	.96	.97	.96
10-factor Seven methods—one for each individual rater and three skills factors	333.95	144	.02	.99	.99	.99	.99	.99

Note: df = Degrees of freedom; RMR = Root-mean-square residual; GFI = Goodness-of-fit index; NFI = Normed fit index; NNFI = Non-normed fit index; CFI = Comparative fit index; RFI = Relative fit index; For each model, increase in χ^2 over 10-factor model is significant at the .01 level.

rater-trait correlations to zero. Furthermore, most of the rater-trait intercorrelations were not significant. Accordingly, all subsequent models were estimated allowing level/rater and trait factors to be correlated among themselves (*intracorrelated*), but not correlated with each other (*intercorrelated*).

Table 2 presents fit statistics which can be used to investigate how well the various factor structures fit the data. The 3-factor and both 7-factor models can be rejected immediately, as each provides a poor fit to the data. These results indicate that both method and trait factors are necessary to account for the data. Further, with the possible exception of bosses (see below), methods that are associated only with level (boss vs. peer vs. subordinate vs. self) do not account for the data as well as individual rater sources. Although the 3- and 7-factor models fit the data poorly, both the 9-factor and 10-factor models fit the data well. Although the 10-factor model fits the data best, because both models surpass most rules of thumb typically used in evaluating the adequacy of a model, further comparisons between the two models are warranted.

Although both the 9- and 10-factor models are adequate on an absolute level, several additional tests reveal that the 10-factor model fits the data relatively better than the 9-factor model. First, because the 9-factor model is nested within the 10-factor model, the difference in χ^2 between these models is itself distributed as χ^2 . Thus, subtracting the χ^2 statistics indicates if the fit of the more restricted model (in this case, the 9-factor model) is significantly worse than the less restricted model (in this case, the 10-factor model). In fact, the difference in χ^2 is significant ($\Delta\chi^2 = 779.41$ with 6 *df*, $p < .01$), indicating that the 10-factor model fits the data significantly better. Second, Jöreskog (1993) suggests that, when selecting one of several a priori specified models, the model with the lowest AIC statistic should be favored. Given the substantially lower AIC value for the 10-factor ($AIC = 472.41$) than the 9-factor ($AIC = 1,275.09$) model, the 10-factor model is preferred under this criterion as well. Third, the standardized fit statistics of the 9- and 10-factor model differed by .03 (see Table 2), also attesting to the superiority of the 10-factor model (Widaman, 1985). Finally, analysis of standardized residuals (Hu & Bentler, 1995; Jöreskog, 1993) revealed clear differences in the models. Inspection of the stemleaf plots indicated that the standardized residuals for both models approximated a normal distribution. However, the Q-plot of standardized residuals revealed that the 10-factor model closely followed a 45° line with no apparent outliers while the standardized residuals from the 9-factor model more closely approximated a 30° line with several outliers. Further, the average absolute standard residual for the 9-factor model was 18% higher than the average residual for the 10-factor model (.93 vs. .79, respectively). In summary, although both models fit the data, these comparisons indicate that the 10-factor model provides the best fit to the data.

Given the relative superiority of the 10-factor model, Table 3 provides the factor loadings of that model. As the table indicates, the method loadings are very strong. In fact, the average rater factor loading is .84. Although all of the trait loadings are nonzero (confidence intervals around the loadings excluded zero), they are considerably weaker in magnitude than the rater level loadings. The average trait loading is .30. That the method loadings are much higher than the trait loadings is consistent with the multitrait-multimethod results presented earlier.¹

¹ In an analysis of multitrait-multimethod matrices of performance appraisal data, Conway (1996) found that correlated uniqueness models (CU; Kenny & Kashy, 1992; Marsh, 1989) often outperformed CFA models. Like CFA, CU models allow investigation of trait and method effects. However, rather than estimating method factors, CU models allow variables measured with common methods to have correlated errors. Because CU models do not allow as fine-grained tests as CFA, though, we used CFA as our primary method of

Discussion

More than 75 years ago Thorndike (1920) observed that when supervisors rated their subordinates, the correlations among performance dimensions were "higher than reality" (p. 25) and "too high and too even" (p. 27). Research conducted since that time has further documented the ubiquitous phenomenon of method effects in performance ratings, and has shown that such effects represent one of the largest sources of error in performance ratings (Cooper, 1981). However, an unanswered question in this body of research is whether such effects are associated with raters' levels or with individual raters. Because previous MTMR studies have included only one rater for each rating perspective, they have not been able to separate these two potential sources of method effects. The procedures in this study, whereby two raters from each level are included, coupled with the use of CFA, allowed us to address this issue directly.

We tested five models that hypothesized different factor structures of rating methods and traits that account for variance in performance ratings. The finding of central interest is that method variance in performance ratings is associated more strongly with individual raters than with the level of the ratings. The possible exception to this is that boss ratings may constitute a separate method factor that can be distinguished from all other levels of ratings (as indicated by the relatively good fit of the 9-factor model). As suggested earlier, it is plausible that bosses might constitute a common factor because they are more likely to share a common frame of reference by virtue of their prior training and experience. (This finding has practical implications for the way feedback is delivered in multirater feedback systems and is discussed below.) However, as well as the 9-factor model fit the data, it did not fit as well as the 10-factor model, which showed that covariation in ratings was associated with each of the seven raters and the three traits. Research reviewed earlier indicated that method effects in MTMM data are commonplace

analysis. This is consistent with Becker and Cote's (1994) recommendation, "... if boundary conditions are not a problem, CFA is generally preferable to alternative methods" (p. 635). Nevertheless, because it has been suggested as an alternative to CFA models, the CU model was estimated. (Due to space limitations, all details on model estimation are not reported here but are available from the authors on request.) Analyzing the data using a correlated uniqueness model provided similar conclusions to the CFA results. Specifically, the CU model that closely approximated the 10-factor CFA model (i.e., allowing both trait and source correlated errors) fit the data well $\chi^2 = 311.05$ with 138 *df*; RMR = .07; GFI = .97; NFI = .98; NNFI = .98, CFI = .99; RFI = .97). Furthermore, like the CFA models, test of alternative CU models suggested the model allowing both source and trait effects provided the best fit to the data. Thus, the CU results, like the CFA results, suggest that both source and trait effects must be included to properly account for the covariance structure.

TABLE 3
Factor Loadings of Ten-Factor Model

Method and skill	Method						Skill			
	Boss #1	Boss #2	Sub. #1	Sub. #2	Peer #1	Peer #2	Self	Hum. rel.	Tech.	Admin.
Boss #1-human relations	.77							.43	.38	
Boss #1-technical	.80									.19
Boss #1-administrative	.94									
Boss #2-human relations		.78						.42	.41	
Boss #2-technical		.80								.19
Boss #2-administrative		.93						.29	.20	
Subordinate #1-human relations			.84							.24
Subordinate #1-technical			.88					.29	.23	
Subordinate #1-administrative			.91		.84					.25
Subordinate #2-human relations					.88					
Subordinate #2-technical					.91			.35	.32	
Subordinate #2-administrative						.81				.20
Peer #1-human relations					.81					
Peer #1-technical					.81					
Peer #1-administrative					.93			.36	.33	
Peer #2-human relations						.80				.15
Peer #2-technical						.81				
Peer #2-administrative						.93				
Self-human relations							.81	.36	.38	.39
Self-technical							.80			
Self-administrative							.85			

Note: Sub. = Subordinate factor; confidence intervals around all factor loadings exclude zero. Correlation between trait factors were as follows: Human Relations-Technical = $-.14$; Human Relations-Administrative = $.09$; and Technical-Administrative = $-.01$. Correlation between same-level rater method factors were: Bosses = $.40$; Peers = $.28$; Subordinates = $.31$. Average correlations between different-level rater method factors were: Boss-Peer = $.29$; Boss-Subordinate = $.19$; Boss-Self = $.12$; Peer-Subordinate = $.21$; Peer-Self = $.12$; Subordinate-Self = $.10$.

in the literature and that their magnitude is typically larger than is considered optimal (Conway, 1996). Our results are consistent with these findings but cast doubt on the inference that is often drawn that method effects are primarily attributable to the level of the rater (boss, peer, subordinates, or self). Our results show that method effects due to individual raters are distinguishable from those attributable to the rater's level, and that the former are more important than the latter.

The results also confirmed the presence of trait effects in the performance ratings. Although these findings are of secondary interest in the study, they deserve further comment. Confirmation of the trait effects was evidenced by the fact that the fit of the 7-factor model which posited only method effects was dramatically improved by the inclusion of the three trait factors (i.e., the 10-factor model which posited separate rating methods for each rater and three trait factors). This indicates that in addition to variance accounted for by individual raters and rating level, variance is also accounted for by managers' traits. As reported earlier, the magnitude of the trait factors was modest compared to the method effects. Further, it should be noted that the rejection of the 3-factor model also indicates that covariation in ratings cannot be accounted for solely by the traits (with no method factors). Nonetheless, the presence of trait factors indicates that rating variance associated with the three traits is distinguishable from rating variance associated with the individual rater and the level of the ratings.

Anonymous reviewers of this paper raised the possibility that the use of Mann's (1965) three dimensional taxonomy constrained our ability to derive trait effects. Although at least four of the five raters agreed on the assignment of the MSP dimensions to Mann's three dimensions, this provides little insight into the underlying structure of the MSP. Had a different taxonomy been used, or had the MSP scales been assigned differently to the dimensions, then the trait effects may have accounted for substantially more variance in the ratings.

In order to address this issue, we conducted a principal components analysis (with varimax rotation) of the 16 MSP dimensions. It revealed a 3-factor structure which corresponded very closely to Mann's three factors. The scales that loaded most highly on each factor were identical to those the raters had assigned to the three factors. The three factors and the corresponding MSP dimensions and factor loadings were as follows: Factor 1-Human Relations: Human Relations (.91), Listening (.90), Personal Adaptability (.75), Motivation (.75), and Conflict Management (.69); Factor 2-Administrative: Personal Organization (.85), Delegation (.81), Planning (.78), Organizing (.73) Informing (.68); Factor 3-Technical Skills: Occupational and Technical Knowledge (.77), Financial and Quantitative (.73), Personal Motivation (.64), Problem Solving (.60),

Oral Communication (.49), Written Communication (.48). In addition, several of the dimensions had cross loadings greater than .40 on other MSP dimensions: Informing (.40) and Problem Solving (.49) loaded on the Human Relations factor; Oral Communication (.44), Written Communications (.44), Problem Solving (.45) and Motivation (.40) loaded on the Administrative factor; and Personal Adaptability (.41) loaded on the Technical factor. The percentage variance accounted for by three factors was 37.5, 36.1 and 26.4, respectively.

These results indicate that Mann's (1965) taxonomy provides an acceptable framework for examining trait affects associated with the ratings on the MSP scales. An important implication of these findings is that the lack of fit of the trait-only factor structure is not due to the inappropriateness of Mann's 3-factor taxonomy. One other implication of our findings is that researchers and practitioners may find Mann's model of managerial performance to be a useful taxonomy for investigating managers' performance.

The finding that method effects were more strongly associated with individual raters than with the rater's level raises several issues for future research. Although we know that there is unique variance associated with each rater's ratings, we cannot necessarily conclude that each is accounting for unique true score variance. Thus, we cannot draw conclusions about the relative accuracy of ratings provided by different raters or about the relative accuracy of ratings from different levels. Another issue is whether the present results would generalize to ratees in occupations other than management (where most of the raters are themselves managers). For example, when the raters are sales representatives it is possible that the rater's level may become more salient, while differences within levels are diminished. That is, if raters are external customers rather than bosses or peers there may be a more pronounced method effect associated with the level of the rater (e.g., customers) rather than for individual raters as we observed in the present study.

The present results have several practical implications. At the most basic level the results provide support for the implicit premise underlying most 360-degree systems. That is, because ratings from each rater, regardless of level, appear to capture unique rating variance, it is important to include multiple raters in the process rather than relying on the results of a single rater, such as the boss. Our results show that each rater's ratings are different enough from those of other raters to constitute a separate method. The implication of this for 360-degree feedback reports is that information should be displayed separately for each individual rater. (In order to protect the identity of raters, it is probably best not to indicate the level associated with the ratings.) Displaying information in this way allows the ratee to examine the pattern of each rater's

ratings across the skills to determine relative strengths and weaknesses. For example, a finding that each of the seven raters assigned their lowest ratings to Human Relations would provide important feedback to a ratee that this skill area is a relative weakness and should be the focus of developmental planning. Because most 360-degree feedback programs are used primarily for developmental purposes where the focus is on the identification of strengths and weaknesses, information about the pattern of individual raters' ratings across traits is of critical importance. Thus, for this important use of 360-degree feedback, our results indicate that it is best to consider each rater's ratings separately.

Another implication of our findings pertains to the widespread practice of aggregating ratings made by raters within the same level (e.g., averaging all peer ratings on a skill such as Human Relations). Our results show that in most cases this practice is inappropriate. As discussed earlier, ratings made by raters within the same level (e.g., two peers or two subordinates) are no more similar to each other than are ratings made by raters from different levels (e.g., a boss and a peer or a peer and a subordinate). Ultimately, the usefulness of ratings provided to ratees depends on their construct validity. For this reason, we believe it is inappropriate in most cases to aggregate ratings within (or across) rating levels.

Although our results suggest that ratings of managerial performance should not be aggregated, it is also true that disaggregated ratings are less reliable than aggregated ratings. In the present study, when ratings from all seven raters are averaged, the resulting reliabilities are more than 60% higher than for any single rater's ratings. Using the Spearman-Brown formula based on seven raters (two bosses, two peers, two subordinates, and self), the reliabilities for the Human Relations, Administrative, and Technical factors are .71, .69, and .71, respectively. Thus, our results suggest that although aggregating performance ratings reduces the construct validity of the ratings, it would at the same time increase the estimated reliability of the ratings.

How might one explain this paradox? A widely acknowledged axiom in personnel psychology is that reliability is a necessary, but not sufficient, condition for validity. The results of the present study clearly illustrate this. The average of the seven raters' ratings results in a measure that is substantially more reliable than a measure based on one or even two raters' ratings. However, the results also show that each rater's ratings are sufficiently different from each other to be considered a separate method. In most areas of research, increases in estimated reliability result in higher construct validity; in fact, reliability bounds estimated validity. However, one cannot therefore logically conclude that increasing reliability will always increase validity. An example of this point

is illustrated by a recent meta-analysis of interview validity (McDaniel, Whetzel, Schmidt, & Maurer, 1994). This study showed that panel interviews (multiple interviewers providing multiple ratings in one setting) resulted in lower validity in predicting job performance than individual interviews, even though McDaniel et al. noted that panel interviews are likely to be more reliable. Our results are consistent with those of McDaniel et al.—although reliability typically improves validity, this is not always the case. Thus, although it is true that averaging seven raters' ratings results in a more reliable measure, it is not true that the resulting measure is more construct valid.

One possible exception to this recommendation is our finding that ratings made by bosses may be similar enough to constitute a common method. This suggests that the practice of aggregating ratings made by two or more bosses may be appropriate. The resulting measure is more reliable than a single rating, and is also likely to be construct valid. For example, considering data in this study, the reliability of two boss ratings is approximately 40% higher than a single boss's ratings for each of the three skills: .61 compared to .44 for Human Relations; .57 compared to .40 for Administrative; .59 compared to .42 for Technical. There is also some evidence that there are meaningful differences between the mean level of boss ratings and that of other raters. Harris and Schaubroeck (1988) showed that ratings made by bosses were over half a standard deviation lower than those made by the ratees themselves. These results support the idea of treating boss ratings as a common method.

There is one additional caveat we would like to point out regarding aggregation practices in 360-feedback reports. Even in those cases where it is appropriate to compute means within the level of raters (as for bosses), it is important to provide a measure of the dispersion of the ratings. This might include reporting the actual ratings made by the raters on the scale in order to illustrate how different individual ratings might be, or it might include the standard deviation of raters' ratings. Clearly, inferences drawn about the meaning of an average rating of 3.0 with a standard deviation of 1.5 (on a 5-point scale) are quite different from those drawn from the same average with a standard deviation of zero. Therefore, we strongly recommend that in those limited situations where it is appropriate to average ratings, information regarding the dispersion of the ratings also be provided to facilitate interpretation of the feedback.

When interpreting the results of this research, several issues should be kept in mind. First, ratings in this study were made for developmental purposes only, which raises a question about the generalizability of our results to ratings provided in the context of administrative decisions. However, Kraiger and Ford (1985) investigated a similar question in

their meta-analytic study, and found that rater-ratee effects were not moderated by purpose of the ratings. In light of their findings the generalizability concerns expressed above may be minor. Another issue pertains to the conditions under which ratings were obtained in the study and their potential effects on the subsequent ratings. Managers' participation in the study was voluntary, indicating that they were actively seeking feedback. As in most multirater feedback systems, managers in this study selected the peers and subordinates who rated them. We do not know how these raters perceived the feedback-seeking behavior of target managers in this study, nor do we know how or if the performance of managers who participate voluntarily differs from the performance raters who do not. Both of these issues appear to be worthwhile topics for future research.

Despite these potential limitations, we believe there are several features of this study that enhance its contribution to the literature. The study used a large sample of ratees who held jobs from the same job family (management) and were rated using a common instrument and for a common purpose, thereby eliminating potential confounding for any of these reasons. Further, managers were rated by seven raters from four levels, two bosses, two peers, two subordinates, and self-ratings for each ratee. This unique sample along with the use of CFA, allowed us to uncouple method effects in performance ratings that are due to individual raters from those that due to the rating level. The major contribution of the paper is the finding that method effects in MTMR data are associated more strongly with individual raters than with the rater's level.

REFERENCES

- Atwater LE, Yammarino FJ. (1992). Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *PERSONNEL PSYCHOLOGY*, 45, 141-163.
- Becker TE, Cote JA. (1994). Additive and multiplicative method effects in applied psychological research: An empirical assessment of three models. *Journal of Management*, 20, 625-641.
- Bentler PM. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.
- Bentler PM. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Bentler PM, Bonnett DG. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Bollen KA. (1989). *Structural equations with latent variables*. New York: Wiley.
- Conway JM. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management*, 22, 139-162.
- Cooper WH. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Furnham A, Stringfield P. (1994). Congruence of self and subordinate ratings of managerial practices as a correlate of boss evaluation. *Journal of Occupational and Organizational Psychology*, 67, 57-67.

- Harris MM, Schaubroeck J. (1988). A meta-analysis of self-boss, self-peer, and peer-boss ratings. *PERSONNEL PSYCHOLOGY*, 41, 43-62.
- Hazucha JF, Hezlett SA, Schneider RL. (1993). The impact of 360-degree feedback on management skill development. *Human Resource Management*, 32, 325-352.
- Hu L, Bentler PM. (1995). Evaluating model fit. In Hoyle RH (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76-99). Thousand Oaks, CA: Sage.
- Idaszak JR., Bottom WP, Drasgow F. (1988). A test of the measurement equivalence of revised Job Diagnostic Survey: Past problems and current solutions. *Journal of Applied Psychology*, 73, 647-656.
- Jöreskog KG. (1993). Testing structural equation models. In Bollen KA, Bollen JS, Long S (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.
- Jöreskog KG, Sörbom D. (1993). *LISREL 8 user's reference guide*. Chicago: Scientific Software.
- Kenny DA, Kashy DA. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, 112, 165-172.
- Klimoski RJ, London M. (1974). Role of the rater in performance appraisal. *Journal of Applied Psychology*, 59, 445-451.
- Kraiger K, Ford JK. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, 70, 56-65.
- 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, 679-685.
- London M, Beatty RW. (1993). 360-degree feedback as a competitive advantage. *Human Resource Management*, 32, 352-373.
- London M, Smither JW. (1995). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance related outcomes? Theory-based applications and directions for research. *PERSONNEL PSYCHOLOGY*, 48, 803-839.
- Mann 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*. San Francisco: Chandler.
- Marsh HW. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, 13, 335-361.
- Marsh HW, Balla JR, McDonald RP. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, 103, 391-410.
- McDaniel MA, Whetzel DL, Schmidt FL, Maurer SD. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599-616.
- McEvoy GM, Beatty RW. (1989). Assessment centers and subordinate appraisals of managers: A seven-year examination of predictive validity. *PERSONNEL PSYCHOLOGY*, 42, 37-52.
- Medsker GJ, Williams LJ, Holahan PJ. (1994). *Journal of Management*, 20, 439-464.
- Mount MK. (1984). Psychometric properties of subordinate ratings of managerial performance. *PERSONNEL PSYCHOLOGY*, 37, 687-702.
- Murphy KR, Anhalt RL. (1992). Is halo error a property of the rater, ratees, or the specific behavior observed? *Journal of Applied Psychology*, 77, 494-500.
- Rothstein HR. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322-327.

- Schmidt FL. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115-129.
- Schmitt N, Noe RA, Gottschalk R. (1986). Using the lens model to magnify raters' consistency, matching and shared bias. *Academy of Management Journal*, 29, 130-139.
- Sevy BA, Olson RD, McGuire DP, Frazier ME, Paajanen G. (1985). *Managerial skills profile technical manual*. Minneapolis: Personnel Decisions, Inc.
- Thorndike EL. (1920). A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.
- Tornow WW. (1993). Perceptions or reality: Is multi-perspective measurement a means or an end? *Human Resource Management*, 32, 221-230.
- Tsui AS. (1983). *Qualities of judgmental ratings by four rater perspectives*. East Lansing, MI: National Center for Research on Teacher Learning. (ERIC Document Reproduction Service No. ED 237 913)
- Tsui AS, Ohlott P. (1988). Multiple assessment of managerial effectiveness: Interrater agreement and consensus in effectiveness models. *PERSONNEL PSYCHOLOGY*, 41, 779-803.
- Viswesvaran C, Ones DS, Schmidt FL. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81, 557-574.
- Widaman KF. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, 9, 1-26.
- Wohlers AJ, Hall MJ, London M. (1993). Subordinates rating managers: Organizational and demographic correlates of self/subordinate agreement. *Journal of Occupational and Organizational Psychology*, 66, 263-275.
- Yukl G, Van Fleet DD. (1990). Theory and research on leadership in organizations. In Dunnette MD, Hough LM (Eds.), *Handbook of industrial and organizational psychology*. Palo Alto, CA: Consulting Psychologists Press.

Copyright of Personnel Psychology is the property of Blackwell Publishing Limited. The copyright in an individual article may be maintained by the author in certain cases. 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.