Are Performance Appraisal Ratings From Different Rating Sources Comparable?

Jeffrey D. Facteau
Auburn University

S. Bartholomew Craig
Kaplan DeVries, Inc.

The purpose of this study was to test whether a multisource performance appraisal instrument exhibited measurement invariance across different groups of raters. Multiple-groups confirmatory factor analysis as well as item response theory (IRT) techniques were used to test for invariance of the rating instrument across self, peer, supervisor, and subordinate raters. The results of the confirmatory factor analysis indicated that the rating instrument was invariant across these rater groups. The IRT analysis yielded some evidence of differential item and test functioning, but it was limited to the effects of just 3 items and was trivial in magnitude. Taken together, the results suggest that the rating instrument could be regarded as invariant across the rater groups, thus supporting the practice of directly comparing their ratings. Implications for research and practice are discussed, as well as for understanding the meaning of between-source rating discrepancies.

Given that formal performance appraisal (PA) systems serve a variety of important functions within organizations (Cleveland, Murphy, & Williams, 1989) and that in most organizations ratings are obtained only from employees' supervisors (Bernardin & Villanova, 1986), it is not surprising that a great deal of research has focused on the psychometric quality of supervisory ratings. Unfortunately, a number of studies conducted over the past several decades indicate that supervisory ratings often are plagued by a host of potential problems, including halo, leniency, intentional manipulation, and race, gender, or age biases (see Cardy & Dobbins, 1994; Cascio, 1991; Landy & Farr, 1980, for reviews). Recognizing these limitations, researchers turned their attention to examining alternative rating sources such as peer, subordinate, and self ratings.

Perhaps one of the most consistent findings in the empirical literature on PA systems is that the ratings obtained from different sources generally do not converge. The intercorrelations among the ratings provided by different types of raters tend to be moderate at best. For example, in a meta-analysis, Harris and Schaubroeck (1988) reported mean self–supervisor, self–peer, and peer–supervisor rating correlations (corrected for unreliability) of .35, .36, and .62, respectively. Similarly, Mount (1984) reported mean supervisor–subordinate and subordinate–self rating correlations of .24 and .19, respectively. Finally, Conway and Huffcutt (1997) reported that the correlations (corrected for unreliability) among ratings made by self, peer, supervisor, and subordinate raters ranged from a high of .79 (supervisor–peer) to a low of .14 (subordinate–self).

The average ratings provided by different rating sources also tend to differ. Thornton (1980) reported that a “preponderance of studies show that individuals rate themselves higher than they are rated by comparison groups” (p. 265). Mount (1984) found that self ratings tend to be higher than supervisor ratings, which in turn tend to be higher than subordinate ratings. Harris and Schaubroeck (1988) found that self ratings averaged .70 SD higher than supervisor ratings and .28 SD higher than peer ratings. Peer ratings averaged .23 SD higher than supervisor ratings. Although consistent with Mount’s findings, none of these differences was significant, however. Harris and Schaubroeck attributed this to the large variance in effects across studies included in their meta-analysis.

A number of different explanations have emerged to account for why different types of raters (e.g., subordinates, supervisors) do not agree in their ratings. Campbell and Lee (1988) argued that different rater groups may have different conceptualizations of what constitutes effective performance in a particular job. Murphy and Cleveland (1995) suggested that raters differ in their opportunity to observe any given individuals’ work behavior and that these differences in perspective may account for disagreements among their ratings. Similarly, Lance, Teachout, and Donnelly (1992) noted that the modest correspondence among ratings from different types of raters may be due to the fact that they are exposed to only moderately overlapping sets of ratee behavior. This “ecological perspective” to rating source differences (Lance & Woehr, 1989) suggests that strong correspondence among ratings from different sources should not be expected. Finally, Campbell and Lee (1988) as well as Cardy and Dobbins (1994) described a number of processes that may lead to disagreements.

Jeffrey D. Facteau, Department of Psychology, Auburn University; S. Bartholomew Craig, Kaplan DeVries, Inc., Greensboro, North Carolina.

We thank Robert T. Ladd, Mike Rush, Keith Widaman, and Nambury Raju for their helpful advice and assistance during the course of this research. We also are grateful to Dianna Corrigan, Mike Moomaw, and Kevin Nolan for the instrumental role that they played in the project. Finally, we thank Carolyn Facteau and John Veres for their thoughtful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Jeffrey D. Facteau, Department of Psychology, Auburn University, 226 Thach Hall, Auburn, Alabama 36849-5214. Electronic mail may be sent to facteauj@mail.auburn.edu.
among the ratings provided by different types of raters. These processes include well-established attributional tendencies, such as the self-serving attributional bias and the actor–observer effect, as well as motivational and informational differences between rating sources such as self raters’ needs for self-enhancement (Farh & Dobbins, 1989b) and differences in social comparison information available to self raters and their supervisors (Farh & Dobbins, 1989a).

Although these are compelling explanations for why PA ratings from alternative sources differ, a more fundamental explanation that may at least partially account for the lack of convergence has received little attention in the literature. Specifically, it may be that performance rating instruments are not equivalent across different groups of raters. This is not to say that the items that raters are asked to complete when evaluating a target individual differ by rating source or that the performance constructs they are intended to measure differ. Instead, it may be that different rater groups have different conceptualizations about the dimensionality of job performance. Alternatively, even if different raters share a common conceptualization of job performance, the items contained on a rating instrument may relate differently to the underlying performance constructs across rater groups.

This issue is referred to in the psychometric literature as measurement invariance or measurement equivalence. To compare different groups with respect to their level (i.e., average rating) on a measurement instrument or to examine the relationships among scores on a measure across groups (e.g., rating sources), it must be assumed that the numeric values under consideration are on the same measurement scale. Measurements are on the same scale and, as a result, are comparable when the empirical relationships between the indicators (e.g., items) contained on an instrument and the constructs they measure are invariant across groups (Reise, Widaman, & Pugh, 1993). If an instrument designed to measure a latent trait is invariant across groups, then individuals with the same standing on the trait will have the same expected observed score. Measurement equivalence does not require that the distributional properties of obtained scores (e.g., means, variances, etc.) be equal across groups; it only requires that the empirical relationships between indicators and the latent constructs they are intended to measure be equivalent (cf. Drasgow, 1984, 1987; Drasgow & Kanfer, 1985; Ildaszak, Bottom, & Drasgow, 1988). If the assumption of measurement invariance does not hold, then mean differences between groups on an instrument are potentially artifactual (Reise et al., 1993). Furthermore, the failure to find convergence in the scores obtained from members of different groups (e.g., self vs. supervisor ratings) may reflect the fact that the measure is not invariant across the groups. Essentially, without measurement invariance, observed scores from different groups are on a different scale and, therefore, are not directly comparable (Drasgow & Kanfer, 1985).

Despite the importance of measurement invariance as a prerequisite for comparing scores across groups, it is most often assumed in multigroup research but rarely if ever tested directly (Byrne, Baron, & Campbell, 1993). With regard to the PA literature, few studies have explicitly tested for the invariance of a performance rating instrument across rating sources. Lance and Bennett (1997) reported several ways in which a measure of interpersonal proficiency was not invariant (i.e., noninvariant) in Air Force samples of supervisor, self, and peer raters. Using multiple-groups confirmatory factor analysis (CFA), they found that self, peer, and supervisor raters held different conceptualizations of the interpersonal proficiency construct in five of eight samples. In addition, they found that self raters’ judgments of their own performance were less variable than peers’ and supervisors’. Finally, they found that self ratings were more reliable, but this was attributed to restriction in the range of individual item ratings. In another study, Maurer, Raju, and Collins (1998) used multiple-groups CFA and item response theory (IRT) techniques to assess the measurement invariance of peer and subordinate ratings of managers’ teambuilding skills. The multiple-groups CFA and IRT analyses converged, showing that the team-building instrument was invariant across the two rating sources.

Lance and Bennett’s (1997) study suggests that performance ratings (of interpersonal proficiency) from different sources may reflect different underlying constructs, whereas Maurer et al.’s (1998) findings suggest that peer and subordinate ratings (of team building) can be treated as equivalent. Because the findings of these two studies conflict, additional research is needed that clarifies whether it is appropriate to treat measures of performance completed by different rater groups as equivalent. Furthermore, these studies are limited in two ways that also suggest a need for additional research. First, in neither study was invariance of the rating instrument tested across all four of the rating sources that have received considerable attention in the PA literature (i.e., supervisor, self, peer, and subordinate ratings). Lance and Bennett tested for invariance across self, supervisor and peer raters, whereas Maurer et al.’s study used only peer and subordinate raters. As a result, no research has been conducted to date that directly tests for measurement invariance across, for example, subordinate and supervisor ratings. Contemporary multisource PA systems often include ratings from all four sources. Thus, it is important to determine whether they are comparable. Second, in both studies, a rating instrument designed to measure only a single performance construct was examined for measurement invariance. In most organizational settings, however, performance rating instruments are designed to measure target individuals’ performance on a number of separate, but related, performance dimensions. As a result, little, if anything, is known about the measurement invariance of these multifaceted performance rating measures. It may be that ratings are invariant on certain dimensions but not others. Conway and Huffcutt (1997) observed higher between-source correlations for performance dimensions that dealt with interpersonal behavior. Mount (1984) observed a similar dimension effect and suggested that there may be something inherent in certain dimensions (e.g., limited opportunity to observe relevant behavior) that makes them difficult to rate. An alternative explanation might be that certain dimensions are interpreted differently by different groups of raters, and, as a result, ratings on those dimensions are not strongly correlated across rater groups. The purpose of this research, therefore, was to test whether a multidimensional performance rating instrument was invariant across supervisor, self, subordinate, and peer raters. By examining a multifaceted rating instrument, it was possible to examine the realistic possibility that measures of certain performance constructs on a rating instrument would be invariant across groups, whereas others would not be invariant.

Whether multifaceted performance rating instruments are invariant across self, peer, subordinate, and supervisor rater groups
is an important issue given the current popularity of multisource PA systems. In a typical multisource PA system, target managers are evaluated along several performance dimensions (e.g., coaching, teamwork, leadership) by their boss, peers, subordinates, and sometimes their customers. Managers typically receive a feedback report that contrasts their ratings from different sources, including their own self ratings (Dunnette, 1993; London & Smither, 1995; Tornow, 1993). If a multisource PA instrument is not invariant across these different sources of information, these comparisons are meaningless because different constructs are being measured or the items are differentially related to the constructs. The importance of this issue is highlighted by the fact that managers are often encouraged to consider reasons for any discrepancies among the different rating sources. If the rating instrument is not invariant across these sources, this effort may be in vain.

Mount, Judge, Scullen, Systma, and Hezlett (1998) recently found that individual raters account for more variance in multirater–multirater (MTMR) performance data than do rater groups such as peers, subordinates, and supervisors. Their study focused on understanding the source of method effects in MTMR data (i.e., are method effects attributable to individual raters or to rater groups?). In this study, we researched whether ratings made by different groups of raters reflect measures of the same performance constructs and, consequently, whether they are comparable.

One might conclude that the question of invariance across rating sources is a moot issue given Mount et al.’s (1998) finding that rater groups do not account for much variance in MTMR data. However, there are several reasons why the issue of invariance across rating sources should be examined despite their conclusions. First, it may be premature to abandon the longstanding practice of differentiating performance ratings by rating source on the basis of a single study. Second, examining whether performance ratings are invariant across different rater groups can illuminate previous research findings (e.g., Harris & Schaubroeck, 1988). Finally, as Mount et al. acknowledged, the practice of aggregating ratings within rater groups and contrasting them is widespread in organizations that use multisource feedback systems. There is little reason to believe that this practice will cease on the basis of a single empirical study (cf. Banks & Murphy, 1985). This research can clarify whether it is meaningful to compare the ratings made by different constituencies.

Studies of this kind typically are concerned with the broad question of whether a particular instrument displays measurement invariance across well-defined groups. As a result, specific hypotheses usually are not offered. For example, Allen and Oshagan (1995) tested for invariance of the UCLA Loneliness Scale across groups defined by a number of demographic variables (e.g., gender, race, marital status, income level). Byrne et al. (1993) tested for invariance of the Beck Depression Inventory across men and women, whereas Byrne (1993) examined the Maslach Burnout Inventory across elementary, intermediate, and secondary school teachers. In each of these studies, the general hypothesis was that the measurement instrument of interest was noninvariant across the different groups, and the null hypothesis was that the instrument was invariant across groups. In keeping with this convention, the null hypothesis of interest in our study was that a performance rating instrument completed by self, peer, supervisor, and subordinate raters would be invariant across these rater groups. The alternative hypothesis was that the performance rating instrument would be noninvariant (i.e., would differ) across these rater groups.

Method

Participants

The sample in this study included employees of a large, southeastern utility company who provided ratings of managers’ performance. The managers represented a variety of functional areas, such as power generation, power delivery, and customer operations. In all, 1,883 self rating forms (i.e., target managers); 12,779 peer rating forms; 3,049 supervisor rating forms; and 2,773 subordinate rating forms were completed and returned. Because of concerns for anonymity, no information regarding raters’ demographic characteristics was collected.

Procedure

The data for this study were collected during the administration of a multisource PA and feedback system. Self, peer, supervisor, and subordinate ratings of managers’ performance were collected on a multisource appraisal form (MAF). The form was identical for each rater group, and it contained 46 behaviorially oriented items that were designed to measure 10 dimensions of managerial performance, as well as 12 experimental items. The performance dimensions included credibility, customer focus, business knowledge, taking ownership, cooperation, empowerment, problem solving, risk taking, responsibility, motivating others, and openness to change. These performance dimensions were identified by the organization as being critical for successful leadership within the company.

Managers who were being evaluated as part of the multisource appraisal system completed self ratings of their performance. In addition, they distributed copies of the appraisal form to their supervisors and to the peers and subordinates from whom they desired performance feedback. Before distributing copies of the rating form to peers and subordinates, managers provided the following information on each form: (a) their own name, social security number, and functional area; (b) the type of rater who would be completing the form (e.g., subordinate); and (c) the date. Each manager was free to distribute as many ratings forms to peers and subordinates as he or she wished. To ensure anonymity, raters were not asked to sign the rating form after completing it, nor were they asked to provide any form of identifying information. After rating forms were processed, managers received feedback about their ratings in the form of a feedback report. These reports summarized the ratings that managers received on each performance dimension as well as on each item by rating source.

Measures

A brief description of each dimension on the original MAF, as well as the number of items comprising each dimension, appears in Table 1. When completing the MAF, raters were encouraged to think about how often the manager they were rating demonstrated the behavior described by each item. They were asked to base their ratings on behaviors that they had observed, not on how they believed a manager might behave. Ratings were made on a 5-point Likert-type scale that ranged from 1 (never) to 5 (always). If a manager had not had the opportunity to demonstrate the behavior described by a particular item, raters were instructed to select an alternative that was labeled “don’t know.”

Analyses

We used both CFA and IRT techniques to test for invariance of the MAF across rater groups. Although both approaches are sufficient for testing measurement equivalence, each provides unique information about the equivalence of a measuring instrument across different groups. Our pur-
pose was not to enumerate the similarities and differences between CFA and IRT approaches. However, the interested reader is referred to Maurer et al. (1998) and Reise et al. (1993) for excellent discussions of the similarities and differences between these two approaches to studying measurement equivalence.

**CFA**

The CFA procedures outlined by Byrne (1994); Byrne, Shavelson, and Muthen (1989); and Reise et al. (1993) were used to test for measurement invariance of the MAP across the self, peer, supervisor, and subordinate rater groups. This consisted of two primary stages. The goal of the first stage was to identify well-fitting measurement models separately within each group. Here, the a priori measurement model hypothesized to underlie the data was tested in each group. The results of this stage typically indicated that a better fitting model could be found. As a result, the MAF was revised (if the revisions made sense substantively) until an acceptable fit to the data was obtained (see below for a description of these revisions).

In the second stage, the data from all rater groups were analyzed simultaneously. An ordered sequence of analyses was conducted to test for measurement invariance. First, in the baseline model, the measurement models underlying the revised MAF from each group were estimated simultaneously, without any equality constraints. The baseline model provided a test of whether the different rater groups shared a common conceptualization of managerial performance vis-à-vis the factor structure. That is, the model tested whether the structure (i.e., the number of latent variables and the items associated with each latent variable) underlying raters’ evaluations was invariant across the different rater groups. Next, factor loadings ($\lambda_{ij}$) were constrained to be equal across the rater groups in the invariant model. This model provided a test of whether the individual items on the revised MAF were equally related to their respective latent variables across the four rater groups.

There is some disagreement in the literature about how one specifies an invariant measurement model. For example, Byrne (1994) suggested that when testing for measurement invariance, a fully invariant measurement model should include cross-group constraints on the following sets of parameters: (a) item factor loadings ($\lambda_{ij}$), (b) factor variances ($\phi_j$), and (c) factor covariances ($\phi_{ij}$). On the other hand, Reise et al. (1993) noted that it is often inappropriate to place cross-group constraints on factor variances and covariances. This is because, in theory, factor loadings should be invariant across groups, whereas factor variances and (as a result) factor covariances are often sample specific. Furthermore, Widaman (personal communication, September 19, 1997) argued that it would be reasonable to expect factor variances and covariances to be equal when groups are randomly and representatively sampled from a single population. However, when groups are sampled from different populations, it is likely that factor variances and, as a result, factor covariances will differ across groups (cf. MacCallum & Tucker, 1991). Because the rater groups that were used in this study were nonrandom samples from different populations, it was expected that there would be differences in their variances and covariances as a function of sampling. As a result, equality constraints in the invariant measurement model were placed on factor loadings but not on factor variances or covariances.

Several fit indices were used to assess the appropriateness of the different models that were tested. The $\chi^2$ statistic is often used to test whether a particular model results in a significantly worse fit to the data relative to a perfectly fitting model. Nonsignificant $\chi^2$ values are indicative of good model fit. A problem, however, is that the $\chi^2$ statistic is very sensitive in large samples and, as a result, can lead to the rejection of well-fitting models. As a result, several other fit indexes were used in this study to evaluate model fit. These included several incremental fit indexes such as the normed fit index (NFI; Bentler & Bonnet, 1980), comparative fit index (CFI; Bentler, 1990), and Tucker–Lewis index (TLI; Tucker & Lewis, 1973), in addition to the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993). Both the NFI and CFI reflect the proportion of improvement in fit achieved by a particular model relative to a null model (in which the observed variables are treated as independent). The CFI, however, is less dependent on sample size than the NFI (Hu & Bentler, 1995). The TLI has the same interpretation as the NFI and CFI; however, it includes an adjustment for model complexity. Values of these three indexes that are greater than or equal to .90 typically are interpreted as representing a good fit to the data. RMSEA is a measure of lack of fit per degree of freedom for a model. Values of RMSEA between 0 and .05 can be interpreted as reflecting a close fit of the model in the population, whereas values of about .08 or less reflect reasonable fit (Browne & Cudeck, 1993). Like the TLI, the RMSEA takes model complexity into account. Given two models with equal fit in the population but with different numbers of parameters, RMSEA will be smaller for the model with fewer parameters (MacCallum, Browne, & Sugawara, 1996).

Of primary importance in this study was whether placing constraints on the factor loadings in the invariant model was appropriate. One way to determine if the constraints are appropriate is to conduct a $\chi^2$ difference test in which the fit of the invariant model is directly compared with the fit of the baseline model. However, this test is problematic for two reasons. First, because of the $\chi^2$ statistic’s sensitivity in large samples, in all likelihood the $\chi^2$ difference test would lead to the rejection of the invariant model in the large samples used here, even if the loss of fit resulting from the constraints on the factor loadings was modest. Second, the $\chi^2$ statistic tests for exact fit—that is, it tests whether a particular model is exactly correct in the population (MacCallum et al., 1996). In the context of testing for measurement equivalence, the $\chi^2$ difference test whether the invariant model fits exactly as well as the baseline model. It was fully expected that constraining the factor loadings to be equivalent across the different rater groups would result in some loss of fit. The key issue, however, was whether the loss of fit was substantial enough to conclude that the invariant model was a poor representation of the data and, as a result, that the revised

---

**Table 1**

**Description of the Scales Included on the Multisource Appraisal Form**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Measures the extent to which a manager . . .</th>
<th>No. items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credibility</td>
<td>. . . is trustworthy and treats other individuals with dignity and respect.</td>
<td>8</td>
</tr>
<tr>
<td>Customer focus</td>
<td>. . . views his/her customers as a priority and tries to improve customer satisfaction.</td>
<td>5</td>
</tr>
<tr>
<td>Business knowledge</td>
<td>. . . is knowledgeable of the company’s business, markets, and competitors.</td>
<td>4</td>
</tr>
<tr>
<td>Taking ownership</td>
<td>. . . makes decisions as if he/she has a personal stake in the success of the organization.</td>
<td>2</td>
</tr>
<tr>
<td>Cooperation</td>
<td>. . . attempts to foster teamwork and cooperation.</td>
<td>3</td>
</tr>
<tr>
<td>Problem solving</td>
<td>. . . effectively makes decisions and overcomes obstacles.</td>
<td>8</td>
</tr>
<tr>
<td>Risk taking</td>
<td>. . . is willing to make difficult decisions.</td>
<td>2</td>
</tr>
<tr>
<td>Responsibility</td>
<td>. . . is willing to accept responsibility for mistakes, as well as the quality of his/her work.</td>
<td>5</td>
</tr>
<tr>
<td>Motivating others</td>
<td>. . . recognizes, rewards, inspires, and motivates others to achieve high levels of performance.</td>
<td>7</td>
</tr>
<tr>
<td>Openness to change</td>
<td>. . . tolerates change and new ideas.</td>
<td>4</td>
</tr>
</tbody>
</table>
MAF was nonequivalent across rating sources. Browne and Cudeck (1993) as well as MacCallum et al. have argued that it is unreasonable to expect models to fit data perfectly and that the reliance on such an overly stringent criterion as exact fit is likely to lead to the rejection of good models when samples are large. They suggest that a more reasonable test is that of close fit, which tests whether a model provides a close approximation to real-world relationships and effects. For these reasons, in this study, the test of close fit was used to evaluate whether the constraints placed on the factor loadings in the invariant model were appropriate. Strong evidence of close fit is obtained when the entire 90% confidence interval of RMSEA falls below .05 (Browne & Cudeck, 1993; MacCallum et al., 1996).

The CFA analyses were conducted with the EQS structural equations modeling program (Bentler, 1995). Only those performance dimensions that were composed of 3 or more items (indicators) were included in the a priori measurement models tested in the first stage. This was done to avoid statistical problems of underidentification (cf. Bollen, 1989). As a result, the Taking Ownership and Risk-Taking scales were not included in the analyses. Thus, the a priori measurement models tested separately within each rater group consisted of 44 indicator variables (i.e., items) that measured the following eight latent variables: credibility (8 items), customer focus (5 items), business knowledge (4 items), cooperation (3 items), problem solving (8 items), responsibility (5 items), motivating others (7 items), and openness to change (4 items). The scale of the latent variables was set by fixing the loading for the first indicator of each latent variable to unity. Moreover, the eight latent variables were allowed to intercorrelate.

Finally, to avoid estimation problems, only cases having complete data on all 44 items were included in the analyses. Elimination of cases with missing values or “don’t know” responses led to the following sample sizes for each rater group: self (n = 1,555), peer (n = 5,704), supervisor (n = 1,795), and subordinate (n = 1,517). The sample sizes for each rater group were different because our focus in this study was on the ratings provided by individual raters, not the ratings received by individual managers. Moreover, not all managers were rated by each of the constituencies studied here. For example, some managers were rated by peers and their boss but not by their subordinates. Other managers, however, were rated by all four rater groups. Thus, different but overlapping sets of managers were the “targets” of the ratings for each group. This does not pose a problem for the analyses conducted here because the focus was on the relationships between items and the constructs they measured rather than mean differences between the different groups of raters. Finally, because such large sample sizes were available, each rater group was randomly split in half to form a calibration sample and a hold-out sample. The calibration sample for each rater group was used to test and revise the a priori measurement model in the first stage of analyses, whereas the hold-out sample for each rater group was used to cross-validate the revised models (cf. Bollen, 1989; Byrne, 1989). Then, the baseline and invariant models were estimated using the full sample of raters from each rater group.

IRT Procedures

Differential functioning of items and tests. Raju, van der Linden, and Fleer (1995) proposed a framework for examining measurement equivalence, at both the item and scale levels, called “differential functioning of items and tests” (DFIT). The DFIT framework, which is based on IRT (Lord, 1952, 1980), defines differential functioning as a difference in the expected item or scale scores for individuals with the same standing on the latent construct (θ) that is attributable to their membership in different groups. In the language of the DFIT framework, the expected score on an item or scale, given θ, is called the “true score,” and it is expressed on the raw metric of the item or scale (e.g., a 5-point scale for a 5-point Likert-type response format).

Analyses that are based on DFIT yield several types of differential functioning indexes. Noncompensatory differential item functioning (NCDIF) is a purely item-level statistic that reflects true score differences for the two groups under examination. As the “noncompensatory” moniker suggests, the NCDIF index considers each item separately, without regard for the functioning of other items in the scale. Mathematically, NCDIF is the square of the difference in true scores for the two groups being compared, averaged across θ. Thus, the square root of NCDIF gives the expected difference in item responses for individuals with the same standing on θ, but belonging to different groups. Differential test functioning (DTF) is analogous to NCDIF, but it is a scale-level index. The square root of DTF is the average expected scale-score difference among individuals from different groups, but with the same standing on θ.

Compensatory differential item functioning (CDIF) is an item-level index that represents an item’s net contribution to DTF. An important concept in DTF is the directionality of differential functioning. Two items may exhibit similar levels of differential functioning, but in opposite directions (e.g., in a peer–subordinate comparison, one item might favor peers and the other might favor subordinates). In such a scenario, the two items could cancel each other out and produce no net differential functioning at the scale level. Conversely, a number of items that have nonsignificant, but nonzero, levels of NCDIF in the same direction can produce a scale with significant DTF because of the cumulative effect of item-level differential functioning at the scale level. Because the CDIF index represents an item’s net contribution to DTF, it takes into account the functioning of other items on the scale, in contrast to NCDIF.

The relative importance placed on NCDIF, CDIF, and DTF depends on the manner in which the scale is to be used. For example, if only scale-level scores are to be used, such as in a research setting, then only DTF and CDIF may be of interest. In most applications of multisource feedback, both scale-level and item-level scores are presented in managers’ feedback reports. Managers are encouraged to use this information as they construct their developmental plans. In this study, therefore, NCDIF and DTF were of primary concern.

A difference in true scores for members of different groups is considered significant when the NCDIF index for an item exceeds a priori specified critical value and its associated χ² statistic is significant. The critical value for significant NCDIF depends on the number of response categories for an item. To conduct the DTF analyses in this study, it was necessary to collapse the two lowest categories of the response continuum, yielding a 4-point scale (see below). Raju (personal communication, March 1999) has recommended that the critical NCDIF value for an item with four response categories be set to 0.054. The critical DTF value for a scale composed of such items is 0.054 multiplied by the number of items on the scale.

Preparation for IRT-based analyses. One constraint imposed by IRT-based analysis of polytomous items involves empty response categories. Because the IRT model for graded responses (Samejima, 1969) treats each response category like a separate, dichotomously scored item, categories that are never endorsed are akin to items with zero variance and, thus, contain no information. As a result, it is impossible to estimate the parameters for an item with empty response categories. In the data analyzed here, the lowest response category for several of the items contained no responses. To compensate, the two lowest categories were collapsed for all IRT-based analyses, producing a 4-point scale. The three highest response categories from the original 5-point scale remained unchanged.

The graded response IRT model (Samejima, 1969) also assumes that scales are unidimensional. Reckase (1979) found that IRT-based parameter estimation techniques are fairly robust to departures from strict unidimensionality, and he suggested that the assumption could be considered met if a dominant first factor, accounting for at least 20% of the variance, is present. Maximum likelihood factor analyses were conducted to determine if this criterion was met. The revised MAF contained seven scales. Thus for each of the four rater groups, 7 separate factor analyses were conducted (i.e., a total of 28 analyses). For each analysis, the criterion recommended by Reckase was satisfied. For all seven scales on the revised MAF, the first factor accounted for more than 20% of the variance. Averaging across the scales, the first factor accounted for 68%, 54%, 66%, and 62% of the
variance in the peer, self, subordinate, and supervisor rater groups, respectively.

**Item parameter estimation.** The implementation of Samejima’s (1969) graded response model in the MULTILOG computer program (Thissen, 1995) was used to estimate item parameters. Each scale on the revised MAF was analyzed separately, nested within rating source, for a total of 28 separate MULTILOG analyses. Under the graded response model, each item’s parameters define its relation to the underlying construct (θ), with three (i.e., the number of response categories minus 1) b parameters indicating the relative difficulty of the response categories, and one a parameter indicating the item’s slope (i.e., the sensitivity with which differences in θ translate into differences in item responses).

**Person parameter estimation.** After item parameters were estimated, those estimates were used in a subsequent MULTILOG (Thissen, 1995) analysis to compute scale scores, or θ estimates, for each person in the data set. Under IRT, test scoring is a conditional probability problem that is solved with a maximum likelihood estimation procedure. That is, the procedure assigns each person to the standing on θ that is most likely to have resulted in the observed pattern of item responses, given the known probabilistic relations between items and θ that are defined by the item parameters. The peakedness of the likelihood function is reflected in the standard error estimates that accompany the θ estimates and, thus, indicate the certainty with which individuals’ standing on the construct have been estimated. By convention, θ estimates are expressed as z scores. In this stage of the analysis, each person’s ratings were scored using the item parameters estimated for his or her own rater group.

**Equating of parameter metrics.** Before IRT-based estimates of person and item parameters can be compared to assess measurement equivalence, the parameter estimates must be placed on the same metric. This task was accomplished with the EQUATE 2.1 computer program (Baker, 1995), which implements the procedure developed by Stocking and Lord (1983) for equating parameter metrics. The outcome of this stage of the analysis is a set of transformation coefficients (i.e., slope and intercept) that allow parameter estimates from one rating source to be linearly transformed to the metric of the parameter estimates from another rating source. Following Candell and Drosg (1988), the transformation coefficients were reestimated when any items demonstrated significant NCDIF after the first estimation. This second estimation procedure excluded from analysis any items that showed differential functioning using the initial transformation coefficients, thus producing more stable coefficients. This process was completed for each possible pairwise comparison of rating sources and separately for each subscale, for a total of 42 separate equating procedures.

**Analysis of differential item functioning.** Differential functioning was examined using a version of the DFFT4P computer program that had been modified to handle the large sample sizes used here (Raju, 1999). DFFT4P reads in the item and person parameter estimates generated earlier, along with the appropriate transformation coefficients, and calculates the differential functioning indexes described above (i.e., NCDIF, CDFI, and DTFI). The reader is referred to Raju et al. (1995) for more detail about the calculation of the DFFT indexes. As with the equating procedures, each subscale was analyzed separately, in each of six possible pairwise comparisons between rating sources, resulting in 42 separate analyses.

**Results**

**CFA Analyses**

**Step 1: Establishing Baseline Models**

In the first stage of analyses, the a priori measurement model was evaluated in the calibration sample of each rater group. Table 2 presents the results of these analyses. As indicated by the fit statistics displayed in the table, the a priori measurement model provided a reasonable fit to the data in each group. However, for each group, at least one of these indexes suggested that improvements to the a priori model could be achieved. As a result, revisions to the a priori measurement model were initiated in each rater group.

Revisions were suggested by a number of factors, including actual parameter values and the Lagrange Multiplier (LM) test option in the EQS package. Essentially, the LM test provides an indication of the gain in model fit (vis-à-vis a reduction in the model χ² value) that can be achieved by estimating a previously unestimated parameter (e.g., a factor loading or error covariance). Freeing parameters suggested by the LM test was done only when it made sense from a substantive standpoint. For example, error covariances between items were estimated for several pairs of items that were worded similarly and that loaded on the same factor. Within each pair, the items referred to similar actions and differed only in terms of the target of the action (e.g., “Removes obstacles to teamwork” and “Removes obstacles to successful completion of work”). Although estimating error covariances poses the risk of overlooking unmodeled latent variables (Gerbing & Anderson, 1984), omitting error covariances can lead to an underestimation of model parameters, particularly factor loadings (Reddy, 1992). Several researchers have argued that error covariances can be estimated if there is a substantive reason for doing so, such as when two items are substantially similar in content (cf. Bollen, 1989; Byrne, 1994; Hoyle, 1991; Kline, 1998). Two other types of revisions were made to the measurement model. First, six items were freed to load on factors they were not initially anticipated to measure. This was done only if it made sense on the basis of an item’s content and the nature of the other items composing the factor. Finally, several items were eliminated from further analysis if they had significant loadings on multiple factors, which suggested that they were not good indicators of any single factor.

All of the revisions made to the MAF in each group were incorporated in the revised measurement model for that group.
Space limitations preclude a detailed discussion of each revision that was made to the a priori model in each group. However, a summary of the revisions and their rationale is provided in Table 3. Overall, the revised measurement models were nearly identical across the different rater groups. In each group, the revised MAF consisted of 39 indicators of seven latent factors (the eighth factor in the a priori model, openness to change, was eliminated). The Ethical Behavior factor consisted of six items that measured the extent to which managers are fair, truthful, and respect the rights of others (e.g., "Treats all people fairly"). The Customer Focus factor included four items that measured the extent to which managers actively interact with customers, resolve customer complaints, and attempt to improve customer satisfaction (e.g., "Identifies solutions which improve customer satisfaction"). The Business Understanding factor consisted of four items that assessed the extent to which managers understand the organization’s business operations as well as the strategic issues and trends in the external business environment (e.g., "Demonstrates an understanding of the company’s business functions"). The Cooperation factor included six items that measured the degree to which managers work cooperatively with others and foster an environment in which effective teamwork can occur (e.g., "Removes obstacles to teamwork"). The Problem-Solving factor included six items that measured the extent to which managers anticipate problems, understand the important dimensions of a problem, and take action to overcome problems by implementing practical solutions (e.g., "Produces practical solutions"). The Responsibility factor consisted of five items that measured the degree to which managers take responsibility for their own decisions, outcomes, and mistakes, as well as the extent to which they fulfill their responsibilities to others (e.g., "Accepts the responsibility for his or her mistakes"). The Motivating Others factor (nine items) measured the extent to which managers created a motivating climate through such things as coaching, providing feedback and encouragement, and employee recognition (e.g., "Recognizes others for their accomplishments"). The average intercorrelation among the seven latent variables for self, peer, supervisor, and subordinate raters were .74, .85, .74, and .87, respectively. Finally, each of the items on the revised MAF served as an indicator of the same latent factor in each of the four rater groups. Finally, the revised measurement models for self and subordinate raters included one parameter (an error covariance between items sharing similar content) that was not estimated in the peer and supervisor models. As a result, the self and subordinate models had 1 df less than the other two models.

The fit of the revised measurement model in each calibration sample is shown in Table 2. Because several revisions were made to the a priori measurement model, the possibility that the revised model reflected capitalization on chance factors in the calibration samples as opposed to the true model underlying the data had to be addressed. One way to address this concern is by cross-validating the revised model on independent samples (Bollen, 1989; Byrne, 1989). As a result, the revised measurement model for each rater group was cross-validated in the hold-out sample for that group. As can be seen in Table 2, the revised measurement model for each rater group cross-validated well in the respective hold-out samples. In each of these hold-out samples, all of the factor loadings were statistically significant (p < .05). Moreover, the average standardized factor loadings in the self, peer, supervisor, and subordinate rater groups were .62, .79, .74, and .77, respectively. Together, these results indicate that the revisions made to the MAF resulted in a measurement model that better represented the underlying structure of the data, as opposed to one that simply capitalized on chance factors in the calibration samples. As a result, the revised measurement models served as the baseline models in the second stage of analyses.

Step 2: Testing Measurement Invariance

The first step in testing for measurement invariance involved estimating the baseline model. In this model, the revised measurement models from each rater group were estimated simultaneously. The covariance matrices from each group were analyzed. The variance of each latent variable was fixed at unity in the self rater

Table 3
Description of Revisions to the A Priori Measurement Models in Each Rater Group

<table>
<thead>
<tr>
<th>Revision</th>
<th>Groups affected</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Eliminate four negatively worded items.</td>
<td>Slf, P, Spv, Sub</td>
<td>Consistently low factor loadings</td>
</tr>
<tr>
<td>2. Drop Item 11.</td>
<td>Slf, P, Spv, Sub</td>
<td>Significant loadings on multiple factors</td>
</tr>
<tr>
<td>3. Item 8 is an indicator of F4 not F8.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
<tr>
<td>4. Item 40 is an indicator of F7 not F8.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
<tr>
<td>5. Drop item 33, leaving no indicators of F8.</td>
<td>Slf, P, Spv, Sub</td>
<td>Item does not appear to be a meaningful indicator of F8, nor F1-F7</td>
</tr>
<tr>
<td>6. Drop F8.</td>
<td>Slf, P, Spv, Sub</td>
<td>No remaining indicators</td>
</tr>
<tr>
<td>7. Item 5 is an indicator of F7 not F5.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
<tr>
<td>9. Item 42 is an indicator of F4 not F1.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
<tr>
<td>10. Free error covariance for Items 24 and 31.</td>
<td>Slf, P, Spv, Sub</td>
<td>Similar item content</td>
</tr>
<tr>
<td>11. Free error covariance for Items 36 and 48.</td>
<td>Slf, P, Spv, Sub</td>
<td>Similar item content</td>
</tr>
<tr>
<td>12. Free error covariance for Items 19 and 40.</td>
<td>Slf, Sub</td>
<td>Similar item content</td>
</tr>
<tr>
<td>13. Free error covariance for Items 32 and 41.</td>
<td>Slf, P, Spv, Sub</td>
<td>Similar item content</td>
</tr>
<tr>
<td>14. Item 42 is an indicator of F4, not F1.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
<tr>
<td>15. Item 7 is an indicator of F4 not F1.</td>
<td>Slf, P, Spv, Sub</td>
<td>Revision deemed appropriate given LM test, item content, and nature of factors</td>
</tr>
</tbody>
</table>

Note. F1 = credibility; F2 = customer focus; F3 = business knowledge; F4 = cooperation; F5 = problem solving; F6 = responsibility; F7 = motivating others; F8 = openness to change; Slf = self; P = peer; Spv = supervisor; Sub = subordinate; LM = Lagrange Multiplier.
group, and the corresponding variances were freely estimated in the remaining three groups. Furthermore, the first indicator of each latent variable was constrained to be equal across the four groups. The baseline model was specified in this way to set the scale for the latent variables, as well as to achieve comparability of the factor loadings across the four groups (Reise et al., 1993). Finally, because the baseline model was specified in this way, the variances of the latent variables in the peer, supervisor, and subordinate rater groups were estimated relative to the unit variances in the self-rater group.

Table 4 presents the goodness-of-fit indexes for both the baseline and invariant models. For the baseline model, the $\chi^2$ statistic was large and significant. Thus, the hypothesis that the baseline model provided an exact fit to the data must be rejected. However, as stated earlier, one can expect a statistically significant $\chi^2$ value when large samples like the ones in this study are used. As a result, other goodness-of-fit measures were relied on to evaluate model fit. The incremental fit indexes (NFI, CFI, and TLI) displayed in Table 4 were all well above .90, indicating that the baseline model provided a good fit to the data. Moreover, the value of RMSEA was .048, and the entire confidence interval around RMSEA was below .05 (lower and upper values of .0473 and .0486, respectively). This means that although the hypothesis that the baseline model provides an exact fit to the data must be rejected, the hypothesis that the model provides a close fit to the data cannot be rejected. The baseline model appears to provide a close approximation to the relationships among the items on the MAF that were included in the model. For each rater group, the same seven latent factors were found to underlie raters’ responses to the items on the revised MAF. Moreover, for each group of raters, each of the seven factors was composed of the same items. Thus, a common structure underlies all raters’ responses to the items on the revised MAF.

Next, the invariant model was tested. This was accomplished by specifying a new model in which the factor loadings in the baseline model were constrained to be equivalent across the four rater groups. The results for the invariant model also are presented in Table 4. Specifying the factor loadings to be equivalent across the four rater groups resulted in a statistically significant loss of fit, as indicated by the significant difference in the $\chi^2$ values for the baseline and invariant models, $\Delta\chi^2(96) = 685.52, p < .05$. However, inspection of the incremental fit indexes presented in the table indicates that even though the invariant model resulted in a statistically significant loss of fit, the drop in fit relative to the baseline model was quite small. The values of the NFI and CFI each dropped .003, and there was no change in the value of the TLI. Even though the constraints on the factor loadings in the invariant model resulted in a statistically significant loss of fit, the difference in fit between the two models was negligible. In all likelihood, the discrepancy between these two results reflects the sensitivity of the $\chi^2$ difference test in large samples like those used here. Thus, the invariant model still provided a good fit to the data. The results for RMSEA support this conclusion. The upper bound of the 90% confidence interval for RMSEA was still below .05. Thus, despite constraining the factor loadings to be equivalent across the four rating scales, the invariant model still provided a close approximation to the data. Therefore, it seems reasonable to conclude that the 39 items on the revised MAF compose a multifaceted measure of managerial performance that is equivalent (with the exception of the unique error covariance in the self and subordinate groups) across the four rating scales.

Finally, one other finding warrants brief mention. As indicated earlier, the factor variances for peer, supervisor, and subordinate raters were estimated relative to the unit variances in the self-rater group. Table 5 presents the variances of the latent variables in each of the four groups. In all but one case, the variances of the latent variables in the peer, supervisor, and subordinate groups were significantly larger than the corresponding variances in the self-rater group. In several cases, these differences were quite large. For example, the variances of the responsibility and credibility factors among subordinate raters were more than 2.8 times larger than the corresponding variances in the self-rater group. Thus, at the level of the latent performance variables, there was much less variability in managers’ self-ratings than in the ratings made by the other rater groups.

**DFIT Analyses**

The DFIT analyses generated 276 differential functioning indexes: [(39 NCDIF indexes (1 for each item) + 7 DTF indexes (1 for each scale)] × [6 comparisons]). The differential functioning indexes, along with their $\chi^2$ test statistics, are shown in Table 6. In the table, it can be seen that many of the indexes met one of the two criteria necessary to show differential functioning. However, despite the performance of 276 separate tests, only five instances of NCDIF and only one instance of DTF were observed. All six instances of differential functioning were constrained to the Motivating Others Scale, and all occurred in comparisons involving subordinates.

As can be seen in Table 6, one item (MO-4) accounted for three of the five instances of NCDIF. Moreover, the CDFI index for this item indicated that if it were removed, the DTF for the Motivating Others scale would become nonsignificant. Thus, four of the six instances of differential functioning that were observed could

---

**Table 4**

**Goodness-of-Fit Statistics for Multiple-Groups Confirmatory Factor Analyses Testing for Measurement Invariance**

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>$df$</th>
<th>$\Delta\chi^2$</th>
<th>$\Delta df$</th>
<th>NFI</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA&lt;sub&gt;L&lt;/sub&gt;</th>
<th>RMSEA</th>
<th>RMSEA&lt;sub&gt;U&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>19,195.05</td>
<td>2,710</td>
<td></td>
<td></td>
<td>.933</td>
<td>.942</td>
<td>.936</td>
<td>.0473</td>
<td>.0480</td>
<td>.0486</td>
</tr>
<tr>
<td>Invariant</td>
<td>19,881.57</td>
<td>2,806</td>
<td>686.52</td>
<td>96</td>
<td>.930</td>
<td>.939</td>
<td>.936</td>
<td>.0474</td>
<td>.0480</td>
<td>.0486</td>
</tr>
</tbody>
</table>

*Note.* All $\chi^2$ values are statistically significant ($p < .05$). NFI = normed fit index; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; RMSEA<sub>L</sub> and RMSEA<sub>U</sub> = lower and upper bounds of the 90% confidence interval around RMSEA, respectively.
be accounted for by a single item. Two other items (MO-3 and MO-6) accounted for the remaining instances of NCDIF that were observed.

It also is important to consider the magnitude of the differential functioning that was observed. This is especially important given the large samples that were used and the sensitivity of the \( \chi^2 \) statistic when this is the case. For the five instances of NCDIF, the average expected difference in observed scores between rater groups was only .29 on a 4-point scale. Similarly, the expected difference between rater groups in scores on the Motivating Others scale was just .81. This is a trivial difference given that scores on this scale had a possible range of 9 to 36.

Together, the results of the DFT analysis suggest that although some instances of statistically significant differential functioning were observed, they were localized to only a few items, and the magnitude of the differential functioning was trivial. This conclusion is consistent with the results of the CFA, which showed that within the limits of close fit, the revised MAF could be regarded as invariant across the four rater groups. Thus, both sets of analyses suggest that raters with equivalent impressions of a manager’s performance can be expected to provide comparable performance ratings, regardless of whether they are peers, subordinates, superiors, or the managers themselves (i.e., self ratings).

Discussion

A consistent finding in the performance appraisal literature is that ratings obtained from different rating sources tend not to converge. This lack of convergence is manifested in mean differences between the ratings provided by different sources, as well as in low to moderate between-source correlations among ratings (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Mount, 1984; Thornton, 1980). Although a number of explanations for this lack of convergence have been offered and investigated, this study examined a more fundamental reason that may explain why ratings from different sources fail to converge. Specifically, the goal of this study was to assess whether a multifaceted performance rating instrument measured the same performance constructs when completed by self, peer, supervisor, and subordinate raters.

The results of the confirmatory factor analytic tests of invariance revealed that there was a minor feature of the measurement model that was not invariant across the four rater groups. Specifically, an error covariance in the self and subordinate models was not included in the peer and supervisor models. This means that self and subordinate raters’ responses to two items were influenced by the similarity in the wording of these items, whereas peer and supervisor raters were not. These two items began with the phrase, “Encourages others to...”. It may be that self and subordinate raters were influenced more by the stem of these items when making their ratings (i.e., “encourages...”), whereas peers’ and supervisors’ ratings were more influenced by the behavior that was being encouraged (e.g., “...improve performance”).

Aside from the unique error covariance in the self and subordinate groups, the multiple-groups CFAs demonstrated that the revised MAF was otherwise invariant across the four rater groups. Although the baseline model did not fit the data exactly, it did provide a close fit. Thus, the different rater groups shared a common conceptualization of the managerial performance dimensions underlying the items on the revised MAF. When the factor loadings were constrained in the invariant model, there was a significant loss of fit. However, only a trivial drop in the values of the incremental fit indexes was observed. Moreover, the results for the 90% confidence interval of RMSEA revealed that the hypothesis of close fit (MacCallum et al., 1996) could not be rejected. This means that the baseline model and the invariant model were nearly indistinguishable in terms of how well they fit the data. This is underscored by the fact that the values of RMSEA for the baseline and invariant models were identical to four decimal places (i.e., .0480). Thus, it is appropriate to conclude that the different rater groups shared a common conceptualization of the dimensions of managerial performance underlying the items on the revised MAF and that the relationships between the items and the constructs they measured, although not precisely identical across the rater groups, were close enough to be regarded as equivalent.

The results of the DFT analysis corroborate the CFA results. Here, for all comparisons among the four rater groups, very little evidence of differential functioning was observed. Although some items and scales met one criterion necessary for establishing differential functioning, only three items and one scale met both of the necessary criteria, and all of those instances were confined to the Motivating Others scale. Further, the differential functioning
that was detected was trivial in magnitude. Thus, on the basis of the DFIT analysis, it seems reasonable to conclude that the revised MAF was invariant across the four rater groups. It is noteworthy that both methods of assessing measurement invariance converged, especially given the fact that CFA procedures model linear relationships between items and the constructs they are intended to measure, whereas IRT procedures model nonlinear relationships between latent traits and the probabilities of specific item responses (Maurer et al., 1998).

These conclusions have several implications. First, each of the items on the revised MAF can be construed as a measure of the same underlying performance construct in each group of raters. Second, the items are equally effective indicators of their respective constructs across the four rater groups. Thus, one could expect

\[ \chi^2 \]

that are underlined. Indexes marked with an asterisk (*) are statistically significant \((p < .01)\) and exceed the critical value for differential functioning. Differential item functioning is indicated by significant \(\chi^2\) values and NCDIF values greater than .054 \(\times\) the number of items in a scale. supv = supervisor; sub = subordinate; NCDIF = noncompensatory differential item functioning; DFIT = differential functioning of items and tests.

Note.

that was detected was trivial in magnitude. Thus, on the basis of the DFIT analysis, it seems reasonable to conclude that the revised MAF was invariant across the four rater groups. It is noteworthy that both methods of assessing measurement invariance converged, especially given the fact that CFA procedures model linear relationships between items and the constructs they are intended to measure, whereas IRT procedures model nonlinear relationships between latent traits and the probabilities of specific item responses (Maurer et al., 1998).

These conclusions have several implications. First, each of the items on the revised MAF can be construed as a measure of the same underlying performance construct in each group of raters. Second, the items are equally effective indicators of their respective constructs across the four rater groups. Thus, one could expect

\[ \chi^2 \]

that are underlined. Indexes marked with an asterisk (*) are statistically significant \((p < .01)\) and exceed the critical value for differential functioning. Differential item functioning is indicated by significant \(\chi^2\) values and NCDIF values greater than .054 \(\times\) the number of items in a scale. supv = supervisor; sub = subordinate; NCDIF = noncompensatory differential item functioning; DFIT = differential functioning of items and tests.

Note.
that a given manager's rating on an item would have the same meaning, regardless of whether the item was rated by a peer, supervisor, subordinate, or the actual target manager (i.e., self-rating). Finally, regardless of the perspective from which an individual makes ratings (i.e., self, peer, supervisor, or subordinate), raters sharing the same impression or evaluation of a manager's performance can be expected to make comparable ratings on the revised MAF.

The findings of this study are congruent with the results of Maurer et al. (1998), who found that a team-building measure was invariant for peer and subordinate raters. However, they diverge from the findings reported by Lance and Bennett (1997), who found that a measure of interpersonal proficiency was not invariant across self, peer, and supervisor raters. They found that raters had different conceptualizations of interpersonal proficiency and that the items on the scale related differently to the underlying factor across the different rating sources. It is unclear what accounts for this discrepancy, especially given the fact that Lance and Bennett's study included several features that should have promoted invariance (e.g., behavioral scale anchors, frame-of-reference rater training) but that were absent from this study and Maurer et al.'s study. One feature that may explain the discrepancy is that the appraisal system in Lance and Bennett's study was used for research purposes (i.e., validation and training evaluation), whereas the rating systems in this study as well as Maurer et al.'s were used for development. In addition, in the latter studies, the items on the rating instrument reflected specific behaviors (e.g., "Shows appreciation for team accomplishments"), whereas the items on the Interpersonal Proficiency factor studied by Lance and Bennett reflected broader behavioral dimensions such as integrity, leadership, and self-control. Although the more abstract items in Lance and Bennett's study were defined behaviorally, perhaps the specificity of the behavioral items in this study and Maurer et al.'s contributed to a common interpretation among the different rater groups, which may have facilitated the finding of measurement invariance. Maurer et al. suggested that a number of rating system variables (e.g., scale type, dimension content) might contribute to the equivalence of rating instruments. Clearly, future research should examine the generalizability of the results reported here to other types of performance rating instruments and contexts.

Another noteworthy finding from this study is the amount of variance in the latent performance variables in the peer, supervisor, and subordinate rater groups compared with self-raters. The variances of the latent variables in the former three groups were estimated relative to the unit variances in the self rater group. In all but one instance (business knowledge for subordinate raters), the variances of the latent variables in these three groups were significantly larger than the corresponding variances in the self rater group. These differences often were substantial. In many cases, the variance of a latent variable in the peer, supervisor, or subordinate group was more than 1.5 times the corresponding variance for self raters. The largest of these differences was for the credibility dimension in the subordinate rater group, which was nearly three times the size of the corresponding variance for self raters. In addition, the same dimensions had the largest variances in each of the peer, supervisor, and subordinate rater groups. These dimensions included credibility, cooperation, problem solving, and responsibility.

This means that, as a group, self raters were far less discriminating in their ratings than raters in the other three groups. Stated another way, self raters perceived fewer differences among themselves than peer, supervisor, or subordinate raters. Lance and Bennett (1997) obtained similar results for self, supervisor and peer ratings. One implication of this finding is that weak correlations between self ratings and others' ratings (Conway & Huffcutt, 1997; Harris & Schaubroeck, 1988; Mount, 1984) may, in part, be an artifact of range restriction. That is, range restriction may attenuate observed correlations between self ratings and others' ratings. This suggestion, however, is contradicted by Harris and Schaubroeck's finding that correcting self ratings for range restriction does not improve self–other correlations. The standard deviation for self ratings in their meta-analysis was 78% of the standard deviation for supervisor ratings and 84% of the standard deviation for peer ratings. However, when averaged across latent variables, the standard deviation for self-ratings in this study was 52% of the standard deviation for peer ratings, 69% of the standard deviation of supervisor ratings, and 57% of the standard deviation for subordinate ratings. Thus, in Harris and Schaubroeck's study, the amount of range restriction in self ratings was not as severe as was observed here. More extreme range restriction, like that observed here, might attenuate self–other rating correlations. Moreover, in the peer, supervisor, and subordinate rater groups, the same four dimensions had the largest variances, compared with other dimensions. This may explain why self–other correlations are higher for some dimensions compared with others.

**Practical Implications**

On a practical level, the findings reported here support a common practice in multisource feedback systems. Specifically, the ratings that managers receive from different rater groups often are compared directly in these systems (Dunnette, 1993; London & Smither, 1995; Tornow, 1993). The results reported here are encouraging in that they suggest that comparisons such as these are legitimate. The results support the conclusion that the revised MAF was invariant across the four rater groups, which means that the same underlying performance constructs were being measured in each group. This implies that differences in observed ratings cannot be attributed to differences between rater groups in what the items measure. Thus, asking managers to resolve, explain, or understand discrepancies between rating sources is not an exercise in futility, as would be the case if the discrepancies were an artifact of the measurement system.

Discrepancies in the ratings provided by different rater groups may arise from a host of other factors, none of which can be ruled out when a rating instrument is found to be invariant. For example, both informational and motivational biases may produce self–other rating discrepancies (cf. Campbell & Lee, 1988; Cardy & Dobbins, 1994; Farh & Dobbins, 1989a, 1989b). Alternatively, rating discrepancies may occur because managers' behavior is conditional on the constituency with which they interact. This ecological perspective (Lance et al., 1992; Lance & Woehr, 1989) suggests that discrepancies can reflect real differences in managers' behavior. Thus, it is important to emphasize that a finding of measurement equivalence or invariance for a rating instrument implies only that the observed scores it produces are on the same scale in each rater group. It means that the ratings can be inter-
interpreted as reflecting the same underlying constructs in each group. It does not imply that the resulting ratings will agree or that they are accurate.

Limitations

There are some possible limitations of this study that should be discussed. First, target managers in the multisource feedback system studied here were allowed to choose which peers and subordinates evaluated them. This may have contributed to the finding of measurement equivalence if managers distributed rating forms only to individuals who shared the conceptualization of managerial performance that was normative within the organization. If individuals with idiosyncratic conceptualizations of managerial performance were excluded because of managers’ choices, the results reported here may not generalize to multisource feedback systems that give managers less control over who evaluates them.

A second limitation of this study is that it focused on a single multisource feedback system. As a result, the results may not generalize to other systems in different contexts that have different features (e.g., rater training, administrative purpose, absolute vs. relative scale anchors). Future research should attempt to replicate this study’s findings in other contexts. Finally, the sample used here included employees in one organization, which may limit generalizability as well. However, individuals from a number of departments and functional areas participated, which, together, represented a broad array of positions. Thus, although it may be difficult to generalize the results of this study to different types of organizations or to different types of multisource feedback systems, the results are likely to generalize to other managerial and supervisory positions.

Conclusion

In conclusion, the purpose of this study was to examine whether a multidimensional rating instrument was invariant across the four rating sources that are commonly used in multisource feedback systems. Although previous research has examined this issue (Lance & Bennett, 1997; Maurer et al., 1998), no other study had examined the invariance of a multifaceted instrument across all four rater groups. The results of the study revealed that the rating instrument could be regarded as measuring the same underlying performance constructs in each rater group, thus supporting the practice of directly comparing the ratings that managers receive from different rating sources. The results also suggest that the failure of observed ratings from different sources to converge may be less a function of the measurement system than of substantive differences between the four rater groups. Thus, researchers should continue in their efforts to understand the differences between rating sources that may account for discrepancies in their ratings.

References


Received February 15, 1999
Revision received March 16, 2000
Accepted March 20, 2000