PERSONNEL PSYCHOLOGY

THE STUDY OF PEOPLE AT WORK

PERSONNEL PSYCHOLOGY
2012, 65, 531–563

EVIDENCE FOR THE EFFECTIVENESS OF AN ALTERNATIVE MULTISOURCE PERFORMANCE RATING METHODOLOGY

BRIAN J. HOFFMAN
University of Georgia

C. ALLEN GORMAN
Radford University

CARRIE A. BLAIR
College of Charleston

JOHN P. MERIAC
University of Missouri, St. Louis

BENJAMIN OVERSTREET
University of Georgia

E. KATE ATCHLEY
University of Tennessee, Knoxville

Despite persistent concerns as to the quality of performance information obtained from multisource performance ratings (MSPRs), little research has sought ways to improve the psychometric properties of MSPRs. Borrowing from past methodologies designed to improve performance ratings, we present a new method of presenting items in MSPRs, frame-of-reference scales (FORS), and test the efficacy of this method in a field and lab study. The field study used confirmatory factor analysis to compare the FORS to traditional rating scales and revealed that FORS are associated with increased variance due to dimensions, decreased overlap among dimensions, and decreased error. The laboratory study compared rating accuracy associated with FORS relative to frame-of-reference training (FORT) and a control group and demonstrated that FORS are associated with higher levels of accuracy than the control group and similar levels of accuracy as FORT. Implications for the design and implementation of FORS are discussed.

Despite the centrality of performance ratings to a cross section of human resource functions, organizational scholars and practitioners have always had a somewhat uneasy relationship with them. Many scholars question whether performance appraisal (PA) ratings provide meaningful

Study 2 was funded, in part, by a Faculty Innovation Grant from Angelo State University. The authors wish to thank Josh Collins, Lauren Felton, Jessica Stoner, and Jennifer Thorndike for their assistance with the data collection for Study 2.

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

information (Murphy, 2008), and others have urged that the practice of performance appraisal be discontinued entirely (Deming, 1986). Given their centrality to management research and practice and the ongoing concerns of their accuracy and value, it is not surprising that management researchers have consistently sought ways to improve the quality of performance ratings. In this vein, rating scale design has received substantial attention (Paterson, 1922, 1923; Smith & Kendall, 1963). Despite a voluminous body of literature on scale design, the results have been interpreted as disappointing (Murphy, 2008). Indeed, since Landy and Farr's (1980) proposed moratorium, PA scale design research has been scarce.

In the years since, few areas of research have enjoyed as much attention as multisource performance ratings (MSPRs; i.e., 360-degree feedback; Campbell & Lee, 1988; Harris & Schaubroeck, 1988; Lance & Woehr, 1989). However, the widespread attention paid to MSPRs has only amplified concerns over the value of performance ratings (Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Murphy, 2008; Viswesvaran, Schmidt, & Ones, 2002, 2005). It is noteworthy that just as MSPR research and practice became more popular, efforts to improve the quality of performance ratings through design interventions began to taper off. Consequently, little research has been directed toward the design of MSPRs.

The goal of this study is to reconsider the value and usefulness of rating scale interventions, with a specific emphasis on the development and evaluation of a scale design approach that is amenable to incorporation into MSPRs. Borrowing from past PA scale design research and rater training efforts, we propose an alternative performance measurement system that can feasibly be incorporated in MSPR systems, labeled frame-of-reference scales (FORS).We then present two studies that outline a psychometric evaluation of this technique. Study 1 uses confirmatory factor analysis (CFA)–based multitrait-multimethod (MTMM) analyses to provide a psychometric evaluation of the instrument using a pre-post design with a managerial sample. Study 2 extends this work in a laboratory setting by investigating the efficacy of the instrument for improving rating accuracy compared to a traditional frame-of-reference training (FORT) program and a control rating scale.

Evaluating Performance Ratings

Over the past century, a variety of methodological and statistical methods have been used to evaluate the quality of performance ratings, including rater "errors" (Saal, Downey, & Lahey, 1980), rater accuracy (Borman, 1974; Cronbach, 1955), scale factor structure (Campbell & Fiske, 1959; Conway, 1996), and nomological network (James, 1973). Operating under the assumption that ratings that were overly high (leniency) or failed

to distinguish between performance constructs (halo) were indicative of poor quality ratings, rating "errors" were the most frequently used criteria when evaluating performance ratings for the better part of the 20th century (Austin & Villanova, 1992), and this trend is especially evident in PA scale design research (Cardy & Dobbins, 1994). Yet, research has indicated that the relationship between accuracy and these psychometric "errors" tends to be weak and can even be positive (e.g., Becker & Cardy, 1986; Cooper, 1981; Murphy & Balzer, 1989). Accordingly, it is now widely recognized that rater "errors" are poor indicators of the quality of ratings (Fisicaro, 1988; Murphy, 2008; Murphy & Balzer, 1989; Nathan & Tippins, 1990). Consequently, the conclusions of the extensive literature base that has "debunked" the usefulness of PA scale redesign efforts using rater errors to index rating quality are dubious.

From the late 1970s through the 1990s, performance appraisal research moved to the laboratory and focused on rater accuracy, where rater accuracy indices were derived by comparing observed ratings to "true" performance levels (Bernadin, Tyler, & Wiese, 2001). The common procedure of this method is to prepare stimulus materials that depict an employee exhibiting performance behaviors, either through a video recording or written description of the performance. Expert raters provide ratings intended to reflect the true levels of performance. Finally, participants provide ratings, and their ratings are compared to the expert-generated true scores using a variety of statistical indices (c.f., Borman, 1974; Cronbach, 1955). Rater accuracy approaches have commonly been used in the evaluation of rater training as well as research aimed at understanding the cognitive processes associated with performance ratings; however, relatively little research examined accuracy as a criterion in scale design interventions (Cardy & Dobbins, 1994). Thus, it is unclear whether scale design influences performance rating accuracy. Importantly, although this paradigm is very useful because it allows for a direct comparison between ratings and actual performance, it is difficult to obtain "true scores" for use as a referent in field settings. Thus, rater accuracy is limited to laboratory applications. Many have questioned whether findings from lab-based accuracy studies generalize to field settings (Sulsky & Balzer, 1988), and performance rating lab studies have declined over the last decade.

In the place of rater error and accuracy approaches, recent research has increasingly emphasized the factor structure and construct validity of performance ratings (Austin & Crespin, 2006). In the context of MSPRs, a common approach has been to use CFA-based MTMM analyses (see Lance, Hoffman, Gentry, & Baranik, 2008, for a review). These tools provide a means to evaluate performance ratings in field settings by evaluating the contribution of multiple sources of variance (e.g., dimensions, source, and error), key information in evaluating the construct validity of

performance ratings (cf. Cambpbell & Fiske, 1959; Lance et al., 2008). Although this type of analytical approach is routinely used to investigate the construct validity of performance ratings (Conway, 1999) and other measurement tools (e.g., assessment centers; Hoffman, Melchers et al., 2011), such approaches have rarely been employed as a criterion of rating scale design. Using this approach, the two rating scales are compared in terms of dimension effects, source effects, error, and correlations among dimension factors (see Lance et al., 2008). As we will discuss below, MSPR instruments are generally designed to measure multiple behavioral competencies. Accordingly, it is important that there is some evidence for the construct validity of these competencies, as indicated by dimension effects. Given that feedback is given on multiple competencies, it is important that the competencies are actually distinguished on the measure. To the degree that raters cannot distinguish competencies, the ability to glean specific feedback from the tool is prohibited, resulting in reduced developmental usefulness (Hoffman & Baldwin, 2012). As in other measurement contexts, it is important to maximize the amount of systematic variance and minimize the amount of error variance. Finally, although source effects are cosmetically similar to method effects, MSPR research often interprets source effects as performance relevant variance rather than rater bias (see Lance et al., 2008). Thus, we do not make specific predictions regarding the influence of scale design on source effects.

The present research uses multiple approaches to evaluating the effectiveness of FORS. Specifically, Study 1 evaluates the efficacy of FORS in a field setting by comparing the results of MTMM-based CFA models across FORS and a control rating scale. These analyses will allow for the determination of the relative influence of performance dimensions, rater source factors, and error variance associated with different scale formats. In Study 2 we evaluate the scale in a laboratory setting in order to evaluate whether the scale is associated with an increase in accuracy. By leveraging the generalizability of the field with the control of the lab, this study minimizes the weaknesses associated with each and facilitates a more comprehensive evaluation of the scale than would be provided by relying on either method in isolation.

Interventions to Improve Performance Ratings

Scale Design Interventions

The search for methods to enhance the quality of ratings has been a persistent theme throughout the history of the management sciences (Kingstrom & Bass, 1981). Despite substantial variations in approaches to rating scale design, many existing approaches (e.g., behaviorally

anchored rating scales [BARS], behavioral observation scales [BOS], and behavioral summary scales [BSS]) are similar in that they focus on providing more concrete, specific, or objectively scored behavioral information that raters can use as a referent when evaluating performance. Such scale design "fixes" are generally viewed as minimally successful (Landy & Farr, 1980; Murphy, 2008). However, DeNisi (1996) observed that "the basis for these comparisons was. . .usually. . .the level of psychometric errors present in the data . . . .Clearly, we were still willing to assume that the absence of rating errors indicated more accurate ratings" (p. 7). It is now recognized that this assumption is unfounded (Fisicaro, 1988; Murphy & Balzer, 1989). In fact, some research that has used more appropriate indices (e.g., rater accuracy, scale standard error) to evaluate scale redesign has actually yielded favorable results (Benson, Buckley, & Hall, 1988; Borman et al., 2001; Goffin, Gellatly, Paunonen, Jackson, & Meyer, 1996; Tziner, 1984).

Rater Training

In contrast to scale design, rater training is generally accepted as an effective method to increase the psychometric soundness of performance ratings (Woehr, 2008). Similar to scale design approaches, popular rater training approaches (e.g., frame-of-reference training [FORT]) involve more clearly defining behavioral categories and relevant behaviors using examples of effective and ineffective performance (Sulsky & Day, 1992; 1994). In contrast to scale design research that typically investigated rater *errors*, numerous studies supported the effectiveness of FORT for improving rating *accuracy* (see Woehr & Huffcutt, 1994, for a review). Thus, evidence has accumulated that providing raters with more concrete behavioral information in the form of rater training and possibly even rater scale alterations can have a beneficial influence on the quality of performance ratings. Yet, the value of rating scale design has rarely been tested in the context of MSPRs.

Multisource Performance Ratings

MSPRs refer to a process in which performance ratings are collected from multiple sources, including supervisors, subordinates, peers, and/or clients/customers (Atwater, Waldman, & Brett, 2002), and this information is typically used as feedback for employee development (Church & Bracken, 1997; Ghorpade, 2000; London & Smither, 1995). Recent research has criticized the psychometric properties and usefulness of MSPRs (Mount et al., 1998; Murphy, 2008; Viswesvaran et al. 2002, 2005).Whereas high levels of interrater reliability are necessary for

adequate measurement in most PA systems, MSPRs are founded on the assumption that raters from different levels provide unique and meaningful performance information (Borman, 1974; Lance et al., 2008). From this perspective, some level of cross-source disagreement is desirable, and source effects are not necessarily an indicator of poor quality ratings (Lance et al., 2008).

More problematic for MSPRs is the strong correlations among dimension factors and the relatively small magnitude of dimension variance (Hoffman & Woehr, 2009; Lance et al., 2008; Mount et al., 1998; Scullen, Mount, & Judge, 2003; Smither, London, & Reilly, 2005). For instance, Hoffman, Lance, Bynum, and Gentry (2010) investigated the structure of two popular MSPR instruments in two large independent samples and found that, on average, dimensions accounted for only 7% of the variance in MSPRs and that dimension factors were strongly correlated. This is a critical issue given that MSPRs are commonly used in developmental settings and that dimensions are the focus when interpreting and acting upon developmental feedback. In other words, rather than providing dimension-specific feedback, MSPRs often reflect a general impression associated with each rating source, making it difficult to focus on specific developmental areas. In fact, weak evidence for dimensions potentially accounts for the modest effect of MSPRs on performance improvement (Smither et al., 2005), leading some to urge more attention to dimension effects and the overlap among dimensions in MSPRs (Hoffman & Woehr, 2009; Lance et al., 2008).

Despite continuing questions concerning the psychometric soundness of MSPRs, existing research has rarely investigated methods to improve the psychometric quality of MSPRs. One potential cause for this disconnect is that the PA literature was moving away from design interventions such as rating scales and rater training just as MSPR research became popular. A second reason may be that the context of MSPRs is different from that of a traditional PA, making it difficult to apply methods originating from traditional PA settings. For instance, whereas traditional PAs are (ideally) based on a job analysis and specific to a particular job, many MSPR instruments are based on broader competency models (cf. Lombardo &McCauley 1994; London, 2001; Rogelberg & Waclawski, 2001) that generalize across many different jobs, professions, departments, organizational levels, and organizations (Lombardo &McCauley, 1994). Consequently, it is difficult to develop concrete scores for each behavioral anchor needed for a BARS instrument and FORT (Bernardin, Tyler, & Wiese, 2001; Hauenstein, 1998). Similarly, FORT is time consuming and expensive (Stamoulis & Hauenstein, 1993), and these costs would exponentially increase in MSPR contexts due to increasing the number of raters to be trained (Timmreck & Bracken, 1995).

However, the central premise of many of these interventions, the need to provide raters a common frame of reference before they provide ratings, seems especially important in the context of MSPRs. MSPRs are characterized by diverse sets of raters with limited experience evaluating others, a lack of familiarity of the demands associated with the job being rated, and different perspectives on the nature of effective performance (Hoffman & Woehr, 2009; Hooijberg & Choi, 2000; Lance et al., 2008). In this context, interventions designed to ensure that different raters rely on the same standard have the potential to be particularly beneficial.

Study 1

Study 1 describes the evaluation of an instrument designed to improve the quality of performance ratings and to be amenable to incorporation into a MSPR framework. The proposed scale, referred to as Frame of Reference Scales (FORS), adds dimension definitions and examples of effective and ineffective behaviors to a set of behavioral items associated with each dimension (see Appendix A). FORS are designed to increase the likelihood that raters evaluate performance using a common definition and distinguish effective from ineffective performance using a common standard. Accordingly, we propose that FORS will result in more favorable psychometric characteristics relative to standard MSPR instruments. To be clear, the FORS approach draws from previous approaches to improve the quality of performance ratings. In other words, despite many variations, prior methods designed to increase rating quality (e.g., BARS, BSS, BOS, and FORT) often involve providing more specific, concrete behavioral information for raters to use as a reference in evaluating coworker performance (Murphy, 2008). FORS is based on similar underlying principles as these approaches; however, there are a few key features that distinguish FORS from past endeavors.

First, FORS differs from FORT by providing performance definitions and behavioral examples on the rating scale itself, rather than through training, and thus, is less onerous than FORT and more amenable to incorporation, to multirater contexts. Second, a primary difference between BARS and FORS is that BARS typically links specific behavioral critical incidents to a specific scale point, whereas FORS provides a few critical incidents labeled more broadly as examples of effective or ineffective performance. As noted by Hauenstein and Foti (1989), raters confuse BARS' behavioral anchors with concrete rating referents, rather than examples of one of many behaviors, and this confusion has been suggested to be detrimental to user reactions of BARS (Tziner, Joanis, & Murphy, 2000). Next, FORS uses multiple behavioral items per competency, whereas BARS and similar approaches (BSS) use a single item per competency

(Smith & Kendall, 1963). The reliance on a single rating per dimension is likely to reduce the reliability of the ratings (Borman, 1979) and potentially reduces the value of tools in developmental settings (Tziner et al., 2000). Given that MSPRs are often used in developmental settings, the potential for feedback based on multiple specific behavioral items is a key advantage of FORS relative to other scale design approaches.

In summary, approaches to evaluate the efficacy of scale design in field settings have been elusive, and the majority of past work has been conducted in lab settings. In addition, little research has attended to the design of multirater systems. Study 1 advances the literature by providing one of the few investigations of scale design in the context of MSPRs. In doing so, we demonstrate the use of CFA-based MTMM results to provide an alternative set of rating scale design criteria. More specifically, this paper contributes to the literature by providing the first analysis of the degree to which adding dimension definitions and examples of effective and ineffective performance to a traditional MSPR instrument improves the quality of MSPRs in terms of factor structure, factor correlations, and the magnitude of dimension and error variance.

Method

Participants

Three hundred twenty-one (19% female) professionals enrolled in an executive MBA program at a large southeastern U.S. university between the years 2004 and 2008 participated in this study. These participants were employed in a wide variety of industries and managerial positions. On average, the participants had 9.54 years supervisory experience ($SD = 7.35$) and supervised 8.00 direct reports ($SD = 8.88$).

Procedure

Prior to enrolling in the program, the participants' immediate supervisors and five direct reports were asked to complete an appraisal of the participants' managerial competencies. The participants were mailed the rating forms to be completed by their coworkers prior to beginning the EMBA program and were instructed to distribute the MSPR forms to their coworkers. To ensure anonymity, those completing the surveys were instructed to mail the MSPR forms directly to the university upon completion.

Between 2004 and 2005, 131 participants were evaluated using the standard MSPR form. Following 2005, the scale was evaluated and redesigned. During this redesign, the FORS aspect of the instrument was

added. Between 2006 and 2008, 190 participants were evaluated using the revised scale with the FORS. An average of 2.7 subordinate raters evaluated each manager. Because of changes made during the scale redesign, the dimensions and items varied across the two forms. However, across the two forms, seven dimensions assessed with 25 items were consistent across the two forms. This study focused on a comparison of these seven dimensions. Consistent with past work (Hoffman & Woehr, 2009), subordinate ratings were averaged prior to analyses.

Measures

Standard MSPR Form

The standard MSPR form was designed specifically for use in the EMBA program. The goal of the MSPR in this setting is to provide feedback on the competencies needed for effective managerial performance. Given the wide range of backgrounds of the EMBA participants, the dimensions were necessarily applicable to participants in different professions, organizations, and organizational levels. For instance, interpersonal sensitivity, problem-solving skills, and motivating followers are components of effective managerial performance, regardless of the context (Borman & Brush, 1993; Hoffman, Woehr, Maldegan, & Lyons, 2011). The original scale consisted of 80 items measuring 19 performance dimensions. Consistent with typical MSPR systems, the items were similar to behavioral observation scales by giving multiple specific behaviors to rate for each competency, and consistent with popular MSPR measures (e.g., Lombardo & McCauley, 1994), the standard scale did not present the rater with performance definitions and examples.

FORS

The revised scale included 17 performance dimensions measured with 71 items. After excluding items that were inconsistent across the forms, seven scales measured with 25 items were retained for this study. Dimensions included are (a) problem solving (5 items), (b) motivating followers (4 items), (c) participative leadership (3 items), (d) stress tolerance (3 items), (e) conflict management (4 items), (f) interpersonal sensitivity (3 items), and (g) performance management (3 items). All items were measured on a five-point Likert-type scale ranging from 1 = *strongly disagree* to 5 = *strongly agree*. However, in addition to the standard behavioral items, a brief definition of the dimension and an example of effective and ineffective performance were included. After reading the description,

participants provide their ratings on the same behavioral items as in the standard MSPR sample.

The FORS development team had approximately 30 years of combined experience in the assessment of managerial performance (through both MSPR systems and managerial assessment centers). The development team was also closely involved with the EMBA program and had close knowledge of the goals of the leadership development aspect of the program. The FORS team followed three primary steps in developing the instrument. First, interviews were conducted with the coordinators of the EMBA program to identify competencies to be included in the measure. Second, FORS definitions and examples of effective and ineffective performance for each dimension were generated independently by three of the members of the development team. The definitions were consistent with the definitions of associated constructs typically found in the broader management literature (e.g., *The Successful Managers Handbook,* Davis, Skube, Hellervik, Gebelein, & Sheard, 1996). In addition to relying on the broader literature, the examples were culled from the scoring procedures from a managerial assessment center. Finally, the definitions and critical incidents were compared to determine the final content of each dimension definition and critical incident with an emphasis on including those examples that are observable and that would be relevant across managerial positions.

Models Tested

To evaluate the structure of the two scales, a set of CFA-based MTMM models were compared based on the models outlined by Widaman (1985). The evaluation of these models is consistent with construct validity research (Campbell & Fiske, 1959) and with CFA-based applications of Campbell and Fiske's MTMM approach to MSPRs (Lance et al., 2008). The first model specified only performance dimensions. This model specifies both supervisor and followers' ratings of the same dimension to load on the same dimension factor, resulting in seven dimensions (Model 1). The next model specified source effects only, such that all items rated by a single source load on a single factor, resulting in a supervisor and subordinate source factor (Model 2). Next, the traditional source and dimension MTMM model (Model 3) was specified. In such models, construct validity evidence for dimensions is provided to the degree that dimension factors are relatively large, larger than source factors, and correlations among dimensions are weak (Lance et al., 2008). In essence, supporting the construct validity and distinguishability of the performance dimensions provides evidence that the performance behaviors are being accurately evaluated and

TABLE 1
Means, Standard Deviations, and Coefficients Alpha for Standard MSPR and FORS

	Standard MSPR			FORS		
Dimension x source	*M*	*SD*	α	*M*	*SD*	α
Supervisor						
Participation	3.93	0.59	0.72	3.88	0.78	0.83
Motivating followers	4.19	0.63	0.68	4.01	0.72	0.76
Stress tolerance	4.02	0.81	0.86	3.85	1.02	0.95
Problem solving	4.04	0.53	0.83	4.05	0.65	0.84
Conflict management	3.84	0.58	0.62	3.98	0.58	0.66
Sensitivity	4.17	0.65	0.77	4.08	0.74	0.78
Performance management	3.86	0.80	0.86	3.83	0.83	0.91
Subordinate						
Participation	4.11	0.40	0.75	4.03	0.48	0.81
Motivating followers	4.20	0.46	0.67	4.06	0.48	0.78
Stress tolerance	4.17	0.61	0.94	4.02	0.71	0.96
Problem solving	4.19	0.36	0.82	4.13	0.43	0.89
Conflict management	3.84	0.38	0.62	3.95	0.46	0.69
Sensitivity	4.23	0.52	0.83	4.16	0.53	0.89
Performance management	4.06	0.51	0.83	3.87	0.58	0.91

distinguished; both of which are critical for useful feedback (Lance et al., 2008). Given that the correlated trait-correlated method model is among the "most widely accepted and implemented models" (Lance et al., 2008, p. 224), we used this parameterization that explicitly models both source and dimension factors, allows source factors to correlate with other source factors, allows dimension factors to correlate with other dimension factors, and sets the source and dimensions correlations equal to zero (Widaman, 1985).

Results

Table 1 presents scale means, scale standard deviations, and coefficients alpha reliabilities for the seven dimensions associated with the FORS and the standard MSPR. Subordinate ratings were aggregated at the item level prior to analyses. Scale means, alphas, and standard deviations were calculated on the average of subordinate ratings.

For the most part, the individual dimensions met accepted standards for reliability. Although the reliability coefficients were similar across the FORS and standard MSPR, the reliability estimates associated with the FORS dimensions were generally higher than those from the standard scale. In addition, although the differences were generally small, the FORS

TABLE 2
Model Fit Statistics for Standard MSPR and FORS

	χ^2	*Df*	CFI	TLI	RMSEA	SRMSR	$\Delta\chi^2$	Δdf
Standard MSPR								
1. 7-dimensions[1]	3474.73	1154	0.793	0.781	0.184	0.168		
2. 2-sources	3257.97	1174	0.814	0.806	0.143	0.123		
3. 7-dimensions + 2 sources	2246.3	1103	0.899	0.886	0.075	0.093	1011.67*	71
FORS								
1. 7-dimensions	5169.38	1154	0.803	0.791	0.181	0.191		
2. 2-sources	5105.75	1174	0.807	0.799	0.167	0.122		
3. 7-dimensions + 2 sources	2800.99	1103	0.917	0.908	0.091	0.098	2304.76*	71

Note. *Significant at $p < .001$.[1]Solution was inadmissible.

dimensions were characterized by a lower mean and a higher degree of variability.

To conduct the confirmatory factor analyses, we input the 50 x 50 item level correlation matrix (25 items x two rater sources) into LISREL version 8.5. Model fit indices for both the standard and FORS are presented in Table 2. The rank order of the three models tested was consistent with the findings of past research investigating the structure of MSPRs (Hoffman et al., 2010; Scullen et al., 2003)[1]. For both scales, the 7-dimension, 2-source model (Model 3) provided the closest fit to the data. Consistent with past construct-related validity research, we focused on the parameter estimates (e.g., factor loadings, latent correlations, error) associated with each rating scale. The standard scale was characterized by sporadic negative and nonsignificant loadings on the dimension factors (30% of the dimension loadings were negative or nonsignificant) and strong and sometimes negative correlations among latent dimension factors (Table 3). For instance, participation was negatively related with interpersonal sensitivity and stress tolerance. Although such anomalous loadings are not out of the ordinary for CFA-based MTMM results (cf. Hoffman, Melchers et al., 2011; Lance et al., 2000), it is noteworthy that all of the factor loadings on the FORS were positive and significant. In short, the

[1]Based on research supporting a general factor in models of performance (Hoffman et al., 2010; Viswesveran et al., 2005), a fourth model, adding a general factor to Model 3, was originally included in analyses. This model provided a closer fit to the data than did the other models but returned a mildly inadmissible solution in the standard scale sample with a negative error term and a path loading of 1.22. Based on the recommendation of an anonymous reviewer, this model was removed from the manuscript. Importantly, the primary conclusions regarding the efficacy of FORS relative to the standard scale do not change based on the inclusion of the general factor.

TABLE 3
Latent Factor Correlations for Standard MSPR and FORS

Standard MSPR ($n = 130$)	1	2	3	4	5	6	7	8	9
1. Participation	1.00								
2. Motivating followers	0.90	1.00							
3. Stress tolerance	−0.29	−0.43	1.00						
4. Problem solving	0.81	0.72	−0.35	1.00					
5. Conflict management	0.99	0.90	−0.35	0.94	1.00				
6. Sensitivity	−0.41	−0.19	0.64	−0.32	−0.23	1.00			
7. Performance management	0.80	0.89	−0.38	0.73	0.75	−0.24	1.00		
8. Subordinate	–[a]	–	–	–	–	–	–	1.00	
9. Supervisor	–[a]	–	–	–	–	–	–	0.28	1.00
FORS ($n = 191$)									
1. Participation	1.00								
2. Motivating followers	0.66	1.00							
3. Stress tolerance	0.15	0.15	1.00						
4. Problem solving	0.75	0.58	0.18	1.00					
5. Conflict management	0.54	0.66	−0.09	0.65	1.00				
6. Sensitivity	0.16	0.10	0.53	0.24	−0.25	1.00			
7. Performance management	0.56	0.76	0.05	0.56	0.59	0.09	1.00		
8. Subordinate	–[a]	–	–	–	–	–	–	1.00	
9. Supervisor	–[a]	–	–	–	–	–	–	0.18	1.00

Note. [a]Latent factor correlations with source factors were set to zero for model identification. For the standard scale, r 0.16, $p < 0.05$; r 0.34, $p < 0.01$; For the FORSs, r 0.20, $p < 0.05$, for r 0.25, $p < 0.01$.

FORS solution was much cleaner than that of the standard MSPR. Given persistent anomalous loadings associated with such MTMM CFA results and the difficulty in interpreting the results of such models, the relatively clean solution associated with FORS is an unexpected advantage of this tool.

The factor loadings and latent factor correlations also point to the superiority of the FORS. Specifically, dimensions explained 60% more variance in the FORS relative to the standard scale (31% and 19%, respectively), suggesting that when measured using the FORS, dimensions were characterized by greater levels of construct-related validity. Similarly, the FORS was associated with a sizeable decrease in the average correlation among dimensions relative to the standard scale (mean latent dimension intercorrelation = 0.40 and 0.58, respectively). Thus, dimension effects were stronger, and the dimensions were more clearly distinguishable when ratings were provided using FORS. Given the importance of concrete, specific feedback in behavioral settings, the stronger support for

differentiated dimensions is a key advantage to the use of FORS in developmental contexts. In addition, the FORS was associated with a 10% reduction in error variance (50% and 40%, respectively), another important psychometric characteristic in scale development (Nunnally & Bernstein, 1994). Finally, source factors accounted for 30% of the variance in FORS relative to 32% of the variance in the standard scale, indicating that the rating scale had little influence on source factors.

Supplemental Analyses

It has recently been argued that reactions to PA systems may be as important as the system's psychometric soundness (DeNisi & Sonesh, 2011). To evaluate raters' reactions to FORS, we included a single item inquiring as to whether the raters found the dimension description and examples of effective and ineffective performance useful in making their ratings. The mean for this item was 4.22, and 87% of raters either agreed or strongly agreed that the dimension definition and examples were helpful. Because this item was added for the purposes of the scale redesign, it was not possible to compare this value to the standard scale. Still, the relatively positive response to this item suggests favorable rater reactions to FORS.

Study 1 Discussion

Despite persistent criticism that measurement scale features offer minimal influence on the quality of performance ratings (Landy & Farr, 1980; Murphy, 2008), the results of Study 1 indicate that the addition of dimension definitions and examples of effective and ineffective performance to a traditional MSPR scale improved the psychometric quality of performance ratings by (a) yielding a cleaner pattern of dimension factor loadings, (b) increasing the magnitude of dimension variance in MSPRs by 60%, (c) decreasing the percent overlap between dimensions from 34% to 16%, and (d) decreasing the amount of error in the measurement of performance by 10%. As noted above, given persistent concerns as to the construct-related validity of MSPR dimensions (Lance et al., 2008), the importance of accurately distinguishing dimensions to the usefulness of feedback, and the importance of reducing measurement error, there reflect important advantages of FORS relative to the standard scale.

In addition to demonstrating an approach for improving the quality of information received from MSPRs, this study contributes to the literature by demonstrating an approach to evaluate rating design interventions that is amenable to use in field settings. Specifically, MTMM-based CFA of the two rating scales showed promise as a means of comparing the efficacy of

scale design alterations in field settings. Importantly, although MTMM-based analyses are a well established approach to evaluate construct validity of performance ratings, this approach has not yet been applied to evaluating the influence of scale design on the quality of MSPRs. Despite these contributions, this study suffers from the same limitation as other rating scale field research; namely, it is impossible to derive true scores of performance in a field setting (Bernardin et al., 2001), prohibiting the evaluation of accuracy. Accordingly, although FORS enhanced the construct related validity of MSPRs, it is important to determine the influence of FORS on rating accuracy.

Study 2

Developing different relationships across organizational levels that are instrumental to the rationale of MSPRs would be a challenge in an artificial setting. Nevertheless, lab studies can shed light on the influence of scale design on rating accuracy. Study 2 supplements the insights from Study 1 by investigating the accuracy of ratings using a laboratory-based target score design. By comparing participants' ratings to "true score" ratings generated by subject matter experts (SMEs) in a lab setting, this approach facilitates inferences as to the capacity of FORS to improve rater accuracy. Given the evidence from Study 1 that FORS increase rating quality, we hypothesize that:

Hypothesis 1: Ratings made using FORS will be more accurate than ratings made using a standard rating scale.

Moreover, given that the FORS were developed by incorporating FORT principles, FORS has potential to serve as a more efficient alternative to FORT. However, in order to draw this conclusion it is necessary to compare FORS and FORT. Although rating scale design (e.g., BARS and FORS) and FORT use similar approaches to improve ratings, FORT has seen consistent support in the literature (Woehr & Huffcutt, 1994), whereas scale redesign is largely viewed as an ineffective intervention (cf. Landy & Farr, 1980; Murphy, 2008). As previously alluded to, one potential reason for the disparate findings is that scale redesign research largely used rater errors as a criterion, whereas FORT research tended to use rater accuracy (DeNisi, 1996). Based on developments in rater training and its impact on rating quality (Woehr & Huffcutt, 1994), it is evident that providing a common frame of reference has a positive influence on rating accuracy. Because FORS is based on the same social information processing foundation as FORT, it is expected that both will have a positive influence on rating accuracy. However, to demonstrate whether this is the case, it is important to investigate the proposed measurement system

relative to alternatives, such as FORT. Accordingly, in addition to extending Study 1 by directly investigating rater accuracy, Study 2 further extends Study 1 by comparing the efficacy of FORS relative to FORT. Given the ubiquitous finding that FORT improves rating accuracy (Roch, Mishra, Kieszczynka, & Woehr, 2010; Woehr & Huffcutt, 1994), we hypothesize:

Hypothesis 2: Participants who receive FORT will have more accurate ratings than individuals in the control training condition.

Finally, although we propose that both FORS and FORT should increase accuracy, it is difficult to derive any rationale with which to hypothesize that either will be more effective than the other. Therefore, we do not offer any specific predictions of the relative effectiveness of FORS relative to FORT.

Research Question 1: Will there be any significant differences in rating accuracy between FORT and FORS?

Method

Participants

One hundred and fifty-one undergraduate students from a regional southwestern U.S. university were solicited to participate in the present study. The mean age of participants was 21.09 years ($SD = 4.53$), and most held part-time jobs (66%). The sample was predominantly Caucasian (65%; 25% Hispanic) and female (55%). Fifty-three percent of the sample reported that they had no experience rating the job performance of another person.

Procedure

Participants were randomly assigned to either a FORS condition ($n = 52$), FORT condition ($n = 49$), or a control training condition ($n = 50$). Before each session, participants received a brief introduction to the session and then received instructions corresponding to their assigned condition. Participants then viewed two videotaped performance episodes (described below) that were presented in random order across individual participants. At the conclusion of each performance episode, participants recorded their ratings in the spaces provided on the rating form. Upon viewing and rating all of the episodes, participants completed a demographic questionnaire.

Stimulus Materials

The two performance episodes that served as the stimuli in the present study have been used in prior FORT research (Gorman & Rentsch, 2009) and consistent with past work (Schleicher & Day, 1998; Sulsky & Day, 1992, 1994) depicted a meeting between a manager and direct report. In this study, the participants were instructed to evaluate the manager's performance (rather than the subordinate), and one performance episode depicted a relatively effective performer and the other a relatively ineffective performer. The exercises were designed to elicit behaviors relevant to the following performance dimensions: analysis, decisiveness, leadership, confrontation, and interpersonal sensitivity.

In order to assess rating accuracy, comparison (target score) performance ratings were collected from SMEs. Using procedures recommended by Sulsky and Balzer (1988), three upper-level industrial- organizational psychology graduate students independently observed and rated the recorded episodes. Each of the SMEs in the present study had previously received intensive 30-hr training over 6 days, followed by annual day-long review training for their roles as assessors and regularly conducted assessments for both administrative and developmental purposes. Thus, SMEs were extremely familiar with the scenario depicted in the performance episode and the dimensions being rated. After independently rating the videos, the SMEs met to discuss rating differences and, through consensus, generated a set of comparison scores.

Conditions

All conditions were conducted by trained graduate students using standard procedures, where participants (a) were told they would be rating performance, (b) were supplied either the FORS and control training (Condition 1), the standard scale and control training (Condition 2), or the standard scale and FORT (Condition 3); (c) observed the performance episodes; and (d) were provided ratings.

FORS

Participants in the FORS condition were given the FORS (See Appendix B) and instructed to read the dimension definitions, examples of effective and ineffective behaviors, and scale anchors as the trainer read them aloud. To ensure that the time frame matched between the training sessions, the trainer presented a video that broadly described the PA process. They then observed the performance episodes and provided ratings on the FORS.

The FORS instrument used in Study 2 differed from the instrument used in Study 1 in two ways. First, the FORS in Study 1 asked participants to provide ratings on each item associated with each dimension. However, because this study centered on observational accuracy (see Noonan & Sulsky, 2001; Woehr & Huffcutt, 1994) and consistent with past work (Gorman & Rentsch, 2009), we adapted this format by incorporating behavioral observation checklists within each dimension. In addition, whereas we used seven performance dimensions in Study 1, we used the five dimensions that were originally included in the rating stimuli episodes.

Control

Participants in the control training were presented with the standard rating scale, which included dimension definitions and behavioral checklist items and were instructed to read along as the trainer read each of the dimension definitions aloud. To maintain consistency in session length with the training session, they were shown a video describing PA. The control training session also lasted approximately 45 minutes. They then observed the performance episode and provided ratings.

FORT

The FORT proceeded according to Pulakos' (1984, 1986) protocol. The participants were provided with the same rating scale as in the control condition and were instructed to read along as the trainer read the dimension definitions aloud. Next, the trainer discussed ratee behaviors that illustrated different performance levels for each scale. Participants were then shown a videotape of a practice vignette and were asked to evaluate the ratee using the scales provided, and the ratings were written on a blackboard and discussed by the group of participants. Finally, the trainer provided feedback to participants explaining why the ratee should receive a particular rating (target score) on a given dimension. The entire training session lasted approximately 45 minutes. Participants then observed the performance episodes and provided ratings using the same scale as used in the control condition.

Rating Form

The rating form for all three conditions provided each dimension name, definitions of each dimension, a set of behavioral checklist items for each dimension, and a set of scale anchors. Thus, the only difference between the conditions in terms of the rating scale used was the inclusion

of the examples of effective and ineffective performance associated with the FORS condition. Each dimension was rated using an 11-point Likert-type rating scale (1.0 = *extremely weak*, 1.7 = *very weak*, 2.0 = *weak*, 2.5 = *moderately weak*, 2.7 = *slightly weak*, 3.0 = *satisfactory*, 3.5 = *effective*, 3.7 = *very effective*, 4.0 = *highly effective*, 4.5 = *extremely effective*, 5.0 = *exceptional*). Although this rating scale is somewhat unconventional, we elected to retain the rating scale that had been previously established for the AC on which the performance episodes were based. Because we were interested in the use of FORS, rather than the rating scale, none of the participants received instruction on behaviors associated with specific scale points. Instead, like raters in a traditional performance rating context, the participants were instructed to use the anchors and their behavioral checklist ratings to provide a rating on each of the five dimensions.

Rating Accuracy

Using the formulae provided by Sulsky and Balzer (1988), rating accuracy (RA) was assessed via Cronbach's (1955) four indices of rating accuracy: (a) elevation accuracy (EA), (b) differential elevation (DE), (c) differential accuracy (DA), and (d) stereotype accuracy (SA). Each index reflects a different conceptualization of the distance between participants' ratings and the target scores derived from the SMEs. EA represents the differential grand mean between the manifest ratings and target ratings and is interpreted as an index of overall accuracy. DE represents the differential main effect of ratees and is interpreted as an index of the accuracy with which a rater distinguishes between ratees across dimensions. SA refers to the differential main effect of dimensions and is interpreted as an index of the accuracy with which a rater discriminates among performance dimensions across ratees. Finally, DA refers to the differential ratee by dimension interaction and is interpreted as an index of the accuracy with which a rater identifies individual patterns of strengths and weaknesses (Jelley & Goffin, 2001; Sulsky & Balzer, 1988). Lower scores on these measures represent higher accuracy, whereas higher scores indicate lower levels of accuracy. Borman's (1977) correlational measure, differential accuracy (BDA), was also calculated. BDA measures the correlation between ratings on each dimension and the corresponding target scores and is typically considered an index of rating validity (Sulsky & Day, 1994).

Results

Means, standard deviations, and intercorrelations among study variables are reported in Table 4.

TABLE 4
Means, Standard Deviations, and Intercorrelations Among Study 2 Variables (N = 151)

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Gender[a]	1.43	0.50	–								
2. Age	20.98	4.56	−0.09	–							
3. GPA	2.91	0.55	−0.07	−0.16	–						
4. Rating experience	2.37	6.36	0.09	0.18	−0.03	–					
5. EA	0.63	0.23	0.09	0.13	0.11	−0.06	–				
6. DE	0.32	0.24	0.05	0.06	0.14	0.04	0.43	–			
7. SA	0.30	0.13	−0.03	0.07	0.01	−0.09	0.35	0.03	–		
8. DA	0.21	0.09	0.01	−0.07	0.14	0.02	0.12	−0.01	0.27	–	
9. BDA	0.86	0.49	−0.13	0.04	0.17	0.16	−0.17	0.04	−0.32	−0.13	–

Note. r 0.17, $p < 0.05$; r 0.21, $p < 0.01$. GPA = grade point average. Rating experience = total number of times having rated the job performance of another person. EA = elevation accuracy. DE = differential elevation. SA = stereotype accuracy. DA = differential accuracy. BDA = Borman's differential accuracy. [a] 1 = *female*, 2 = *male*.

FORS and Rating Accuracy

Hypothesis 1 predicted that participants in the FORS condition would produce more accurate ratings than would control participants (Table 5). Consistent with previous rating accuracy research (Gorman & Rentsch, 2009; Schleicher, Day, Mayes, & Riggio, 2002), we used a multivariate framework to test the hypotheses. Planned contrast MANOVA, examining the difference between FORS versus control on the five rating accuracy indices (i.e., the set of dependent variables) revealed that participants in the FORS condition provided more accurate ratings than participants in the control condition, $F(5, 96) = 4.08$, $P < 0.01$; Wilks's $\Lambda = 0.83$, partial $\eta^2 = 0.18$. A discriminant analysis revealed one significant eigenvalue ($P < 0.01$), with condition accounting for 100% of the variance in the accuracy composite. The structure coefficients from this analysis indicated that BDA, DA, and EA were driving the discrimination between the two conditions (0.93, -0.35, and -0.33, respectively). Follow-up univariate ANOVAs supported this finding (Table 5). Together, these results support Hypothesis 1.

FOR Training and Rating Accuracy

Hypothesis 2 predicted that participants in the FORT condition would produce more accurate ratings than would the control. Planned contrast MANOVA, testing the difference between FORT versus control and the five rating accuracy indices as the multiple dependent variables, confirmed this prediction, $F(5, 92) = 4.32$, $p < 0.01$; Wilks's $\Lambda = 0.81$, partial

TABLE 5
Means, Standard Deviations, and Rating Accuracy Test Results Across Training Conditions in Study 2

	Condition		
Accuracy index	FOR scales	FOR training	Control
	($n = 52$)	($n = 49$)	($n = 50$)
Elevation accuracy			
M	0.63_a	0.56_b	0.71_c
SD	0.24	0.19	0.25
Differential elevation			
M	0.31	0.25_a	0.36_b
SD	0.24	0.21	0.24
Stereotype accuracy			
M	0.28	0.29	0.31
SD	0.11	0.14	0.14
Differential accuracy			
M	0.20_a	0.21	0.24_b
SD	0.09	0.08	0.12
Borman's differential accuracy			
M	1.03_a	0.92_a	0.62_b
SD	0.46	0.37	0.53

Note. Lower values on elevation, differential elevation, stereotype accuracy, and differential accuracy denote greater accuracy. Higher values on Borman's differential accuracy represent greater accuracy. Means with different subscripts are significantly different at $p < 0.01$. FOR = frame of reference.

$\eta^2 = 0.19$. A discriminant analysis revealed one significant eigenvalue ($p < 0.01$), with condition accounting for 100% of the variance in the accuracy composite. The structure coefficients from this analysis indicated that BDA, EA, and DE were driving the discrimination between the two conditions (0.69, -0.69, and -0.47, respectively). Follow-up univariate ANOVAs supported this finding. These results support Hypothesis 2.

FOR Scales Versus FOR Training

To address Research Question 1, we used planned contrast MANOVA, examining the difference between FORS versus FORT on the five rating accuracy indices as the dependent variables. Results revealed no significant difference between the two conditions, $F(5, 94) = 1.18$, *ns*; Wilks's $\Lambda = 0.94$, partial $\eta^2 = 0.06$. Follow-up univariate ANOVAs indicated no significant differences for each of the rating accuracy indices except EA, $t(1.72)$, $P < 0.01$. Together, these results indicate that FORS and FORT are characterized by similar levels of accuracy.

Study 2 Discussion

Consistent with the enhanced psychometric quality revealed in Study 1, Study 2 showed that FORS were effective at improving rating accuracy compared to a control group. We also supplemented this investigation by comparing accuracy gains stemming from FORS to accuracy increases associated with FORT. The effect size for the influence of FORS on accuracy (partial $\eta^2 = 0.18$) was roughly the same as the effect size for the FORT program on accuracy (partial $\eta^2 = 0.19$), suggesting that FORS might be as useful as FORT in improving rating accuracy. Furthermore, the effect size for both FORS and FORT were similar to the multivariate effect size reported in Schleicher et al.'s (2002) study of FORT effectiveness (partial $\eta^2 = 0.24$), supporting the generalizability of our findings. Implications of these findings are outlined in the general discussion.

Although these results provide important insights regarding the quality of FORS-derived ratings, conducting the study in a lab setting hinders the ability to generalize these results to field settings. However, the support for the efficacy of FORS across lab and field settings bolsters the generalizabiltiy of the results. In addition, FORS were specifically designed to be used in multisource rating contexts; however, because multisource evaluations are predicated on the assumption that a preexisting relationship with the rater fundamentally alters the performance information (Hoffman & Woehr, 2009; Lance et al., 2008), it is difficult to study multirater systems in lab settings. Nevertheless, the supportive evidence from the single rater lab study shows that the value of FORS is not contingent on a multirater system. Instead, FORS are potentially useful in single rater systems as well. Future research applying FORS to settings where single raters provide ratings for research/developmental purposes (e.g., ratings of leadership or organizational citizenship behaviors) has the potential to enhance the quality of information collected in a variety of research areas. DeRue, Nahrgang, Wellman, & Humphrey (2011) directed to "revise existing measures of leader behavior such that we can better capture the conceptual distinctions among leader behaviors" (p. 38). FORS provide a useful first step in this direction.

General Discussion

Our results demonstrate a method to improve the quality of information gained from multisource performance assessments. This conclusion is bolstered by converging evidence from both a field and laboratory setting, using two distinct criteria of rating quality, and two different versions of the rating instrument. In doing so, this study demonstrates the application of a novel method and criterion to judge the efficacy of performance rating

design interventions. (i.e., the comparison of rating scales using parameter estimates from CFA-based MTMM models). More generally, this study contributes to the literature by presenting one of the first examinations of scale design approaches to enhance the quality of MSPR instruments.

Integration of Key Findings

Beyond the general support for FORS, several noteworthy trends emerged when comparing the results across Study 1 and Study 2. First, the finding in Study 2 that FORS were associated with higher levels of elevation accuracy (or overall accuracy) parallels the results of Study 1 that FORS were characterized by less error variance. Specifically, elevation accuracy is an index of overall accuracy, indicating that, across dimensions, performance was rated more accurately. Similarly, the amount of error in the CFA models reflects the influence of systematic (dimension plus source) relative to random (error) variance. Accordingly, the results of Study 1 suggesting that, overall, FORS are associated with less error are consistent with the results of Study 2 that overall, FORS were associated with more accurate ratings. Thus, the results of both studies indicate that ratees are ranked more accurately with FORS relative to the standard scale. These findings are important in circumstances requiring the use of overall performance measures (e.g., research or administrative settings) because they indicate that, across dimensions, performance is rated more accurately and with less error when using FORS. More concretely, evidence that FORS were characterized by a decrease in non systematic (error) variance clearly supports the value of this method.

Next, the finding in Study 2 that FORS were associated with higher levels of DA and BDA (accurately distinguishing strengths and weaknesses of a given ratee) relative to the control scale is consistent with the finding of larger dimension loadings and weaker dimension correlations for the FORS in Study 1. Collectively, these findings indicate that FORS are a particularly useful tool for capturing dimensional information. Support for dimension effects is fundamental to the construct validity of measures (Campbell & Fiske, 1959; Lance et al., 2008) and is particularly important to MSPR settings where dimensions are often the focus of feedback and subsequent development (Hoffman & Woehr, 2009). Although DE is proposed to be the most important form of accuracy in administrative settings (Murphy, Garcia, Kerkar, Martin, & Balzer, 1982), in developmental settings where dimensional feedback is the focus, DA is arguably more important (Cardy & Dobbins, 1994). Given that MSPRs are used in developmental contexts, the ability to give specific dimensional feedback and accurately distinguish strengths and weaknesses is a key strength of FORS.

Finally, the finding in Study 2 that FORS did not enhance SA is consistent with the CFA findings in Study 1 that the FORS were associated with a similar portion of variance due to source effects relative to the traditional scale. As noted by Werner (1994), "if the intercorrelation among dimensions is higher. . .then this should also influence stereotype accuracy, that is, the ability to accurately capture performance on each dimension across each source." (p. 100). Similarly, source factors reflect covariance among all ratings provided by a given rater source. In this way, the consistency in results across Study 1 and Study 2 is not surprising. To the degree that source factors are interpreted as substantively meaningful source-specific perspectives on performance (Hoffman & Woehr, 2009), rater general impression has been shown to introduce valid variance to performance ratings (Nathan & Tippins, 1994), and the usefulness of MSPRs are predicated on source-based differences in performance ratings (Lance et al., 2008), the lack of impact on source effects is not necessarily a limitation of FORs. In any case, the positive influence of FORS appears to be specific to enhancing the support for dimensions rather than altering the influence of source-specific general impression. Together, the consistency in the pattern of results across the two settings and across different criterion variables strengthens the generalizability of the findings.

Comparison With Prior Research and Theoretical Implications

These findings have several key implications for PA research. First, the idea that the design of rating scales does little to enhance the quality of performance ratings must be reevaluated. Although advances in scale development have been sparse since 1980 (Borman et al., 2001), the present study joins a handful of studies in showing that pessimism regarding the value of rating scale design may be overstated (cf., Borman et al., 2001; Wagner & Goffin, 1997; Woehr & Miller, 1997). Although we do not necessarily advocate a renaissance in scale design research, as PA research has moved in other important directions, we do advocate a reevaluation of the efficacy of rating scale design.

In addition, the FORS instrument seems particularly useful by compensating for many of the criticisms of BARS. For instance, because the anchors used in BARS are very specific, raters may have difficulty matching the behavioral anchors with the observed performance (Borman, 1979). As a result, the rater is forced to at least partially base their ratings on inferences (Murphy & Constans, 1987) or choose between two anchors that both describe the ratee (Bernardin & Smith, 1981). FORS addresses these limitations by not directly linking specific examples to a rating scale.

A common criticism against PA design research has been the reliance on artificial laboratory settings. Study 1 advances the literature by demonstrating the use of an alternative to laboratory-based indices by examining the factor structure of FORS in a sample of managers from diverse industries. Although MTMM-based analyses are well established as a means of evaluating the construct validity of performance ratings, this method has rarely been applied to evaluating scale design differences. The application of MTMM-based CFA to rating scale evaluation has several benefits. First, this method provides a means to examine scale design changes in field settings. Second, it allows for the isolation of specific components of rating quality that have been improved. As demonstrated above, interpreting the proportion of variance attributable to dimensions, sources, and error has the potential to offer a richer understanding of the influence of scale design alterations than would be provided by more traditional scale design criterion variables (e.g., rater error or scale reliability).

Although our results generally support the influence of FORS on rating quality, the improvement above traditional scales is not straightforward, and the magnitude of effects varied from weak to moderate. For instance, in Study 1, although FORS accounted for twice the dimension variance as the standard MSPR scale, dimension variance only improved by around 10% in absolute terms. Similarly, although we saw increased accuracy in the laboratory setting, the largest improvements were still relatively moderate in terms of effect size. Nevertheless, the FORS approach is less complex and less expensive than the development and implementation of BARS and FORT. For instance, because the FORS were designed around general competencies relevant to most managerial jobs, a job analysis was not necessary; we instead relied on the voluminous literature on managerial effectiveness (Borman & Brush, 1993). In addition, given that the formal training of multiple raters from multiple levels is cost prohibitive in MSPR contexts, the FORS approach reflects a cost-effective solution relative to other popular approaches such as FORT. Given the centrality of PA to organizational functioning and the prominence of MSPRs in modern organizations, even relatively small gains in the quality of performance information have the potential to make a meaningful impact on a variety of functions. Thus, from a utility perspective, it might reasonably be argued that the gains associated with FORS outweigh the costs.

Limitations and Avenues for Future Research

Although the generalizability of our findings is bolstered by support across two settings, it is important to remember that these findings are based on only two samples. However, the standard MSPR instrument was generally consistent with MSPR instruments found in other settings

by assessing general managerial competencies using behavioral items. In addition, the consistency in results between the standard MSPR scale in Study 1 and other CFA-based investigations of MSPRs is encouraging (cf. Hoffman et al., 2010). Nevertheless, research replicating these findings in different settings with different performance instruments is needed. A related limitation is that the development of FORS was constrained by organizational goals associated with a larger scale development. In other words, we were only able to use a subset of items and scales that were actually measured by the MSPR instrument. Although this is certainly a limitation, the items we did analyze were consistent across the standard and FORS administrations, and thus, this did not impact the comparison of the rating scales.

Next, our findings only pertain to the use of FORS in laboratory and developmental settings. It is unclear whether the benefits of FORS will also be seen in administrative settings. In administrative PA, rater motivation to distort ratings certainly plays a greater role than in the laboratory and likely a greater role than in the context of developmental ratings (Murphy, 2008). In addition, although the efficiency of FORS development is a strength in terms of cost, it is unclear whether providing general examples of effective and ineffective performance, rather than tying them to a specific rating, as with BARS and FORT, will withstand legal scrutiny. However, this is not a key limitation to the use of FORS in their intended setting, employee development. In addition, it is possible that FORT will be more efficacious than FORS in field settings, where time between behavior observation and performance evaluation is more staggered. We recommend that organizations that have the resources available should invest in FORT programs for single-rater rating systems, at least until additional research can verify the value of FORS. However, for organizations that are deterred by the practical limitations of a full training program for a MSPR system, FORS represent a practical, yet methodologically rigorous option for improving rating quality.

Future research on FORS should also consider alternative criteria, such as rater and ratee reactions (MacDonald & Sulsky, 2009) and other psychometric considerations such as rating validity (Bartram, 2007). Although the supplemental analyses from Study 1 suggest that raters have generally positive reactions to FORS, much more work is needed in this area. In addition, further work might also examine the cognitive effects of FORS. Gorman and Rentsch (2009), for example, showed that FORT results in improved rating accuracy by influencing raters' mental models, or schemas, of performance. Future studies should incorporate this cognitive model to isolate the mechanisms that account for the effects of FORS.

Conclusions

Overall, the FORS approach shows promise as a method for improving the psychometric properties of performance ratings. Although it is clear that the inclusion of FORS increased the quality of performance ratings, the effects were generally small to moderate. However, given the central role that accuracy plays in the value of performance ratings and the ubiquity of MSPRs in modern organizations, the higher quality performance information associated with FORS outweigh the costs. More generally, these findings suggest that one's choice in rating scales can make a difference in performance rating quality.

REFERENCES

Atwater LE, Waldman DA, Brett JF. (2002). Understanding and optimizing multisource feedback. *Human Resource Management*, *41*, 193–208.

Austin JT, Crespin TR. (2006). Problems of criteria in industrial and organizational psychology: Progress, pitfalls, and prospects. In Bennett W, Jr., Lance CE, Woehr DJ (Eds.), *Performance measurement: Current perspectives and future challenges* (pp. 9–48). Mahwah, NJ: Erlbaum.

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

Bartram D. (2007). Increasing validity with forced-choice criterion measure formats. *International Journal of Selection and Assessment*, *15*, 263–272.

Becker BE, Cardy RL. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology*, *71*, 662–671.

Benson PG, Buckley MR, Hall S. (1988). The impact of rating scale format on rater accuracy: An evaluation of the mixed standard scale. *Journal of Management*, *14*, 415–423.

Bernardin HJ, Smith PC. (1981). A clarification of some issues regarding the development and use of behaviorally anchored rating scales (BARS). *Journal of Applied Psychology*, *66*, 458–463.

Bernardin HJ, Tyler CL, Wiese DS. (2001). A reconsideration of strategies for rater training. In Ferris GR (Ed.), *Research in personnel and human resources management* (pp. 221–274). Stamford, CT: JAI Press.

Borman WC. (1974). The rating of individuals in organizations: An alternate approach. *Organizational Behavior & Human Performance*, *12*, 105–124.

Borman WC. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior & Human Performance*, *20*, 238–252.

Borman WC. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology*, *64*, 410–421.

Borman WC, Brush DH. (1993). More progress toward a taxonomy of managerial performance requirements. *Human Performance*, *6*, 1–21.

Borman WC, Buck DE, Hanson MA, Motowidlo SJ, Stark S, Drasgow F. (2001). An examination of the comparative reliability, validity, and accuracy of performance ratings made using computerized adaptive rating scales. *Journal of Applied Psychology*, *86*, 965–973.

Campbell DJ, Fiske DW. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell DJ, Lee C. (1988). Self-appraisal in performance evaluation: Development versus evaluation. *The Academy of Management Review*, *13*, 302–314.

Cardy R, Dobbins GH. (1994). *Performance appraisal: A consideration of alternative perspectives*. Cincinnati, OH: South-Western.

Church AH, Bracken, DW (Eds.). (1997). Special issue: 360-degree feedback systems. *Group and Organization Management*, *22*, 149–161.

Conway JM. (1996). Analysis and design of multitrait-multirater performance appraisal studies. *Journal of Management*, *22*, 139–162.

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

Cooper WJ. (1981). Ubiquitous halo: Sources, solutions, and a paradox. *Psychological Bulletin*, *59*, 177–193.

Cronbach L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin*, *52*, 177–193.

Davis BL, Skube CL, Hellervik LW, Gebelein SH, Sheard JL. (1996). *Successful manager's handbook*. Minneapolis, MN: Personnel Decisions International.

Deming WE. (1986). *Out of the crisis*. Cambridge: MIT Institute for Advanced Engineering Study.

DeNisi AS. (1996). *Cognitive approach to performance appraisal: A program of research*. New York, NY: Taylor & Francis.

DeNisi AS, Sonesh S. (2011). The appraisal and management of performance at work. In Zedeck S (Ed.), *APA handbook of industrial and organizational psychology* (Vol. 1, pp. 255–281). Washington DC: APA Press.

DeRue DS, Nahrgang JD, Wellman N, Humphrey SE. (2011). Trait and behavioral theories of leadership: An integration and meta-analytic test of their relative validity. PERSONNEL PSYCHOLOGY. *64*, 7–52.

Fisicaro SA. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology*, *73*, 239–244.

Ghorpade J. (2000). Managing five paradoxes of 360-degree feedback. *Academy of Management Executive*, *14*, 140–150.

Goffin RD, Gellatly IR, Paunonen SV, Jackson DN, Meyer JP. (1996). Criterion validation of two approaches to performance appraisal: The behavioral observation scale and the relative percentile method. *Journal of Business and Psychology*, *11*, 23–33.

Gorman CA, Rentsch JR. (2009). Evaluating frame-of-reference rater training effectiveness using performance schema accuracy. *Journal of Applied Psychology*, *94*, 1336–1344.

Harris MM, Schaubroeck J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. PERSONNEL PSYCHOLOGY, *41*, 43–62.

Hauenstein NMA. (1998). Training raters to increase accuracy of appraisals and the usefulness of feedback. In Smither JW (Ed.), *Performance appraisal: State of the art in practice*. (pp. 404–411). San Francisco, CA: Jossey-Bass.

Hauenstein NMA, Foti RJ. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. PERSONNEL PSYCHOLOGY, *42*, 359–379.

Hoffman BJ, Baldwin SP. (2012). Modern managerial assessment: A comparison of assessment centers and multisource feedback (143–162). In G Thornton & N Povah (Eds.), *Assessment centers and global talent management*. Burlington, VT: Gower.

Hoffman BJ, Lance C, Bynum B, Gentry B. (2010). Rater source effects are alive and well after all. PERSONNEL PSYCHOLOGY, *63*, 119–151.

Hoffman BJ, Melchers K, Blair CA, Kleinmann M, Ladd, R. (2011). Exercises and dimensions are the currency of assessment centers. PERSONNEL PSYCHOLOGY, *64*, 351–395.

Hoffman BJ, Woehr DJ. (2009). Disentangling the meaning of multisource performance rating source and dimension factors. PERSONNEL PSYCHOLOGY, *62*, 735–765.

Hoffman BJ, Woehr DJ, Maldegan R, Lyons B. (2011). Great man or great myth? A meta-analysis of the relationship between individual difference and effective leadership. *Journal of Occupational and Organisational Psychology*, *84*, 347–381.

Hooijberg R, Choi J. (2000). Which leadership roles matter to whom? An examination of rater effects on perceptions of effectiveness. *The Leadership Quarterly*, *11*, 341–364.

James LR. (1973). Criterion models and construct validity for criteria. *Psychological Bulletin*, *80*, 75–83.

Jelley RB, Goffin RD. (2001). Can performance-feedback accuracy be improved? Effects of rater priming and rating-scale format on rating accuracy. *Journal of Applied Psychology*, *86*, 134–144.

Kingstrom PO, Bass AR. (1981). A critical analysis of studies comparing behaviorally anchored ratings scales (BARS) and other rating formats. PERSONNEL PSYCHOLOGY, *34*, 263–289.

Lance CE, Hoffman BJ, Gentry B, Baranik LE. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. Human Resource Management Review, *18*, *223–232*.

Lance CE, Newbolt WH, Gatewood RD, Foster MR, French NR, Smith DE. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance*, *13*, 323–353.

Lance CE, Woehr DJ. (1989). Statistical control of halo: Clarification from two cognitive models of the performance appraisal process. *Journal of Applied Psychology*, *71*, 679–685.

Landy FJ, Farr JL. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107.

Lombardo MM, McCauley CD. (1994). *BENCHMARKS® : A manual and trainer's guide*. Greensboro, NC: Center for Creative Leadership.

London M. (2001). *How people evaluate others in organizations*. Mahwah, NJ: Erlbaum.

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.

MacDonald HA, Sulsky LM. (2009). Rating formats and rater training redux: A context-specific approach for enhancing the effectiveness of performance management. *Canadian Journal of Behavioral Science*, *41*, 227–240.

Mount MK, Judge TA, Scullen SE, Sytsma MR, Hezlett SA. (1998). Trait, rater, and level effects in 360-degree performance settings. PERSONNEL PSYCHOLOGY, *51*, 557–576.

Murphy KR. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology: Perspectives on Research and Practice*, *1*, 148–160.

Murphy KR, Balzer WK. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *746*, 19–624.

Murphy KR, Constans JL. (1987). Behavioral anchors as a source of bias in rating. *Journal of Applied Psychology*, *72*, 573–577.

Murphy KR, Garcia M, Kerkar S, Martin C, Balzer WK. (1982). Relationship between observational accuracy and accuracy in evaluating performance. *Journal of Applied Psychology*, *67*, 320–325.

Nathan BR, Tippins N. (1990). The consequences of halo "error" in performance ratings: A field study of the moderating effect of halo on test validation results. *Journal of Applied Psychology*, *75*, 290–296.

Noonan LE, Sulsky LM. (2001). Impact of frame-of-reference and behavioral observation training on alternative training effectiveness criteria in a Canadian military sample. *Human Performance*, *14*, 3–26.

Nunnally JC, Bernstein IH. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Paterson DG. (1922). The Scott Company graphic rating scale. *Journal of Personnel Research*, *1*, 361–376.

Paterson DG. (1923). Methods of rating human qualities. *The Annals of the American Academy of Political and Social Science*, *110*, 81–93.

Pulakos ED. (1984). A comparison of rater training programs: Error training and accuracy training. *Journal of Applied Psychology*, *69*, 581–588.

Pulakos ED. (1986). The development of training programs to increase accuracy with different rating tasks. *Organizational Behavior and Human Decision Processes*, *38*, 78–91.

Roch SG, Mishra V, Kieszczynka U, Woehr DJ. (2010, April). *Frame of reference training: An updated meta-analysis*. Paper presented at the 25th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Rogelberg SG, Waclawski J. (2001). Instrumentation design. In Bracken DW, Timmreck CW, Church AH (Eds.), *The handbook of multisource feedback: the comprehensive resource for designing and implementing MSF processes*. San Francisco, CA: Jossey-Bass.

Saal FE, Downey RG, Lahey MA. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413–428.

Schleicher DJ, Day DV. (1998). A cognitive evaluation of frame-of-referencetraining: Content and process issues. *Organizational Behavior and Human Decision Processes*, *38*, 78–91.

Schleicher DJ, Day DV, Mayes BT, Riggio RE. (2002). A new frame for frame of reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology*, *87*, 735–746.

Scullen SE, Mount MK, Judge TA. (2003). Evidence of the construct validity of developmental ratings of managerial performance. *Journal of Applied Psychology*, *88*, 50–66.

Smith PC, Kendall LM. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*, 149–155.

Smither JW, London M, Reilly RR. (2005). Does performance improve following multisource feedback? A theoretical model, meta-analysis and review of empirical findings. PERSONNEL PSYCHOLOGY, *58*, 33–66.

Stamoulis DT, Hauenstein NMA. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of AppliedPsychology*, *78*, 994–1003.

Sulsky WK, Balzer WK. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497–506.

Sulsky LM, Day DV. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology*, *77*, 501–510.

Sulsky, LM, Day DV. (1994). Effect of frame-of-reference training on rateraccuracy under alternative time delays. *Journal of Applied Psychology*, *79*, 535–543.

Timmreck CW, Bracken DW. (1995, May). Upward feedback in the trenches: Challenges and realities. In Tornow WW (Chair), *Upward feedback: The ups and downs of it*.

Symposium conducted at the 10th Annual Conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Tziner A. (1984). A fairer examination of rating scales when used for performance appraisal in a real organizational setting. *Journal of Organizational Behavior*, 5, 103–112.

Tziner A, Joanis C, Murphy KR. (2000). A comparison of three methods of performance appraisal with regard to goal properties, goal perception, and ratee satisfaction. *Group and Organization Management*, *25*, 175–190.

Viswesvaran C, Schmidt FL, Ones DS. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, *87*, 345–354.

Viswesvaran C, Schmidt, FL, Ones DS. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131.

Wagner SH, Goffin RD. (1997). Differences in accuracy of absolute and comparative performance appraisal methods. *Organizational Behavior and Human Decision Processes*, *70*, 95–103.

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

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

Woehr DJ. (2008). On the relationship between job performance ratings and ratings of job performance: What do we really know? *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 161–166.

Woehr DJ, Huffcutt AI. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, *67*, 189–205.

Woehr, DJ, Miller MJ. (1997). Distributional ratings of performance: More evidence for a new rating format. *Journal of Management*, *23*, 5, 705–721.

APPENDIX A

Sample Dimensions from Frame of Reference Scales used in Study 1

Problem Solving					
Problem solving involves understanding problems and making appropriate decisions to resolve these problems. Effective problem solving entails gathering pertinent information, recognizing key issues, basing decisions on sound rationale, and considering the implications of one's actions. Ineffective problem solving occurs when a manager does not attempt to gather relevant information, makes premature decisions, or confuses details of a given problem.					
At work, he/she					
1. Searches for additional information in order to identify the cause of problems.	1	2	3	4	5
2. Considers multiple solutions to problems.	1	2	3	4	5
3. Explicitly provides rationale for his/her decisions	1	2	3	4	5
Interpersonal Sensitivity					
Interpersonal sensitivity is defined as an individual's concern for the feelings and needs of others. Effective interpersonal sensitivity occurs when a person works to build rapport with others, is attentive to others' thoughts and feelings, and shows concerns for coworkers' personal issues. Ineffective interpersonal sensitivity occurs when one is inattentive or alienates others.					
At work, he/she					
4. Treats others with dignity and respect	1	2	3	4	5
5. Responds appropriately to the feelings of others	1	2	3	4	5
6. Avoids interrupting others when they are speaking	1	2	3	4	5

APPENDIX B

Sample Dimension From Frame of Reference Scales Used in Study 2

1.0 Extremely Weak	1.7 Very Weak	2.0 Weak	2.5 Moderately Weak	2.7 Slightly Weak	3.0 Satisfactory	3.5 Effective	3.7 Very Effective	4.0 Highly Effective	4.5 Extremely Effective	5.0 Exceptional

Interpersonal Sensitivity

Interpersonal sensitivity is defined as an individual's concern for the feelings and needs of others. Effective interpersonal sensitivity occurs when a person works to build rapport with others, is attentive to others' thoughts and feelings, and shows concerns for coworkers' personal issues. Ineffective interpersonal sensitivity occurs when one is inattentive or alienates others.

In the video, the manager…	**Check all that apply**		
Displayed attentive behaviors (e.g., eye contact, nodding)	□ Yes	□ No	□ Not applicable
Tried to establish rapport with the role player (make small talk)	□ Yes	□ No	□ Not applicable
Used humor	□ Yes	□ No	□ Not applicable
Exchanged social pleasantries	□ Yes	□ No	□ Not applicable
Acknowledged the contributions of the role player	□ Yes	□ No	□ Not applicable
Did not interrupt the role player	□ Yes	□ No	□ Not applicable

	Circle one rating										
Overall	1.0	1.7	2.0	2.5	2.7	3.0	3.5	3.7	4.0	4.5	5.0