

Running Head: RATING DIFFERENCES IN MULTI-RATER FEEDBACK

Rating Differences in Multi-Rater Feedback:

A New Look at an Old Issue

P. Gail Wise

The University of Georgia

Paper presented at the meeting of the International Personnel Management Association
Assessment Council's (IPMAAC) Conference on Professional Personnel Assessment,
Chicago, IL, 1998

Mailing Address: Irwin & Browning

100 Galleria Parkway, Suite 1000

Atlanta, GA 30339

Telephone: (770) 952-2000

E-mail: *chier.wise@mci2000.com*

Rating Differences in Multi-Rater Feedback: A New Look at an Old Issue

Abstract

Understanding rater disagreement in multi-rater (360) feedback efforts is important to both scientists and practitioners. This study used structural equations modeling to test for the presence of (a) construct definition differences and (b) rating scale point differences in each of sixteen performance dimensions. Data used were from the Center for Creative Leadership's "Benchmarks[®]" instrument.

In this sample, the construct validity of fourteen of the sixteen Benchmarks[®] dimensions was upheld across groups. Because construct validity is necessary for comparing mean ratings of one rater group to those of another, these results provide evidence that well-constructed scales can exhibit similar factor structure across groups.

The remaining fourteen dimensions exhibited rating scale point definition differences. This is problematic, because in multi-rater feedback efforts, each rater group needs to define rating scale points similarly in order to compare mean ratings across groups. Implications of this finding are discussed.

"Three hundred and sixty-degree feedback" ("360") is the popular name for performance feedback collected from multiple raters. In the typical 360 process, supervisor(s), subordinates, peers, and (less frequently) internal or external customers provide feedback on performance for each target ratee, using some type of standardized instrument (London & Smither, in press; Tornow, 1993a). The ratee then is expected to use the data, along with his/her self-ratings, to make appropriate behavioral changes to improve performance (London & Smither, in press).

Multi-rater feedback has assumed a substantial role in U.S. organizations in the last decade (O'Reilly, 1994). In spite of its rapid growth, however, much remains to be learned about its application and interpretation (London & Smither, in press). While there have been a number of studies documenting the lack of agreement among rater groups (London, 1995), few have focused at the level of dimensions rather than overall ratings. Additionally, few have attempted to determine why a lack of agreement exists (Cardy & Dobbins, 1994).

The purpose of this research, therefore, is to contribute to a broader understanding of 360 and rating differences

among different rater groups. This paper will begin by exploring the 360 process. The research on the level of agreement that has been exhibited among rater groups will be presented. Finally, a study to explore rater agreement at the dimension level will be reviewed.

Multi-Rater Feedback: An Overview

Process of Multi-Rater Feedback

360 is a complex, multi-step process. Based upon a literature review, a model (Figure 1) was developed to illustrate key components of the 360 process, and to provide structure to the broad spectrum of literatures with relevance to 360. This process model will now be explored.

Purpose. The first component in Figure 1's process model is the purpose for which multi-rater feedback is to be used. 360 has been used by researchers and practitioners to address a variety of individual and organizational goals. They include:

1. To improve the subjective measurement of performance. Although supervisory ratings are widely used for measuring performance, they are subject to a variety of intentional and unintentional errors (Cardy & Dobbins, 1994). 360 has

been used to improve performance measurement by supplementing supervisory ratings with those of multiple raters. Some researchers argue that by using multiple raters, 360 allows triangulation on a ratee's psychometric "true score" (Tornow, 1993b). Others argue that 360 is valuable precisely because there is no single true score on which to triangulate (Lance, Teachout, & Donnelly, 1992; London, 1995).

2. For self-development. The most widespread use at present is for self-development (Hollenbeck, Sorcher, & Moses, 1994).

3. For organizational development. In this application, the 360 instrument is structured to assess attributes viewed as essential for success (Hollenbeck et al., 1994; O'Reilly, 1994).

4. For administrative decisions (e.g., salary increases, promotion, or layoff). While such use is rare at present, it is an area of increasing interest to organizations (Lancaster, 1996).

What is to be measured and how.

The second critical area of interest in Figure 1 is the instrumentation. 360 instruments may be custom-designed, or may be purchased from a variety of commercial providers (Hollenbeck et al., 1994). They may consist of single- or multiple-item dimensions. Both quantitative and qualitative data may be gathered.

Ratees. The third block in Figure 1 addresses the target ratee(s). Participation in 360 may be voluntary or mandatory, depending upon the intervention's purpose (Lancaster, 1996).

Raters. Selection of raters and identifiability of individual raters' feedback is the fourth block in Figure 1. The number of raters from each source may vary. Some organizations require raters to be anonymous to minimize rating distortion (London, 1995), while others identify raters to encourage discussion (Milliman,

Zawacki, Norman, Powell, & Kirksey, 1994).

Instrument Scoring and Preparation of Feedback. 360 instrument scoring and the content of the resulting feedback report (block 5 in Figure 1) differ among organizations. Some retain outside firms or designate a neutral internal resource to provide the feedback (Milliman et al., 1994), while others require the ratee or the supervisor to consolidate all feedback.

Providing Feedback to the Ratee.

Figure 1's sixth block addresses the process of providing feedback to the ratee. 360's complexity places significant cognitive demands on the ratee. Research has shown that the more complex the feedback, the more likely recipients will distort it by focusing on results that "match their self-perceptions" and ignoring contradictory ones (London, 1995, p. 221). Consequently, while some organizations rely on each ratee to interpret 360-degree feedback, others provide one-on-one coaching by professionals (Kaplan, 1993).

Intended outcomes. The final block in Figure 1 focuses on the expected outcomes of multi-rater feedback interventions. Research evidence for 360's effectiveness shows mixed results for criteria such as (a) ratee acceptance, (b) rater comfort, (c) actual individual behavior change, and (d) improvements in work group relationships and performance (London, 1995).

Summary. This overview of the process of multi-rater feedback underscores its complexity. For the ratee, in particular, interpreting the output of a 360 feedback effort is an undertaking with significant cognitive demands. To make sense of the data, the ratee must interpret feedback from each rating source, and each source may have a different evaluation. To examine this issue of rater agreement, the literature on interrater agreement will now be explored.

Research on Rater Agreement

Why Ratings Differ Among Raters

There are many general reasons why differences may exist among raters. While random error certainly accounts for some interrater disagreement, a variety of informational, cognitive, affective, and contextual factors may also lead to systematic differences in rater evaluations (Cardy & Dobbins, 1994). Two sources of rater differences are the focus of this study.

Lack of construct validity. In order to compare performance ratings across raters, the appraisal measures must be "construct valid" -- that is, the measures must be reflective of variance in the underlying dimensions of performance (Feldman, 1986). If raters are defining the dimensions of performance differently, then the measures are not construct valid, and the ratings will not exhibit convergent validity across raters (Riordan & Vandenberg, 1994). "Leading Employees" may be used as an example. Suppose that subordinates view an effective leader as one who encourages self-direction and provides minimal guidance, while peers and supervisors expect a leader to provide close monitoring. This lack of a common definition will contribute to rating differences, making comparisons across rater groups invalid.

Different calibration of the rating scales used. A second source of rating differences is the way raters define the points on the rating scale (Riordan & Vandenberg, 1994). Raters may define scale points differently, creating systematic differences among ratings. Riordan and Vandenberg use the example of a 3 ("neither agree nor disagree") on a 5-point Likert scale (1994, p. 644). Some raters interpret this as "no opinion," while others interpret it as "mild agreement."

Agreement Among Rating Sources

Even though there are a variety of causes of disagreement among rater groups, agreement is desirable for at least two reasons. From a psychometric perspective, high interrater agreement is desirable to ensure that the dimensions of performance assessed are done so with scientific soundness (Borman, 1991). Second, high interrater agreement eases the cognitive demands on the ratee, making it simpler to understand, accept, and act on the feedback (London, 1995).

Despite the desirability of rater agreement in 360, the plethora of informational, cognitive, contextual, and affective issues affecting ratings mentioned above means that disagreement is to be expected. Multitrait-multimethod studies, with different rater groups constituting different methods, have demonstrated the presence of rater group effects (e.g., Holzbach, 1978; Klimoski & London, 1974; Lance et al., 1992; London & Wohlers, 1991; Mount, 1984). A number of authors have noted a different factor structure for self ratings than for those by others, including Holzbach (1978), Lawler (1967), and Thornton (1980). Correlations among ratings from different sources have varied across studies, but none have been exceptionally high (Harris & Schaubroeck, 1988; Landy & Farr, 1980).

Note that all the studies mentioned up to this point focused on overall (not dimensional) ratings of performance. While overall ratings provide general information to the ratee, they provide no clues as to what specific elements of performance are an issue. Dimension-level ratings provide such focus. Unfortunately, dimension-level research in 360 is sparse. In the only study of self-subordinate ratings located that reported correlations, a mean correlation of .24 was found across the two dimensions

assessed (London & Wohlers, 1991). The other study found that explored dimension-level agreement suffers from the use of single item, trait-based dimensions (Wohlers & London, 1989). Neither of these studies used analytic techniques that could directly test for causes of interrater disagreement. In short, opportunity exists for exploration of interrater agreement at the dimension level, using psychometrically sound, multi-item instruments and analytic techniques that could help explain observed disagreement.

Current Study

Multi-rater feedback efforts are rooted in the assumption that all rater groups define each performance dimension similarly (i.e., that the dimensions exhibit construct validity), and that raters calibrate rating scale points similarly. The purpose of this research is to test these two assumptions. Because of its exploratory nature, no hypotheses will be offered as to which dimensions will be subject to each type of interrater disagreement. However, it is expected that those dimensions which are least observable, and for which an objective performance standard is least likely to exist (e.g., Harris & Schaubroeck, 1988; Shore, Shore, & Thornton, 1992), will exhibit the least construct validity and the most susceptibility to scale calibration differences.

METHOD

Instrument

The data used for this study were obtained from the Center for Creative Leadership (CCL), a non-profit institution devoted to the study and practice of leadership. The 360 data were derived from Benchmarks® -- a widely used, commercially-available instrument that was developed from CCL studies of how successful managers develop (Lombardo &

McCauley, 1995). Sixteen performance dimensions (assessed by 106 items) were used in this study, and include (a) "Resourcefulness"; (b) "Doing whatever it takes"; (c) "Being a quick study"; (d) "Decisiveness"; (e) "Leading employees"; (f) "Setting a developmental climate"; (g) "Confronting problem employees"; (h) "Work team orientation"; (i) "Hiring talented staff"; (j) "Building and mending relationships"; (k) "Compassion and sensitivity"; (l) "Straightforwardness and composure"; (m) "Balance between personal life and work"; (n) "Self-awareness"; (o) "Putting people at ease"; and (p) "Acting with flexibility." Coefficients alpha, calculated by CCL researchers on a validation sample of 336 supervisors, ranged from .75 for the "Decisiveness" scale to .97 for the "Hiring Talented Staff" scale (Lombardo & McCauley, 1995).

Procedure

Upon registering at CCL to receive feedback, ratees received twelve copies of Benchmarks® for distribution to supervisory, peer, subordinate, and self-raters. Anonymity was guaranteed to peer and subordinate raters. All raters returned the completed forms directly to CCL, where they were collated and analyzed to produce feedback reports.

Subjects

The subjects of this study were all recipients of Benchmarks® feedback in 1996. Sample size for the study was 1173 ratees. One randomly-selected subordinate, peer, and supervisor rating was used for each ratee, to minimize problems of non-independence of data. Final sample sizes of each rater group, due to missing data, were (a) 1055 subordinates, (b) 1139 peers, and (c) 1013 supervisors. Demographic information was available only for the ratees. Six hundred ninety-eight (59.5%) were male, and 287 (24.5%) were female

(16% did not identify gender). The mean age was 42.9, and mean years of schooling were 17.3. Eight hundred fifty-one (72.5%) were white and 97 (8.3%) were minorities; the remainder did not identify race.

Analyses

Structural Equations Modeling

As discussed, a variety of factors may underlie rater group differences, including (a) lack of construct validity, and (b) different calibrations of the rating scale used. An application of structural equations modeling (SEM), commonly used in longitudinal studies to assess group-level alpha/beta/gamma (ABG) change (e.g., Chrobot, Lance, Gowan, & Gatewood, 1995; Schaubroeck & Green, 1989), was used to test for these two sources of group differences.

With ABG-type SEM, the following analytical steps were performed for each of the sixteen Benchmarks[®] scales with a series of seven nested structural equations models. This was accomplished using LISREL-8's multi-sample analysis feature (Jöreskog & Sörbom, 1996).

1. Model 1 was an omnibus test in which the variance/covariance matrices for each group were constrained to be equal. If Model 1 was not shown to be a good representation of the data, this indicated that measurement differences existed among the rater groups for the Benchmarks[®] scale in question, and a series of increasingly restrictive models were used to localize the source of the measurement differences (Riordan & Vandenberg, 1994).

2. The first more restrictive model (Model 2) constrained the groups to have identical factor structures. Here, if all groups contained a single factor and the same pattern of factor loadings, evidence for construct validity was established across the four rater groups for the Benchmarks[®] scale

in question. It was then appropriate to test the next more restrictive model. However, if Model 2 was not supported, further analysis was unwarranted because the scale was interpreted differently by different rater groups (Riordan & Vandenberg, 1994).

3. Assuming that Model 2 was supported, Models 3 and 4 were used to examine the equality of scale points across groups. Model 3 constrained factor loadings on the underlying latent variable to be equivalent across groups (Schmitt, Pulakos, & Lieblein, 1984). Model 4 constrained the latent variable's variance to be equal across groups. If either Models 3 or 4 were not supported for a given scale, it indicated that different rater groups defined the rating scale points differently, and further analyses were unwarranted for that scale (Chrobot et al., 1995).

4. Model 5 (an add-on to traditional ABG analyses) constrained intercepts to be equal across groups. This model examined whether one or more groups had a "consistent tendency to answer higher or lower" than the other groups (Bollen, 1989, p. 368).

5. Model 6 (also an addition to traditional ABG analyses) constrained error variances to be equal across groups. This tested the equivalence of item-level reliabilities, to pinpoint items that might be functioning differently (C. E. Lance, personal communication, December 3, 1996).

6. Lastly, assuming that Models 1, 2, 3, and 4 were supported, Model 7 constrained the factor means for each scale to be equivalent across groups. Here, a loss of fit from Model 4 indicated the presence of simple mean differences (Riordan & Vandenberg, 1994).

A variety of goodness-of-fit indices (GFIs) were used to assess model fit,

including (a) the chi-square statistic (Bollen, 1989); (b) the chi-square statistic divided by its degrees of freedom (Medsker, Williams, & Holahan, 1994); (c) the Root Mean Square Residual (RMSR); (d) the Normed Fit Index (NFI) (Bentler, 1990); (e) the goodness of fit indicator (GFI) produced by LISREL-8 (Medsker et al., 1994); (f) the Tucker-Lewis reliability index (TLI) (Marsh, Balla, & McDonald, 1988), and (g) the chi-square difference test (Steiger, Shapiro, & Browne, 1985).

Other Analyses

In the event that the tests of nested models described above revealed either (a) lack of construct validity or (b) different calibration of the rating scales, two different analyses were conducted to try to understand the cause of these rater group differences.

The first analysis was undertaken when a lack of construct validity was found for a given Benchmarks® scale. In this event, exploratory factor analysis (EFA) was used to explore how each rater group was conceptualizing the construct(s) contained in what had been constructed to be a unidimensional scale (Kim & Mueller, 1978).

Second, all sixteen scales were assessed by an independent panel of subject matter experts as to (a) their "observability" by peers, subordinates, and supervisors (e.g., Shore et al., 1992) and (b) the extent to which an objective standard of performance exists for the variable represented by the scale ("definability") (e.g., Feldman, 1986). Both observability and existence of objective standards of performance have been hypothesized to contribute to rater agreement. The subject matter experts (SMEs) were practicing managers in 8 U.S. for-profit organizations. Fifty-five surveys were distributed, and 43 were returned, for a response rate of 78.18%.

RESULTS

Table 1 contains descriptive statistics (mean, standard deviation, and range) for each of the sixteen dimensions within each of the four rater groups in this sample. Surprisingly, paired sample correlated t-tests conducted between the column totals in Table 1 revealed no differences in overall mean ratings. Self-ratings ($\bar{M}=3.816$) did not differ significantly from peer ratings [$\bar{M}=3.713$; $t(30)=.124$; $p>.05$], nor did they differ significantly from subordinate ratings [$\bar{M}=3.768$; $t(30)=.065$; $p>.05$] or supervisor ratings [$\bar{M}=3.767$; $t(30)=.076$; $p>.05$].

Structural Equations Modeling

As discussed, a series of seven nested SEM models was used to examine sources of observed rater group differences. Results of these analyses for all sixteen dimensions appear in Table 2. "NA" appears in each block where, due to the hierarchical nature of the nested models, it would be inappropriate to conduct analyses on all four rater groups.

As shown in Table 2, different factor structures were found in two of the Benchmarks® dimensions: "Doing whatever it takes" and "Straightforwardness and composure." For these dimensions, then, comparisons of rater group results would be inappropriate, because the rater groups were defining the dimension differently.

Rating scale point differences were prevalent in these data. All of the remaining fourteen dimensions showed evidence of this difference. As discussed, for these dimensions, comparisons of simple mean differences across rater groups would be inappropriate because different groups were defining the response scale points differently. Of the fourteen dimensions exhibiting this difference, four were attributable solely to the "self-rating" group,

as indicated by the footnote in Table 2. The poor fit in the remaining ten dimensions could not be attributed to a single group.

Other Analyses

Two additional types of analyses were undertaken to understand the SEM results: (a) EFA and (b) subjective analysis of each dimension by a sample of practicing managers.

Exploratory Factor Analysis

The two dimensions exhibiting different factor structures ("doing Whatever it Takes" and "straightforwardness and composure") were subjected to EFA, in order to understand how different rater groups conceptualized the items (Kim & Mueller, 1978). Parallel analysis was used as the EFA decision rule (Lautenschlager, Lance, & Flaherty, 1989).

For "doing whatever it takes," the peer, subordinate, and supervisory groups in this sample each displayed a single underlying factor. The self group, however, displayed two underlying factors: (a) "doing whatever it takes *to get the job done*" and (b) "doing what it takes *to manage your career*." These results indicate that the self group did conceptualize the scale differently than the three other rating groups. For "straightforwardness and composure," however, EFA supported a two factor solution for all four rater groups. These two factors represent a simple split in the scale between "straightforwardness" and "composure."

Subject Matter Expert (SME) Survey

It was hypothesized that those dimensions (a) for which an objective standard of performance is least likely to exist, and (b) which are least observable, would be most susceptible to (a) construct validity problems and (b) rating scale

calibration differences across rater groups. This hypothesis was not supported.

Table 2 maps the means and standard deviations for each dimension in the SME survey against the SEM results reviewed earlier. The overall mean and standard deviation for each category were calculated, and are noted at the bottom of the appropriate column in Table 2. Individual item means that differed from their corresponding category mean by one standard deviation or more are marked with a star. Unfortunately, the dimensions exhibited differing factor structures across groups were not singled out by low ratings in the SME survey.

DISCUSSION

Multi-rater feedback efforts have two underlying assumptions: (a) that all rater groups define each performance dimension similarly (i.e., that the dimensions exhibit construct validity), and (b) that raters calibrate rating scale points similarly. Using a rigorously developed, commercially available 360-degree feedback instrument with multiple items in each dimension, this study found solid support for the first assumption, but none for the second.

This study's results will be presented in the following sequence. First, the question of whether the rater groups defined the sixteen Benchmarks® constructs similarly will be explored. Second, the issue of whether the rater groups defined the rating scale points equivalently will be investigated. Finally, this study's implications for scientists and practitioners will be discussed.

Did Different Rater Groups Define Constructs Similarly?

In this sample, the construct validity (factor structure) of 14 of the 16 Benchmarks® dimensions was upheld across rater groups. Only two dimensions ("Doing whatever it takes" and "Straightforwardness and composure") displayed evidence of multi-dimensionality across groups. Two different issues were at play in these two dimensions, however. For "straightforwardness and composure," all four rater groups demonstrated the same two-factor EFA solution. Hence, the exhibited construct invalidity could be seen as a scale construction issue, easily remedied by splitting the scale in two.

The second dimension, "doing whatever it takes," represents a different scenario. Here, self-raters made distinctions in their answers that suggest two underlying constructs, while "other" raters (supervisor, subordinate, and peer) saw only one. Self-raters distinguished between "doing whatever it takes *to manage your career*" and "doing whatever it takes *to get the job done*." This distinction seems reasonable for self-raters to make. Unlike "other" raters, self-raters are aware of the moves they undertake for career preservation and enhancement (Hall & Associates, 1991). Here, then, construct invalidity is a more profound issue than in the "straightforwardness and composure" dimension. Perhaps "other" raters cannot distinguish "career management activities" from activities performed simply to "ensure quality performance in the face of obstacles," so that including such a dimension in 360 is inappropriate.

With the exception of the two dimensions just discussed, the Benchmarks® instrument displayed solid factor structure. This strong showing was somewhat surprising, given reports in the literature of

differential factor structures in multi-rater performance data (e.g., Holzbach, 1978; Lawler, 1967; and Thornton, 1980). The implication of this study's results is that well-constructed scales can exhibit construct validity across different rater groups.

Did Different Rater Groups Define Rating Scale Points Similarly?

The assumption that rating scale points are defined similarly across rater groups was not upheld in this sample. All of the dimensions that didn't display different factor structures exhibited scale point definition differences across rater groups. This study's results are disturbing, because in order to compare ratings of one group to those of another, each needs to be defining the rating scale points in the same way (Riordan & Vandenberg, 1994). Because of the prevalence of this problem in these data, additional research may be justified to examine possible reasons for such rating scale point differences. At least three possible explanations may exist.

1. Differences in "observability" and "definability" of each dimension: Previous researchers have postulated that differences in "observability" and "definability" may underlie rater group differences (e.g., Harris & Schaubroeck, 1988; Shore et al., 1992). Unfortunately, but perhaps not surprisingly, SME ratings of "observability" and "definability" in this study were not helpful in identifying dimensions exhibiting rating scale point differences. SMEs have not been able to single out items that function differentially across race and gender groups in biodata research, nor have they been able to describe characteristics that contribute to differential item functioning (Stennett, Stokes, Thompson, & Wise, 1995). It is possible, then, that relying on SME judgment to identify dimensions subject to scale point differences is not a fruitful endeavor.

However, it is also possible that the composition of the SME group sampled contributed to the poor results. Factors affecting SME judgments of "observability" and "definability" may be organization-specific, so that these results may have been very different if collected from a single organization. While the survey was not intended to tap only the respondents' current organizational realities, many respondents made their ratings based only on current opportunity to observe. Should future researchers attempt to replicate this survey, perhaps even more effort should be made to help respondents think beyond their current organizational circumstances.

2. Likert-type scale anchor differences: It is possible that some Likert-type scales are more subject to beta-type definitional differences across rater groups than others. Benchmarks® used a 1-5 rating scale, ranging from "not at all" to "to a very great extent." Perhaps other scales, such as those requiring judgments of frequency, effectiveness, or importance, are less subject to different scale point definitions than this one. Landy and Farr (1980) note that scaling has been one of the weakest aspects of performance appraisal development. These results seem to support the desirability of further research on standardizing scale point definitions across rater groups.

3. Organizational differences in rating scale point definitions: It is also possible that rating scale point definitions may vary by organization or function (just as SME judgments, as discussed above). Note that the Benchmarks® data used in this research were collected across a number of organizations, so that possible organizational differences were not explored in this study. It is possible that definitions of these Likert-type scale points may vary less across individuals within a given organization than across organizations, due

to the influence of culture or shared experiences. Perhaps training efforts within a single organization could help create alignment around scale point definitions, so that peer, supervisor, subordinate, and self-ratings could indeed be compared. If additional research shows that rating scale calibration is not amenable to training, 360 instrumentation may need to be customized to individual organizations to help ensure that only items which exhibit common scale point definitions are used.

Implications of Study for Scientists and Practitioners

A process model of multi-rater feedback, synthesizing the literature on 360, was presented earlier. The results of this study suggest two additions to the "what is to be measured and how" portion of the model (box 2 in Figure 1). As noted earlier, both scientists and practitioners need to be concerned with item content. Giving all raters an opportunity to assess their "opportunity to observe" or "comfort in providing a rating" for each item may be a useful addition to 360 instruments. For scientists, such assessments may help them to identify problematic items or dimensions. For practitioners, such a rating could help ratees evaluate the feedback received. If, for instance, subordinates identified a category as being difficult to observe or uncomfortable to evaluate, ratees could focus more closely on item ratings given by peers or subordinates.

The second addition to this part of the model is to ensure that the instrument is pre-tested for construct validity and scale calibration across rater groups. Rating differences will be interpreted by ratees as simple mean differences, and their search for explanations for those differences will center on ratee-controllable issues, not artifacts of the test development process.

Summary

This study adds to our understanding of multi-rater feedback in several ways. First, it extends the research on interrater agreement by applying structural equations modeling techniques to probe the reasons for observed rater differences. Second, it explores dimension- level 360 feedback in a large sample, using a well-constructed commercial instrument with multiple items per dimension. Third, it finds evidence for two key types of differences in ratings across rater groups: (a) differing factor structures, and (b) differing scale point definitions.

It is clear, both from the literature review and this study's findings, that understanding 360 is a difficult undertaking. Because "the real work of 360-degree feedback really begins when people receive their results" (Kaplan, 1993, p. 300), it is imperative that instrument design and development efforts seek to minimize psychometric anomalies in the feedback data.

There are three implications for researchers and practitioners. First, design of a 360 instrument is a scientific endeavor, requiring application of sound psychometric techniques. Second, selection or design of a 360 instrument for use in a given organization requires analysis of the dimensions, items, and rating scale points used to ensure construct validity and rating scale equivalence across groups. Lastly, interpretation of 360 should not focus solely on simple mean differences across rater groups without testing for the possibility of construct validity differences and rating scale calibration differences.

REFERENCES

- Bentler, P. M. (1990). Comparative fit indices in structural models. Psychological Bulletin, 107(2), 238-246.
- Bollen, K. A. (1989). Structural equations with latent variables. New York: John Wiley.
- Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of Industrial and Organizational Psychology (2nd ed., Vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.
- Cardy, R. L., & Dobbins, G. H. (1994). Performance appraisal: Alternative perspectives. Cincinnati: South-Western.
- Chrobot, D. L., Lance, C. E., Gowan, M. A., & Gatewood, R. D. (1995). Extended Unemployment and Re-Employment: Alpha, Beta, Gamma Changes in Employment Commitment. Paper presented at the meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. Research in Personnel and Human Resources Management, 4, 45-99.
- Hall, D. T. & Associates. (1991). Career development in organizations. San Francisco, CA: Jossey-Bass.
- Harris, M. M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.
- Harris, M. M., Smith, D. E., & Champagne, D. (1995). A field study of performance appraisal purpose: Research-

versus administrative-based ratings. Personnel Psychology, 48, 151-160.

Hollenbeck, G. P., Sorcher, M., & Moses, J. (1994). 360 degree feedback. Workshop conducted at the meeting of Society for Industrial and Organizational Psychology, Nashville, TN.

Holzbach, R. L. (1978). Rater bias in performance ratings: Superior, self-, and peer ratings. Journal of Applied Psychology, 63(5), 579-588.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL-8 user's reference guide. Mooresville, IN: Scientific Software International.

Kaplan, R. E. (1993). 360-degree feedback plus: Boosting the power of co-worker ratings for executives. Human Resource Management, 32(2&3), 299-314.

Kim, J. O., & Mueller, C. W. (1978). Introduction to factor analysis: What it is and how to do it. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-013. Beverly Hills, CA: Sage.

Klimoski, R. J., & London, M. (1974). Role of the rater in performance appraisal. Journal of Applied Psychology, 59(4), 445-451.

Lancaster, H. (1996, July 9). Performance reviews are more valuable when more join in. The Wall Street Journal, p. B1.

Lance, C. E., Teachout, M. S., & Donnelly, T. M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. Journal of Applied Psychology, 77(4), 437-452.

Landy, F. J., & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87(1), 72-107.

Lautenschlager, G., Lance, C. E., & Flaherty, V. (1989). Parallel analysis: Revised equations for estimating the latent roots of random data correlation matrices. Educational and Psychological Measurement, 45, 339-345.

Lawler, E. E., III. (1967). The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 51(5), 369-381.

Lombardo, M. M., & McCauley, C. (1995). Benchmarks®: Developmental reference points -- A manual and trainer's guide. Greensboro, NC: Center for Creative Leadership.

London, M. (1995). Self and interpersonal insight: How people gain understanding of themselves and others in organizations. New York: Oxford University Press.

London, M., & Smither, J. W. (In press). Can multi-source feedback change perceptions of goal accomplishment, self-evaluations, and performance-related outcomes? Theory-based applications and directions for research. Personnel Psychology.

London, M., & Wohlers, A. J. (1991). Agreement between subordinate and self-ratings in upward feedback. Personnel Psychology, 44, 375-390.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Medsker, G. J., Williams, L. J., & Holahan, P. J. (1994). A review of current practices for evaluating causal models in organizational behavior and human resources management research. Journal of Management, 20(2), 439-464.

Milliman, J. F., Zawacki, R. A., Norman, C., Powell, L., & Kirksey, J. (1994, November). Companies evaluate employees from all perspectives. Personnel Journal, pp. 99-103.

Mount, M. K. (1984). Psychometric properties of subordinate ratings of managerial performance. Personnel Psychology, 37, 687-702.

Murphy, K. R., & Cleveland, J. N. (1991). Performance appraisal: An organizational perspective. Englewood Cliffs, NJ: Prentice-Hall.

O'Reilly, B. (1994, October 17). 360 feedback can change your life. Fortune, 93-100.

Riordan, C. M. & Vandenberg, R. J. (1994). A Central Question in cross-cultural Research: Do Employees of different Cultures interpret Work-related Measures in an Equivalent Manner? Journal of Management, 20(3), 643-671.

Schaubroeck, J., & Green, S. G. (1989). Confirmatory Factor Analytic Procedures for assessing Change During Organizational Entry. Journal of Applied Psychology, 74(6), 892-900.

Schmitt, N., Pulakos, E. D., & Lieblein, A. (1984). Comparison of three techniques to assess group-level beta and gamma change. Applied Psychological Measurement, 8, 249-260.

Shore, T. H., Shore, L. M., & Thornton, G. C., III. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. Journal of Applied Psychology, 77(1), 42-54.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. Psychometrika, 50(3), 253-264.

Stennett, R. B., Stokes, G. S., Thompson, K. E., & Wise, P. G. (1995). The usefulness of prescreening biodata items for adverse impact. Paper presented at the meeting of the American Psychological Association, New York, New York.

Thornton, G. C., III. (1980). Psychometric properties of self-appraisals of job performance. Personnel Psychology, 33, 263-271.

Tornow, W. W. (1993a). Editor's note: Introduction to special issue on 360-degree feedback. Human Resource Management, 32(2&3), 211-219.

Tornow, W. W. (1993b). Perceptions or reality: Is multi-perspective measurement a means or an end? Human Resource Management, 32(2&3), 221-229.

Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. Personnel Psychology, 42, 235-261.

Figure 1

Process Model of Multi-Rater Feedback

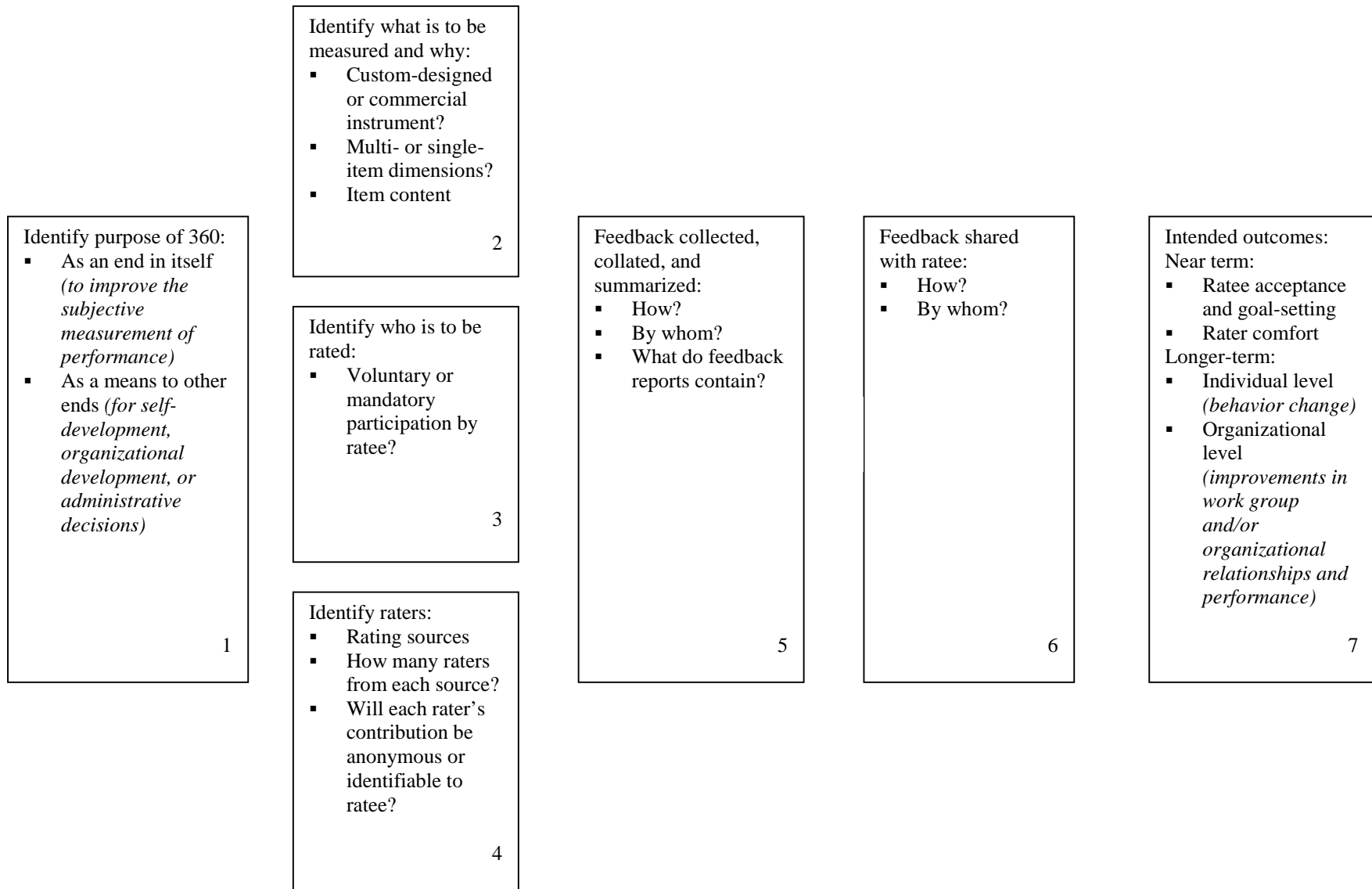


Table 1

Descriptive Statistics for Each Rater Group for Benchmarks® Scales

Scale	Peer Ratings *				Self Ratings **				Subordinate Ratings ***				Supervisor Ratings ****			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Resourcefulness	3.750	0.522	1.59	5.00	0.799	0.414	1.71	5.00	3.827	0.578	1.41	5.00	3.730	0.529	1.59	5.00
Doing whatever it takes	3.834	0.565	1.46	5.00	3.940	0.452	1.43	5.00	3.941	0.589	1.71	5.00	3.834	0.559	1.36	5.00
Being a quick study	3.917	0.681	1.00	5.00	3.847	0.585	1.75	5.00	3.972	0.718	1.00	5.00	3.986	0.645	1.67	5.00
Decisiveness	3.593	0.762	1.25	5.00	3.771	0.657	1.25	5.00	3.670	0.777	1.00	5.00	3.614	0.762	1.00	5.00
Leading employees	3.630	0.595	1.62	5.00	3.824	0.432	2.00	5.00	3.689	0.661	1.08	5.00	3.696	0.548	1.46	5.00
Setting a developmental climate	3.711	0.650	1.20	5.00	3.946	0.474	1.40	5.00	3.771	0.735	1.00	5.00	3.834	0.565	1.60	5.00
Confronting problem employees	3.477	0.757	1.00	5.00	3.423	0.671	1.25	5.00	3.437	0.812	1.00	5.00	3.426	0.745	1.00	5.00
Work team orientation	3.737	0.718	1.25	5.00	3.934	0.603	2.00	5.00	3.871	0.711	1.00	5.00	3.704	0.691	1.25	5.00
Hiring talented staff	3.685	0.693	1.33	5.00	3.955	0.598	2.00	5.00	3.871	0.702	1.00	5.00	3.714	0.651	1.50	5.00
Building and mending relationships	3.680	0.664	1.45	5.00	3.775	0.458	1.36	5.00	3.743	0.677	1.18	5.00	3.724	0.635	1.64	5.00
Compassion and sensitivity	3.580	0.727	1.00	5.00	3.801	0.544	1.50	5.00	3.577	0.798	1.00	5.00	3.752	0.621	1.25	5.00
Straightforwardness and composure	3.960	0.682	1.00	5.00	4.033	0.499	1.33	5.00	4.047	0.685	1.17	5.00	4.159	0.588	1.17	5.00
Balance between personal life and work	3.737	0.790	1.00	5.00	3.535	0.794	1.00	5.00	3.685	0.892	1.00	5.00	3.818	0.703	1.00	5.00
Self-awareness	3.563	0.706	1.00	5.00	3.800	0.514	1.50	5.00	3.559	0.777	1.00	5.00	3.646	0.705	1.00	5.00
Putting people at ease	3.871	0.838	1.00	5.00	3.823	0.650	1.00	5.00	3.920	0.842	1.00	5.00	3.917	0.769	1.50	5.00
Acting with flexibility	3.685	0.663	1.40	5.00	3.844	0.475	1.80	5.00	3.703	0.708	1.20	5.00	3.723	0.585	1.60	5.00
<u>MEAN</u>	3.713	0.688	1.22	5.00	3.816	0.551	1.52	5.00	3.768	0.729	1.11	5.00	3.767	0.644	1.35	5.00

Notes.

* N = 1139

** N = 1173

*** N = 1055

**** N = 1013

Table 2
 Benchmarks[®] Dimensions: Summary of Study's Analyses

Dimension	SEM Results from Benchmarks [®] Data					SME Survey Results ²							
	Different Factor Structure	Different Scale Point Definition	Different Intercepts	Differing Item Reliabilities	Simple Mean Differences	Definition of performance?		Subordinate's performance observable?		Peer's performance observable?		Supervisor's performance observable?	
						Mean	SD	Mean	SD	Mean	SD	Mean	SD
Resourcefulness		**	NA	NA	NA	3.47	0.93	3.93	0.59	3.37	0.90	3.37	1.02
Doing whatever it takes	**1A	NA	NA	NA	NA	3.63	1.02	3.98	0.89	3.56	0.91	3.60	1.03
Being a quick study		**	NA	NA	NA	3.53	1.05	3.77	0.92	3.07	1.01	3.19	1.14
Decisiveness		**1	NA	NA	NA	3.63	0.85	3.81	0.76	3.53	0.91	4.07	0.80
Leading employees		**	NA	NA	NA	4.14	0.71	3.84	0.78	3.56	0.88	4.09	0.92
Setting a developmental climate		**	NA	NA	NA	3.52	0.59	3.26	0.90	2.86	0.89	4.14	0.83
Confronting problem employees		**1	NA	NA	NA	3.79	0.91	3.67	0.92	2.86	0.94	3.42	1.16
Work team orientation		**	NA	NA	NA	3.79	0.72	3.53	0.67	3.33	0.81	3.81	0.85
Hiring talented staff		**	NA	NA	NA	3.74	0.85	3.79	1.04	3.51	0.83	3.93	0.89
Building and mending relationships		**1	NA	NA	NA	3.35	0.87	3.53	0.80	3.21	0.86	3.35	0.95
Compassion and sensitivity		**	NA	NA	NA	2.93	1.03	3.28	0.91	2.93	0.88	3.79	0.94
Straightforwardness and composure	**	NA	NA	NA	NA	3.70	0.83	3.86	0.60	3.37	0.85	3.95	0.75
Balance between personal life and work		**	NA	NA	NA	2.47	0.96	2.77	1.00	2.58	1.10	2.70	1.12
Self-awareness		**1	NA	NA	NA	3.07	1.08	3.23	0.78	2.81	0.93	2.88	0.96
Putting people at ease		**	NA	NA	NA	3.12	0.93	3.84	0.81	3.49	0.83	3.95	0.72
Acting with flexibility		**	NA	NA	NA	3.44	0.93	3.74	0.69	3.56	0.83	3.98	0.74
SME Survey category means and standard deviations						3.4559	0.4962	3.6148	0.4574	3.2253	0.6058	3.6395	0.6694

Notes.

SEM Results from Benchmarks[®] Data Section:

1 = Self ratings responsible for difference; other three groups exhibited good fit when tested without self.

1A - Self-ratings responsible for difference, as determined through exploratory factor analysis.

No footnote = Cannot attribute fit difference to a single rater group.

NA = Analysis is meaningless due to hierarchical nature of SEM test.

SME Survey Results Section:

2 - N=43

Bold/italicized - Dimension mean is greater than one standard deviation from category mean.