Interrater Agreement in Multi-Source Performance Appraisal: A Commentary

Dennis P. Bozeman


Stable URL: http://links.jstor.org/sici?sici=0894-3796%28199707%2918%3A4%3C13%3AIAIMPA%3E2.0.CO%3B2-F

*Journal of Organizational Behavior* is currently published by John Wiley & Sons.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/jwiley.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
Interrater agreement in multi-source performance appraisal: a commentary

DENNIS P. BOZEMAN
Department of Management, College of Business, Florida State University, Tallahassee, FL 32306-1042, U.S.A.

J. Organiz. Behav. 18: 313–316 (1997)
No. of Figures: 0 No. of Tables: 0 No. of References: 14

For a number of years, scholars have touted the benefits of using performance ratings from multiple organizational rates to appraise individual performance (e.g. Buckner, 1959; Campbell, Dunnette, Lawler and Weick, 1970; Latham and Wexley, 1982). According to the performance appraisal literature, multi-rater evaluation provides an integrated assessment of individual performance that maximizes the strengths and minimizes the weakness of individual ratings (Latham and Wexley, 1982), a fuller conceptualization and measurement of the job performance domain (Borman, 1974; Henderson, 1984), an improved legal defensibility over single-source ratings (Bernardin and Beatty, 1984), and an increased use of performance feedback for individual improvement and development (Cardy and Dobbins, 1994). Multi-rater evaluation also is an attractive prospect to individual ratees, in that raters tend to perceive multi-rater evaluation as a fairer and more acceptable method of performance appraisal than traditional single-source evaluation (Latham and Wexley, 1982; McEvoy, 1990). The purposes of this paper are to: (1) discuss some of the key issues surrounding the use of multi-rater performance assessments in organizations, and (2) introduce some new ideas concerning how the use and application of multi-source performance appraisals in organizations might be improved.

The Validity of Multi-Source Performance Appraisal Ratings

Despite the purported benefits of multi-source performance appraisal cited previously, the extant empirical literature shows that different rater groups (e.g. supervisors, subordinates) frequently do not agree concerning an individual’s job performance (e.g. Bernardin and Beatty, 1984; Borman, 1974; Harris and Schaubroeck, 1988; Mount, 1984; Zalesny and Kirsch, 1989). Based on traditional conceptions of reliability and validity, low interrater agreement indicates unreliability and, therefore, invalidity. Accordingly, the validity of multi-source performance appraisal has been questioned. In response to this charge, some scholars have asserted that ratings obtained from different rater groups may be valid, even if they do not exhibit high levels of
agreement with one another (Borman, 1991; Landy and Farr, 1980). Further, because raters
within various groups have different opportunities to observe an individual’s job performance
behaviors, disagreement among rater groups might well be expected (Murphy and Cleveland,
1995). Moreover, the performance ratings provided by different rater groups also are likely role-
related (Borman, 1991). That is, various rater groups likely evaluate the aspects of the focal
individual’s performance that are most relevant to the raters themselves. For example, super-
visors likely evaluate an individual’s job performance quite differently than his or her sub-
ordinates would, in that supervisors rate the focal individual in his or her role as a subordinate,
and subordinates rate the focal individual in his or her role as a supervisor. These two situations
would appear to constitute different domains of job performance. Further, it seems quite possible
that one could perform well in a subordinate role, but not in a supervisory role. According to the
preceding rationale, then interrater ratings from different ratings sources should not necessarily
be in agreement, in that they are not assessing the same, but different, aspects of job performance.
Stated a bit more directly, the various rater groups are not rating the same thing.

Therefore, the issue of interrater agreement across rater groups does not appear to be directly
relevant in the context of multi-source performance appraisal. It does not seem logical to expect
or require agreement between separate performance ratings that do not clearly measure the same
phenomenon. The fact that performance ratings provided by different rater groups do not refer to
the same phenomenon, and, thus, do not often agree with one another does not itself render the
process of multi-source performance appraisal invalid. Accordingly, reliability and validity in
this context are essentially non-issues, and interrater agreement across different rater groups
should not be considered a prerequisite to ratings validity. To use a well-worn analogy, compar-
ing ratings across rating groups is a bit like comparing apples and oranges: both are fruits, but
fruits of a different kind.

A Revised Perspective: Role-Based
Performance Appraisal

Based on the preceding discussion and analysis, one potential direction for multi-source
performance appraisal might be to recognize the role-related nature of performance ratings from
various rater perspectives and use this characteristic as an advantage for individual performance
assessment. That is, a role-based multi-rater method of performance appraisal might well be used
to extend and improve the prospect of multiple-rater evaluation in organizations. The issues and
implications of such an approach are described in the following sections.

Methodological issues and implications

First, from a methodological perspective, several important issues emerge. Although it does not
seem appropriate to require interrater agreement across rater groups, it does seem logical to
require agreement between observers within rater groups as an indication of the reliability and
validity of performance ratings (Borman, 1991). Clearly, if the evaluations of raters within a given
rater group (e.g. subordinates, peers) exhibit little systematic agreement or convergence con-
cerning the performance of a focal individual, then the reliability, and, hence, validity, of the
ratings should be seriously questioned. Interrater agreement within rater groups, then, would
seem an appropriate requirement that would enable assessments of rating validity to be conducted. Second, the preceding discussion and analysis also has implications for the creation of performance rating instruments for use by different rater groups. Certainly, individual performance in peer, subordinate, and supervisor roles consists of different behaviors, duties, and responsibilities, yet current multi-rater formats (e.g. 360-degree feedback systems) treat the rating criteria linked to these job facets as universal. It does not seem logical to expect supervisors, subordinates, and peers to respond similarly to identical, generic questions about an individual’s performance; each rater group experiences the focal individual’s performance behaviors in a different manner and capacity. It seems a bit naive to expect that a common rating instrument across all of these roles and rater groups is appropriate. A possible source of contamination and confusion in the analysis and interpretation of performance ratings across rater groups stems from the use of overall self-ratings as bases for paired interrater comparisons. Although, as argued, the evaluations provided by subordinates, peers, and supervisors are very likely role-specific, in contrast, focal individuals typically provide one global self-assessment across their various organizational roles. Thus, the comparison of others’ role-specific ratings with global self-ratings constitutes an analysis of different phenomena, and, as such, negates any requirements of interrater agreement as precondition to validity. One way to facilitate a role-based appraisal process would be to develop specific, role-related questions that could be targeted to each rater group for the evaluation of focal individuals in each of their various role-related capacities. If, for example, focal individuals were asked to rate themselves within specific roles, interrater agreement between self–other rating pairs (e.g. self-ratings of supervisory performance and subordinate ratings of supervisory performance) might be more accurately and appropriately assessed.

**Evaluative issues and implications**

In addition, by acknowledging explicitly what raters do implicitly (i.e. that rater evaluations are role-related), it may be possible to evaluate personnel more specifically and effectively. For example, following the rationale provided in this paper, individuals might be required to earn certain minimal rating levels across all non-self rater groups as a prerequisite when making important personnel decisions (e.g. raises, promotions). As such, this form of multi-source evaluation would constitute an improved method of evaluation in organizations by providing a more comprehensive assessment of individual performance that would enable more accurate personnel decisions to be made.

**Developmental issues and implications**

Finally, a role-based performance appraisal method also could provide greater specificity of feedback for developmental purposes. It would seem that if performance ratings were delineated and differentiated in a role-related manner, an improved yield of usable information for individual development would be provided. Such a method would allow one to pinpoint more precisely the locus of problem areas in individual performance and suggest possible avenues for remediation. For example, a focal individual might well exhibit performance deficiencies in his or her role as a subordinate, but not in his or her other roles. Because areas in need of improvement and personal development might well be specific to a particular role performed by an individual, the more context-specific feedback provided by such a method would provide superior feedback to that afforded by a more general appraisal system.
I am interested in receiving feedback from academic colleagues on this proposed appraisal method. Do you find it to be a viable method and alternative? Are there some critical problems or shortcomings that I have overlooked? I welcome your responses, feedback, and input.

References