Construct Validity Evidence for Multisource Performance Ratings: Is Interrater Reliability Enough?

Jisoo Ock
Seoul, Republic of Korea

As organizations become decentralized and work becomes team based, organizations are adopting performance management practices that integrate employees' performance information from multiple perspectives (e.g., 360-degree performance ratings). Both arguments for and against the use of performance ratings presented in the focal article focused on rater agreement (or lack thereof) as evidence supporting the position that multisource ratings are a useful (or not a useful) approach to performance appraisal. In the argument for the use of multisource ratings, Adler, Campion, and Grubb (Adler et al., 2016) point out that multisource ratings are advantageous because they lead to increased interrater reliability in the ratings. Although Adler and colleagues were not explicit about why this would be true, proponents of multisource ratings often cite the measurement theory assumption that increasing the number of raters will yield more valid and reliable scores to the extent that there is any correlation in the ratings (Shrout & Fleiss, 1979). In the argument against the use of multisource performance ratings, Colquitt, Murphy, and Ollander-Krane argued that because multisource ratings pool together ratings from raters who are systematically different in terms of their roles and perspectives about the target employee's performance, the increased number of raters is not expected to resolve the low level of interrater agreement that is typically observed in performance ratings (Viswesvaran, Ones, & Schmidt, 1996).

The focus on agreement (or disagreement) among raters as the key issue in the argument for or against the use of multisource performance ratings is not surprising given that reliability is a focal index of the quality of measurement scores. Reliability can be estimated in various ways depending on the relevant source of measurement error (Cortina, 1993), but interrater reliability in particular has emerged as *the* reliability index of choice in defining the psychometric quality of performance ratings (Murphy, 2008). Interrater reliability considers rater idiosyncrasies as a source of random measurement error (Schmidt & Hunter, 1996). However, an underlying

Jisoo Ock, Seoul, Republic of Korea.
Correspondence concerning this article should be addressed to Jisoo Ock, 355 Achasanro Gwangjin Acrotel A-1506, Seoul, Republic of Korea. E-mail: jisoo.ock@gmail.com

https://doi.org/10.1017/iop.2016.19 Published online by Cambridge University Press

assumption in the use of multisource ratings is that different rating sources provide unique performance-relevant information (Borman, 1974, 1997), meaning that different rating sources are actually expected to disagree with respect to their perception of target performance (Hoffman, Lance, Bynum, & Gentry, 2010). To the extent that each rating source provides uniquely meaningful information about a target employee's job performance, collapsing all variance not shared across raters into error would result in collapsing meaningful variance into error, which in turn is expected to produce inappropriate inferences regarding the construct validity of multisource performance ratings (Murphy & DeShon, 2000). Then, whether a systematic source effect in multisource ratings represents undesirable source-specific bias or independently valid performance-relevant information is an important issue that has implications for what multisource ratings represent. This commentary seeks to supplement the focal article's discussion of multiple source ratings by providing more detailed discussion of the construct validity of multiple source ratings. Specifically, I briefly review previous studies that have examined the validity and meaning of multisource ratings to assess the value of interrater reliability evidence in arguments for or against the use of multisource performance ratings.

Source Effect in Multisource Performance Ratings: Meaningful Variance Versus Bias

Consistent with the underlying assumption in the use of multisource ratings that different rating sources provide unique performance-relevant information, previous studies that examined the internal structure of multisource performance ratings have consistently found that rating source usually accounts for a significant proportion of variance in multisource performance ratings (Hoffman et al., 2010; Lance, Hoffman, Gentry, & Baranik, 2008; Woehr, Sheehan, & Bennett, 2005). From a traditional psychometric perspective, variance attributed to the rating source represents undesirable construct-irrelevant bias that should be reduced (Podsakoff, MacKenzie, Podsakoff, & Lee, 2003), but in multisource performance ratings, the underlying assumption is that different rating sources are expected to provide source-specific information about the target employee's job performance. Based on this perspective, there should be rating disagreements between rating sources, but the source of rating disagreement should represent reliable performance-relevant information in each of the rating sources.

An internal structure approach to examining the construct validity of multisource performance ratings may be supplemented with a nomological network approach that examines the patterns of covariance among source effects and external measures of performance to derive the extent

https://doi.org/10.1017/iop.2016.19 Published online by Cambridge University Press

to which rating source effects in multisource ratings represent substantively meaningful source-specific variance (Cronbach & Meehl, 1955). Specifically, source-specific variance should correlate with relevant externally measured constructs to the extent that they represent substantively meaningful performance-relevant variance. In addition, consistent with the theoretical explanation that different rating sources capture different aspects of performance, rating source effects should be differentially related to external measures of job performance to the extent that each rating source relies on different performance information to provide the ratings (Hoffman & Woehr, 2009).

Previous research that implemented a nomological network approach provided support to the assumption that rating source effects represent substantively meaningful performance-relevant variance. Namely, Hoffman and Woehr (2009) collected multisource ratings of managers enrolled in an executive master of business administration program (ratings collected from supervisors, peers, and subordinate employees of the participants) and asked them to also participate in an assessment center that measured different managerial skills (decision making, judgment, influencing others, persuasiveness, and coaching). Consistent with previous studies that examined the internal factor structure of multisource ratings, Hoffman and Woehr (2009) found clear support for the factor structure that modeled each rating source as a separate rating source factor (supervisor, peer, and subordinate). Correlations of the rating source factors with external variables provided further support to the assumption that source effects represent substantively meaningful performance-relevant variance. Specifically, all three factors showed weak to moderately significant correlations with the relevant measured managerial skills (e.g., $r = .29$ between subordinate latent factor and leadership skills). Furthermore, each rating source effect showed differential relationships with the measured managerial skills (i.e., confidence intervals around the correlation difference did not include zero; Meng, Rosenthal, & Rubin, 1992). For example, subordinate source factor showed a stronger correlation with the leadership skill factor ($r = .29$) than the peer (difference in $r = .13$) or manager (difference in $r = .16$) source factors.

Taken together, Hoffman and Woehr's (2009) results indicate that not only do rating source factors represent substantively meaningful variance but also each factor can provide source-specific information that is uniquely related to performance. These findings provide a more in-depth perspective into the meaning of the rating source effect in multisource ratings and what they represent that cannot be derived from internal structure or interrater reliability based approaches to examining the construct validity of multisource ratings.

https://doi.org/10.1017/iop.2016.19 Published online by Cambridge University Press

Implications for Validity of Multisource Ratings

As multisource ratings have become an increasingly common performance measurement in practice organizations, there has been a corresponding increase in the amount of research attention paid to investigating the psychometric properties of multisource ratings (e.g., Conway, 1996; Conway & Huffcutt, 1997; Mount, Judge, Scullen, Systma, & Hezlett, 1998). Much of this research has relied on an internal approach that examines the covariance of ratings made by different sources, including interrater reliability evidence that was briefly discussed in the focal article. Interestingly, the contrast between assumptions underlying the use of multisource ratings and assumptions regarding what represents *true* variance in interrater reliability elicits questions regarding the information that interrater reliability estimates can provide about the construct validity of multisource performance ratings. That is, interrater reliability considers rater idiosyncrasies as a source of random measurement error, but the use of multisource ratings is based on the assumption that each rating source provides a unique perspective about a target employee's performance. As a result, different rating sources are expected to have a low level of agreement, but each source is expected to provide source-specific valid performance information.

In addition to the consistent stream of research evidence that has shown that rating source factors represent a reliable source of variance in multisource performance ratings, Hoffman and Woehr's (2009) findings with respect to the relationship between different rating source factors and measures of job performance provide evidence supporting the underlying assumption in the multisource performance ratings that rating source represents a meaningful source of specific variance as opposed to bias. Although the authors in the focal article focused on the interrater reliability evidence to support their argument for or against the use of multisource ratings, the literature reviewed in this commentary suggests that interrater reliability alone is not sufficient as evidence for (or against) the construct validity of multisource performance ratings.

References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly? A debate. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(2), 219–252.

Borman, W. C. (1974). The rating of individuals in organizations: An alternative approach. *Organizational Behavior and Human Performance, 12*, 105–124.

Borman, W. C. (1997). 360 ratings: An analysis of assumptions and research agenda for evaluating their validity. *Human Resource Management Review, 7*, 299–315.

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

https://doi.org/10.1017/iop.2016.19 Published online by Cambridge University Press

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331–360.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.

Hoffman, B. J., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). Rater source effects are alive and well after all. *Personnel Psychology, 63*, 119–151.

Hoffman, B. J., & Woehr, D. J. (2009). Disentangling the meaning of multisource performance rating source and dimension factors. *Personnel Psychology, 62*, 735–765.

Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*, 223–232.

Meng, X. L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*, 172–175.

Mount, M. K., Judge, T. A., Scullen, S. E., Systma, M. R., & Hezlett, S. A. (1998). Trait, rater, and level effects in 360-degree performance ratings. *Personnel Psychology, 51*, 557–576.

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

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology, 53*, 873–900.

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., & Lee, J. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879–903.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Woehr, D. J., Sheehan, M. K., & Bennett, W. (2005). Assessing measurement equivalence across rating sources: A multitrait–multirater approach. *Journal of Applied Psychology, 90*, 592–600.