Comparing Rater Groups: How To Disentangle Rating Reliability From Construct-Level Disagreements

Chockalingam Viswesvaran
Florida International University

Deniz S. Ones
University of Minnesota

Frank L. Schmidt
University of Iowa

In this commentary, we build on Bracken, Rose, and Church's (2016) definition stating that 360° feedback should involve "the analysis of meaningful comparisons of rater perceptions across multiple ratees, between specific groups of raters" (p. 764). Bracken et al. expand on this component of the definition later by stressing that "the ability to conduct *meaningful comparisons of rater perceptions* both between (inter) and within (intra) groups is central and, indeed, unique to any true 360° feedback process" (p. 767; italicized in their focal article). Bracken et al. stress that "This element of our definition acknowledges that 360° feedback data represent rater perceptions that may contradict each other while each being true and valid observations" (p. 767).

Bracken et al. (p. 768) present six questions, three of which stress intergroup comparisons: Question 2, which reads, "Is the feedback process conducted in a way that formally segments raters into clearly defined and meaningful groups?"; Question 4, which reads, "Is the feedback collected . . . to establish reliability, *which can vary by rater group*?" [emphasis added]); and Question 5, which reads, does "the feedback process . . . provide the user with sufficiently clear and *reliable* [emphasis added] insights into inter- and intragroup perceptions?" The original definition, as well as the three questions, clearly emphasizes the need for delineating distinct groups of raters. Finally, in discussing how we can facilitate evolution of 360° feedback, Bracken et al. call for a more accurate description of how group membership is operationalized.

Chockalingam Viswesvaran, Department of Psychology, Florida International University; Deniz S. Ones, Department of Psychology, University of Minnesota; Frank L. Schmidt, Department of Management, University of Iowa.

Correspondence concerning this article should be addressed to Chockalingam Viswesvaran, Department of Psychology, Florida International University, 1200 SW 8th Street, Miami, FL 33199. E-mail: vish@fiu.edu

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

Bracken et al.'s definition and subsequent discussions focus on how 360° feedback can provide insights into differences among raters belonging to different groups (organizational hierarchy, projects, etc.). In this commentary, we emphasize that practice and research need to distinguish between differences due to construct-level disagreements and rater reliability (Viswesvaran, Schmidt, & Ones, 2002). Agreement between raters (within the same group) is reduced by interrater unreliability. Agreement between raters belonging to two distinct groups is lowered by both (a) interrater unreliability and (b) any disagreement on the nature of what is rated (i.e., construct-level disagreement).

The observed correlation between raters belonging to two distinct groups is reduced if raters from the two groups are rating two different constructs (or raters from the two groups have different perceptions of what is rated). This difference could result because of differences in their understanding of the exact nature of what is rated (i.e., construct-level disagreements). Conversely, even when the raters from the two distinct groups are rating the same construct (i.e., have similar perceptions of what is rated), the observed correlation is attenuated because of idiosyncratic perceptions that vary among raters from within the same group. The rater-specific (within raters from the same group) idiosyncratic component parallels item-specific variance assigned to measurement error in computations of coefficient alpha and other similar reliability indicators (Charles, 2005; Salgado, Moscoso, & Lado, 2003; Schmidt & Hunter, 1996; Schmidt, Viswesvaran, & Ones, 2000). We emphasize that the rater-specific variance component in ratings is substantial in most areas of research (cf. Huffcutt, Culbertson, & Weyhrauch, 2013; Viswesvaran, Ones, Schmidt, Le, & Oh, 2014; Voskuijl & van Sliedregt, 2002) and especially so in performance assessments (Viswesvaran, Ones, & Schmidt, 1996).

Thus, examination of the observed correlation between raters from different groups does not provide sufficiently clear and *reliable* insights into inter- and intragroup perceptions (Bracken et al.). What is needed is an examination of the interrater reliability for raters within each group, as well as the observed correlation between raters from different groups. Given the observed correlation between raters from different groups and the interrater reliability in each group, we can disentangle true construct-level differences from rater unreliability.

Diagnostic Steps in Disentangling Construct-Level Differences and Rater Reliability

For 360° feedback to provide users with clear and reliable insights into intra- and intergroup perceptions, first the interrater reliability of raters within

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

each well-defined group should be examined. Interrater reliability estimates for the two groups of raters can be compared, and this comparison can be the basis for focused training programs, if needed.

Second, to diagnose whether there are true differences between predefined rater groups, estimated true-score correlations between raters from the different groups can be computed. Specifically, the observed correlations between raters from different groups are corrected with interrater reliability estimates from each group. If raters from different groups are rating the same construct, the estimated true-score correlation is expected to be 1.0 (within sampling error). That is, once the attenuating effects of measurement error are eliminated, agreement will be perfect. If the corrected correlation is less than 1.0, an examination of whether the associated confidence intervals include 1.0 should ensue. To do this, the confidence intervals around the *observed* correlation should be constructed, and then *the end points of the interval should be corrected for the attenuating effects of measurement error*—the resulting values are the end points of the confidence interval around the corrected correlation. (It is important to note that it is inappropriate to form the confidence intervals around the corrected correlation using the standard formula, as the sampling error associated with the correlation is increased due to reliability corrections.)

If the confidence intervals include 1.0, the conclusion is that any lack of convergence between the groups of raters may not be due to failure of the two groups of raters to assess the same construct. We wish to highlight that even if the confidence intervals include 1.0 (or if the estimated true-score correlation is 1.0), it is a good practice to include raters from distinct groups in the 360° feedback assessments for other reasons. Doing so will (a) increase the reliability of the measurement of 360° feedback constructs, (b) provide a broader coverage of the content domain of the constructs assessed, and (c) enhance user acceptability.

If the confidence intervals do not include 1.0, then the inference is that the observed correlation is lowered due to (a) the biasing effects of measurement error in each group of raters (rater reliability) as well as (b) underlying perceptions of what is rated differing across the two groups of raters (i.e., construct-level disagreement). If the estimated true-score correlation is .60 and the observed correlation is .20, we infer that rater unreliability reduces the observed correlation by .40.

A reviewer raised the concern that there is no reason why construct-level disagreement cannot affect ratings by raters from the same group. The reviewer argued that it is entirely possible that a focal leader might behave differently toward raters from different groups (e.g., providing helpful feedback to peers but unhelpful feedback to direct reports). The reviewer continued that if that happens, we should expect raters from different groups to

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

disagree when rating the focal leader and the true-score correlation between different groups will not be 1.0.

If the process described above occurs, it does not show up in the ratings, because the true-score correlation between ratings by different groups is in fact essentially 1.00 (cf. Mount, Judge, Scullen, Sytsma, & Hezlett, 1998; Viswesvaran et al., 2002). Thus, although this process is plausible, it is not consistent with the empirical evidence. We do agree with the reviewer that if this process occurs, the true-score correlation will not be 1.0. However, the observed correlation will be lowered much more (than the true-score correlation) because there will be disagreements among peers (and among direct reports) on the helpful feedback. The unreliability among peers and among direct reports should be removed from the observed correlation before concluding that peers and direct reports disagree on ratings of feedback behaviors.

The reviewer continued, "The same question can be asked about raters from within the same group. Leader–member exchange research shows that a leader can be expected to behave differently toward different direct reports (e.g., in-group versus outgroup members). If this happens, we would expect that direct reports will not agree when rating the focal leader because they are observing different behaviors by that focal leader."

Although plausible on surface, if this process occurs, it will be reflected in low interrater reliability among subordinate ratings. We know of no evidence showing clustering of subordinate ratings, with high agreement within clusters and low agreement between clusters, where clusters are homogeneous on relationships with the supervisor. Therefore, it is reasonable to take the pooled (averaged) ratings by subordinates as reflecting the construct (see also Viswesvaran et al., 2005).

In essence, the reviewer is "asking about a circumstance where there is no construct-level disagreement but where raters from different groups (or even raters from the same group) observe different behaviors from the focal leader and therefore provide different ratings of the focal leader." If different raters see different behaviors (consistently/reliably), then they are rating different constructs. However, the observed correlation will be lowered by differences in behaviors observed and rated as well as by unreliability in ratings. The main thrust of this comment is to stress the need to disentangle the unreliability in the individual ratings from true differences.

Construct-Level Disagreements and Rater Reliability in Supervisory and Peer Ratings of Job Performance

Bracken et al. stress in their focal article the distinction between 360° feedback and "alternate forms of feedback (AFF)." Bracken et al. (p. 765), despite academic research (Mount et al., 1998) clearly showing that it might be

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

better from a true-score measurement perspective to combine ratings from naturally occurring work groups (peers, supervisors, etc.), argue that "a true 360° feedback assessment, under our definition, must be designed in such a way that differences in rater perception are clearly identified and meaningful comparisons can be made between perceptions of different rater groups, agreement (or lack thereof) within a rater group, and even individual raters (e.g., self, manager) where appropriate."

There are two different issues here that need to be clarified. First, even accepting Bracken et al.'s assertion that 360° feedback must be designed in such a way that differences in rater perceptions are clearly identified, the procedures outlined here need to be followed to disentangle rater reliability and construct-level disagreements. That is, an empirical basis is essential to support the hypotheses that the conceptually distinct, carefully formed rater groups are meaningfully different and that raters within a group share perceptions among themselves to a greater degree than with raters from other groups. The procedures outlined here provide a systematic process to evaluate the validity of the groups formed.

Second, in performance appraisal and job performance ratings, a common assertion is that peers and supervisors are rating different constructs. Elaborate process mechanisms have been postulated to explain why peers and supervisors are rating different constructs even when presented with the same rating instrument (Borman, 1979; Wohlers & London, 1989). Hypotheses have been advanced that the opportunity to observe differs across peers and supervisors, which results in construct-level differences. The behavioral ambiguity in defining a performance dimension has been hypothesized to differ between peers and supervisors (what is construed as counterproductive behaviors differs between peers and supervisors, what is construed as effective leadership differs between peers and supervisors, etc.).

Against this backdrop, however, empirical data have accumulated (e.g., Mount et al., 1998; Viswesvaran et al., 2002) to show that peers and supervisors are rating the same underlying constructs and that the observed between group correlations are attenuated due to interrater unreliability in peer and supervisory ratings (Viswesvaran et al., 1996, 2002). Facteau and Craig (2001), using confirmatory factor analysis (CFA) and item response theory (IRT) with a large dataset of 360° feedback ratings, found construct equivalence between peers, supervisors, and subordinates. Maurer, Raju, and Collins (1998), employing both CFA and IRT, found evidence of construct-level agreements between peer and subordinate ratings of performance. Viswesvaran et al. (2002), using meta-analytic techniques, found evidence of construct-level agreement between peer and supervisor ratings. Thus, multiple studies using large datasets and different analytic techniques (CFA, IRT, meta-analysis) have found that *at the true-score level, supervisor*

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

and peer ratings of job performance and its dimensions correlate 1.00, showing that they are rating the same constructs. This refutes the idea that different groups are perceiving and rating different aspects of performance.[1]

Conclusion

An examination of (a) the observed correlation between raters from different groups, (b) the interrater reliability within each group, and (c) the observed correlation corrected for interrater unreliability in each group thus serves as a diagnostic tool to assess where the disagreements occur. This approach also provides a basis for identifying empirically distinct rater groups. Bracken et al. lament how the term "peers" represents a very heterogeneous group. A similar comment can be made about "customers" as a group. The process outlined here will help in empirically testing and validating hypothesized unique groups of raters. Disentangling rater reliability and construct-level disagreements is essential.

References

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

Bracken, D. W., Rose, D. S., & Church, A. H. (2016). The evolution and devolution of 360° feedback. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 9*(4), 761–794.

Charles, E. P. (2005). The correction for attenuation due to measurement error: Clarifying concepts and creating confidence sets. *Psychological Methods, 10*(2), 206–226.

Facteau, J. D., & Craig, S. B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86*, 215–227.

Huffcutt, A. I., Culbertson, S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21*(3), 264–276.

Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83*, 693–702.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R., & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51*(3), 557–576.

Salgado, J. F., Moscoso, S., & Lado, M. (2003). Test–retest reliability of job performance dimensions in managers. *International Journal of Selection and Assessment, 11*, 98–101.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199–223.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology, 53*, 901–912.

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.

[1] In this section we focused on the construct of job performance because most 360° feedback applications target job performance constructs (i.e., the empirical results we present are specific to the job performance domain). However, the diagnostic tools we have described are relevant for 360° assessments for other domains such as personality, job analysis, assessment centers, interviewer ratings, and so on.

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analyses. *Industrial and Organizational Psychology, 7*, 507–518.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (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, F. L., & Ones, D. S. (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.

Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*(1), 52–62.

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

Why the Qualms With Qualitative? Utilizing Qualitative Methods in 360° Feedback

Adam Kabins
Korn Ferry Hay Group

Although the authors of the focal article provide a comprehensive definition of 360° feedback, one exclusionary criterion results in an overly narrow definition of 360° feedback. Specifically, Point 3 in their definition described the criticality of strictly using quantitative methods in collecting 360° feedback. The authors provided a brief rationale by stating, "Data generated from truly qualitative interviews would not allow comparisons between rater groups on the same set of behaviors" (Bracken, Rose, & Church, 2016, p. 765). Although there is little doubt about the value in taking a quantitative approach for gathering 360° feedback, it is not clear why this has to be the sole approach. Below, I outline three issues with taking this constricted methodology. That is, first, excluding qualitative methods is not in line with the purpose of 360° feedback, which is directed at minimizing criterion deficiency. Second, qualitative methodologies (in conjunction with quantitative methodologies) are more equipped to provide and inspire a call to action (supporting the change component addressed by the authors). Finally, there are qualitative methods that allow for rigorous quantitative analysis and can provide an additional source of macro organizational-level data.

Adam Kabins, Korn Ferry Hay Group, Dallas, Texas.

Correspondence concerning this article should be addressed to Adam Kabins, Korn Ferry Hay Group, Suite 1450, 2101 Cedar Springs Road, Dallas, TX 75201. E-mail: adamkabins@gmail.com

https://doi.org/10.1017/iop.2016.88 Published online by Cambridge University Press