Industrial and Organizational Psychology (2024), **17**, 309–313
doi:10.1017/iop.2024.16

COMMENTARY

On putting the horse (raters and criteria) before the cart (variance components in ratings)

Pengda Wang, Hwayeon Myeong, and Frederick L. Oswald

Department of Psychological Sciences, Rice University, Houston, TX, USA
Corresponding author: Pengda Wang; Email: pw32@rice.edu

In their focal article pertaining to the nature of job performance ratings, Foster et al. (2024) decompose rating variance into ratee main effects, rater main effects, and rater–ratee interaction effects. In doing so, the authors emphasize that ratee main effects, ideally reflecting actual ratee performance, tend to constitute only 20%–30% of the total variance in job performance ratings. Under this assumption, they then claim that predicting ratee main effects variance, rather than total variance, will provide a more precise (and higher) reflection of a predictor's utility.

Although we understand the central argument, it steers us toward a broader issue on which this argument critically depends: significant flaws within conventional performance ratings demand continued serious research and practice efforts toward improving them. Continuing to focus on improvements in defining and measuring performance (i.e., the "criterion problem"; Austin & Villanova, 1992) is a necessary prerequisite for the variance components of performance ratings to be interpretable and useful. The nature of performance criteria, the types of jobs, raters, and rater training, the number of raters, and the choice of selection-relevant predictors are just a few of the critical concerns to consider in tandem with the concerns raised in the focal article. Statistically estimating the effectiveness of selection measures based on performance measures and raters that are known a priori to be highly deficient can of course be done mechanically but obviously would be less illuminating. In this commentary, we provide some big-picture comments on performance ratings helpful for interpreting any statistics derived from them.

Enhance interrater reliability and validity (rather than adjusting leniency bias)

The biases of raters, or rater main effects, are a significant source of systematic error in traditional performance ratings. The focal article suggests the classical approach of gathering evaluations from several raters and averaging them to diminish rater main effects and boost the variance attributed to ratee main effects. This approach only works well if certain assumptions are met, such as (a) the shared variance among raters is largely job relevant (e.g., not irrelevant due to shared-but-incorrect assumptions about ratee performance) and (b) the variance unique to raters is truly random error (e.g., not a rater's unique observation of ratee performance). Psychometric error estimates indicate but do not guarantee the quality of ratings. Therefore, an important a priori approach is to improve interrater reliability and agreement through developing and implementing rater training on the basis of well-developed performance standards and well-understood administrative or developmental goals, as found in prior research (McIntyre et al., 1984; Murphy & Cleveland, 1995). That way, whether rater agreement statistics and validity coefficients are low or high, we have a better understanding of *why* that might be the case (vs. merely assuming that ratings are job relevant and assuming that raters know how to conduct ratings). Although the focal article argues that using the "leniency measure" can be a promising method to correct the rater bias, it could often be infeasible, for one because it requires a sufficient number of rating instances per rater.

Further, given what we know about the multidimensionality of job performance, performance ratings are surely influenced by the type of performance being assessed, whether it involves task-specific behaviors, organizational citizenship behaviors (OCBs), or counterproductive work behaviors (CWBs). Even these broad performance domains have their own important distinctions; for example, CWB has been found to encompass 11 distinct behavioral dimensions (Gruys & Sackett, 2003). Thus, rater training will stand to enhance the quality of ratings by focusing on the correct performance dimensions while diminishing the influence of subjectivity found in more general ratings of performance. Instituting a standardized training program for raters, one that includes rater retraining over time, can foster greater consistency and accuracy in evaluations across different raters, ensuring they are more firmly rooted in a thorough understanding of the job's varied performance dimensions. Importantly, the goals of raters when conducting performance ratings (e.g., focusing on fairness, strengths, and weaknesses) must be aligned during rater training as well (Murphy et al., 2004). Psychometrically speaking, when raters are better trained, they are more reliable, and you therefore need fewer of them per ratee (McIntyre et al., 1984).

Job relevance is the key

The essence of performance appraisal content is determined by its relevance to the job. An accurate job analysis provides an in-depth understanding of the core responsibilities and requirements of the position, thus providing a solid foundation for the development of performance evaluation criteria (Morgeson et al., 2019). Combined with the aforementioned rater training, this will stand to increase ratee variance beyond the 20% and 30% mentioned in Foster et al. (2024) and decrease rater main effects, and rater × ratee interaction effects. If one decides not to factor in these latter effects when estimating performance ratings, as Foster et al. (2024) have done, that does not make these effects go away. Rather than attempting statistical adjustments, they should be minimized by design, via rater training combined with high-quality, job-relevant performance assessments.

Suggestions for improving performance ratings (and their variance components)

To summarize and extend our points above, we propose the following suggestions to address the inherent limitations in traditional performance ratings:

Rater training

Rater training establishes a shared understanding of performance standards as established by a job analysis, leading to uniform criteria across evaluations for more accurate and reliable assessments. Unlike post hoc adjustments for leniency biases in statistical models, which fail to address the root cause of varying standards among raters, job-analysis-based rater training is a proactive approach that directly improves the quality of rating data. Such training programs would ensure that evaluations are not only consistent across different raters but also grounded in a comprehensive understanding of a job's multifaceted performance dimensions. It also reduces the influence of statistical artifacts that if corrected for in estimating operational validities (correcting for criterion reliability) will come with standard errors that are inflated by that correction factor (Oswald et al., 2015).

https://doi.org/10.1017/iop.2024.16 Published online by Cambridge University Press

Mechanical combination of human ratings

A supplementary approach is to apply a mechanical combination of performance data (ratings and otherwise) to achieve more precise performance evaluations and better assessment systems. For decades, scholars have highlighted the advantages of the mechanical combination of human ratings over traditional human judgment across diverse domains of research and practice (e.g., organizations, health and medicine, and college admissions; Dawes et al., 1989; Grove & Meehl, 1996; Kuncel et al., 2013; Meehl, 1954). The superiority of the mechanical combination lies in the elimination of the human tendency to apply inconsistent weights to information when making judgments (Kuncel et al., 2014). Mechanical (algorithmic) combination of data applies fixed rules that are transparent and subject to scrutiny and modification, thus increasing the fairness and reliability of the final scores, while aiming to reduce sources of human subjective biases.

Technology-based measurement (e.g., AI)

Technologies can measure human performance as well as combine those measures algorithmically. Of course, this comes with the important problem of ensuring technologies are measuring the "right" behaviors at work, not just any behaviors; and it also should not be implemented essentially as a surveillance tool that decreases job satisfaction and morale (Ravid et al., 2022). Compared with the process and data associated with human ratings, the technological data gathering process, and the data themselves, have the potential to be more transparent and subject to expert input to increase its fairness and accuracy (Woo et al., 2024). For example, in AI-based algorithmic interviews, the content and dimensions of interviews can be combined mechanically in a consistent manner to all applicants and evaluate their performance (Hickman et al., 2022); biases can be monitored while measuring job-relevant KSAOs (Putka et al., 2022). We are realists here, not idealists: we are not saying technological assessment is perfect, only that it has the potential to provide improvements over the idiosyncrasies and subjectivity often found in human raters who are overworked and under trained. Technological assessments can be designed based on detailed job analyses that identify the essential tasks and competencies required for a role, which can ensure that performance evaluations are grounded in concrete, job-specific criteria.

It could be quite beneficial to incorporate automated or semiautomated (e.g., AI-assisted) technical solutions to performance evaluation of desired aspects of task performance, OCBs, and CWBs. Even if not all aspects of performance could be evaluated in this manner, so long as appropriate bias-reducing guardrails are in place, a technology-driven approach could offer significant employee insights while reducing time and expense in performing evaluations. Such technological approaches to performance evaluation are not widespread and are not cost free; however, in the long term, they would stand to be more cost effective and simpler to deploy, minimizing or even eliminating the need for human raters and bypassing common logistical hurdles, such as scheduling conflicts. By contrast, traditional methods aimed at minimizing errors often come with high costs with no guarantees. Just as more items can improve the reliability of a test, gathering evaluations from multiple trained raters can reduce rater main effects but also can incur substantial expenses. Dilchert (2018) indicates that the average cost for each human rater ranges from $37 to $110 per hour. Sometimes investing in involving multiple raters for each ratee might only slightly elevate the proportion of ratee main effects, such as from 25% to 30% in Jackson et al. (2020), and thus we would urge practitioners to engage in a cost–benefit analysis of this approach before wholesale implementation.

Thus, we still promote the use of job analysis for defining performance criteria, we advocate for rater training in traditional performance evaluations, and we also anticipate the transition to technology-based evaluations that should be developed with the same care. Although technology-based performance assessment, combined with mechanical combination methods, might offer many advantages in the ways noted, they are not without potential shortcomings. Biases need to

https://doi.org/10.1017/iop.2024.16 Published online by Cambridge University Press

be monitored in traditional performance assessments involving human raters. Biases also remain of high concern in technology-based assessments, perhaps more so to the extent these assessments are shrouded in the messiness of big data combined with the opaqueness of machine learning algorithms applied to those data. Another significant issue might be the lack of adaptability of technological assessments of human performance in new or unexpected performance scenarios (LeCun et al., 2015) and in new and emerging occupations. This limitation may call for hybrid strategies of performance appraisal that merge the precision of the mechanical method with trained raters' understanding of performance and context. Machine learning algorithms can provide ratings based on past data and outcomes; human experts can provide input on novel situations on which algorithms may not be adequately trained; and the two can be considered and combined in a hybrid approach. This hybrid approach could operate in an iterative loop, where human judgments help to refine and adjust the algorithm's predictions based on new information or contexts that emerge, and algorithmic ratings based on past data inform human ratings made in the current job context. This dual approach would hope to leverage the systematic accuracy and consistency of mechanical methods while incorporating the depth and flexibility of human evaluators.

At the end, we appreciate the opportunity in this commentary to provide some broader considerations and context around the Foster et al. (2024) focal article.

Competing interests. None.

References

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology*, **77**(6), 836–874. https://doi.org/10.1037/0021-9010.77.6.836

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, **243**(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Dilchert, S. (2018). Cognitive ability. In D. S. Ones, N. Anderson, C. Viswesvaran, & H. K. Sinangil (Ed.), *SAGE handbook of industrial, work and organizational psychology: Personnel psychology and employee performance* (2nd ed. pp. 248–275). Sage Publications.

Foster, J., Steel, P., Harms, P., O'Neill, T., & Wood, D. (2024). Selection tests work better than we think they do, and have for years. *Industrial and Organizational Psychology*, **17**(3), 269–282.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, **2**(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Gruys, M. L., & Sackett, P. R. (2003). Investigating the dimensionality of counterproductive work behavior. *International Journal of Selection and Assessment*, **11**(1), 30–42. https://doi.org/10.1111/1468-2389.00224

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, **107**(8), 1323–1351. https://doi.org/10.1037/apl0000695

Jackson, D. J. R., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S. (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, **105**(3), 312–329. https://doi.org/10.1037/apl0000434

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, **98**(6), 1060–1072. https://doi.org/10.1037/a0034156

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). A meta-analysis of letters of recommendation in college and graduate admissions: Reasons for hope. *International Journal of Selection and Assessment*, **22**(1), 101–107. https://doi.org/10.1111/ijsa.12060

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, **521**(7553), 436–444. https://doi.org/10.1038/nature14539

McIntyre, R. M., Smith, D. E., & Hassett, C. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology*, **69**(1), 147–156. https://doi.org/10.1037/0021-9010.69.1.147

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. https://doi.org/10.1037/11281-000.

Morgeson, F. P., Brannick, M. T., & Levine, E. L. (2019). *Job and work analysis: Methods, research, and applications for human resource management*. Sage Publications. https://doi.org/10.4135/9781071872536

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage Publications.

Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters who pursue different goals give different ratings. *Journal of Applied Psychology*, **89**(1), 158–164. https://doi.org/10.1037/0021-9010.89.1.158

Oswald, F. L., Ercan, S., McAbee, S. T., Ock, J., & Shaw, A. (2015). Imperfect corrections or correct imperfections? Psychometric corrections in meta-analysis. *Industrial and Organizational Psychology*, **8**(2), e1–e4. https://doi.org/10.1017/iop.2015.17

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2022). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, **38**(2), 385–410. https://doi.org/10.1007/s10869-022-09824-0

Ravid, D. M., White, J. C., Tomczak, D. L., Miles, A. F., & Behrend, T. S. (2022). A meta-analysis of the effects of electronic performance monitoring on work outcomes. *Personnel Psychology*, **76**(1), 5–40. https://doi.org/10.1111/peps.12514

Woo, S. E., Tay, L., & Oswald, F. L. (2024). *Artificial intelligence, machine learning, and big data: Improvements to the science of people at work and applications to practice*. Personnel Psychology. https://doi.org/10.1111/peps.12643

Cite this article: Wang, P., Myeong, H., & Oswald, F. L. (2024). On putting the horse (raters and criteria) before the cart (variance components in ratings). *Industrial and Organizational Psychology* **17**, 309–313. https://doi.org/10.1017/iop.2024.16

https://doi.org/10.1017/iop.2024.16 Published online by Cambridge University Press