Of Babies and Bathwater: Don't Throw the Measure Out With the Application

David J. Woehr
University of North Carolina Charlotte

Sylvia G. Roch
University at Albany, State University of New York

Adler et al. (2016) provide a discussion of the pros and cons surrounding the issue of "Getting Rid of Performance Ratings." Yet neither the pro nor the con side of the debate appears to fully consider the central role of performance ratings outside the realm of performance management. In 1949, Robert L. Thorndike wrote,

> The key to effective research in personnel selection and classification is *an adequate measure of proficiency on the job.* Only when proficiency measures can be obtained for the individuals who have been tested is it possible to check the effectiveness of test and selection procedures. (Thorndike, 1949, p. 6, italics added)

This statement remains as true today as it was in 1949. For better or worse, performance ratings have been the most frequently used measure of "proficiency on the job" for nearly 100 years (Austin & Villanova, 1992). And if performance rating in organizations is truly a "failed experiment," does this call into question all of the research for which performance ratings have served as the criteria? Performance ratings are the criterion of choice not only for validating selection measures but also for evaluating training interventions (Goldstein & Ford, 2002).

So before admitting defeat with respect to performance ratings, we believe it important to consider the evidence suggesting performance ratings do indeed capture performance. One such piece of evidence often not discussed is the extent to which performance ratings are correlated with conceptually relevant predictors. In general, predictors that should be related to individual job performance do indeed predict performance ratings. As noted above, in studies investigating the relationship between predictors and job performance, job performance is most often assessed using supervisory ratings (Schmidt & Hunter, 1998). The literature clearly demonstrates that a variety of predictors frequently utilized for selection and assessment

David J. Woehr, Department of Management, University of North Carolina Charlotte; Sylvia G. Roch, Department of Psychology, University at Albany, State University of New York.

Correspondence concerning this article should be addressed to David J. Woehr, Department of Management, University of North Carolina Charlotte, 9201 University City Boulevard, Charlotte, NC 28223-0001. E-mail: dwoehr@uncc.edu

https://doi.org/10.1017/iop.2016.25 Published online by Cambridge University Press

purposes demonstrate substantial relationships with job performance as typically assessed. For example, cognitive ability has a corrected validity of approximately $\rho = .50$ (e.g., Bertua, Anderson, & Salgado, 2005), even though estimates range as high as $\rho = .62$ (Salgado, Anderson, Moscoso, Bertua, & De Fruyt, 2003) and as low as $\rho = .45$ (Hunter, 1983). Job knowledge has a validity coefficient of $\rho = .48$ (Hunter & Hunter, 1984), $\rho = .36$ for assessment centers (Arthur, Day, McNelly, & Edens, 2003; Gaugler, Rosenthal, Thornton, & Bentson, 1987), $\rho = .32$ for biodata (Rothstein, Schmidt, Erwin, Owens, & Sparks, 1990), $\rho = .37$ for interviews (Huffcutt & Arthur, 1994; McDaniel, Whetzel, Schmidt, & Maurer, 1994), and $\rho = .34$ for situational judgment tests (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001). Given that job performance is most often assessed via supervisory performance ratings, these studies provide an indication of the general level of predictability of these ratings. Even though these correlations do not necessarily provide incontrovertible evidence with respect to the construct validity of performance ratings, they are certainly consistent with theoretical expectations and inconsistent with the notion that performance ratings do not work.

Moreover, several studies have specifically examined differences across various criterion methods with respect to criterion-related validity. These studies provide a direct comparison of the level of predictability between rating-based measures and more objective measures such as production records, sales records, and output, often referred to as hard criteria. Schmitt, Gooding, Noe, and Kirsch (1984), for example, investigated published validation studies between 1964 and 1982. They examined validity coefficients as a function of the type of criterion used. They report remarkably similar validity coefficients among some criteria. In particular, the average r was .26 for performance ratings, $r = .25$ for turnover, and $r = .21$ for productivity. It should be noted that other criteria had higher validities ($r = .36$ for status change, and $r = .40$ for wages). Similarly, Schmidt and Rader (1999) investigated phone interviews and found almost the same validity coefficients for performance ratings ($\rho = .40$) as for production records ($\rho = .40$) and job tenure ($\rho = .39$). Interestingly, they found higher validity coefficients for performance ratings than for sales performance ($\rho = .24$). In general, research indicates that supervisory performance ratings typically demonstrate criterion-related validity levels as good as, if not better than, those of other criterion measures. Predictability has long been viewed as a desirable criterion characteristic (e.g., Blum & Naylor, 1968).

It may be argued that subjective measures such as supervisory performance ratings are likely to demonstrate bias to a greater degree than more objective criteria. However, in a meta-analysis investigating to what extent race influences performance evaluations, McKay and McDaniel (2006)

https://doi.org/10.1017/iop.2016.25 Published online by Cambridge University Press

found that race influenced subjective and objective task-related ratings to essentially the same degree ($d = 0.18$ versus $d = 0.20$). Interestingly, race influenced subjective estimates of absenteeism ($d = -0.01$) to a lesser degree than more objective estimates of absenteeism ($d = 0.11$). It is not possible to determine to what extent these effect sizes represent bias versus true performance differences. Nevertheless, these findings do indicate that performance ratings demonstrate no more race-based differences than do objective measures.

Although it is important not to overly confound construct and method when making comparisons among predictor and criterion relationships (Arthur & Villado, 2008), the literature to date suggests that supervisor ratings of job performance are consistently predicted by those constructs expected to be related to job performance. Rating-based measures of performance appear to be as, if not more, predictable than nonrating measures. It is also notable that supervisor ratings of performance tend to show no more race-based differences than do other more objective criteria.

Performance ratings are the criterion of choice not only for selection measures but also when evaluating whether training has influenced employees' on-the-job behavior or, in other words, training transfer (Goldstein & Ford, 2002). Taylor, Russ-Eft, and Taylor (2009) in their meta-analysis of the literature investigating the transfer of management training found that supervisor ratings were the most frequently used criterion when evaluating training transfer versus self, peer, and subordinate rating. Even more important, Taylor et al. found larger effect sizes for training when the performance ratings targeted the training content. This is in line with Kraiger's (2002) suggestion that when determining the organizational payoff of training, it is important to focus on changes in behavior on the job. Performance ratings allow the organization to assess to what extent training has influenced on-the-job behavior, information that is not available in bottom line performance measures, which may be influenced by factors outside of the employees' control.

Of course, one could argue that the performance ratings used for personnel research are not the same as those used for administrative purposes in organizations. There is certainly a good bit of research focusing on this "purpose of appraisal" effect (Jawahar & Williams, 1997). Yet it has been widely noted that the same performance ratings are regularly used for multiple purposes, and much research utilizes operational performance ratings. Even if one accepts that research-oriented ratings are different from administration-oriented ratings, this suggests that the problem is not with the performance ratings themselves but with the way in which they are used. Of course, much has been written about problems in the performance management process. Both sides of the debate seem to agree that performance

https://doi.org/10.1017/iop.2016.25 Published online by Cambridge University Press

management practices are almost universally not well implemented. But if we suddenly had a perfect measure of job performance, would these problems be alleviated? We think not. So while performance management in organizations may be a messy, poorly managed, and poorly implemented process, we should be cautious not to lay the blame on the quality of performance ratings. Let's not throw the performance rating baby out with the performance management bathwater.

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.

Arthur, W., Day, E. A., McNelly, T. L., & Edens, P. S. (2003). A meta-analysis of the criterion-related validity of assessment center dimensions. *Personnel Psychology, 56*, 125–154.

Arthur, W., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology, 93*, 435–442.

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

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387–409.

Blum, M., & Naylor, J. (1968). *Industrial psychology*. New York, NY: Harper & Row.

Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Goldstein, I. L., & Ford, K. J. (2002). *Training in organizations: Needs assessment, development, and evaluation* (4th ed.). Belmont, CA: Wadsworth.

Huffcutt, A. I., & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184–190.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization analysis to the general aptitude test battery*. Washington, DC: Department of Labor, Employment Services.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–95.

Jawahar, I. M., & Williams, C. R. (1997). Where all the children are above average: The performance appraisal purpose effect. *Personnel Psychology, 50*, 905–925.

Kraiger, K. (2002). Decision-based evaluation. In K. Kraiger (Ed.), *Creating, implementing, and managing effective training and development: State of the art lessons for practice* (pp. 331–375). San Francisco, CA: Jossey-Bass.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730–740.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews—A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616.

https://doi.org/10.1017/iop.2016.25 Published online by Cambridge University Press

McKay, P. F., & McDaniel, M. A. (2006). A reexamination of Black–White mean differences in work performance: More data, more moderators. *Journal of Applied Psychology, 91*, 538–554.

Rothstein, H. R., Schmidt, F. L., Erwin, F. W., Owens, W. A., & Sparks, C. P. (1990). Biographical data in employment selection: Can validities be made generalizable. *Journal of Applied Psychology, 75*, 175–184.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & De Fruyt, F. (2003). International validity generalization of GMA and cognitive ability: A European community meta-analysis. *Personnel Psychology, 56*, 537–605.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L., & Rader, M. (1999). Exploring the boundary conditions for interview validity: Meta-analytic validity findings for a new interview type. *Personnel Psychology, 52*, 445–464.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Taylor, P. J., Russ-Eft, D., & Taylor, H. (2009). Transfer of management training from alternative perspectives. *Journal of Applied Psychology, 94*, 104–112.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: Wiley.

The Relationship Between the Number of Raters and the Validity of Performance Ratings

Matt C. Howard
University of South Alabama and Pennsylvania State University

In the focal article "Getting Rid of Performance Ratings: Genius or Folly? A Debate," two groups of authors argued the merits of performance ratings (Adler et al., 2016). Despite varied views, both sides noted the importance of including multiple raters to obtain more accurate performance ratings. As the pro side noted, "if ratings can be pooled across many similarly situated raters, it should be possible to obtain quite reliable assessments" (Adler et al., p. 236). Even the con side noted, "In theory, it is possible to obtain ratings from multiple raters and pool them to eliminate some types of

Matt C. Howard, Mitchell College of Business, Department of Management, University of South Alabama, and Department of Psychology, Pennsylvania State University.

Thanks to Rick Jacobs and Alex McKay for their comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Matt C. Howard, Mitchell College of Business, Department of Management, University of South Alabama, 5811 USA Drive South, Room 346, Mobile, AL 36688-0002. E-mail: mhoward@southalabama.edu

https://doi.org/10.1017/iop.2016.25 Published online by Cambridge University Press