Industrial and Organizational Psychology
PERSPECTIVES ON SCIENCE AND PRACTICE

.80 and Beyond: Recommendations for Disattenuating Correlations

MICHAEL J. BURKE
Tulane University

RONALD S. LANDIS AND MAURA I. BURKE
Illinois Institute of Technology

Correspondence concerning this article should be addressed to Michael J. Burke.
E-mail: mburke1@tulane.edu
Address: Freeman School of Business, 7 McAlister Place, Tulane University, New Orleans, LA 70118

LeBreton, Scherer, and James (2014) concluded a commendable focal article on correcting correlations for unreliability in supervisory ratings with three possible outcomes for our field: (a) the disappearance of supervisory ratings, (b) a renewed emphasis

https://doi.org/10.1017/S1754942600006830 Published online by Cambridge University Press

on improving the psychometric properties of supervisory ratings, or (c) the use of more accurate and reasonable estimates of the reliability of supervisory ratings when correcting correlations. In this commentary, we address the logic, necessity, and value of each option to stakeholders of industrial–organizational (I–O) research and practice as well as offer suggestions for moving forward with the third option.

First, the disappearance of supervisory ratings is not necessarily consistent with the logic of LeBreton et al.'s work, which is to not denigrate a method due to researchers' application of a statistical value (i.e., an estimate of interrater reliability) associated with the method. It would seem that their logic and argument applied to meta-analysis and structural equation modeling as statistical methods should similarly apply to supervisory ratings as a measurement method. Concerning this point, we emphasize that the criterion reliability estimate used to correct correlations in many validity generalization/meta-analytic studies is not directly tied to the observed correlations in the meta-analysis. That is, the statistical value referred to in many of LeBreton et al.'s arguments, the criterion reliability estimate of .52, is an *assumed population estimate* of criterion reliability used for correcting the average observed correlation in an artifact distribution meta-analysis (i.e., a meta-analysis that uses hypothetical reliability distributions). As such, we do not believe a concern with a *hypothetical statistical value* should necessarily lead to the demise of a measurement procedure, in this case supervisory ratings, within test validation studies or otherwise. One need only examine any basic reference in our field to appreciate the necessity and value of the use of supervisory ratings in research and practice (e.g., see Wildman, Bedwell, Salas, & Smith-Jentsch, 2011). Therefore, we do not believe Option 1 is a viable outcome in maintaining/advancing the integrity of our field.

The second possible option concerned a renewed emphasis on improving the psychometric properties of supervisory ratings. The research on supervisory ratings has certainly led to some important advances (Landy & Farr, 1980). These advances, however, have largely stalled for the last several decades. Although we are not opposed to future research aimed at improving the psychometric properties of supervisory ratings, the utility of such an effort is not clear to us. So, without further clarification, we do not believe Option 2 is tenable.

The third possible option focused on the use of more accurate and reasonable estimates of the reliability of supervisory ratings. From a scientific perspective, this option is potentially valuable in enhancing our understanding of relationships between true scores on measures, in reference to the classical test theory model (Lord & Novick, 1968), and to construct relationships more generally. Following, we suggest three fruitful avenues for researchers and practitioners to pursue.

First, for those who need to use an *assumed* estimate of the reliability of supervisory ratings when making corrections to correlations, we encourage a wider view of the literature when estimating such a *hypothetical value or values*. In this regard, we emphasize that possible *assumed population estimates of interrater reliability* that approach a desired reliability of .90 for interrater reliability, noted by LeBreton et al. and others, can be made from performance information reported in the broader science literature. For decades, researchers have been reporting estimates of interrater reliabilities, often based on assessments of interrater agreement between paired evaluators, of approximately .80 to .90 for the criterion of work performance. This reporting of reliability spans studies with adult workers in a broad cross-section of fields of inquiry including applied psychology, ergonomics, industrial hygiene, industrial relations, management, nursing, occupational medicine, occupational therapy, and occupational safety (see Burke et al., 2006, 2011).

In summarizing a body of literature on interrater reliability estimates, Burke et al. (2011) reported that the average interrater reliability for assessments of performance

https://doi.org/10.1017/S1754942600006830 Published online by Cambridge University Press

was approximately .88 with the majority of the 38 estimates exceeding a value of .80. These interrater reliability estimates were gathered from a review of the literature evaluating the efficacy of worker safety and health training with respect to work performance as the criterion. We note that the consistency of performance ratings in this literature is often expressed as percent agreement between raters. As shown by Burke and Dunlap (2002), percent agreement between raters can be equated with the average correlation between raters (e.g., a percent agreement of 83% is equal to an average correlation of .80 between raters). What is different about these criterion reliability estimates in comparison to those reported within the test validation literature is that they are typically based on two observers whose sole focus, *at the same point in time*, is the evaluation of worker performance. As a consequence, these studies, in many ways, control for extraneous factors (such as opportunities to observe) that plague efforts to obtain accurate sample-based or assumed population estimates of the reliability of supervisory ratings in primary and meta-analytic test validation studies, respectively.

The second avenue that we encourage researchers and practitioners to pursue when correcting correlations for unreliability in supervisory ratings, with either a sample-based or an assumed population criterion reliability estimate, is the use of appropriate correction formulae. This point applies to computing not only the corrected correlation but also for estimating the standard error (the square root of the sampling variance) of the corrected correlation for use in establishing confidence intervals. In the case of correcting individual correlations for unreliability in supervisory ratings, we refer the reader to equations presented in Raju and Brand (2003) and Raju, Burke, Normand, and Langlois (1991) for obtaining appropriate estimates of corrected correlations and their standard errors. Importantly, the equations of Raju and colleagues not only take into account sampling error due to sample size (*N*), but they also account for sampling error in sample-based estimates of predictor and criterion reliability. Following these procedures will, in large part, address the critical issue discussed by Lord and Novick (1968) almost 50 years ago—to assure that essentially no more or less error variation is introduced into the estimate of the population predictor reliability, the population criterion reliability, and the population correlation (see p. 138).

In the case of correcting a mean correlation for unreliability in supervisory ratings, we also encourage the use of a formula for estimating the standard error of the mean corrected correlation that accounts for sampling error in the sample-based reliabilities (see Burke & Landis, 2003). To date, only a few meta-analyses compute the standard error for the mean corrected correlation with the formula presented in Burke and Landis (2003) (e.g., Bing, Davison, Minor, Novicevic, & Frink, 2011; Christian, Garza, & Slaughter, 2011; Lang, Kersting, Hulsheger, & Lang, 2010). As a result, the majority of confidence intervals for mean corrected correlations in the literature do not account for sampling error in the sample-based reliabilities. Accounting for sampling error in the sample-based reliabilities is critical for establishing confidence intervals that capture the mean population correlation with a known probability.

Third, we advocate for greater attention to how corrections for supervisory ratings of performance are being made when estimating relationships between constructs at the group or business unit level of analysis. A recent trend in meta-analyses is the use of aggregated data to estimate relationships between constructs at the group or business unit level of analysis (e.g., Harter, Schmidt, Asplund, Killham, & Agrawal, 2010; Hong, Liao, Hu, & Jiang, 2013; Wallace et al., 2013; Whitman, Caleo, Carpenter, Horner, & Bernerth, 2012; Whitman, Van Rooy, & Viswesvaran, 2010). In many cases, aggregated ratings of performance are employed as criteria, and correlations involving these data are corrected for criterion unreliability. In

https://doi.org/10.1017/S1754942600006830 Published online by Cambridge University Press

all studies where corrections have been made for criterion unreliability, either assumed criterion reliabilities were used or sample-based criterion reliabilities were employed without accounting for sampling error in the sample-based reliability estimates. As such, this nascent literature is beginning to mirror the problems noted by LeBreton et al. when it comes to making corrections for criterion unreliability. More recently, Burke, Landis, and Burke (2014) provided guidance for how to appropriately correct individual and meta-analytic correlations based on aggregated data along with how to establish confidence intervals for such corrected correlations. Importantly, in relation to LeBreton et al.'s call for using more accurate and reasonable estimates of criterion reliability for correcting correlations, the average criterion reliability for aggregated performance data reported by Burke et al. (2014) was .76 for ratings of performance and .83 for performance outcomes (e.g., sales, accidents, injuries).

Finally, LeBreton et al. reiterated a long-standing belief in our field that a corrected correlation reflects the relationship between measures that are lengthened indefinitely or that are of indefinite length. This belief has persisted despite efforts to appropriately describe the meaning of different types of corrected correlations. In the case of correcting a correlation for assumed population predictor and criterion reliabilities, the corrected correlation reflects the relationship between true scores on measures of particular lengths in the unrestricted population. This type of corrected correlation follows from a specific correlation model (see equation 1 of Raju et al., 1991 and discussions within Lord & Novick, 1968) and only has basic sampling error due to sample size associated with it. In the case of correcting an observed correlation for sample-based predictor and criterion reliabilities, the corrected correlation also reflects the relationship between true scores on measures of particular lengths. However, this type of corrected correlation has basic sampling error and sampling error in the sample-based reliability estimates associated with it. As such, this type of correlation follows a different correlation model (see equation 2 of Raju et al., 1991) and equation for estimating its sampling variance and the square root of this quantity, the standard error of the corrected correlation.

In conclusion, we strongly agree with LeBreton et al.'s third option for addressing issues concerning correcting correlations for unreliability in supervisory ratings and present suggestions for how this option can be extended in several ways. First, we suggest that researchers and practitioners use information available in the broader scientific literature for obtaining more accurate and reasonable estimates of *assumed* interrater reliability for such corrections. A provisional estimate of **.80** is a reasonable literature-based estimate of the reliability of supervisory ratings based on individual-level data. Second, we encourage researchers and practitioners to use appropriate formulae for correcting individual and meta-analytic correlations, and for computing sampling variances (and their square roots, the standard errors) associated with these corrected correlations when establishing confidence intervals. Finally, we advocate for extending best practices concerning correcting correlations based on individual-level data to corrections for measurement unreliability when estimating relationships between constructs at the group or business unit level of analysis with aggregated data. In the latter case, a provisional estimate of **.80** is a reasonable literature-based estimate of the reliability of performance assessments based on aggregated data. Together, these recommendations are offered for improving estimates of relationships between true scores on measures and relationships between constructs more generally.

References

Bing, M. N., Davison, H. K., Minor, I., Novicevic, M. M., & Frink, D. D. (2011). The prediction of task and contextual performance by political skill: A meta-analysis and moderator test. *Journal of Vocational Behavior*, *79*, 563–577.

https://doi.org/10.1017/S1754942600006830 Published online by Cambridge University Press

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, *5*, 159–172.

Burke, M. J., & Landis, R. (2003). Methodological and conceptual issues in applications of meta-analysis. In K. Murphy (Ed.), *Validity generalization: A critical review* (pp. 287–310). Hillsdale, NJ: Lawrence Erlbaum.

Burke, M. I., Landis, R., & Burke, M. J. (2014, May). Conducting meta-analysis at the group-level of analysis. In S. B. Morris (Chair), *Meta-analysis methods for messy, incomplete, and complex data*. Symposium at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Burke, M. J., Salvador, R., Smith-Crowe, K., Chan-Serafin, S., Smith, A., & Sonesh, S. (2011). How workplace hazards and training influence learning and performance. *Journal of Applied Psychology*, *96*, 46–70.

Burke, M. J., Sarpy, S. A., Smith-Crowe, K., Chan-Serafin, S., Islam, G., & Salvador, R. (2006). Relative effectiveness of safety and health training methods. *American Journal of Public Health*, *96*, 315–324.

Christian, M. S., Garza, A. S., & Slaughter, J. E. (2011). Work engagement: A quantitative review and test of its relations with task and contextual performance. *Personnel Psychology*, *64*, 89–136.

Harter, J. K., Schmidt, F. L., Asplund, J. W., Killham, E. A., & Agrawal, S. (2010). Causal impact of employee work perceptions on the bottom line of organizations. *Perspectives on Psychological Science*, *5*(4), 378–389.

Hong, Y., Liao, H., Hu, J., & Jiang, K. (2013). Missing link in the service profit chain: A meta-analytic review of the antecedents, consequences, and moderators of service climate. *Journal of Applied Psychology*, *98*(2), 237–267.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*, 72–107.

Lang, J. W. B., Kersting, M., Hulsheger, U. R., & Lang, J. (2010). General mental ability, narrower cognitive abilities, and job performance: The perspective of the nested factors model of cognitive abilities. *Personnel Psychology*, *63*, 595–640.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(4), 478–500.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Raju, N. S., & Brand, P. A. (2003). Determining the significance of correlations corrected for unreliability and range restriction. *Applied Psychological Measurement*, *27*, 52–71.

Raju, N. S., Burke, M. J., Normand, J., & Langlois, G. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, *76*, 432–446.

Wallace, J. C., Edwards, B. D., Paul, J., Burke, M. J., Christian, M., & Eissa, G. (2013). Change the referent? A meta-analytic investigation of direct and referent-shift consensus models for organizational climate. *Journal of Management*, *39*, 1156–1183.

Whitman, D. S., Caleo, S., Carpenter, N. C., Horner, M. T., & Bernerth, J. B. (2012). Fairness at the collective level: A meta-analytic examination of the consequences and boundary conditions of organizational justice climate. *Journal of Applied Psychology*, *97*(4), 776–791.

Whitman, D. S., Van Rooy, D. L., & Viswesvaran, C. (2010). Satisfaction, citizenship behaviors, and performance in work units: A meta-analysis of collective construct relations. *Personnel Psychology*, *63*(1), 41–81.

Wildman, J. L., Bedwell, W. L., Salas, E., & Smith-Jentsch, K. A. (2011). Performance measurement at work: A multilevel perspective. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology: Volume 1 building and developing the organization* (pp. 303–341). Washington, DC: American Psychological Association.

https://doi.org/10.1017/S1754942600006830 Published online by Cambridge University Press