Industrial and Organizational Psychology
PERSPECTIVES ON SCIENCE AND PRACTICE

Increasing Interrater Reliability Using Composite Performance Measures

MICHAEL P. WILMOT, BRENTON M. WIERNIK, AND JACK W. KOSTAL
University of Minnesota

LeBreton, Scherer, and James (2014) raise important questions about the implications of unreliable criterion measurement in industrial–organizational (I–O) psychology. Although we do not share the authors' bleak outlook concerning the state of cumulative knowledge in our field, we do agree that continuing to rely on a criterion with an average reliability of .52 (i.e., single-supervisor rating of overall job performance) is undesirable for the continued advancement of the science and practice of I–O psychology. However, contrary to the conclusions of LeBreton and colleagues, improving criterion reliability need not require abandoning supervisory ratings. Rather, by applying a contemporary understanding of the structure of job performance (Campbell, 2012), reliable and construct-valid supervisory ratings can indeed be obtained. In this commentary, we estimate the interrater reliability of a multidimensional job performance measure and demonstrate that the composite ratings derived from such a measure may ameliorate many of the concerns that LeBreton et al. raise.

Correspondence concerning this article should be addressed to Brenton Wiernik.
E-mail: wiernik@workpsy.ch
Address: Department of Psychology, University of Minnesota, Minneapolis, MN 55455

Order of authorship is arbitrary. All authors contributed equally.

Single Versus Multiple Measures of Overall Job Performance

Assessments of the construct of overall job performance can be obtained either using a *single scale* or by compositing *several distinct scales* measuring more specific dimensions. Although both methods tap the same construct, the latter is likely to result in more reliable estimates. The .52 reliability estimate at the center of debate refers to overall job performance ratings of the single-scale variety. Specifically, it is for "ratings on statements (or ranking of individuals on statements) referring to overall performance, overall effectiveness, overall job performance, overall work reputation, or the sum of all individual dimensions rated" (Viswesvaran, Ones, & Schmidt, 1996, p. 561). Such single-scale measures ask raters to make overall summary judgments of employees' performance or combine a limited number of items assessing specific dimensions in an unstructured manner. The vast majority of validation research has used criterion measures of this type, and their associated interrater reliability estimates have left something to be desired.

As a remedy to the weak interrater reliability of overall performance measures, we recommend that I–O psychologists develop and use construct-valid measures of specific job performance dimensions, each of which are assessed by distinct scales. Job performance is multidimensional in nature,

https://doi.org/10.1017/S1754942600006854 Published online by Cambridge University Press

Table 1. *Reliability and Intercorrelations Among Ratings of Job Performance Dimensions*

	Performance dimension	1	2	3	4	5	6	7	8
1	Quality	*.65*							
2	Quantity/productivity	.27	*.57*						
3	Job knowledge	.38	.46	*.56*					
4	Interpersonal competence	.10	.29	.17	*.53*				
5	Compliance/acceptance of authority	.18	.31	.29	.33	*.60*			
6	Effort	.30	.28	.22	.24	.34	*.56*		
7	Leadership	.29	.26	.34	.21	.22	.29	*.55*	
8	Administrative competence	.22	.34	.23	.22	.29	.31	.20	*.59*

Note. Values from Viswesvaran et al. (2005) tables 3 and 7. Values on diagonal are interrater reliabilities for supervisory ratings. Values below the diagonal are observed supervisor–peer correlations.

and it should be modeled and measured accordingly (Campbell, 2012). To illustrate the benefits of using a multidimensional performance assessment approach, we used meta-analytic estimates of interrater reliability and intercorrelations of job performance dimensions, along with familiar psychometric formulae, to estimate the reliability of several hypothetical performance measures.

Composite Performance Measures: A Worked Example

The following analyses are based on meta-analytic data reported by Viswesvaran, Schmidt, and Ones (2005). Authors reported interrater reliabilities of supervisory ratings for each of eight dimensions of job performance (their table 3). Due to a lack of available estimates, the authors did not report the intercorrelations among performance dimensions as rated by different supervisors. As a result, we used the observed correlations between supervisor and peer ratings to approximate these intercorrelations (their table 7). All of these values are shown in Table 1.

We used these values to estimate the reliability of composite measures of several different performance composites. First, we composited ratings of "quality," "quantity/productivity," and "job knowledge" to create an index of *technical performance*. Similarly, we combined ratings of "interpersonal competence" and "compliance/acceptance of authority" to create an index of *contextual performance*. Finally, we created an index of *supervisory performance* by compositing ratings of "administrative competence" and "leadership." In addition, we also computed reliability estimates for indices of overall performance for a hypothetical *nonmanagerial* position (composite of technical performance, contextual performance, and effort) and a hypothetical *managerial* position (composite of technical performance, contextual performance, supervisory performance, and effort). Because the mean reliabilities for dimensions reported by Viswesvaran et al. (2005) did not differentiate by job type of ratee, these overall performance composites should be viewed as illustrative; to the degree that interrater reliabilities for performance dimensions differ across job types, composite reliabilities may also differ. Estimates of interrater reliability for the five composite measures of performance are shown in Table 2.

We are persuaded by Campbell (2012) that job performance is a multidimensional construct and that composites constructed from specific job performance dimensions should be conceptualized as *aggregate* in nature. That is, they are a sum of their component parts and are intended to be an index of an employee's overall positive contributions to the organization. This overall contribution (not a latent variable consisting of the shared variance across dimensions) is the real construct of interest in performance measurement. Therefore,

https://doi.org/10.1017/S1754942600006854 Published online by Cambridge University Press

Table 2. *Reliability Estimates for Single-Rater Composite Job Performance Measures*

	Performance composite	Mean r_{yy}	Mosier r_{YY}
1	Technical performance	.59	.77
2	Contextual performance	.60	.67
3	Supervisory performance	.57	.64
4	Overall performance—nonmanagerial	.58	.82
5	Overall performance—managerial	.58	.85

the unique information provided by each performance dimension (e.g., *effort*, *leadership*, and *quality*) is important, even though it is not shared with other dimensions. With this in mind, we contend that Mosier's (1943) formula is the appropriate procedure to estimate the reliability of an overall job performance index based on a composite of dimensional ratings. Mosier's (1943) formula assumes that the specific factor variance of each component in a composite contains *relevant and reliable* variance that should also be included in the reliability estimate alongside the shared factor variance. The formula is provided below:

$$\text{Mosier } r_{YY} = \frac{\Sigma rel_y + \left(2 * \Sigma r_{yy'}\right)}{n + \left(2 * \Sigma r_{yy'}\right)}$$

where r_{YY} is the reliability of the composite, n is the number of components in the composite, $r_{yy'}$ is the correlation between two components, and rel_y is the reliability of one component.

As shown in Table 2, the average interrater reliability of the components contributing to overall performance estimates is .58 for both the hypothetical managerial and nonmanagerial positions. However, when the reliability of these overall job performance composites are computed using the formula above, the estimated interrater reliabilities for the composites are $r_{YY} = .85$ for the managerial position and $r_{YY} = .82$ for the nonmanagerial position. (Note that these values may be underestimates if the supervisor–peer correlations used to estimate dimension intercorrelations are smaller than the correlations between dimensions assessed by two supervisors). These reliability estimates are not only higher than the average interrater reliability of $r_{yy} = .52$ for single-scale ratings of overall performance, but they are also based on more theoretically defensible grounds, namely, the multidimensionality of job performance. Such composite measures of performance may also serve additional practical purposes, such as facilitating feedback based on evaluations of individual performance dimensions.

Although a reliability of .82 is much higher than .52, single-rater ratings of performance may still contain more error than desirable, especially if performance ratings are intended to make key decisions about individual employees, such as promotions, rather than simply to estimate the relationship between a predictor construct and performance. As with single-rater, single-scale measures of overall performance, reliability can be further enhanced by using multiple raters. When composite performance is computed as the sum of average dimension ratings for two raters, reliability rises to .90. This value rises to .95 with only four raters. Such levels of reliability would take 8 and 16 raters, respectively, if single-scale measures of overall job performance were used.

Conclusions

Historically, I–O psychologists have not properly modeled and measured our criterion space and instead have lamented the "criterion problem" without truly seeking to address it. Fortunately, the criterion problem is not really a problem anymore. We understand how to model and measure the multidimensionality of job performance (Campbell, 2012; Campbell, McCloy, Oppler, & Sager, 1993; Viswesvaran & Ones, 2000). Therefore, the "criterion problem" is no longer *the* problem; *the* problem now is the more banal one of actually

https://doi.org/10.1017/S1754942600006854 Published online by Cambridge University Press

developing and applying reliable and valid measures of the dimensions of performance. Single-supervisor ratings of brief overall assessments of job performance have dominated our field because they are inexpensive, fast, easy to collect, and frequently used in organizations. Developing performance measures assessing several specific performance dimensions is not substantially more difficult. What is more, doing so can dramatically increase criterion reliability, despite still being based on supervisor ratings.

In future meta-analytic efforts, researchers should consider computing estimates of validity for overall job performance using composites of validity estimates for specific performance dimensions; the formula used in this article can be easily applied to estimate composite performance reliabilities in most meta-analytic situations. When only overall validities are reported, the structure and content of the criterion instrument should be carefully evaluated to determine the appropriate meta-analytic estimate to use to approximate reliability. Such practices will result in more conservative corrections that, nevertheless, accurately represent the true operational validity of predictors.

In summary, we should not be content with the low reliability of single ratings of undifferentiated overall job performance. However, improving reliability does not necessitate abandoning more than a century of research and practice on job performance ratings. Application of our contemporary understanding of the multidimensional nature of job performance and associated psychometric procedures can accomplish this goal. We urge I–O psychologists to develop and use construct-valid measures of specific job performance dimensions, each of which is assessed by a distinct scale. Only then should these scales score be combined into one overall job performance score. Rating a handful of single-item scales and then summing them to form an overall metric does *not* constitute the kind of composite measure we advocate in this paper. Composite performance scales will allow for increased reliability of overall performance ratings and will more faithfully represent the job performance construct, thereby contributing to the continued advancement of the science and practice of I–O psychology.

References

Campbell, J. P. (2012). Behavior, performance, and effectiveness in the twenty-first century. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (Vol. 1, pp. 159–196). New York, NY: Oxford University Press.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmitt, & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 35–70). San Francisco, CA: Jossey-Bass.

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.

Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, *8*, 161–168.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment*, *8*, 216–226.

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.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in job performance ratings? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*, 108–131.

https://doi.org/10.1017/S1754942600006854 Published online by Cambridge University Press