Industrial and Organizational Psychology
PERSPECTIVES ON SCIENCE AND PRACTICE

When Correcting for Unreliability of Job Performance Ratings, the Best Estimate Is Still .52

WINNY SHEN
University of South Florida

JEFFREY M. CUCINA
U.S. Customs and Border Protection

PHILIP T. WALMSLEY
U.S. Office of Personnel Management

BENJAMIN K. SELTZER
Washington & Jefferson College

In this commentary we answer three questions that are often posed when debating the usefulness and accuracy of correcting criterion-related validity coefficients for unreliability: (a) Is .52 an inaccurate estimate? (b) Do corrections for criterion unreliability lead us to choose different selection tools? (c) Is too much variance explained?

Is .52 an Inaccurate Estimate?

LeBreton, Scherer, and James (2014) argue that the statistical value our field typically uses for corrections for attenuation of performance ratings appears to be "too low" (p. 497). However, as with any empirical finding, Viswesvaran, Ones, and Schmidt's (1996) meta-analytic estimate of .52 ($k = 40$, $N = 14{,}650$) to represent the interrater reliability for a single supervisor's rating of job performance is both testable and falsifiable. Accordingly, we are unaware of any evidence that directly contradicts this estimate. In fact, .52 is consistent with previous estimates, including Conway and Huffcutt's (1997) separate meta-analytic estimate of .50 ($k = 69$, $N = 10{,}369$); Rothstein's (1990) asymptotic estimate of .55 (for duty ratings); Hunter's (1983) estimate of .60; King, Hunter, and Schmidt's (1980) estimate of .60; and Scullen, Mount, and Sytsma's (1996) estimate of .45. Given that LeBreton et al. do not appear to take issue with the use of interrater reliability (rather than intrarater reliability) as the conceptually appropriate estimate for unreliability corrections for performance ratings (we refer interested readers to the exchange between Murphy & DeShon, 2000 and Schmidt, Viswesvaran, & Ones, 2000 for more information on this issue), the current evidence is robust regarding the accuracy of the .52 estimate.

Some critics of the .52 estimate state that even if this value is a correct estimate of the relationship between two individual raters' ratings, two raters may reliably assess different aspects of an individual's performance. Regardless, the .52 estimate is the most relevant one for an organization; it focuses

Correspondence concerning this article should be addressed to Winny Shen.
E-mail: winny.shen@uwaterloo.ca
Address: University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1
The views expressed in this paper are those of the authors and do not necessarily reflect the views of U.S. Customs and Border Protection, the U.S. Office of Personnel Management, or the U.S. Federal Government.

https://doi.org/10.1017/S1754942600006805

on the shared variance—or what two different supervisors have in common—when they think of an employee's performance.[1] Because selection systems are put into place to benefit an organization (rather than a given supervisor), it makes conceptual sense to focus on predicting the variance in performance ratings shared by multiple "possible" supervisors from the population of potential supervisors.

Do Corrections for Criterion Unreliability Lead Us to Choose Different Selection Tools?

LeBreton et al. seem to suggest that the use of corrections for criterion unreliability can alter the choice of selection instruments that an organization chooses to implement when they speak of the "practical policy decisions affecting industrial–organizational (I–O) psychologists working in organizations" (p. 492). This statement warrants inspection. The corrected operational validity coefficient can be expressed by dividing the correlation coefficient by the square root of the reliability, as follows:

$$\hat{r}_{12} = \frac{r_{12}}{\sqrt{r_{11}}\sqrt{r_{22}}}$$

where $\hat{r}_{12}$ = the reliability corrected validity coefficient, r_{12} = the observed validity coefficient, r_{11} = the reliability of the test, and r_{22} = the reliability of the criterion.

When computing operational validity coefficients, the value for the reliability of the test (r_{11}) is removed from the equation, and the formula can be rearranged as:

$$\hat{r}_{12} = \frac{1}{\sqrt{r_{22}}}\left(r_{12}\right)$$

When this formula is applied to different tests that all use the same criterion (e.g., supervisory ratings with a reliability of .52), the validities are all multiplied by a constant (i.e., $1/\sqrt{.52} = 1.39$) and the correction for criterion unreliability is a linear transformation, as follows:

$$\hat{r}_{12} = \frac{1}{\sqrt{.52}}\left(r_{12}\right) = 1.39\left(r_{12}\right)$$

As a result, the rank ordering of the operational validities of different predictors remains unchanged by the correction for criterion unreliability. In Table 1, we present Schmidt and Hunter's (1998) job performance predictor validity table alongside the uncorrected validities and validities corrected using a higher criterion reliability (i.e., .80). In all cases, general mental ability (GMA) and work sample tests have the highest validity, graphology and age the lowest. Thus, an organization choosing between predictors to use (based on Schmidt & Hunter's meta-analytic data) will make the same decisions regardless of the value used to correct for criterion unreliability. The same finding occurs in local validation studies when the organization compares the validity of several different predictors to a single criterion.[2]

[1] Even if multiple supervisors rate each employee, the unique variance measured by each supervisor will cancel out in favor of the shared variance when computing a composite criterion variable. To demonstrate, Jensen (1998, pp. 103–104) and Spearman (1927, Appendix, pp. xix–xxi) provide a formula for computing the *g*-loading of a composite score, which we adapt for job performance:

$$r_{tp} = \left(1 + \left\{\sum \frac{r_{sp}^2}{\left(1 - r_{sp}^2\right)}\right\}^{-1}\right)^{-1/2}$$

where r_{tp} = correlation between total composite rating score and latent performance variable, r_{sp} = performance rating's correlation with the latent performance variable (i.e., performance loading of the ratings; analogous to the *g*-loading of the subtest).

The percentage of variance in a unit-weighted composite of supervisors accounted for by the shared variance increases from .52 (or a loading of .72 on performance) with one supervisor to .68 (or a loading of .83) of with two supervisors to .76 (or a loading of .87) with three supervisors.

[2] In LeBreton et al.'s (2014) table 2, the rank ordering of the validity of different assessments changes

Table 1. *Schmidt and Hunter's (1998) Criterion-Related Validities Using Different Types of Corrections for Unreliability*

Predictor	1. Corrected validity (using .52 reliability)	2. Corrected validity (using .80 reliability)	3. Validity uncorrected for unreliability	4. Percent of total possible validity ($\sqrt{.52} = .72$)
General mental ability	.51	.41	.37	51%
Work sample	.54	.44	.39	54%
Integrity	.41	.33	.30	41%
Conscientiousness	.31	.25	.22	31%
Structured interviews	.51	.41	.37	51%
Unstructured interviews	.38	.31	.27	38%
Job knowledge tests	.48	.39	.35	48%
Job tryout procedure	.44	.35	.32	44%
Peer ratings	.49	.40	.35	49%
T&E[a] behavioral consistency	.45	.36	.32	45%
Reference checks	.26	.21	.19	26%
Job experience (years)	.18	.15	.13	18%
Biodata	.35	.28	.25	35%
Assessment centers	.37	.30	.27	37%
T&E[a] point method	.11	.09	.08	11%
Years of education	.10	.08	.07	10%
Interests	.10	.08	.07	10%
Graphology	.02	.02	.01	2%
Age	−.01	−.01	−.01	1%

Note. Both Spearman's rho and the Pearson r correlation between each combination of values in Columns 1, 2, 3, and 4 are identical ($r = 1.00$ and $\rho = 1.00$), meaning that the rank order of predictors is preserved. Adapted with permission from the American Psychological Association from table 1 of Schmidt and Hunter (1998).
[a] T&E: Training and Experience.

If we do not correct for unreliability in performance ratings, the maximum correlation between a predictor and performance is then no longer 1.0. To explain, we can decompose the variance of supervisory ratings using classical test theory. Nunnally and Bernstein (1994, p. 237) note that the reliability coefficient is the proportion of variance in an observed score that is due to the true score. Thus, based on an interrater reliability of .52, 52% of the variance in supervisory ratings is due to true score and 48% is due to measurement error; in other words:

$$\text{reliability} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{observed}}} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{error}}}$$

$$\text{reliability} = \frac{.52}{.52 + .48} = \frac{.52}{1.00}$$

Taking the square root of the reliability (i.e., the square root of the variance accounted for by true score) gives the correlation between the observed score and the true score (.72). This value is also known as the reliability index, which Nunnally and Bernstein define as "the correlation between a set of scores on a given test (x_1) and corresponding true scores" (p. 222). The key point to note here is that if we were to locate a variable that perfectly predicts

when comparing the uncorrected and the operational validity matrices. Work samples demonstrated the highest validity in the uncorrected matrix and structured interviews demonstrated the highest validity in the operational validity matrix. However, this change in rank ordering is due to corrections for range restriction, which were estimated to be different across the assessments (see p. 484) rather than corrections for criterion unreliability.

one's true score for performance, it would have a correlation of .72 with the observed supervisory rating (recall that measurement error is random error and thus cannot correlate with the predictor). Using a peer's rating of performance, which has a lower interrater reliability (Viswesvaran et al., 1996), as the criterion, the highest possible observed correlation would be .65. Given that both peer and supervisor ratings assess the construct of job performance, we think that it causes unnecessary confusion to rely on observed correlations where the maximum validity coefficients differ as a function of the reliability of the criterion. This also makes it difficult for psychologists to compare the magnitude of raw validity coefficients, particularly across different criteria.

One solution might be to express uncorrected validity coefficients as the percentage of the maximum possible validity coefficient. For example, an observed validity of .25 for a predictor in predicting supervisory ratings is 35% of .72. However, this is exactly equivalent to the correction for criterion unreliability (see Table 1, Column 4). In other words, the corrected validity coefficient can also be interpreted as the ratio of the observed validity coefficient to the maximum possible validity coefficient.

Is Too Much Variance Explained?

It is interesting that historically the I–O literature has bemoaned the presence of a "validity ceiling," and the field seemed to be unable to make large gains in the prediction of job performance (Highhouse, 2008). In contrast, LeBreton et al. appear to have the opposite concern—that we may be able to predict too much, perhaps even all, of the variance in job performance once accounting for statistical artifacts. In addition to their four focal predictors (i.e., GMA, integrity, structured interview, work sample), LeBreton et al. list an additional 24 variables that have been shown to be related to job performance meta-analytically. However, we believe that many of the variables LeBreton et al. included in their list are variables that Sackett, Borneman, and Connelly (2009) would argue are likely unknowable at time of hire. For example, Sackett et al. specifically argue that support from one's supervisor and colleagues may be determinants of performance unknowable at time of hire, and coworker support was, in fact, included in LeBreton et al.'s list ($\rho = .24$; Chiaburu & Harrison, 2008).

Furthermore, in contrast to LeBreton et al.'s assertion that organizational variables, such as procedural justice, are likely unrelated to their focal predictors, our belief is that many of these variables are likely to be at least moderately correlated–limiting the incremental validity we could expect with the inclusion of these additional variables. For example, research has shown that integrity tests mostly tap into Conscientiousness, Agreeableness, and Emotional Stability (Ones & Viswesvaran, 2001), and a recent meta-analysis of organizational justice shows that all three personality traits are moderately related to one's experience of procedural justice ($\rho = .19–.23$; Hutchinson et al., 2014), suggesting that even apparently unrelated variables can share a surprising amount of construct-level variance. In support of this perspective, Paterson, Harms, and Crede (2012) conducted a meta-analysis of over 200 meta-analyses and found an average correlation of .27, suggesting that most variables we study are at least somewhat correlated and validating the first author's long-held personal assumption that the world is correlated .30 (on average; see also Meehl's, 1990, crud factor)!

Another difficulty in accurately estimating intercorrelations among LeBreton et al.'s four focal selection tools of interest is that they included both constructs (e.g., GMA, integrity) and methods (e.g., structured interview, work sample). Although this practice is common in the literature, Arthur and Villado (2008) provide a strong rationale regarding how confounding constructs and methods make these comparisons largely uninterpretable. Conceptually and practically, it is difficult to estimate the relationship between structured interviews and integrity and intelligence because in practice interviews could be used to assess both constructs, just one, or neither (Huffcutt, Conway, Roth, & Stone, 2001).

One final point to remember is that each meta-analytic estimate is an *estimate* of the average effect or relationship between two variables in the population of interest, and there is often meaningful variation (reflected in *SD*ρ and the credibility interval) around these estimates after statistical artifacts are accounted for. This variation could be meaningfully predicted by organizational variables (e.g., Shen et al., 2012), including those listed by LeBreton et al. Thus, we are not arguing that these variables are not important correlates of performance, but rather that they are likely nontrivially correlated with constructs commonly included in selection batteries (i.e., intelligence, personality traits) and/or may influence variability in effects in ways that are not easily captured in meta-analytic matrices, which focus on average effects. In addition, some of these variables (e.g., organizational culture, human resource policies) may have a constant value for all participants in a single organization's validation study. When this occurs, these variables will not explain individual-level differences in job performance, thus reducing the size of the large multiple correlation that LeBreton et al. described.

Conclusion

Based on our review of the evidence, the .52 estimate of the interrater reliability of supervisor ratings of job performance is an appropriate estimate; corrections for unreliability do not appear to change our decisions regarding the choice of one selection tool over another; and most variables may be more strongly correlated than people expect, making it difficult to demonstrate continued incremental validity in predicting job performance when adding additional predictors. We agree with LeBreton et al. that psychologists need to be careful when applying and interpreting corrections, and we are thankful that they sponsored a discussion on the topic. Corrections are critical for both basic science (i.e., estimating population parameters) and practice (i.e., recognizing artifacts attenuating estimates on which our work may be evaluated by stakeholders, courts, and other third parties). Ultimately, the appropriate use of corrections depends on the purpose of the project. If the goal is to explain variation among a sample of incumbents on observed criterion scores, then no corrections need to be made. If the goal is to explain variation among incumbents on a true score for job performance, then a correction for unreliability is not only desirable but necessary. Finally, if the goal is to estimate how much variation among applicants is explained by a predictor for a true score on job performance, then corrections for range restriction and unreliability are indispensible. This goal represents the target validity inference that was included in Binning and Barrett's (1989) figure, but (rather interestingly) is omitted from LeBreton et al.'s reproduction of that figure. We believe that the target validity inference is the most important inference in personnel selection; it provides the critical link from the observed predictor to the criterion construct (see also Putka & Sackett, 2010).

References

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

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*(3), 478–494.

Chiaburu, D. S., & Harrison, D. A. (2008). Do peer make the place? Conceptual synthesis and meta-analysis of coworker effects on perceptions, attitudes, OCBs, and performance. *Journal of Applied Psychology*, *93*, 1082–1103.

Conway, J. M., & Huffcutt, A. I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance*, *10*(4), 331–360.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *1*, 333–342.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, *86*, 897–913.

Hunter, J. E. (1983). *Test validation for 12,000 jobs: An application of job classification and validity generalization to General Aptitude Test Battery* (U.S. Employment Service Test Research Report No. 45). Washington, DC: U.S. Department of Labor.

Hutchinson, D., Shen, W., Telford, B., Andel, S., Jang, S., & Ramsay, S. (2014, May). *Personality and justice perceptions: An updated meta-analysis*. Paper presented at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

Jensen, A. R. (1998). *The* g *factor: The science of mental ability*. Westport, CT: Praeger.

King, L. M., Hunter, J. E., & Schmidt, F. L. (1980). Halo in a multidimensional, forced-choice performance evaluation scale. *Journal of Applied Psychology*, *65*, 507–516.

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.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*, 195–244.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Ones, D. S., & Viswesvaran, C. (2001). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel *selection*. *International Journal of Selection and Assessment*, *9*, 31–39.

Paterson, T., Harms, P., & Crede, M. (2012, April). *The meta of all metas: 30 years of meta-analysis reviewed*. Paper presented at the 27th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Putka, D. J., & Sackett, P. R. (2010). Reliability and validity. In J. L. Farr, & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 9–49). New York, NY: Routledge.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*, 322–327.

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2009). High stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227.

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*(2), 262–274.

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

Scullen, S. E., Mount, M. K., & Sytsma, M. R. (1996, April). *Self, peer, direct report, and boss ratings of managers' performance*. Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Shen, W., Sackett, P. R., Kuncel, N. R., Beatty, A. S., Rigdon, J. L., & Kiger, T. B. (2012). All validities are not created equal: Determinants of variation in SAT validity across schools. *Applied Measurement in Education*, *25*, 197–219.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557–574.