Industrial and Organizational Psychology (2024), **17**, 296–302
doi:10.1017/iop.2024.13

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Operational validity/correlation coefficients are still valid for evaluating selection procedure effectiveness

In-Sue Oh[1] and Huy Le[2]

[1]Department of Management, Fox School of Business, Temple University, Philadelphia, PA, USA and [2]Department of Management, Alvarez College of Business, University of Texas at San Antonio, San Antonio, TX, USA
Corresponding author: In-Sue Oh; Email: insue.oh@gmail.com

In their focal article, Foster et al. (2024) propose reconsidering the current practice of evaluating selection procedure effectiveness. In particular, they recommend squaring observed correlations/validities and then multiplying the squared correlations by a correction/disattenuation factor of 4 (or dividing the squared correlations by ¼) to provide a more accurate index of selection procedure effectiveness. They argue that the correction factor of 4 appropriately accounts for the fact that variance specific to the individuals being rated (ratee main effects) represents only about ¼ of the total variance in job performance ratings. They argue that their recommendation leads to the conclusion that "selection tests work better than we think they do, and have for years." Although we agree with their conclusion (i.e., their article's title), we have some significant conceptual and methodological issues with their recommendation. Ultimately, we advocate for the continued use of operational validity/correlation coefficients in evaluating the effectiveness of selection procedures.

Before elaborating on our issues, we provide a simple example that helps the readers understand the gist of Foster et al.'s (2024) recommendation. Per their recommendation, a selection procedure with observed *r*/validity = .20 can be said to account for 16% of the ratee-relevant variance in job performance ratings. To obtain the value of 16%, we first square the observed validity ($.20^2 = .04$; 4% of the *total* variance in job performance ratings) and multiply it by 4 or divide it by ¼ ($.20^2 \times 4$ or $.20^2/[¼] = .16$; 16% of the *ratee-relevant* variance in job performance ratings), assuming that ¼ of the total variance in job performance ratings is attributed to ratee main effects. Following Foster et al.'s (2024) recommendation, another selection procedure with $r = .40$ would be reinterpreted as accounting for 64% ($.40^2 \times 4$ or $.40^2/[¼] = .64$) of the ratee-relevant variance in job performance ratings.

Do selection researchers regularly square validity/correlation coefficients?

Foster et al.'s (2024) recommendation mentioned above is premised on the following statement: "researchers *regularly* evaluate the effectiveness of selection measures based on the variance they account for in overall job performance ratings" (italic added). As early as 1965, "the validity coefficient is usually defined as the *correlation* of test score with outcome or criterion score" (Cronbach & Gleser, 1965, p. 31). Although we understand that (multiple) correlations (*r*, *R*) are habitually squared and interpreted as variance accounted for (perhaps due to the field's heavy reliance on regression), this is not the *regular* case among scholars and professional practitioners in personnel selection (e.g., Society for Industrial and

Organizational Psychology, 2018).[1] Instead, most of them usually quantify the effectiveness of a selection procedure using an operational validity/correlation coefficient, which is the observed validity/correlation corrected for measurement error in job performance ratings and range restriction on the predictor and, if possible, the criterion (e.g., Schmidt & Hunter, 1998; Viswesvaran et al., 2014, figure 1). Foster et al.'s (2024) recommendation (squaring observed validities/correlations as the first step) also runs counter to their Footnote 1—"see Funder and Ozer (2019) for a discussion about how the widespread practice of squaring correlations to estimate variance explained is actually a highly misleading metric."

Why is it problematic to square validity/correlation coefficients and compare them?

Jensen (1998) argued, "the common habit of squaring the validity (correlation) coefficient to obtain the proportion of variance in the criterion accounted for by the linear regression of criterion measures on test scores, although not statistically incorrect, is an uninformative and misleading way of interpreting a validity (correlation) coefficient" (p. 301; also see Funder & Ozer, 2019; Schmidt, 2017, p. 35). Schmidt and Hunter (2015) similarly argued that "variance-based indexes of effect size are virtually always deceptive and misleading and should be avoided, whether in meta-analysis or in primary research" (pp. 215–216; also see pp. 213–216 for more details). In particular, a serious problem emerges when we compare selection procedures using squared correlations. Using the two observed validity/correlation coefficients (.20 vs. .40) provided above, we can say the validity/correlation coefficient of .40 is twice as high as that of .20 because correlation coefficients are effect sizes that can be directly compared. If we square these correlations as the first step of their recommendation, then we have .16 and .04 (16% vs. 4% of the total variance in job performance ratings), and their difference is now fourfold ($= .16/.04$) instead of twofold ($= .40/.20$). More relevant to their recommendation, if we adjust these squared correlations for the ratee-relevant effects ($16\% \times 4$ vs. $4\% \times 4$), then their difference remains the same at fourfold albeit with each value quadrupled (64% vs. 16% of the *ratee-relevant* variance in job performance ratings). More drastically, if we apply Foster et al.'s (2024) recommendation to two other selection procedures with observed r/validity $= .20$ vs. .60, their difference of threefold is squared and amplified to ninefold: 16% vs. 144% (yes, over 100%!) of the *ratee-specific* variance in job performance ratings. We are unsure whether any researcher in personnel selection would agree with it. Some jokingly noted that if the goal is to maximize the difference between the two predictors, we should cube, rather than square, their validities/correlations.

The problem lies in the fact that squared correlations are related in a very nonlinear way to the corresponding effect sizes (correlations). Thus, comparing squared correlations, whether adjusted for the ratee main effect or not, is problematic because they are not effect sizes and, thus, they are neither compatible nor comparable. This is why validity/correlation coefficients are used, instead of squared correlations, as input to validity generalization (meta-analytic) studies. In fact, Funder and Ozer (2019) further clarified the reasoning behind the incompatibility issue.

> Squaring the r changes the scale of the effect from the original units to squared units. . . . the original, unsquared r reflects the size of the effect on the metric of the original measured units. . . . Similarly, a correlation of .4 reveals an effect twice as large as a correlation of .2; moreover, half of a perfect association is .5, not .707 [its square is .50]. Squaring the r is not merely uninformative; for purposes of evaluating effect size, the practice is actively misleading. (p. 158, bracket added)

[1]In fairness to Foster et al. (2024), we note that some used squared correlations when making pessimistic statements about self-reported measures of personality (never more than 10% of the variance accounted for).

Other issues with squaring correlation/validity coefficients include masking the sign of the relationship and making relatively modest yet still meaningful correlations (e.g., *r*s up to .23) less distinguishable or indistinguishable (e.g., all never more than 5% of variance accounted for; see Funder & Ozer, 2019 for more details). Furthermore, the validity (r) of a selection procedure, not its square (r^2), is a direct determinant of its utility and, if the cost is zero, is proportional to its utility (Brogden, 1949).

What is the basis for the correction factor of 4? Is it trustworthy?

As noted above, Foster et al.'s (2024) recommendation is not just to square an observed validity/correlation coefficient but also to multiply it by 4 assuming the proportion of the total variance attributable to the ratee main effects is ¼ or 25%. Below is an example from the focal article.

> Consider Hunter's 1986 meta-analysis, which reported an observed correlation with job performance ratings of .32 for complex jobs, corresponding to an R^2 of 10.24% ... A conservative estimate of performance relevant ratee main effects may be 25% of the variance in performance ratings ... cognitive ability's R^2 of roughly 10% actually predicts 40% of the variance in ratee main effects in complex jobs.

The ratee-relevant variance represents the total variance in job performance ratings attributable to true ratee/incumbent performance factors (the shared variance by all indicators of all performance dimensions plus the shared variance by all indicators of each performance dimension) versus other nonperformance (error) factors such as rater/supervisor idiosyncrasies and random measurement error (Scullen et al., 2000). In fact, the ratio of the variance due to the ratee main effects to the total variance represents a somewhat narrowly defined interrater reliability of job performance ratings, given that reliability equals true variance divided by total variance—that is, reliability = true variance/(true variance + error variance). That is, the proportional value of .25 can be regarded as their interrater reliability estimate of job performance ratings. They recommend using its inverse (derived from 1/.25) as the correction/disattenuation factor of 4 for squared correlations (i.e., multiplying squared correlations by 4). Below is our issue with this correction factor.

We wonder about the accuracy of the value of .25 mentioned above, as it is not based on any systematic research synthesis (e.g., meta-analysis). It seems to be derived from several articles cited in the focal paper, but each article is based on different study designs, methods, and measures. Moreover, many meta-analyses indicate that the mean interrater reliability of job performance ratings is in the .50–.60 range (Speer et al., 2024; Viswesvaran et al., 1996); Foster et al.'s (2024) suggested value of .25 appears very low, despite their claim that it is "a conservative estimate." Therefore, even if we were to adopt their recommended procedures, the accuracy of the correction factor of 4 (derived from 1/.25) remains a significant question mark.

Is their recommendation psychometrically new, unique, and useful?

As noted above, squaring a validity (correlation) coefficient, although uninformative and misleading in most cases, is not statistically incorrect (Jensen, 1998). Therefore, we may be able to psychometrically evaluate Foster et al.'s (2024) recommendation, particularly given that the .25 value represents their suggested reliability of job performance ratings, as discussed above. Using the example above, a selection procedure's observed r/validity is .3, which gives us an r^2 of .09, indicating that the selection procedure accounts for 9% of the total variance in job

performance ratings. Multiplying the squared *r* of .09 by the correction factor of 4 or dividing the squared *r* of .09 by ¼, we can say the selection procedure explains 36% of the *ratee-relevant* variance in job performance ratings.

Let's use a psychometric disattenuation procedure to estimate the operational validity of the same selection procedure while assuming no range restriction (operational validity = observed validity/SQRT of r_{YY}, the interrater reliability for job performance ratings). We first correct the *r* of .30 for measurement error in job performance ratings using the interrater reliability of .25. This yields the operational validity of .60 (= .30/SQRT of .25 = .30/.50). Per Foster et al.'s (2024) recommendation, if we square it, we have .36, which is the same as the 36% mentioned above.[2] Now, it should be clear that adjusting the squared observed correlation/validity for the ratee main effects as per their recommendation—that is, (r^2/reliability) or (r^2 × [1/reliability]), is identical to correcting the observed correlation/validity for measurement error in job performance ratings using the well-established disattenuation formula and squaring the resulting correlation—that is, (*r*/SQRT of reliability)^2. For example, using the example above with the observed *r* of .30 and the reliability of .25 for job performance ratings, Foster et al.'s (2024) suggestion of correcting a squared observed correlation for ratee main effects—that is, .30^2/.25 or .30^2 × 4, is the same as squaring the corresponding operational validity—that is, (.30/SQRT of .25)^2 or (.30 × 2)^2.

In summary, our point is that Foster et al.'s (2024) recommendation is essentially reduced to squaring operational validity (corrected for measurement error in job performance ratings alone in this case), casting doubt on the psychometric uniqueness and usefulness of their recommendation. In fact, this idea was already suggested by Guion (1965), who noted that "with an unreliable criterion, evaluation should be based upon *predictable* rather than *total* criterion variance" (p. 142, italics original). As noted earlier, their suggested reliability for job performance ratings (.25 for a single rater, .30 for aggregated ratings) is much lower than the currently available meta-analytic evidence in the .50–.60 range (e.g., Speer et al., 2024). Moreover, instead of using their suggested value of .30 for aggregated performance ratings, it is more psychometrically sound to use the Spearman-Brown prophecy formula—that is, .40 for two raters and .50 for three raters, assuming the interrater reliability for a single rater is .25.

Is comparing a squared correlation before and after corrections meaningful?

Foster et al. (2024) discuss the following as an "unrealistically high" correlation/validity.

> A combination of corrections for various forms of [indirect] range restriction and [un]reliability can produce correlations that seem unrealistically high (Schmitt, 2007). For example, Schmidt et al. (2008) estimated that the average correlation between GMA and job performance was as high as .88 for some [highly complex] jobs after applying multiple corrections to an observed correlation of .23, representing over a 1400% increase in the variance accounted for in the criterion measure (i.e., R^2 increasing from 5.5% to 77.4%). (brackets added)

To clarify, the .88 value represents the operational validity estimate corrected for unreliability in job performance ratings and substantial indirect range restriction (low selection ratios associated with hiring for complex jobs) on the predictor. The example above clearly illustrates

[2]If we use another correction factor of 3 while assuming that 1/3 (= 33.3%) of the total variance in job performance ratings is due to ratee main effects—that is, the reliability of the criterion measure is .333, then the selection procedure accounts for 27% (= 9% × 3 or .30^2/[1/3] or .30^2/.333) of the ratee-relevant variance in job performance ratings. Using the reliability of .333, its operational validity is estimated at .52 (= .30/SQRT of .333 = .30/.58). When we square it, we have .27, which is the same as the 27% mentioned above.

the same issue of comparing squared correlations detailed earlier (see Question 2), when comparing a given predictor's validity before and after psychometric corrections per Foster et al.'s (2024) recommendation. As reported earlier, the difference in validity before and after psychometric corrections is approximately four times (3.83 = .88/.23). However, when these two values are squared ($.23^2 = .053$, approximately their value of "5.5%" vs. $.88^2 = .774$, their value of "77.4%"), their difference is also squared to 14.64 times ($= .88^2/.23^2$, approximately their value of "over a 1,400% increase" [which should read as a 1,346% increase]). In fact, when we compare the correlations (.23 vs. 88) instead, it represents a 283% increase in validity due to the psychometric corrections mentioned above. So, their mention of "over a 1,400% increase" is a statistical exaggeration, highlighting a major problem associated with Foster et al.'s (2024) recommendation; as discussed earlier (see Funder & Ozer, 2019), squared correlations are not effect sizes, and, thus, they should not be directly compared.

Are psychometrically corrected validities untrustworthy?

Foster et al.'s (2024) description of Schmidt et al.'s (2008) operational validity estimate of .88 for cognitive ability tests for highly complex jobs as "unrealistically high" appears to suggest that their intuition tells them the real value cannot be as large as .88. We have several issues as follows. First, Foster et al.'s (2024) skepticism seems to be based solely on Schmitt (2007) or their own intuition rather than evidence. One of the issues that Schmitt had with Le et al. (2007) and Schmidt et al. (2008) is that the input artifact values used in their psychometric corrections are old and low, suggesting a possibility of overcorrections (also see Sackett et al., 2022, and Oh et al., 2023, for different views). As noted by Oh and Schmidt (2021), scholars who raised such concerns, however, failed to offer credible evidence that such artifact values (see Schmidt et al., 2008, Appendix B for details) are incorrect because they are low or old. "Low is not necessarily inaccurate (as high is not necessarily accurate), and old is not necessarily inaccurate (as new is not necessarily accurate)" (p. 171). Moreover, they fail to cite and discuss the other side of the debate (Schmidt et al., 2007).[3]

Second, as noted earlier, the .88 value represents a fully corrected operational validity estimate. As such, this value applies to the unrestricted/applicant population, not the restricted/incumbent sample. As is well known, corrections for indirect range restriction can be very substantial in certain conditions—such as high selectivity (low selection ratios) and the high correlation between scores on the actual (yet often unknown) selection criterion and scores on the predictor of interest—to the extent that a negatively observed validity changes sign after proper corrections. As Ree et al. (1994; also see Oh et al., 2023) advised, "this should not be interpreted as discouragement of the application of [range restriction] corrections but, rather, as a caution for use of the proper formula" (p. 301, brackets added). It has been long known that "if any intelligent use is to be made of validity statistics from a restricted group, some statistical [range restriction] correction procedures are necessary to estimate what validity coefficients would have been obtained if it had been possible to obtain test and criterion data from a representative sample of all those to whom the selection devices were applied" (Thorndike, 1949, pp. 171–172). However, there is uncertainty about how corrections for range restriction can and should be done alongside Foster et al.'s (2024) recommendation.

Third, it is ironic that Foster et al. (2024) seem to be uncomfortable or skeptical about psychometrically correcting an observed correlation/validity coefficient for artifacts while simultaneously advocating for correcting a squared observed correlation/validity coefficient for ratee main effects (i.e., multiplying it by the correction factor of 4 [= 1/.25]); as noted

[3]To conserve space, we do not (need to) rehash Schmidt et al.'s (2007) rejoinder to Schmitt (2007) and other commentators; interested readers are referred to pp. 69–74 of the Schmidt et al.'s 2007 article.

earlier (see Question 4), the latter is essentially the same as psychometrically correcting an observed validity coefficient for measurement error in job performance ratings using the reliability of .25 (the very basis of their correction factor of 4). In fact, if an observed *r*/validity is close to .50 or greater, their recommendation of squaring the observed *r*/validity and multiplying it by 4 can also yield very high values (even over 100%, which are implausible values). Conceivably, with additional corrections for range restriction, their recommendation would yield even higher values. It raises the question of why Foster et al. (2024) are not skeptical about these high values while they are so skeptical about the operational validity of .88 mentioned above.

In conclusion, considering all the issues discussed above regarding squaring observed validity/correlation coefficients and adjusting them for ratee main effects (by multiplying the squared correlations by 4), it is difficult to see any real benefit of adopting Foster et al.'s (2024) recommendation, especially in personnel selection where a selection procedure's operational validity, not its square, directly determines its economic value (e.g., utility). Therefore, instead of adopting their recommendation, we suggest we continue to use the operational validity coefficient (i.e., the observed validity corrected for measurement error in job performance ratings and range restriction on the predictor and, if possible, the criterion) as an index of selection procedure effectiveness, as has been done for years.

Acknowledgements. We thank Brian Holtz for his excellent suggestion for the title and the late Frank Schmidt for some of the insights provided in this commentary.

References

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, **2**, 171–183.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. University of Illinois Press.

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.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, **2**, 156–168.

Guion, R. M. (1965). *Personnel testing*. McGraw Hill.

Jensen, A. R. (1998). *The factor*. Westport, CT: Prager.

Le, H., Oh, I.-S., Shaffer, J., & Schmidt, F. (2007). Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis. *Academy of Management Perspectives*, **21**, 6–15.

Oh, I.-S., Le, H., & Roth, P. (2023). Revisiting, Sackett, et al. 's, 2022 rationale behind their recommendation against correcting for range restriction in concurrent validation studies. *Journal of Applied Psychology*, **108**, 1300–1310.

Oh, I.-S., & Schmidt, F. L. (2021). Suggestions for improvement in psychometric corrections in meta-analysis and implications for research on worker age and aging. *Work, Aging and Retirement*, **7**, 167–173.

Ree, M. J., Carretta, T. R., Earles, J. A., & Albert, W. (1994). Sign changes when correcting for range restriction: A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology*, **79**, 298–301.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2022). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, **107**, 2040–2068.

Schmidt, F., Le, H., Oh, I.-S., & Shaffer, J. (2007). General mental ability, job performance, and red herrings: Responses to Osterman, Hauser, and Schmitt. *Academy of Management Perspectives*, **21**, 64–76.

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology*, **5**, 32–41.

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., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, **61**, 827–868.

Schmitt, N. (2007). The value of personnel selection: Reflections on some remarkable claims. *Academy of Management Perspectives*, **21**, 19–23.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, **85**, 956–970.

Society for Industrial and Organizational Psychology (2018). *Principles for the validation and use of personnel selection procedures* (5th ed.). Author.

Speer, A. B., Delacruz, A. Y., Wegmeyer, L. J., & Perrotta, J. (2024). Meta-analytical estimates of interrater reliability for direct supervisor performance ratings: Optimism under optimal measurement designs. *Journal of Applied Psychology*, **109**, 456–467.

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

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., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I.-S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analysis. *Industrial and Organizational Psychology*, **7**, 507–518.

Cite this article: Oh, I-S. & Le, H. (2024). Operational validity/correlation coefficients are still valid for evaluating selection procedure effectiveness. *Industrial and Organizational Psychology* **17**, 296–302. https://doi.org/10.1017/iop.2024.13