*Industrial and Organizational Psychology* (2024), **17**, 303–308
doi:10.1017/iop.2024.21

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

# Same as it ever was: A clarification on the sources of predictable variance in job performance ratings

Paul R. Sackett[1], Dan J. Putka[2], and Brian J. Hoffman[3]

[1]University of Minnesota Twin Cities, Minneapolis, MN, USA, [2]Human Resources Research Organization, Alexandria, VA, USA and [3]University of Georgia, Athens, GA, USA
**Corresponding author:** Paul R. Sackett; Email: psackett@umn.edu

Foster et al. (2024) offer a new perspective on the validity of the predictors used for personnel selection. The heart of their argument is: (a) there are multiple sources of variance in the ratings that are widely used as criteria in estimating validity, (b) generalizability theory gives us a tool for partitioning these sources of variance, (c) person (e.g., ratee) main effects are the only source of reliable between-ratee variance in performance ratings that is predictable, (d) existing work on partitioning variance gives estimates of person main effect variance as accounting for about 25% of rating variance, (e) we should rescale our validity estimates as a percentage of this 25% possible explainable variance, (f) and by doing so we find that our predictors explain a larger proportion of explainable variance in job performance ratings.

This is a novel and clever argument. However, we believe that there are several problematic assumptions made by Foster et al. (2024) that lead us to the conclusion that the proportion of explainable variance in performance ratings in typical validation research is in fact *far* higher than the 25% estimate used by Foster et al. (2024) Consequently, we have not underestimated the value of our predictors. Below we offer the series observations that lead to our conclusion.

## Faulty assumptions about sources of consistent-reliable, between-ratee variance

Much of the argument made by Foster et al. (2024) is premised on the claim that person main effects are the only source of variance "specific to the person rated" (pg. 3) or more specifically, the only source of consistent between-ratee variance that can be predicted by the predictors commonly used in personnel selection. In this case, by "consistent" we mean variance that would be consistent (reliable) across indicators of a given performance dimension of interest such as raters and/or items. Unfortunately, this premise appears to reflect a fundamental misunderstanding of sources of consistent between-ratee variance in ratings, regardless of whether one is dealing with a multi-item, multisource rating assessment designed to assess multiple dimensions or a simpler, behaviorally anchored rating assessment where there is only one rating scale (item) per dimensions assessed.

To set the stage we make three key points. First, sources of variance can viewed as a function of five components—person (p), item (i), dimension (d), rater (r), and source(s)—and all possible interaction between them. Second, a subset of these components reflects between-person variance that is consistent across raters (i.e., reliable variance from a traditional perspective), namely the person main effect and interaction effects that involve persons but that do not include rater. Any consistent between-person component that does not include raters is reliable from a traditional perspective and, thus, is potentially predictable. Third, not all effects may be uniquely estimated depending on one's measurement design. To illustrate, consider Figure 1, which provides a

© The Author(s), 2024. Published by Cambridge University Press on behalf of Society for Industrial and Organizational Psychology.

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press

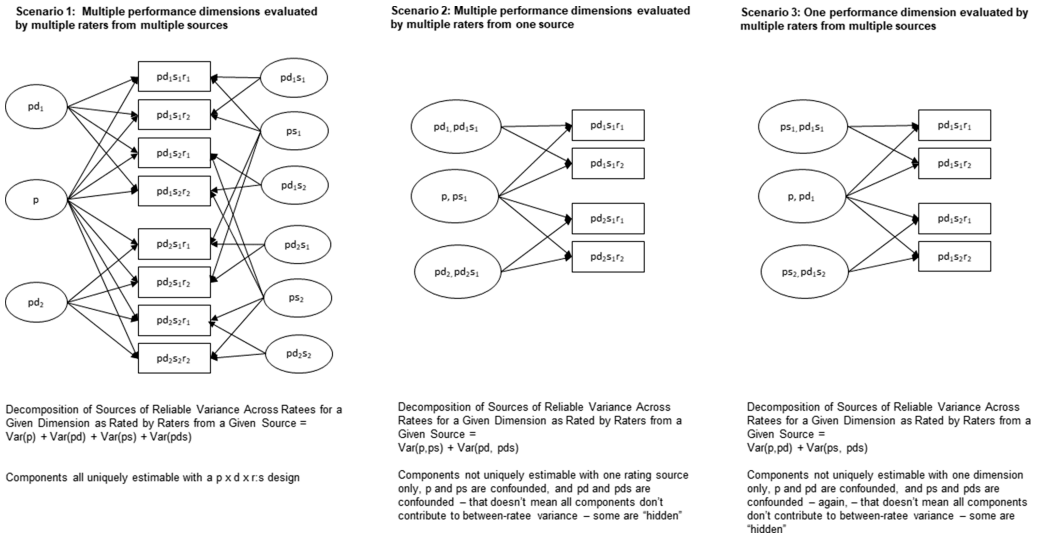



**Figure 1.** Example of decomposition of reliable variance in performance ratings. For the sake of parsimony, we've omitted person x item (pi) and person x item x source (pis) effects, which would also contribute to reliable variance if one was only interested in generalizing ratings across raters (i.e., a traditional interrater reliability perspective on ratings). In the scenarios above, pi would be confounded with pd effects, and pis would be confounded with pds effects, as we are assuming one rating scale (effectively, one item) per dimension.

structural depiction of how consistent between-ratee variance in ratings of a given performance dimension are decomposed into a set of variance components. For the sake of this example, assume each dimension is rated using a single rating scale tailored to that dimension and all raters within a source rate all persons (ratees).

Assume each of the boxes represents observed ratings for ratees on a given dimension as assessed by a given rater within a given rating source. The sources of consistent between-ratee variance can viewed as being a function of the person main effect (p) and three interactions between persons and the other nonrater facets (i.e., person x dimension [pd], person x source [ps], and person x dimension x source [pds]).[1] Not all of these components may be uniquely estimated depending on one's measurement design. For illustration, Scenario 1 describes a situation where all highlighted components are uniquely estimable, and the other scenarios reflect situations where some effects become hidden within or confounded with other effects as a result of design limitations. Clearly, to the extent that the nonperson main effects in Figure 1 are nonzero, they all represent sources of between-person variance that are consistent for raters within a given source. The key takeaway here is that any consistent between-ratee variance is potentially predictable, not simply the person main effect.

Claiming the only variance predictable in ratings of a given performance dimension stems from person main effects is problematic on logical grounds as well. For example, let's say we have a set of two performance constructs of interest: task and contextual performance. By focusing on the person main effect only, we would be claiming the only valid variance in task and contextual performance is that which is common between them. Obviously, that's problematic as they are distinct constructs whose valid variance will reflect both a common (person main effect) and construct-specific factors (person by dimension interaction).

[1]Again, we omit pi and pis effects here for the sake of parsimony but assume under the scenarios depicted in Figure 1, they contribute to pd and pds effects, respectively.

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press

## An empirical illustration of how Foster et al. (2024) underestimate the amount of predictable variance in performance ratings

The person main effects reflect variance common across all dimensions, sources, raters, and items and is akin to a general performance factor. According to Foster et al. (2024) the person effect accounts for the largest proportion of variance in performance ratings and is the only source of variance predictable by selection tools. Their exemplar study is Jackson et al. (2020), which is a sophisticated partitioning of variance in multisource ratings and estimates that the person main effect explains approximately 25% of the variance in performance ratings. In the previous section, we noted how other sources beyond person main effects can contribute to reliable variance in ratings, and here we argue such systematic effects are predictable, which argues against using the person main effect as the upper limit for validity. As an example, we focus below on person x source variance (the second largest source of variance) and briefly review evidence that this source of variance can indeed be predicted by predictors commonly used in selection settings.

The person x source interaction reflects variance that is common across multiple raters from a given rating source and is akin to a general performance factor that is shared across raters from one level. For instance, all the ratings provided by one peer converge with all the ratings provided by another peer but not those provided by supervisors or subordinates. In the exemplar study relied on by Foster et al. (2024) the person x source interaction explained 20%–27%, nearly as large as the person main effect. By focusing on the person main effect only, Foster et al. (2024) imply that the only valid variance in a given performance dimension (e.g., contextual performance) is that which is common between supervisors and peers. This seems problematic from an ecological perspective (Lance et al., 2008) in that supervisors and peers are likely exposed to different elements of contextual performance, so that valid variance in performance will reflect both a common (person main effect) and source specific factor (person by source interaction). Indeed, Hoffman and Woehr (2009) have shown that source effects are differentially correlated with predictors commonly used in selection settings (e.g., assessment centers, cognitive ability, and personality). More plainly, source effects reflect variance that can be predicted in selection contexts, and thus, this variance should be included with the person main effect as a predictable source of variance. Indeed, as stated above, if one were to decompose the variance of two supervisors' ratings (the traditional design of selection systems), then the person x source effect would actually be estimated as part of the person effect. In addition to the person x source interaction, there are several other variance sources that would also contribute to the predictable variance (e.g., person x dimension and person x source x dimension effects,), as these constitute reliable sources of variance.

To help empirically illustrate our points, we provide two tables of pre-aggregated variance component estimates from Jackson et al.'s (2020) Table 2. In our first table, we categorize their variance components into: (a) components of consistent-reliable between-ratee variance, (b) components of inconsistent-unreliable between-ratee variance, and (c) components of total variance that don't contribute to between-ratee variance, under the assumption that one wishes to treat raters and items as sources of error in ratings (i.e., a multifaceted error perspective). In our second table, we provide a comparable categorization of variance components under the assumption that one wishes to treat raters as the only source of error in ratings (i.e., a traditional interrater reliability perspective). In each table, we sum the reliable (and hence predictable) components and unreliable (and hence unpredictable) components and find that the reliable components account for 52.5% of the between-person variance under the assumptions in Table 1 and 67.7% of the between-person under the assumptions in Table 2. Both of these percentages are much larger than Foster et al. (2024)'s estimate of 25% of the between person variance as potentially predictable. Note also that the 67.7% estimate for reliable between-person variance in Table 2 is highly similar to revised estimates for single-rater reliability of job performance ratings published in two recent meta-analyses (Speer et al., 2023; Zhou et al., 2024).

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press

**Table 1.** Variance Decomposition of Pre-Aggregated Multisource Performance Ratings Based on Jackson et al (2020) Table 2 (Treating Raters and Items as Sources of Measurement Error-Multifaceted Error Perspective)

| VC | Raw VC | % of total variance` | % of between-ratee variance |
|---|---|---|---|
| Components of consistent-reliable between-ratee variance | | | |
| P | .064 | 17.2 | 26.4 |
| pd | .003 | 0.7 | 1.1 |
| ps | .060 | 16.2 | 24.8 |
| pds | .001 | 0.1 | 0.2 |
| Total | | | **52.5** |
| Components of inconsistent-unreliable between-ratee variance | | | |
| r:s | .014 | 3.7 | 5.7 |
| pi:d | .033 | 8.8 | 13.4 |
| pr:s | .019 | 5.2 | 7.9 |
| dr:s | .001 | 0.1 | 0.2 |
| ir:ds | .004 | 1.1 | 1.6 |
| pdr:s | .004 | 1.0 | 1.5 |
| pis:d | .004 | 1.2 | 1.8 |
| pir,ds,e | .038 | 10.1 | 15.4 |
| Total | | | **47.5** |
| Components of total variance that don't contribute to between-ratee variance | | | |
| S | .065 | 17.4 | |
| ds | .001 | 0.2 | |
| D | .010 | 2.8 | |
| i:d | .047 | 12.7 | |
| i:ds | .006 | 1.5 | |
| Grand total | .372 | 100.0 | |

*Note*. Jackson et al. (2020) misclassified various components as sources of between-person variance; however, there were additional components that would not contribute to between-person variance in that they did not involve an interaction with persons (p) and reflected factors that we assume were fully crossed with persons (i.e., items, sources, dimensions). We correct this misclassification above when calculating percentages of between-person variance.

## So has the field underestimated validity?

Foster et al. (2024) rescale validity coefficients by (a) squaring them and (b) dividing by .25. If validity were .20, r-squared is .04; dividing by .25 gives a value of .16: In short, we would explain four times as much variance as we previously thought (.16 vs. .04). In contrast, with our estimate of reliable and predictable between-person variance at 67.7%, dividing .04 by .677 gives a value of .059: a modest increase rather than the dramatic fourfold increase asserted by Foster et al. (2024) Importantly, our value is quite comparable to the value obtained using the traditional correction for attenuation formula with the best current meta-analytic estimate of interrater reliability of .65 as the estimate of reliable variance in the criterion measure (Zhou et al., 2024; see also Speer et al., 2023). Our conclusions are that person main effects cannot be used to rescale our existing validity estimates, there are many sources of reliable and hence potentially predictable sources of variance

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press

**Table 2.** Variance Decomposition of Pre-Aggregated Multisource Performance Ratings Based on Jackson et al (2020) Table 2 (Treating Raters as the Only Source of Measurement Error-Traditional Interrater Reliability Perspective)

| VC | Raw VC | % of total variance` | % of between-ratee variance |
|---|---|---|---|
| Components of consistent-reliable between-ratee variance | | | |
| p | .064 | 17.2 | 26.4 |
| pd | .003 | 0.7 | 1.1 |
| ps | .060 | 16.2 | 24.8 |
| pds | .001 | 0.1 | 0.2 |
| pi:d | .033 | 8.8 | 13.4 |
| pis:d | .004 | 1.2 | 1.8 |
| Total | | | **67.7** |
| Components of inconsistent-unreliable between-ratee variance | | | |
| r:s | .014 | 3.7 | 5.7 |
| pr:s | .019 | 5.2 | 7.9 |
| dr:s | .001 | 0.1 | 0.2 |
| ir:ds | .004 | 1.1 | 1.6 |
| pdr:s | .004 | 1.0 | 1.5 |
| pir,ds,e | .038 | 10.1 | 15.4 |
| Total | | | **32.3** |
| Components of total variance that don't contribute to between-ratee variance | | | |
| s | .065 | 17.4 | |
| ds | .001 | 0.2 | |
| d | .010 | 2.8 | |
| i:d | .047 | 12.7 | |
| i:ds | .006 | 1.5 | |
| Grand total | .372 | 100.0 | |

*Note.* Jackson et al. (2020) misclassified various components as sources of between-person variance; however, there were additional components that would not contribute to between-person variance in that they did not involve an interaction with persons (p) and reflected factors that we assume were fully crossed with persons (i.e., items, sources, dimensions). We correct this misclassification above when calculating percentages of between-person variance.

in job performance ratings beyond the person main effect, and the amount of explainable and explained variance in validation settings is the same as it ever was.

## References

**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.

**Hoffman, B. J., & Woehr, D. J.** (2009). Disentangling the meaning of multisource performance rating source and dimension factors. *Personnel Psychology*, **62**(4), 735–765.

**Jackson, D. J., Michaelides, G., Dewberry, C., Schwencke, B., & Toms, S.** (2020). The implications of unconfounding multisource performance ratings. *Journal of Applied Psychology*, **105**(3), 312.

**Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E.** (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review*, **18**(4), 223–232.

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press

**Speer, A. B., Delacruz, A. Y., Wegmeyer, L. J., & Perrotta, J.** (2023). Meta-analytical estimates of interrater reliability for direct supervisor performance ratings: Optimism under optimal measurement designs. *Journal of Applied Psychology*, **109**(3), 456–467. https://doi.org/10.1037/apl0001146

**Zhou, Y., Shen, W., Beatty, A. S., & Sackett, P. R.** (2024). An updated meta-analysis of the interrater reliability of supervisory performance ratings. *Journal of Applied Psychology*, **109**, 949–970. https://doi.org/10.1037/apl0001174

**Cite this article:** Sackett, P. R., Putka, D. J., & Hoffman, B. J. (2024). Same as it ever was: A clarification on the sources of predictable variance in job performance ratings. *Industrial and Organizational Psychology* **17**, 303–308. https://doi.org/10.1017/iop.2024.21

https://doi.org/10.1017/iop.2024.21 Published online by Cambridge University Press