Industrial and Organizational Psychology (2023), **16**, 341–343
doi:10.1017/iop.2023.27

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Interpreting validity evidence: It is time to end the horse race

Kevin Murphy

Department of Work and Employment Studies, University of Limerick, Limerick, Ireland
Email: krm10@me.com

For almost 25 years, two conclusions arising from a series of meta-analyses (summarized by Schmidt & Hunter, 1998) have been widely accepted in the field of I–O psychology: (a) that cognitive ability tests showed substantial validity as predictors of job performance, with scores on these tests accounting for over 25% of the variance in performance, and (b) cognitive ability tests were among the best predictors of performance and, taking into account their simplicity and broad applicability, were likely to be the starting point for most selection systems. Sackett, Zhang, Berry, and Lievens (2022) challenged these conclusions, showing how unrealistic corrections for range restriction in meta-analyses had led to substantial overestimates of the validity of most tests and assessments and suggesting that cognitive tests were not among the best predictors of performance. Sackett, Zhang, Berry and Lievens (2023) illustrate many implications important of their analysis for the evaluation of selection tests and or developing selection test batteries.

Discussions of the validity of alternative predictors of performance often take on the character of a horse race, in which a great deal of attention is given to determining which is the best predictor. From this perspective, one of the messages of Sackett et al. (2022) might be that cognitive ability has been dethroned as the best predictor, and that structured interviews, job knowledge tests, empirically keyed biodata forms and work sample tests are all better choices. In my view, dethroning cognitive ability tests as the best predictor is one of the least important conclusions of the Sackett et al. (2022) review. Although horse races might be fun, the quest to find the best single predictor of performance is arguably pointless because personnel selection is inherently a multivariate problem, not a univariate one.

First, personnel selection is virtually never done based on scores on a single test or assessment. There are certainly scenarios where a low score on a single assessment might lead to a negative selection decision; an applicant for a highly selective college who submits a combined SAT score of 560 (320 in Math and 240 in Evidence-Based Reading and Writing) will almost certainly be rejected. However, real-world selection decisions that are based on any type of systematic assessments will usually be based on multiple assessments (e.g., interviews plus tests, biodata plus interviews). More to the point, the criteria that are used to evaluate the validity and value of selection tests are almost certainly multivariate. That is, although selection tests are often validated against supervisory ratings of job performance, they are not designed or used to predict these ratings, which often show uncertain relationships with actual effectiveness in the workplace (Adler et al., 2016; Murphy et al., 2018). Rather, they are used to help organizations make decisions, and assessing the quality of these decisions often requires the consideration of multiple criteria.

Virtually all meta-analyses of selection test validity take a univariate perspective, usually examining the relationship between test scores and measures of job performance (as noted above, usually supervisory ratings, but sometimes objective measures or measures of training outcomes). Thus, validity if often expressed in terms of a single number (e.g., the corrected correlation

between cognitive ability test scores and job performance is .51 in Schmidt and Hunter, 1998). Sackett et al. emphasize the importance of also considering the variability of validity estimates, but their analysis still focuses on univariate validities. These validity estimates are only partially useful for making decisions about which predictors should be used to select among applicants.

Overall job performance is certainly important, but it is misleading to act as if the goal of personnel selection is to select applicants with the highest estimated performance. For example, organizations might consider both estimates of job performance and assessments of the effects of choosing among alternative assessments on diversity when evaluating selection tests (Sackett et al., 2022 illustrate how revised estimates of test validity change the validity–diversity trade-off often encountered when using cognitive tests in selection). More broadly, they might consider different aspects of job performance (e.g., task proficiency and citizenship behavior) as well as other criteria that are distinct from individual performance and might weight them differently rather than only focusing on estimates of overall performance levels.

Murphy and Shiarella (1997) examined the implications of using multivariate estimates of validity rather than relying on univariate estimates. They showed, for example, that differences from organization to organization in the relative emphasis given to task proficiency versus organizational citizenship might have important implications for evaluations of alternative selection test batteries, and that the differences in multivariate validity when there are multiple predictors and multiple criteria could be substantial, depending on the relative emphasis given to both tests and criteria. Murphy (2019) presented formulas and **R** code for evaluating validity when there are multiple tests and multiple criteria, each of which might be weighted differently. Although at one-time assessments of multivariate validity might have been challenging because they are based on correlation matrices rather than on single pairs of variables, the widespread availability of matrix-friendly languages (e.g., **R**) make assessments of multivariate validity and changes in multivariate validity as the composition of selection test batteries or the range and importance of different criteria change.

The "horse race" orientation encourages researchers and practitioners to focus on a single number (the average validity), or at most the distribution of a univariate validity estimate. We have the tools to easily evaluate multivariate validity and to examine the effects of organizational decisions (e.g., how important are different criteria of criterion dimensions, how much weight should be given to different tests of assessments) on validity. This suggests the days of asking which predictors are best should be, and hopefully are, over. Instead of asking which predictor wins the horse race, we should be asking which combinations of predictors are most appropriate for which purposes and how the values and preferences of users should be taken into account when defining criteria and determining the relative weights to give to both tests and criteria. We have learned a great deal from studies of univariate validity, but it is time to move on from a univariate focus. Serious consideration of the multivariate nature of personnel selection should lead to new and important insights about how to create and evaluate personnel selection systems.

Moving from a univariate to a multivariate perspective on validity has important implications. First, univariate studies are likely to underestimate the validity of personnel selection systems. The validity of personnel decisions based on multiple valid predictors is typically (but not always; see Sackett et al., 2017) higher than the validity of individual predictors, particularly if the correlations among predictors are small. Second, multivariate validity can be substantially affected by the way different organizations define the criterion domain. For example, the validity of a selection system that includes measures of cognitive ability and conscientiousness for predicting a performance domain that includes individual task performance and organizational citizenship is generally highest when the design of the selection system mirrors the emphasis the organization places on task performance versus citizenship (Murphy & Shiarella, 1997). That is, in organizations where citizenship is emphasized, multivariate validity increases when more weight it given to conscientiousness, whereas in organizations that place more emphasis on individual task performance, multivariate validity increases when more weight is given to cognitive ability. This

https://doi.org/10.1017/iop.2023.27 Published online by Cambridge University Press

also implies that an old idea that has largely been discarded, the specificity of validity, might need to be reexamined. Even if jobs are identical in two organizations, differences in the way these organizations define a good performing employee might lead to substantial variation in the validity of predictions from the same selection system. Old ideas about the specificity of validity rarely articulated *why* the validity of predictors or selection systems might vary across organizations, but a fully multivariate perspective suggests that the validity and value of a selection system depend on: (a) how predictors are combined and weighed, (b) how the facets of the criterion domain are combined to define the organization's definition of good performance, and (c) the match of mismatch between selection system design and the organization's emphasis on different aspects of the criterion domain.

References

Adler, S., Campion, M., Colquitt, A., Grubb, A., Murphy, K. R., Ollander-Krane, R., & Pulakos, E. D. (2016). Getting rid of performance ratings: Genius or folly. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **9**, 219–252.

Murphy, K., Cleveland, J. & Hanscom, M. (2018). *Performance appraisal and management*. Sage.

Murphy, K. R. (2019). Understanding how and why adding valid predictors can decrease the validity of selection composites: A generalization of Sackett, Dahlke, Shewach, and Kuncel (2017). *International Journal of Selection and Assessment*, **27**, 249–255.

Murphy, K. R. & Shiarella, A. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, **50**, 823–854.

Sackett, P. R., Dahlke, J. A., Shewach, O. R., & Kuncel, N. R. (2017). Effects of predictor weighting methods on incremental validity. *Journal of Applied Psychology*, **102**, 1421–1434.

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.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **16**(3), 283–300.

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.

Cite this article: Murphy, K. (2023). Interpreting validity evidence: It is time to end the horse race. *Industrial and Organizational Psychology* **16**, 341–343. https://doi.org/10.1017/iop.2023.27