Industrial and Organizational Psychology (2024), **17**, 314–318
doi:10.1017/iop.2024.14

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

E Pluribus Unum? Why criteria should be multimethod and multirater

Jeffrey M. Cucina[1] and Theodore L. Hayes[2]

[1]U.S. Customs and Border Protection, Washington, DC, USA and [2]U.S. Customs and Border Protection, Arlington, VA, USA
Corresponding author: Jeffrey M. Cucina; Email: jcucina@gmail.com

Foster et al.'s (2024) focal article provides a bracing reminder of a central tenet of applied psychology. Individual difference traits, such as general mental ability (GMA) or personality, and trait measurement methods, such as interviews or selection tests, are consistently related to on-the-job performance. We take issue with Foster et al. (2024) because some of the information they offer is limiting and, in our opinion, inaccurate in the service of expediency. This is particularly the case in their table, which attempts to assemble a range of predictor and job performance correlations for use as a handy reference. However, their table and other parts of their argumentation omit consideration of multimethod measurement of performance and the sources of multirater variance on which organizations should focus.

The importance of multimethod criteria: moving beyond sole reliance on supervisory ratings

Foster et al. (2024) write that supervisory ratings of job performance "are widely considered the primary criterion" for personnel selection measures (p. 1). This is not the case nor should it be. Selection research has a rich history with other criterion measures, and multimethod approaches to performance measurement are preferable. Hunter's (1983b) original validity generalization study criteria included both training (typically measured using knowledge tests) and job performance (typically measured using supervisory ratings) criteria. Later, meta-analytic summaries list validities separately for these two criteria (Schmidt, 2013; Schmidt & Hunter, 1998). Many entry-level jobs have a formal training period and training performance is a critical criterion. Indeed, the U.S. military uses training as a criterion for military selection tests (Brown et al., 2006).

As Hunter (1983b) mentioned, training performance is a measure of job knowledge, which is another key criterion for validation studies including the U.S. Army's Project A (McHenry et al., 1990) and civilian federal government studies (Paullin et al., 2010; Schmidt et al., 1986; van Rijn & Payne, 1980). Work sample measures of job performance, including hands-on-performance tests (HOPTs) and low-fidelity job simulations, are also often used as criteria. Work sample criterion measures were used as criteria for General Aptitude Test Battery validation studies (Salgado & Moscoso, 2019). Cucina et al. (2024) meta-analyzed Armed Services Vocational Aptitude Battery validities using HOPTs, which were lauded as the "gold standard" criterion for job performance (Abrahams et al., 2015, p. 45). Low-fidelity job simulations, such as walk-through performance tests in which incumbents walk trained raters through how they would perform job tasks (Ree

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. Federal Government, or any agency of the U.S. Federal Government.

et al., 1994) or paper-and-pencil work simulations (Hayes et al., 2002), are also excellent criteria. There are other criteria beyond supervisory ratings with validity evidence that should be considered when evaluating how well selection tests work. Causal path modeling involving GMA as a predictor indicates that supervisory ratings are a distal outcome, whereas job knowledge and objective measures of task performance (e.g., work samples, HOPTs) are more proximal criteria (Hunter, 1983a, 1986). In fact, the relationship between GMA and supervisory ratings is entirely mediated by the proximal criteria.

Other recent reviews have also, incorrectly, focused on supervisory ratings as the sole criterion measure (Sackett et al., 2022). Perhaps supervisory ratings provide an aura of an independent and external third-party outcome. After all, from a naïve perspective, who would better known an employee's performance than their supervisor? Yet many issues and biases can impact supervisory ratings, including opportunity to observe (MacLane et al., 2020), the use or nonuse of behaviorally anchored ratings scales (BARS), frame-of-reference training, leader–employee relationships (e.g., leader–member exchange; Martin et al., 2016), data collection procedures and proctoring (Grubb, 2011), rating adjustment policies (Al Ali et al., 2012), and social and goal-related issues (Murphy & Cleveland, 1995). Foster et al. (2024) offer suggestions for reducing some biases such as statistically eliminating leniency and strictness. However, these methods can have measurement side effects. If BARS are used, statistically adjusting the ratings may remove the link between the numerical ratings and the behavioral anchors.

From a practical perspective, supervisory ratings for research purposes can be collected easily and cheaply (e.g., using an unproctored online survey platform) and archival administrative ratings may be available. However, sometimes one gets what one pays for. Administrative ratings often lack variance compared to research-based ratings. This was noted in a U.S. General Accounting Office study of the promotions process for special agents at the Drug Enforcement Administration. Supervisors rated the job performance of candidates for promotions to supervisory positions. U.S. General Accounting Office (2003, p. 27) reported that the average ratings were "uniformly exceptional—almost a perfect 5," which calls into question the "critical importance in other HR decisions, such as promotions" of supervisory job performance ratings (Foster et al., 2024 p. 5). Research-based ratings can also suffer quality issues (Grubb, 2011).

The importance of multirater criteria: measuring ratee variance shared across supervisors

Foster et al. (2024) hypothesize that criterion-related validity will improve by predicting a supervisor's unique perspective on an employee's performance (e.g., after removing variance in performance ratings shared with other supervisors). They suggest "re-establishing the importance (i.e., the weights) that individual supervisors give to the different performance dimensions" (p. 13) and matching applicants to supervisors based on this. At one point, they almost go as far as to equate "individual supervisors" and "sole proprietors" (p. 14).

Here we present an alternative view. From the organization's perspective, it is primarily desirable to predict the ratee variance, not ratee × rater variance, shared across supervisors. Organizations typically hire I-O psychologists to work for the organization's benefit rather than for an individual supervisor. Aspects of performance that all supervisors agree upon should be the criterion of interest rather than idiosyncratic viewpoints of performance from a particular supervisor. Matrixed teams are common in many organizations with employees reporting to multiple supervisors, and it is rarely the case that employees are selected to work only for one supervisor throughout their tenure. For frontline and hourly positions involving shiftwork, employees may report to multiple supervisors in one shift or different ones on different days. Supervisors and employees also transfer to other positions within an organization as their careers progress. Thus, organizations are typically interested in hiring the best employees for the

organization (i.e., those who will be viewed by multiple supervisors as high performers) rather than staffing fiefdoms for individual supervisors filled with employees having the supervisors' pet competencies or subject to idiosyncratic supervisor weighting of competencies. Asking individual supervisors to weight the importance of different competencies for one vacancy for a larger job in an organization negates the critical role of conducting a thorough job analysis. This is paramount to conducting separate job analyses for each supervisor with a sample size of one, focusing explicitly on individual differences in supervisory job analysis ratings and using that disagreement (or error) to improve validity. This approach could introduce any number of biases into selection systems, such as the similar-to-me bias or self-serving bias (Cucina et al., 2012).

Collecting, maintaining, and using this data will prove challenging in practice. Rather than having one test for a larger job (e.g., cashier), an organization would need to have separate tests or cutoff scores for each vacant position for that job, and these would need to be tied to individual supervisors. Different selection standards for the same job title within the same organization would be applied and applicants may view this as unfair. It also presents business necessity and job-relatedness issues if a particular supervisor's selection system yields adverse impact whereas another's does not.

We sympathize with the motivation to maximize criterion-related validity coefficients. However, fully embracing Foster et al. (2024)'s methodology could lead to an instance of Kerr's (1975) folly, whereby we reward ourselves for increased validity coefficients while hoping for an increase in job performance and utility which does not actual materialize. Instead, we recommend matching employees to the job via selection systems that are validated to predict ratee variance in job performance ratings that is shared across multiple raters, not idiosyncratic ratee × rater variance, and are based on a job analysis for the job itself rather than a single supervisor's job analysis ratings. Doing so leads to predictors that are standardized and that have improved construct validity. Furthermore, path analytical research indicates that the ratee variance that is shared across supervisors largely depends on job knowledge and task performance, which should also be considered as criteria in validation efforts.

A call for a comprehensive meta-analytic intercorrelation table

The second problem perpetuated by Foster et al. (2024) is the presentation of a meta-analytic intercorrelation table of predictor traits and methods. We understand the goals of providing this table. However, the complexities of putting such a table together are significant because the meta-analysis literature contains differing levels of correction for measurement error, differing levels of correction for range restriction, different predictor reliabilities and construct validities, and different research populations. Additionally, some selection measures, such as interviews, are methods combining different sources of true score variance rather than tests of traits. Ideally, such a table would include the full range of predictors covered by Schmidt and Hunter (1998) and Schmidt (2013). It would also include corrections for direct (Case II) and indirect (Case IV) range restriction in both the criterion-related validities and predictor correlations in addition to observed correlations that Foster et al. (2024) presented. Correlations with other types of criteria besides just supervisory ratings are needed, as are incorporation of moderators (e.g., job complexity). Examples of meta-analytic tables with consistent corrections include O'Boyle et al. (2011) and Schmidt and Hunter (1998).

Foster et al. (2024) start an important conversation. What is needed now is a table with a broader range of predictors and criteria and consistent degrees of corrections for biases and error. Presenting result as validity coefficients (r) instead of squared validities (r^2) would avoid the issues described by Funder and Ozer (2019) and provide a metric linearly related to utility. Finally, a greater realization that supervisory ratings include multiple sources of variance could lead to better validation studies that describe validity for the work role rather than for a single rater.

Unfortunately, a table such as the one that Foster et al. (2024) present results in unfair comparisons (Cooper & Richardson, 1986) of apples and potatoes.

References

Abrahams, N. M., Mendoza, J. L., & Held, J. D. (2015). Interpreting the correlation (validity) coefficient. In J. D. Held, T. R. Carretta, J. W. Johnson, & R. A. McCloy (Eds.), *Technical guidance for conducting ASVAB validation/standards studies in the U.S. Navy* (pp. 37–49). Navy Personnel Research, Studies, and Technology (Technical Report NPRST-TR-15-2).

Al Ali, O. E., Garner, I., & Magadley, W. (2012). An exploration of the relationship between emotional intelligence and job performance in police organizations. *Journal of Police and Criminal Psychology*, **27**(1), 1–8. https://doi.org/10.1007/s11896-011-9088-9

Brown, K. G., Le, H., & Schmidt, F. L. (2006). Specific aptitude theory revisited: Is there incremental validity for training performance? *International Journal of Selection and Assessment*, **14**(2), 87–10.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology*, **71**(2), 179–184. https://doi.org/10.1037/0021-9010.71.2.179

Cucina, J. M., Burtnick, S. K., De la Flor, M. E., Walmsley, P. T., & Wilson, K. J. (2024). Meta-analytic validity of cognitive ability for hands-on military job proficiency. *Intelligence*, **104**, 101818. https://doi.org/10.1016/j.intell.2024.101818

Cucina, J. M., Martin, N. R., Vasilopoulos, N. L., & Thibodeaux, H. F. (2012). Self-serving bias effects on job analysis ratings. *Journal of Psychology: Interdisciplinary and Applied*, **146**(5), 1–21.

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**(2), 156–168.

Grubb, A. D. (2011). Promotional assessment at the FBI: How the search for a high-tech solution led to a low-tech simulation. In S. Adler & N. T. Tippins (Eds.), *Technology-enhanced assessment of talent* (pp. 293–306). Jossey-Bass.

Hayes, T. L., McElreath, J., & Reilly, S. M. (2022). The criterion-related validity of logic-based tests of reasoning for personnel selection. In T. L. Hayes (chair), *The validity of logic-based measurement for selection and promotion decisions*. Symposium at the 17th Annual SIOP Conference.

Hunter, J. E. (1983a). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Ed.), *Performance measurement and theory* (pp. 257–266). Erlbaum.

Hunter, J. E. (1983b). *Test validation for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB)*. U.S. Department of Labor.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, **29**, 340–362.

Kerr, S. (1975). On the folly of rewarding A, while hoping for B. *Academy of Management Journal*, **18**(4), 769–783.

MacLane, C. N., Cucina, J. M., Busciglio, H. H., & Su, C. (2020). Supervisory opportunity to observe moderates criterion-related validity estimates. *International Journal of Selection and Assessment*, **28**(1), 55–67. https://psycnet.apa.org/doi/10.1111/ijsa.12267

Martin, R., Guillaume, Y., Thomas, G., Lee, A., & Epitropaki, O. (2016). Leader-member exchange (LMX) and performance: A meta-analysis. *Personnel Psychology*, **69**(1), 67–121. https://doi-org.proxygw.wrlc.org/10.1111/peps.12100

McHenry, J. J., Hough, L. M., Toquam, J. L., Hanson, M. A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. *Personnel Psychology*, **43**(2), 335–354. https://doi.org/10.1111/j.1744-6570.1990.tb01562.x

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Sage.

O'Boyle, E. H.Jr, Humphrey, R. H., Pollack, J. M., Hawver, T. H., & Story, P. A. (2011). The relation between emotional intelligence and job performance: A meta-analysis. *Journal of Organizational Behavior*, **32**(5), 788–818. https://doi.org/10.1002/job.714

Paullin, C., Putka, D. J., Tsacoumis, S., & Colberg, M. (2010). Using a logic-based measurement approach to measure cognitive ability. In C. Paullin (chair), *Cognitive ability testing: Exploring new models, methods, and statistical techniques*. Symposium conducted at the 25th Annual SIOP Conference.

Ree, M. J., Earles, J. A., & Teachout, M. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, **79**, 518–524.

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. https://doi.org/10.1037/apl0000994

Salgado, J. F., & Moscoso, S. (2019). Meta-analysis of the validity of general mental ability for five performance criteria: Hunter and Hunter, 1984 revisited. *Frontiers in Psychology*, **10**, 476186.

Schmidt, F. L. (2013). *The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 100 years of research*. Invited presentation at the 2013 Fall Event of the Personnel Testing Council/ Metropolitan Washington.

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. https://doi.org/10.1037/0033-2909.124.2.262

Schmidt, F. L., Hunter, J. E., Outerbridge, A. N., & Trattner, M. H. (1986). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force: An empirically based demonstration. *Personnel Psychology*, **39**, 1–29.

U.S. General Accounting Office. (2003). *Equal employment opportunity: Hiring, promotion, and discipline processes at DEA* (report #GAO-03-413). Author.

van Rijn, P., & Payne, S. S. (1980). *Criterion related validity research base for the D.C. firefighter selection test*. Personnel Research and Development Center, U.S. Office of Personnel Management.

Cite this article: Cucina, J. M. & Hayes, T. L. (2024). E Pluribus Unum? Why criteria should be multimethod and multirater. *Industrial and Organizational Psychology* **17**, 314–318. https://doi.org/10.1017/iop.2024.14