Industrial and Organizational Psychology (2024), **17**, 288–291
doi:10.1017/iop.2024.17

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

External practitioner perspectives on validating selection tools against performance ratings

Chase A. Winterberg and Greg Haudek

Data Science, Hogan Assessments, Tulsa, OK, USA
Corresponding author: Chase A. Winterberg; Email: cwinterberg@hoganassessments.com

Foster et al. (2024) revisit a conversation that is well overdue for reconsideration in our field. We add to the conversation by considering how the focal authors' recommendations translate into practice, particularly for external consultants. In the wake of the debate sparked by Sackett et al.'s (2023) focal article that aired concerns with corrections for indirect range restriction via artifact distributions, many practitioners are confused and overwhelmed in attempting to keep up with the academic literature to provide the best selection advice and validation research for their clients. Foster et al. (2024) present critical considerations for moving our field past the confusion, especially within the selection context.

In this comment, we first consider implications for managing client expectations on validity effect sizes. Next, we discuss practical concerns involved with prediction of performance ratings instead of true performance. Then, we challenge the focus on the selection system without addressing flaws in the performance rating forms. We conclude with proposed practically oriented revisions to Foster et al. (2024)'s three-hurdle system.

Explaining the unexplained

The focal authors' insight into invisible ceilings on the variance in performance ratings that selection procedures can explain will help calibrate expectations of clients and even the research consultants themselves. We, thus, strongly encourage practitioners to make serious attempts to educate clients about what is and is not realistically possible with their selection systems based on Foster et al. (2024) We offer a few suggestions for informing clients and their legal counsel.[1]

When we present validation results to clients, we often start with discussion of the distributions of focal criteria, such as performance ratings and the implications of such for validity estimates. It would be fairly straightforward to walk them through the many sources of variance in performance ratings at this point as well. Because of the practical concerns we discuss in the remaining sections of this comment, it is unlikely to be able to present exact estimates of these variance components for the specific sample at hand. However, the practitioner could explain that academic research (e.g., Scullen et al., 2000) observes that we can only attribute about 20%–30% (Foster et al., 2024) of the information contained in task performance ratings to candidate differences in characteristics, for example. The practitioner could then present converted variance explained metrics from the amount the selection system explains out of total variance to amount explained out of possible variance, as the focal article demonstrates. The practitioner can indicate that, of the differences in performance ratings that we can link to any difference in candidates'

[1]Thanks to an anonymous reviewer for recommending we make these suggestions.

characteristics, scores on the current selection tool overlaps with the resulting percentage of differences in performance scores.

Courts and client legal counsel, on the other hand, tend to overemphasize statistical significance rather than effect size. Thus, we have more ground to cover in convincing the legal community. We first need to discuss the limitations of relying on statistical significance, such as the dependence on sample size, the binary decision-making approach, and the reverse logic of hypothesis testing. This might naturally lead to conversations of effect size and how such metrics can provide a more complete picture of validity, which should begin with calibrating their expectations on what is possible according to Foster et al. (2024)'s guidance.

Predicting ratings versus true performance in practice

Foster et al. (2024)'s recommendations focused on predicting job performance ratings are likely to spark important streams of research that will benefit practice. It would be helpful for future researchers to establish a taxonomy of the important sources of variance in performance ratings and their implications. Such a framework would guide practitioners in designing advanced validation work and managing expectations. Existing models distinguishing among the effects of ratee, rater, item, and their interactions may not be granular enough. For example, as Foster et al. (2024) acknowledged, ratee main effects can capture sources of variance beyond that due to true performance (e.g., criterion contamination). However, the need for external practitioners to develop and validate selection systems that account for performance *ratings* instead of *true performance* poses several challenges.

Fairness and legal concerns

Establishing links to true performance should be the priority but the bare minimum. This primary focus not only makes the most practical sense given the added cost in analyzing multiple sources of variance in ratings, as we discuss below, but it also supports the legal defensibility of selection systems. The purpose of validation from a legal perspective is to demonstrate job relevance (i.e., a relationship with *true performance*). Ratings include variance due to factors unrelated to job requirements. Along these lines, attempting to tailor selection processes to better align candidates to specific manager rating biases could get employers into trouble. Practitioners would need to separate illegal biases, such as those differentially impacting members of protected groups, from acceptable biases, such as a preference toward one performance dimension over another, before they attempt to align selection systems to ratings. Further, a job analysis needs to support any additional factors that influence employment decisions. In addition, holding candidates accountable for certain preferences, stereotypes, and biases held by their future coworkers and supervisors would raise fairness concerns. Although such perspectives will certainly impact their future performance ratings, it will do so in some ways that are beyond the future employee's control.

Once validity in predicting true performance has been established, organizations and practitioners, with the right resources, could further tweak selection systems, with caution, to account for important sources of variance in ratings. As another alternative, practitioners and employers could incorporate consideration of nuanced criteria associated with a candidate's specific position and future performance raters into onboarding and development efforts. Rather than screening the candidate out because they may not align with their supervisors' biases, the organization could provide guidance on how to manage the impending nuances associated with their position.

Practical constraints to implementation

Unfortunately, Foster et al. (2024)'s recommendations toward accounting for the multitude sources of variance in performance ratings are not feasible in most professional validation studies in practical contexts. We highlight a few challenges found consistently in applied settings next.

Situational constraints on validation studies

Foster et al. (2024) contend that "when we only have data representing our predictors and limited performance ratings provided by one rater, usually a manager, at one point in time. . . . , it might be better to not conduct a criterion-related validity study at all" (p. 279). This situation describes virtually every criterion-related validation study we conduct for our clients. Clients are reluctant to give up supervisor and incumbent work time to participate in research. Additionally, it is the exception to find multiple raters who can provide accurate ratings (i.e., having an adequate opportunity to observe performance) of an incumbent's true performance. In other words, it is extremely difficult to obtain meaningful data from multiple raters or from raters across multiple time points. Furthermore, a core objective of criterion-related validation research is to generalize to future applicant pools. Accounting for rater–ratee interactions in validation risks overfitting to current incumbent samples.

However, foregoing a validation study altogether is not advisable, especially if you work with selection instruments likely to cause adverse impact. As practitioners know, validation research in practice requires many compromises between best practices and practical constraints. Rather than giving up when these constraints threaten best practice, we do our best to give the client the best quality guidance within these constraints. Otherwise, many more organizations, than already do, would be using selection procedures, for which they have no validity information, potentially causing harm to candidates (e.g., unjustifiably preventing employment) and organizations (e.g., hiring individuals who cannot perform).

Fortunately, certain design considerations can help improve the quality of the limited data available in practice. Foster et al. (2024) underscore the utility of considering specific raters' preferences, such as to account for person–supervisor fit. We always begin criterion-related validation studies with a job analysis. Job analysis typically includes relevant supervisors and peers for a given role. Leveraging this information appropriately (i.e., avoiding incorporating counterproductive biases) could help incorporate future performance raters' idiosyncratic expectations. However, at a certain point, accounting for rater idiosyncrasies in selection decisions may not be worth the effort. Trying to align candidates to specific individuals in the organization could lead to us chasing moving targets. There is no guarantee that the supervisors and peers that exist while a selection system is being validated will remain in the long term after candidates have been aligned to them.

Organizations' performance data are a mess

In our experience, the performance ratings collected by organizations for administrative purposes are overwhelmingly problematic. Specifically, our clients' internal, administrative performance data, if they even have any available, are commonly obscured by organizational politics and hidden agendas, are negatively skewed with excessive leniency biases (i.e., from a majority of the raters) and are based on psychometrically and conceptually flawed performance rating forms. For this reason, in addition to prefacing with frame of reference and rater error training, we implement our own performance rating form developed specifically for validation research purposes. This likely further raises the ceiling on the variance we can explain in true performance because it minimizes certain rater biases. Thus, this design feature should help prioritize true performance variance.

Avoid perpetuating the criterion problem

We agree that organizations would benefit from the ability to predict performance ratings beyond ratee main effects. However, the poor quality of organizations' performance rating systems may need to be improved prior to validation. We also use job analyses to orient to the critical performance dimensions for the role. Depending on the relationship with the client, this could involve an evaluation of the current performance rating system. In the event major flaws exist in the organization's rating systems, we recommend addressing them before embarking on validation efforts, especially those aimed at tailoring to specific sources of variance. Although following Foster et al. (2024)'s recommendations in the research context will likely guide these improvements, using selection systems to accommodate the problems in performance ratings seems "bass ackwards," with the potential to perpetuate the criterion problem. We should consider whether there is more benefit in reshaping behaviors or systems in the organization to better align with more socially responsible employment decision processes, as opposed to reshaping employment decision processes to align with all of the organizations' biases and weaknesses. However, improving performance ratings will likely take a separate effort, beyond validation, focused on making changes in the current organization and its incumbent raters.

Practically oriented recommendations

In sum, we recommend the following practically oriented clarifications and revisions to Foster et al. (2024)'s three-hurdle system. Although implicit in their propositions, we explicitly argue the broad to narrow approach should characterize the focal criteria throughout the process. A preliminary stage should involve a job analysis with special attention to rater preferences and the quality of current performance ratings. For Stage 1, consideration of person–organization fit should account for true performance, to the extent possible, in terms of organization-wide competencies and performance dimensions. In Stage 2, consideration of person–job fit involves accounting for true performance in terms of job-specific standards. In practice, concerning Stage 3, many organizations may not have multiple positions from which candidates may choose. Rather than finding the best position for the candidate, as Foster et al. (2024) suggest, it may make more sense to find the best candidate for the specific open position, accounting for the specific individuals who may interact with the position holder. Because it does not make much sense to hire candidates before you know where to place them, we further depart from Foster et al. (2024) to recommend that practitioners complete Stage 3 prior to hiring decisions if it is included in selection. This stage should also be considered optional and could be incorporated into onboarding or development efforts instead. We further argue that consideration of adverse impact and how the selection system differentially impacts protected groups warrants its own stage, Stage 4.

References

Foster, J. F., Steel, P., Harms, P., O'Neil, 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.

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*, **16**, 283–300.

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

Cite this article: Winterberg, C. A. & Haudek, G. (2024). External practitioner perspectives on validating selection tools against performance ratings. *Industrial and Organizational Psychology* **17**, 288–291. https://doi.org/10.1017/iop.2024.17