Industrial and Organizational Psychology (2023), **16**, 349–352
doi:10.1017/iop.2023.32

CAMBRIDGE UNIVERSITY PRESS

COMMENTARY

Rearranging the deck chairs on the Titanic: What are practitioners to do?

Gerald V. Barrett and Dennis Doverspike

Barrett and Associates Inc, Silver Lake, OH, USA
Corresponding author: Gerald V. Barrett; Email: gbarrett@barrett-associates.com

We are responding to Sackett et al. (2023) from the perspective of practitioners who have specialized in high-stakes testing situations, particularly for safety forces, for the last 50 years. Sackett et al. used their data to analyze the results for a six-test battery, which included an integrity measure, conscientiousness, biodata, a structured interview, a situational judgment test (SJT), and a cognitive test. In this reply, our primary intent is not to dispute their findings or analysis but to suggest that the use of such a battery would not be feasible for practitioners, especially in safety force and high-stakes situations, regardless of the results of a regression analysis applied to the meta-analytic findings. Furthermore, we will argue that some of the offered estimates of validity are questionable within the context of safety force selection.

Problems with predictors

In entry-level safety force selection, predictors of future malfeasance, including polygraphs and the MMPI, are frequently administered, although such measures are rarely used in promotional testing. Although an integrity test might be feasible for entry-level safety force selection, the information obtained from an integrity assessment could be seen as redundant with that obtained from the polygraph and MMPI and, thus, not cost effective. In addition, although Sackett et al. (2022) estimated the validity of integrity tests to be .31, convincing evidence for the predictive validity of integrity tests in practical settings does not exist. The primary validation study did not report correlations with individual job performance, and the test's adverse impact was artificially reduced by race norming (Paajanen, 1988). This illustrates why there is a need to return to the original, primary studies when interpreting the results of a meta-analysis.

Conscientiousness and other personality tests have and could be used in entry-level safety force selection, although they are rarely used in promotional testing. However, self-report personality tests can be easily faked, and individuals can be taught how to obtain high scores in high-stakes testing situations (Miller & Barrett, 2008). The validity of personality measures also falls when moving from concurrent to predictive designs. Most critically, in order to use a conscientiousness test, the practitioner would need to conduct a local, criterion-related study, which introduces a range of practical and test security-related problems.

Biodata is a vague term that does not correspond to one specific construct. In high-stakes testing, local validation of the biodata measure would be crucial. It would take approximately 400 incumbents and 6 months to develop a keyed biodata instrument using a concurrent validation design. Thus, biodata instruments are not feasible for many high-stakes testing situations due to the time, money, and effort it would take to develop such a test, as well as the limited time span for which such a measure might remain valid.

© The Author(s), 2023. Published by Cambridge University Press on behalf of the Society for Industrial and Organizational Psychology.

https://doi.org/10.1017/iop.2023.32 Published online by Cambridge University Press

Some type of interview is often used in public safety and high-stakes testing, although usually at a later stage in the hiring process due to the cost and administrative hurdles encountered in administering interviews to large numbers of applicants. Structured interviews can measure a wide range of attributes from job knowledge to likeability (Arthur & Villado, 2008). Terms such as "assessment centers" and "interviews" are often used as if they refer to a single construct or fixed set of constructs, but that is not the case in practice. Nevertheless, structured interviews do have a place and are often used in high-stakes testing for both entry level and promotion.

SJTs have been used for over 60 years (Millard, 1952). Again, SJTs do not measure any one particular construct, and their usefulness depends on the construct being assessed (Arthur & Villado, 2008). As with biodata instruments, SJTs can be costly to develop and require local development and validation studies, and the use of SJTs does not necessarily increase the validity or decrease the adverse impact of a selection battery.

The criterion problem

Sackett et al., made an excellent point that "our widely used predictors, on average, exhibit lower operational criterion-related validity with regard to overall performance than has previously been reported," but "overall job performance" is an imprecise term. Mount and Barrick (1995) defined overall job performance as "overall ratings of performance, overall suitability, likelihood for promotion or termination, occupational level achieved, turnover or tenure, and productivity" (Mount & Barrick, 1995, p. 171), which covers a fairly large range of different possible criteria. When overall performance is so ill defined, it leads to imprecision in the validity estimates.

Real world combinations

Sackett et al., much like Schmidt and Hunter (1998), used a simulation to determine both *d*-scores and validity. We have never encountered a primary predictive validation study where a combination of predictors generated a multiple correlation approaching .71 for complex jobs. Real world studies with actual job applicants often result in very different validity coefficients and adverse impact ratios than do those obtained from simulated studies.

This is equally true of the results of actual test administrations following concurrent studies. For example, in a study with police, which investigated the potential for combining a personality test and a cognitive test, a concurrent validation design indicated that a custom designed fire personality test was valid and had no adverse impact. However, when administered to an applicant sample, both the cognitive test and personality test resulted in adverse impact, and this adverse impact increased when the two tests were combined (Barrett & Bishop, 2000).

Sackett et al. emphasized the simulated combination of cognitive ability and traits such as integrity. However, after 40 years, we cannot locate even one archival peer-reviewed predictive validation study that successfully combined a cognitive test with an integrity test to increase validity and reduce adverse impact. In contrast, a meta-analysis of the literature on the prediction of entry-level firefighter performance reported that a combination of cognitive and mechanical skills had relatively high validity compared to other combinations of predictors (Barrett et al., 1999).

Conclusion and suggestions for practitioners

Sackett et al. suggested 18 things a practitioner should do to estimate operational validity. This advice does not correspond with the practical realities of high-stakes selection. We are not reassured by the speculation that combining six predictors, including integrity and conscientiousness, would result in a "true" operational validity and reduced adverse impact.

https://doi.org/10.1017/iop.2023.32 Published online by Cambridge University Press

We do agree with Sackett et al., that custom designed selection systems are optimal if they are based on accurate validation studies. We also agree that scientific truth changes with new information. There is no question Schmidt and Hunter's "classic" Table 1 (Sackett et al., 2017) has been perceived as scientific fact that does not need to be critically examined. Sackett and colleagues have taken a necessary step toward examining this received doctrine (Barrett, 1972). However, rearranging the deck chairs by critiquing restriction of range corrections does not solve the fundamental problems.

Our short set of recommendations for practitioners:

1. If we are misinterpreting Sackett et al., we apologize, and we do not want to minimize the contribution that meta-analysis makes to scientific knowledge and progress, however, we read Sackett et al., as finding that validity estimates based on meta-analytic studies may have limited value in specific situations due to the substantial variance in the validity estimates, as well as the complex effects of restriction of range, and the associated difficulty in accurately estimating restriction of range.
2. Determining the best combination or set of predictors for high-stakes selection situations, including those involving safety forces, requires expert and seasoned judgment by practitioners who can analyze the practical constraints and demands and then use their knowledge of the literature to select the optimal set of assessments for the situation.
3. As noted by Sackett et al., the types of predictors that might be appropriate in entry-level selection are quite different from those typically used in promotional settings; again, this is particularly true with safety forces. Thus, validity estimates and simulation studies are an insufficient guide in identifying an appropriate set of assessment instruments.
4. High-stakes selection scenarios often involve difficult decisions regarding tradeoffs imposed by the need for securing exam materials and questions, which can then limit the choice of validation designs. In addition, the need to limit costs and development time, while producing tailor-made, locally developed tests, further constrains the available choices of both tests and validation designs.
5. As tests used in high-stakes hiring are frequently subject to challenge and litigation, it is critical that assessments be developed in a manner consistent with the *Uniform Guidelines* and that consideration be given to the critiques of plaintiffs' experts.

References

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, **93**(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435.

Barrett, G. V. (1972). Research models of the future for industrial and organizational psychology. *Personnel Psychology*, **25**(1), 1–17. https://doi.org/10.1111/j.1744-6570.1972.tb01086.x

Barrett, G. V., & Bishop, N. B. (2000). *Report on the firefighter selection process and on the development and validation of the Firefighter Personal Preference Inventory (Report prepared for the City of Columbus)*. Barrett & Associates, Inc.

Barrett, G. V., Polomsky, M. D., & McDaniel, M. A. (1999). Selection tests for firefighters: A comprehensive review and meta-analysis. *Journal of Business and Psychology*, **13**(4), 507–513. https://doi.org/https://doi.org/10.1023/A:1022966820186

Millard, K. A. (1952). Is how supervise? An intelligence test? *Journal of Applied Psychology*, **36**(4), 221–224. https://doi.org/10.1037/h0057153

Miller, C. E., & Barrett, G. V. (2008). The coachability and fakability of personality-based selection tests used for police selection. *Public Personnel Management*, **37**(3), 339–351. https://doi.org/10.1177/009102600803700306

Mount, M. K., & Barrick, M. R. (1995). The big five personality dimensions: Implications for research and practice in human resources management. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. **13**, pp. 153–200). JAI Press Inc.

https://doi.org/10.1017/iop.2023.32 Published online by Cambridge University Press

Paajanen, G. E., (1988). The prediction of counterproductive behavior by individual and organizational variables. (Doctoral dissertation, University of Minnesota). Retrieved from https://www.proquest.com/openview/f059fc693a238ec38e3acbebb94594a4/1?pq-origsite=gscholar&cbl=18750&diss=y

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**(10), 1421–1434. https://doi.org/10.1037/apl0000235

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**(11), 2040–2068. https://doi.org/10.1037/apl0000994

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), 262–274. https://doi.org/10.1037/0033-2909.124.2.262

Cite this article: Barrett, G. V. & Doverspike, D. (2023). Rearranging the deck chairs on the Titanic: What are practitioners to do? *Industrial and Organizational Psychology* **16**, 349–352. https://doi.org/10.1017/iop.2023.32

https://doi.org/10.1017/iop.2023.32 Published online by Cambridge University Press