Industrial and Organizational Psychology (2023), **16**, 332–335
doi:10.1017/iop.2023.30

COMMENTARY

Interpreting the magnitude of predictor effect sizes: It is time for more sensible benchmarks

Scott Highhouse[1] and Margaret E. Brooks[2]

[1]Department of Psychology, Bowling Green State University, Bowling Green, Ohio, USA and [2]Department of Management, Bowling Green State University, Bowling Green, Ohio, USA
Corresponding author: Scott Highhouse; Email: shighho@bgsu.edu

Sackett et al. (2021) published a disruptive piece that is summarized in Sackett et al. (2023) focal article. As the authors explain in the focal article, not only did their 2021 paper show that range-restriction overcorrection has led to inflated estimates of validity for selection devices, but that the new corrections actually alter the rank ordering of predictors established in Schmidt and Hunter (1998). Many are celebrating that structured interviews have supplanted general mental ability for the top validity spot. Others, however, were deflated by the generally shrunken effect sizes associated with the new corrections. According to Sackett et al. (2023), many practitioners feel that these revised estimates do not help our traditional predictors compete for success in the marketplace, leaving many wondering how to effectively communicate the relative efficacy of predictors to key stakeholders. We believe that many scientists and practitioners hold unrealistic standards of success. It is time, therefore, for I–O psychologists to adopt and communicate new benchmarks for evaluating predictor effect sizes.

New benchmarks effect sizes

Campbell (1990) argued that unrealistic standards of success cause a sense of hopelessness among I–O psychologists engaged in predicting future job performance of job applicants. This dismal view of predictor validity has led some of them to pursue holistic methods where prediction error is less transparent. Major culprits in this sense of hopelessness are Cohen's (1988) benchmarks for classifying correlations as small (.1), medium (.3), and large (.5). We would argue, however, that these benchmarks have set us up for failure when it comes to communicating magnitude.

So how do we get more realistic effect size benchmarks? We believe the answer is in a paper by Bosco et al. (2015) examining nearly 150,000 effect sizes reported in *Journal of Applied Psychology* and *Personnel Psychology* between the years 1980 and 2010. They examined the distribution of effect sizes for various relation types. We focused only on the nearly 8,000 effect sizes for attitude–behavior relations, as these effect sizes involving actual behavioral outcomes are more relevant to the performance-prediction context than those involving attitude–attitude or attitude–intention relations (Bosco et al., 2015, p. 436). The authors found that, with rudimentary meta-analytic corrections, one can classify small (25th-percentile) effect sizes as $r = 0.07$, medium (50th-percentile) effect sizes as $r = 0.16$, and large (75th-percentile) effect sizes as $r = 0.29$. Table 1 applies these new small, medium, and large effect-size categories to the current state-of-the-science predictor effect sizes reported in Sackett et al. (2023).

Table 1. Effect sizes for common predictors using the Bosco et al. (2015) benchmarks

	Effect size		
	Large $r > .28$	Medium $r > .15$	Small $r > .07$
Predictor	• Structured interview • Job knowledge test • Biodata (keyed empirically) • Assessment center • Work sample • Integrity test • Emotional intelligence (trait)	• SJT • Conscientiousness • Interests • GMA test • Emotional stability • Emotional intelligence (ability) • Biodata (keyed rationally) • Extraversion • Traditional interview • Agreeableness	• Openness • Job experience (years)

Note: Predictors are shown in order of effect size reported in Sackett et al. (2023). Effect sizes for the five-factor model traits are based on contextualized personality items. Situational judgment test (SJT), general mental ability (GMA).

As you can see, these benchmarks are considerably smaller than the Cohen ones and, as such, change our definition of which predictors have small, medium, and large effects. We point out some implications of using these revised effect size benchmarks for communicating predictor effect size.

Implications for communicating our value to stakeholders

Our predictors are powerful

Many of the predictors developed by I–O psychologists for employee selection are among the strongest behavioral predictors in all of applied psychology. Moreover, as Lievens (2013; Lievens et al., 2020) pointed out, the efficacy of selection procedures for predicting job success is often equal to or greater than the efficacy of things thought to be highly useful in medicine, including antihistamines for alleviating allergies and ibuprofen for reducing back pain.

Predictor rank can be deceiving

We believe that it is better to use the labels "small," "medium," and "large" from our Table 1 than it is to use meta-analytic correlations to compare predictors to one another. This is because, as Sackett et al., points out in the focal article, the standard deviations for many of the meta-analytic effects can be quite wide. In addition to this, Sackett et al. (2017) showed that directly comparing validities requires holding constant sample and criterion—something that is rarely done in meta-analyses.

Smaller does not equal worse

It is unwise to rule out predictors because they have "merely" medium or small effect sizes. Barrick and Mount (2000) argued that conscientiousness and emotional stability are necessary for almost every job. Although these may not be among the biggest predictors, the fact that they do not correlate with ability predictors make them especially useful for incremental prediction. Moreover, being categorized among small effect sizes means that a predictor significantly explains variance above chance levels. In contrast, things like handwriting analysis predict no better than chance (Rafaeli & Klimoski, 1983).

Perfect prediction should not be our standard

Perfect prediction in employee selection should not be the standard by which we judge our value, and it is not a relevant ceiling for use in communicating efficacy. We should focus communication on what value we add, not on how far we are from perfection. Consider including meaningful context, such as what the validity would be if one hired at random, as a baseline, and use a reasonable upper bound based on our current understanding of the validity ceiling.

We can do a better job communicating our value

Researchers have recently demonstrated that consumers of effect size information are more persuaded by correlations presented in the form of probabilities and/or frequencies (Brooks et al., 2014; Zhang et al., 2018). Graphical visual aids (e.g., icon arrays, expectancy charts) have also been shown to be useful for communicating efficacy (Garcia-Retamero, & Cokely, 2013; Zhang et al., 2018). If you want to present traditional validity coefficients, context can improve information evaluability of otherwise hard-to-evaluate relations (Childers et al., 2022; Zikmund-Fisher, 2019).

Concluding thoughts

As we have noted elsewhere (Highhouse & Brooks, 2017, 2023), effective hiring requires assessing what is foreseeable at the time of hire, recognizing that the ultimate outcome may be influenced by various things outside of the employer's control. This means that, considering all life, workplace, and random factors that may influence performance, the theoretical ceiling of predictive validity of pre-employment tests is necessarily limited. The view of selection as probabilistic and prone to error is rejected by some I–O psychologists who believe that near perfect prediction is theoretically possible (e.g., Hollenbeck, 2009; Silzer & Jeanneret, 2011). Einhorn (1986) wisely observed, however, that good prediction requires "accepting error to make less error" (p. 387). It is necessary, therefore, that I–O psychologists be unabashed in advocating for our predictors that are proven to be powerful forecasters of future performance. We should focus on competing with those who pedal inferior alternatives to our selection tools rather than competing with impossible standards of success.

References

Barrick, M. R., & Mount, M. K. (2000). Select on conscientiousness and emotional stability. In E. A. Locke (Ed.), *The Blackwell handbook of principles of organizational behavior* (pp. 15–28). Blackwell Publishing.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, **100**(2), 431.

Brooks, M. E., Dalal, D. K., & Nolan, K. P. (2014). Are common language effect sizes easier to understand than traditional effect sizes? *Journal of Applied Psychology*, **99**(2), 332.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organization psychology. In M. D. Dunnette & L. M. Hough (Eds,), *Handbook of industrial and organization psychology*. Consulting Psychologists Press.

Childers, M., Highhouse, S., & Brooks, M. E. (2022). Apples, oranges, and ironing boards: Comparative effect sizes influence lay impressions of test validity. *International Journal of Selection and Assessment*, **30**(2), 230–235.

Cohen, J. (1988). The effect size. In J. Cohen, *Statistical power analysis for the behavioral sciences* (pp. 77–83). Routledge.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment*, **50**(3), 387–395.

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, **22**(5), 392–399.

Highhouse, S., & Brooks, M.E. (2017). Straight talk about selecting for upper-management. In D. G., Collings, K., Mellahi, & W. F., Cascio (Eds.) *The Oxford handbook of talent management* (pp. 268–280). Oxford.

Highhouse, S., & Brooks, M. E. (2023). Improving workplace judgments by reducing noise: Lessons learned from a century of selection research. *Annual Review of Organizational Psychology and Organizational Behavior*, **10**, 519–533.

Hollenbeck, G. P. (2009). Executive selection—What's right . . . and what's wrong. *Industrial and Organizational Psychology*, **2**(2), 130–143.

https://doi.org/10.1017/iop.2023.30 Published online by Cambridge University Press

Lievens, F. (2013, May). Off the beaten path! Towards a paradigm shift in personnel selection research. Invited keynote presented at the EAWOP Conference, Munster, Germany.

Lievens, F., Sackett, P. R., & Zhang, C. (2020). Personnel selection: A longstanding story of impact at the individual, firm, and societal level. *European Journal of Work and Organizational Psychology*, **30**, 1–12.

Rafaeli, A., & Klimoski, R. J. (1983). Predicting sales success through handwriting analysis: An evaluation of the effects of training and handwriting sample content. *Journal of Applied Psychology*, **68**(2), 212.

Sackett, P. R., Shewach, O. R., & Keiser, H. N. (2017). Assessment centers versus cognitive ability tests: Challenging the conventional wisdom on criterion-related validity. *Journal of Applied Psychology*, **102**(10), 1435.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). 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.

Silzer, R., & Jeanneret, R. (2011). Individual psychological assessment: A practice and science in search of common ground. *Industrial and Organizational Psychology*, **4**(3), 270–296.

Zhang, D. C., Highhouse, S., Brooks, M. E., & Zhang, Y. (2018). Communicating the validity of structured job interviews with graphical visual aids. *International Journal of Selection and Assessment*, **26**(2-4), 93–108.

Zikmund-Fisher, B. J. (2019). Helping people know whether measurements have good or bad implications: increasing the evaluability of health and science data communications. *Policy Insights from the Behavioral and Brain Sciences*, **6**(1), 29–37.

Cite this article: Highhouse, S. & Brooks, M. E. (2023). Interpreting the magnitude of predictor effect sizes: It is time for more sensible benchmarks. *Industrial and Organizational Psychology* **16**, 332–335. https://doi.org/10.1017/iop.2023.30