COMMENTARY

Hocus-pocus and hydraulics functions: Anything not worth doing is not worth doing well

Jeremy L. Schoen

College of Business, Tennessee Technological University, Cookeville, TN, USA
Email: jeremy.schoen@gmail.com

It was nice to see a step back from the inflated validities (e.g., Sackett, Zhang, Berry, & Lievens, 2022) promoted by many meta-analysts. Still, the use of corrected validities for purposes of selection is a dubious practice. Although the mistake of correcting for range restriction of unrestricted samples is now apparent, other problems—both technical and legal—still abound. I briefly review works (most of which are more than 40 years old) that describe these challenges. I then provide suggestions for practitioners and researchers. Ultimately, the quote from Robert Fulghum seems appropriate: "Anything not worth doing is worth not doing well."

Legal concerns

The focal article specifically mentions the idea of updating thoughts and practices currently used by selection experts based on their earlier findings (Sackett et al., 2022). The premise of this argument is problematic. The very simple answer is that selection experts should not rely on corrected validities when defending choices about their selection tools.

The solo use of meta-analytic results for purposes of selection is inadvisable. This view is based on concerns that result from outcomes of employment discrimination cases. The solo use of meta-analytic evidence has never won in the Supreme Court of the United States. Instead, the court wants to know that selection tools are valid and free of bias (Biddle, 2010; Landy, 2003; Outtz, 2011) in your context (local validity). The courts appear to view corrections as "hydraulics functions," a bit of hocus-pocus where a meta-analyst can apply various corrections to achieve a validity they desire (Seymour, 1988) regardless of the observed relationships.

Corrected validity represents the maximum *theoretical relationship* that might be observed in a world without error. Corrections shift the relationship studied from the real world to a domain without measurement error or range restriction. Practically, corrected validities are unusable as the real world contains measurement error and range restriction. Operational validity, which applies corrections asymmetrically and mixes the real world with the theoretical realm, is also problematic. Or, as one colleague states, "there is nothing operational about 'operational validity'" (DeSimone, 2015, p. 40). It is not clear what operational validity theoretically represents (LeBreton, Schoen & James, 2017) and should be avoided in practice.

Technical concerns

Artifact quality and estimation

Before applying corrections, it is first important to consider whether or not the artifacts used when making corrections are, indeed, good estimates. To start, what is the unrestricted population? As noted, (Sackett et al., 2022, 2023) it is difficult to find *any* estimate of the variance of

unrestricted populations. But we must first define what we mean by the "applicant population." Is this population everyone who received our recruitment message? Does it include those who were qualified but may have self-selected out and never applied? All people who applied? Qualified applicants? Only those given serious consideration? The variance of each population results in different estimates for the unrestricted variance and subsequently corrected validity. Use of different estimates in a catch-as-catch-can way compares apples to oranges.

Second, it is common for meta-analysts to report corrected validity where measurement error is corrected in both the predictor and criterion. Typically, predictor measurement error is estimated via coefficient alpha. Alpha is the lower bound of internal consistency reliability and is the actual reliability only when items are essentially tau equivalent (a rarely tested assumption). Thus, alpha typically underestimates (Sijtsma, 2009) measure reliability.

Third, ICC estimates of reliability used to correct supervisor ratings of performance as the dependent variable in many meta-analyses, are greatly affected by between rater (supervisor) variation. As an example, it is possible to have an ICC of 0 when there is 100% agreement (no variation) of raters' scores. Variance in job performance ratings is restricted because those who performed poorly were fired or they quit. Even the .6 estimate is downwardly biased because it is developed from ratings of performance of remaining employees. Thus, ICCs should be corrected for range restriction before they are used to correct validities (LeBreton et al., 2017). The three misestimates described in this section all result in overcorrection of validity coefficients.

Statistical assumptions

The application of corrections in meta-analysis relies on a large number of statistical assumptions as well. When validity generalization was first presented, Schmidt and Hunter argued that most of the assumptions could be ignored. Many of these assumptions remain untested, yet, when tested, we find that they do not hold.

First, and at a generic level (for space considerations) meta-analysts must assume that errors are independent, artifacts are independent, and errors and artifacts are independent. These assumptions must hold both within each study and between studies included in the meta-analysis. If these assumptions are false then corrections may "double dip"—correct for the same artifact more than once—and *overcorrect* (James, Demaree, & Mulaik, 1986; James, Demaree, Mulaik, & Ladd, 1992). The assumption that within study artifacts are independent is one of the few assumptions tested and has been found to be false. Some artifact correlations are quite large (Köhler et al., 2015; Yuan et al., 2020).

Second, assumptions about distributions of the sample data used to compute various summary statistics must also be met. For space concerns, I cover just one example. To apply the range restrictions correction, one must assume homoscedasticity of error around the regression line. If the bivariate distribution is football shaped (like most in selection contexts) then we can expect heteroscedasticity. Based on the selection ratio and degree of heteroscedasticity, the use of the correction for range restriction can both under- and overcorrect (Novick & Thayer, 1969). Information about homoscedasticity and selection ratios are rarely available.

Recommendations

A first recommendation for both research and practice is: do not rely on corrected validity. The technical issues noted above, on average, result in overcorrection meaning that corrected validities are overestimates of the theoretical relationship. Aside from legal challenges, difficulties attaining the needed information (often item- level data) to know if we have met various assumptions that would allow us to apply corrections means that it is unlikely that we will ever be able to do so or have faith in the outcome of that process.

Frustratingly, and in contrast to the apparent surprise of some, all of this (and more) has been known for 40 years or more. The use of corrections in applied settings was long seen as objectionable (McNemar, 1962; Novick & Thayer, 1969; Womer, 1968). Yet, Industrial Organizational Psychology, Organizational Behavior, and Human Resources (IO/OB/HR) either did not know about these objections or turned a blind eye to them. Importantly, outside of IO/OB/HR, other fields using meta-analysis do not use psychometric corrections. Those fields typically have deeper training in statistics (or employ statisticians) and understand that various assumptions are not met in practice. Those other fields made the sound decision that it was better to be slightly conservative in estimation rather than engage in objectionable practice.

Notably, researchers in IO/OB/HR did sound the alarm bells roughly 40 years ago on the general notion of validity generalization and the use of corrections as part of that process (c.f. Algera et al., 1984; James et al., 1986, 1992; Kemery et al., 1987). So why have IO/OB/HR researchers continued to so heavily rely on meta-analysis and psychometric corrections? One answer is that meta-analysts told IO/OB/HR what we wanted to hear; research findings are easier to understand than hoped and validities are large. Validity generalization promised to be the salve that healed our wounds. But it is false promise.

As a second recommendation, targeted primarily to researchers, instead of spending so much time and journal space with meta-analyses and discussion of corrections, maybe it is time to turn our attention to improving measurement. As a field, we cannot be satisfied such that 40% of performance measures and often 20% or more of our predictor measures are error. Three-item homemade or modified self-report scales have likely gone about as far as they can go in our work. A bonus is that improved measurement is likely to result in improved observed validity. This then requires psychometrics classes in business PhD programs and journal reviewers' and editors' willingness to reject manuscripts with shortened and unvalidated scales.

A third recommendation, also targeting researchers, as a domain we need to increase our focus on study quality. Although suggestions for the use and interpretation of meta-analysis are great (see DeSimone et al., 2019 for a broad list for both practitioners and researchers), our field is not going to progress with better meta-analyses. The field of medicine, which also suffered from an overreliance on the use of meta-analysis, has returned to their gold standard of double-blind clinically controlled trials. But this complaint is also more than 40 years old as Eysenck (1978) lamented the lack of concern for study quality when conducting meta-analysis referring to the process of meta-analysis as "mega-silliness."

Regardless of arguments to the contrary (c.f. Schmidt & Hunter, 2015), averaging masses of poorly conducted studies is unlikely to improve any of them. Instead, bad studies contaminate the good ones. Although double-blind clinically controlled trials seem out of reach for IO/OB/HR scholarship we can certainly increase our methodological sophistication by engaging in better study design. Ultimately, the field needs to shift from a focus on solo authored and theoretically novel research to work that is published by teams (possibly large teams who may have some authors whose only contribution is access to data so that we can conduct field experiments) and work that is practically meaningful. This demands a major change in our reward structures so that those supporting these changes are able to attain tenure under these new expectations.

Conclusion

Meta-analysis can be a useful tool when used appropriately (DeSimone et al., 2019). As Larry James, frequent critic of meta-analysis said many times before his unexpected passing, "being against meta-analysis is like being against the letter B in the alphabet" (a useful tool). But our field needs to consider where we are going if we wish to remain relevant. Fooling ourselves with corrections does not seem like the right direction. Or, as Larry James also warned, "no meta-analyst will ever win a Nobel Prize." I honestly feel sad in that I fully expect our field to continue to

reward theory over practice, mass produced meta-analyses over better measurement and study design, and a continued belief in the unbelievable results from corrections.

References

Algera, J. A., Jansen, P. G. W., Roe, R. A., & Vijn, P. (1984). Validity generalization: Some critical remarks on the Schmidt-Hunter procedure. *Journal of Occupational Psychology*, **57**, 197–210.

Biddle, D. A. (2010). Should employers rely on local validation studies or validity generalization (VG) to support the use of employment tests in Title VII situations? *Public Personnel Management*, **39**, 307–326.

DeSimone, J. A. (2014). When it's incorrect to correct: A brief history and cautionary note. *Industrial and Organizational Psychology*, **7**, 527–531.

DeSimone, J. A., Köhler, T., & Schoen, J. L. (2019). If it were only that easy: The use of meta-analytic research by organizational scholars. *Organizational Research Methods*, **22**, 867–891.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, **33**(5), 517.

James, L. R., Demaree, R. G., & Mulaik, S. A. (1986). A note on validity generalization procedures. *Journal of Applied Psychology*, **71**, 440–450.

James, L. R., Demaree, R. G., Mulaik, S. A., & Ladd, R. T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology*, **77**, 3–14.

Kemery, E. R., Mossholder, K. W., & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology*, **72**, 30–37.

Köhler, T., Cortina, J. M., Kurtessis, J. N., & Gölz, M. (2015). Are we correcting correctly? Interdependence of reliabilities in meta-analysis. *Organizational Research Methods*, **18**, 355–428.

Landy, F. J. (2003). Validity generalization: Then and now. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 155–195). Lawrence Erlbaum Associates.

LeBreton, J. M., Schoen, J. L., & James, L. R. (2017). Situational specificity, validity generalization, and the future of psychometric meta-analysis. In J. L. Farr & N. T. Tippins (Eds.), *Handbook of employee selection* (2nd ed., pp. 93–114). Routledge.

McNemar, Q. (1962). *Psychological statistics* (3rd ed.). John Wiley and Sons

Novick, M. R., & Thayer, D. T. (1969). *An investigation of the accuracy of the Pearson selection formulas*. Educational Testing Service.

Outtz, J. L. (2011). Abolishing the uniform guidelines: Be careful what you wish for. *Industrial and Organizational Psychology*, **4**, 526–533.

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.

Sackeett, P. R., Zhang, C., Berry, C. M., & Lievins, 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**(3), 283–300.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed). Sage.

Seymour, R. T. (1988). Why plaintiffs' counsel challenge tests, and how they can successfully challenge the theory of "validity generalization". *Journal of Vocational Behavior*, **33**, 331–364.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, **74**, 107–120.

Womer, F. B. (1968). *Basic concepts in testing*. Houghton Mifflin Company.

Yuan, Z., Morgeson F. P., & LeBreton, J. M. (2020). Maybe not so independent after all: The possibility, prevalence, and consequences of violating the independence assumptions in psychometric meta-analysis. *Personnel Psychology*, **73**, 491–516.