*Industrial and Organizational Psychology*, **4** (2011), 554–557.


# About Babies and Bathwater: Retaining Core Principles of the *Uniform Guidelines*

ALAN D. MEAD AND SCOTT B. MORRIS
*Illinois Institute of Technology*

McDaniel, Kepes, and Banks (2011) make a compelling case that the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) are not in line with current scientific and professional practice. However, we argue that the fundamental principles embodied in the *Uniform Guidelines* are still relevant. We discuss the importance of assessing adverse impact and job relatedness in the context of current practice.

## Misconceptions About Adverse Impact Analysis

We disagree with the focal article authors when they assert that, ''an implicit assumption of the *Uniform Guidelines* is that adverse impact is an indication of a flawed test.'' The *Uniform Guidelines* simply stated that a ''procedure having adverse impact constitutes discrimination unless justified,'' and subsequent Equal Employment Opportunity Commission (EEOC) guidance clarified that the 4/5ths rule, ''is not intended as a legal definition, but is a practical means of keeping the attention of the enforcement agencies on serious discrepancies in rates of hiring, promotion and other selection decisions'' (Equal Employment Opportunity Commission, Office of Personnel Administration, Department of Justice, Department of Labor, & Department of Treasury, 1979, Q&A #11). Thus, adverse impact analysis can be viewed as nothing more than a way to identify situations that merit further attention.

The Civil Rights Act requires ''removal of artificial, arbitrary, and unnecessary barriers to employment'' (*Griggs v. Duke Power*, 1971, p. 431). This is far from a claim that tests with adverse impact are flawed; it is a more limited requirement that when an employment practice has a negative impact on a disadvantaged group, the employer bears responsibility to ensure that the practices are indeed necessary to the operation of business.

The authors of the focal article note that almost all selection methods show some degree of adverse impact and argue that, ''given the pervasiveness of adverse impact, the presence of adverse impact should not result in federal interference in employment practices when such interference is based on regulations inconsistent with scientific knowledge.'' Although there is no question that validation requirements should be consistent with scientific knowledge, we suggest that it is reasonable for federal regulators to give greater scrutiny to selection systems that produce larger amounts of adverse impact.

Correspondence concerning this article should be addressed to Alan D. Mead.
E-mail: mead@iit.edu
Address: College of Psychology, Illinois Institute of Technology, 3105 S. Dearborn, Chicago, IL 60616



https://doi.org/10.1111/j.1754-9434.2011.01391.x Published online by Cambridge University Press

The focal article barely touches on a change that we see as most needed in the *Uniform Guidelines*: more thoughtful guidelines on discerning adverse impact. The emphasis in the *Uniform Guidelines* on the 4/5ths rule for assessing adverse impact leads to several problems. First, this approach promotes dichotomous thinking about adverse impact. If the impact ratio is above .8, the test is fine, and no further attention to subgroup differences is needed. If the impact ratio falls below this value, regardless of whether it is .7 or .2, then there is a substantial disparity, and validity evidence must be provided. Although cutoff scores are to some extent unavoidable as decision aids, a consequence is that small differences near the cutoff are exaggerated (i.e., .85 is much better than .75) and large differences in other regions of the scale are largely ignored (i.e., an impact ratio of .7 is not much better than .2).

Our recommendation is that greater attention should be given to the impact ratio as a continuous measure of effect size, rather than strict reliance on the 4/5ths rule of thumb. This will minimize the artificial categorization of adverse impact as present or absent and will allow recognition of efforts to reduce adverse impact, even when they do not fully eliminate group differences. This approach will only be possible if practitioners and federal regulators know how to interpret effect size measures. Thus, guidelines are needed on the magnitude of the adverse impact ratio. This might be in terms of ranges representing varying levels of adverse impact, such as 1.0 to .8 for trivial, .6 to .8 for small, .6 to .4 for moderate, and less than .4 for large adverse impact. Reaching consensus on the exact range for small, medium, and large categories is likely to prove difficult but not impossible—such values have been adopted for effect size measures in other contexts (e.g., Cohen, 1988).

Another area where guidance is needed is how to evaluate both practical and statistical significance. The 4/5ths rule is a test of practical significance and ignores the impact of sampling error and small sample size on the statistical results. Consequently, results can differ substantially across settings, simply due to the sample of individuals who showed up for each test administration. These differences can be quite large when sample sizes are small (Morris & Lobsenz, 2000; Roth, Bobko, & Switzer, 2006).

An alternative is to test for statistical significance, an approach increasingly used by federal enforcement agencies (Cohen & Dunleavy, 2009) and the courts (Esson & Hauenstein, 2006). However, relying solely on statistical significance tests is also problematic, due to the sensitivity of these tests to sample size. When applied to large-scale selection programs, with thousands of applicants, statistical significance is almost guaranteed, even when group differences are trivial. In contrast, when sample size is small, significance tests have low statistical power and will often fail to detect true cases of adverse impact (Collins & Morris, 2008). Unfortunately, naive users can mistake the test statistic or *p*-value as an indicator of the magnitude of effect when in fact it largely reflects sample size. This is of particular concern in the adverse impact context, where significance levels are often discussed in terms of standard deviation units (Dunleavy & Gutman, 2011).

Ideally, both practical and statistical significance should be considered together. One way to achieve this is to construct a confidence interval around the adverse impact ratio, which would convey both the magnitude and precision of the statistical evidence (Morris & Lobsenz, 2000).

## Replacing Local Validation With Meta-Analytic Validity Evidence or Alternatives

The focal article authors do not acknowledge important caveats when it comes to replacing local validation with meta-analytic evidence, nor do they recognize alternatives to validity generalization (VG). We concur with McDaniel and his colleagues that VG is a type of evidence that

https://doi.org/10.1111/j.1754-9434.2011.01391.x Published online by Cambridge University Press

should be given due weight, but we should not lose sight of *job relatedness*, a fundamental principle of the *Uniform Guidelines*. Both psychologists (McDaniel, 2007; Rothstein, 2003; Sackett, 2003) and courts (Biddle, 2010) have raised job-relatedness concerns about VG because VG analyses lack direct links between the primary studies and the local job.

Job-relatedness has many dimensions when applied to VG meta-analyses (McDaniel, 2007; Sackett, 2003): Can the jobs in the meta-analysis be shown to be similar to the local job? Are the predictor and criterion measures included in the meta-analysis similar to the predictor(s) and criterion used locally? In addition, a validity generalization analysis replacing local validation must avoid judgment calls that would be difficult to defend locally.

Further, use of VG results may present complications in practice. It may be difficult to defend the representativeness of the sample of validity coefficients used in the VG study, particularly for a VG study done by others (see McDaniel, Rothstein, & Whetzel, 2006). Concerns about judgment calls and sampling bias are not trivial issues (see Wanous, Sullivan, & Malinak, 1989). For example, VG studies of work samples disagreed about which studies to include and reached substantially different results (Hunter & Hunter, 1984; Roth, Bobko, & McFarland, 2005). Similarly, the National Academy of Sciences endorsed the GATB VG system but disagreed about artifact corrections to such a degree that they presented their own, substantially different VG analysis (Hartigan & Wigdor, 1989). Finally, application of VG findings may also need to address the handling of predictor reliability corrections (see Hunter & Schmidt, 2004, pp. 158–159).

Two alternative approaches may offer stronger evidence of job-relatedness than VG. The Bayesian approach originally proposed by Schmidt and Hunter (1977), (see also Brannick, 2001; Brannick & Hall, 2003) elegantly combines information from prior research with local validation evidence to produce estimates that are more accurate than the local validation results alone, applicable even with small local samples, and yet clearly related to the current job.

Synthetic validation was introduced by Lawshe (1952) as a solution to the concerns laid out in the focal article regarding small employers and jobs at large employers that have small incumbent populations (see Johnson et al., 2010). The foundation of the synthetic validation system is a job analysis, so synthetic validity estimates are clearly related to the local job. However, synthetic validation also has important similarities to VG. For example, synthetic validation also rejects the presence of undetectable ''situational'' moderators, instead assuming that job analysis can detect and quantify levels of moderating variables. In addition, synthetic validation assumes that relationships between job elements and predictor tests generalize across jobs.

Both Bayesian analysis and synthetic validation could be viewed as ''more work'' than simply justifying a predictor based on VG research. In return, these methods offer *increased evidence of job relatedness*, and we note that although the *Atlas* court (*EEOC v. Atlas Paper*, 1989) viewed VG unfavorably, two judgments have been favorable toward synthetic validation (*McCoy v. Willamette Industries*, 2001; *Taylor v. James River Corporation*, 1989) because synthetic validation is so tied to job-analysis.

## Conclusion

Current scientific and professional practice in industrial and organizational (I–O) psychology employs methods that are considerably more complex than those described in the *Uniform Guidelines*. Although these methods require a higher level of methodological sophistication to use and interpret appropriately, they will ultimately provide more useful information about the fundamental issues involved in justifying employee selection systems.

## References

Biddle, D. (2010). Should employers rely on local validation studies or validity generalization (VG)

https://doi.org/10.1111/j.1754-9434.2011.01391.x Published online by Cambridge University Press

to support the use of employment tests in Title VII situations? *Public Personnel Management, 39*, 307–326.

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86*, 468–480.

Brannick, M. T., & Hall, S. M. (2003). Validity generalization from a Bayesian perspective. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 339–364). Mahwah, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, D., & Dunleavy, E. M. (2009). *A review of OFCCP enforcement statistics: A call for transparency in OFCCP reporting*. Washington, DC: The Center for Corporate Equality.

Collins, M. W., & Morris, S. B. (2008). Testing for adverse impact when sample size is small. *Journal of Applied Psychology, 93*, 463–471.

Dunleavy, E. M., & Gutman, A. (2011). On the legal front: An update on the statistical versus practical significance debate: A review of *Stagi v Amtrak* (2010). *The Industrial-Organizational Psychologist, 48*. Retrieved from http://www.siop.org/tip/april11/17gutman.aspx.

*EEOC v. Atlas Paper*. (1989). 868 F.2d. 487, 6th Cir., cert. denied, 58 U.S. L.W. 3213.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290–39315.

Equal Employment Opportunity Commission, Office of Personnel Administration, Department of Justice, Department of Labor, & Department of Treasury. (1979). Adoption of questions and answers to clarify and provide a common interpretation of the uniform guidelines on employee selection procedures. *Federal Register, 44*, 11996–12009.

Esson, P. L., & Hauenstein, N. M. (2006, May). *Exploring the use of the four-fifths rule and significance tests in adverse impact court case rulings*. Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

*Griggs v. Duke Power Co*. (1971). 401 U.S. 424.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Hunter J. E., & Hunter, R. F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Johnson, J. W., Steel, P., Scherbaum, C. A., Hoffman, C. C., Jeanneret, P. R., & Foster, J. (2010). Validation is like motor oil: Synthetic is better. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 305–328.

Lawshe, C. H. (1952). Employee selection. *Personnel Psychology, 6*, 31–34.

*McCoy et al. v. Willamette Industries, Inc*. (2001). U.S. District Court for the Southern District of Georgia, Savannah Division, Civil Action No. CV401-075.

McDaniel, M. A. (2007). Validity generalization as a test validation approach. In S. M. McPhail (Ed.), *Alternative validation strategies: Developing new and leveraging existing validity evidence* (pp. 159–180). New York, NY: Wiley.

McDaniel, M. A., Kepes, S., & Banks, G. C. (2011). The *Uniform Guidelines* are a detriment to the field of personnel selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 4*, 494–514.

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology, 59*, 927–953.

Morris, S. B., & Lobsenz, R. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology, 53*, 89–111.

Roth, P., Bobko, P., & McFarland, L. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology, 58*, 1009–1037.

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology, 91*, 507–522.

Rothstein, H. R. (2003). Progress is our most important product: Contributions of validity generalization and meta-analysis to the development and communication of knowledge in I/O psychology. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 115–154). Mahwah, NJ: Erlbaum.

Sackett, P. R. (2003). The status of validity generalization research: Key issues in drawing inferences from cumulative research findings. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 91–114). Mahwah, NJ: Erlbaum.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

*Taylor v. James River Corporation*. (1989). CA 88-0818-T-C (TC) (S.D. AL).

Wanous, J. P., Sullivan, S. E., & Malinak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74*, 259–264.

https://doi.org/10.1111/j.1754-9434.2011.01391.x Published online by Cambridge University Press