Industrial and Organizational Psychology, **4** (2011), 558–561.
Copyright © 2011 Society for Industrial and Organizational Psychology. 1754-9426/11

Adverse Impact Is Far More Complicated Than the *Uniform Guidelines* Indicate

RICK JACOBS AND PAIGE J. DECKERT
The Pennsylvania State University

JAY SILVA
EB Jacobs, LLC

McDaniel, Kepes, and Banks (2011) pointed out many critical flaws that were either contained in the *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice, 1978) when they were promulgated or that have emerged as problematic over more than 30 years since they have been part of the legal landscape. Much of their criticism focused on the assumed situational specificity hypotheses, the lack of recognition for validity generalization, and the *Uniform Guidelines* adherence to the outdated notion of a tripartite validity model. Addressed in a more limited fashion is the issue of adverse impact: the underlying definition, the prevalence in many selection situations, and the methods by which it is measured. Our goal is to add to the dialogue started by McDaniel et al., by pointing out three issues that have been front and center in the controversy surrounding adverse impact and the need to update the *Uniform Guidelines* to better address each.

Units of Standard Deviation Versus More Traditional Difference Measures

In many court cases, evidence of adverse impact is presented in what has come to be accepted as ‘‘standard deviation’’ analysis. This evidence has produced results that appear to strongly support a conclusion of unfair discrimination. For example, in *Vulcan Society v. NYFD*, White examinees ($N = 13{,}495$) passed a test at a 97.2% rate, whereas Black examinees ($N = 1{,}190$) passed it at an 85.4% rate. This was described as a difference of over 21 standard deviations. To put these standard deviation figures in context, the difference between the highest and lowest scores in a normal distribution is approximately six standard deviations and, even with outliers, would virtually never reach the level of 10 or 12, making differences of 20 or more standard deviations appear like statistical nonsense.

How can a difference in pass rates that does not even violate the 80% rule (e.g., in *Vulcan*, the relative passing rates for one test were 85.4 and 97.2 for Black and White candidates, respectively, yielding an adverse impact ratio of .87), produce a *21 standard deviation* difference? The simple answer is that ‘‘standard deviation’’ analysis does not express the difference between

Correspondence concerning this article should be addressed to Rick Jacobs.
E-mail: rrj@psu.edu
Address: Department of Psychology, Penn State University, 115 Moore Building, University Park, PA 16802

https://doi.org/10.1111/j.1754-9434.2011.01392.x Published online by Cambridge University Press

group means or selection rates in terms of standard deviations. Rather, ''standard deviation'' analysis (i.e., significance tests) expresses the observed group mean difference in terms of a *standard error.* The use of standard errors ignores two important facts; first, a standard error applies to an imaginary sampling distribution that is only relevant when the null is true. Next, and more directly to our point, a standard error is highly impacted by large sample size, and it is the size of the samples that drives up the difference in ''standard deviation'' units. The group difference in true standard deviation units in these same situations typically does not exceed two standard deviations in a worst case scenario.

The real problem is that experts often confuse judges and other individuals in a position to decide upon the evidence. These professionals are typically not as familiar with statistics as they are with the law. By referring to ''standard errors'' as ''units of standard deviation,'' confusion is created between a statistical test of a relatively meaningless hypothesis (that the difference between the groups is exactly 0; more on that later) and our more commonly encountered measures of group differences found in the literature and upon which much is known. There are several well-known, widely accepted, and easily interpreted measures of the size of the difference between two groups that are truly based on standard deviations rather than standard errors, most notably the *d* statistic. The *d* statistic describes the difference between groups in terms of actual standard *deviation* units, either the pooled standard deviation of the two groups (Cohen *d*) or, in studies comparing treatment and control groups, the standard deviation of the control group (Glass *d*). Unlike statistical significance tests and the associated ''standard deviation analysis,'' *d* actually expresses the difference between groups in terms of the number of standard deviations that separate their average scores. For example, White–Black differences in average scores on cognitive ability tests are typically described as large ($d = .80$ to 1.00), whereas White–Black differences in measures of job performance are typically described as small to medium in size ($d = .20–.30$; Roth, Huffcutt, & Bobko, 2003). Unlike measures of statistical significance, the value of *d* is generally not affected by sample size. We *must* stop confusing others with analyses that sound alike but are very different, and we must focus on analyses that actually tell us something about the magnitude of the difference.

Most vexing is the fact that the same test, used in two different situations, can show adverse impact in one setting but not in the other. If we were to use the same test battery to assist in selecting police officers in both New York City and Scranton, PA and that test battery showed the same difference between the majority and minority group across cities, the fact that there might be over 10,000 candidates in New York and only 300 in Scranton could easily result in the conclusion that the test is problematic in New York but perfectly acceptable in Scranton. This makes no sense, the test is exactly the same in both locations.

Hypothesis Testing and the Simplistic Idea of No Difference

It has been suggested by many scholars that hypothesis testing be eliminated. Although we are not adopting that radical of a position, we want to advocate the use of research findings in the areas of cognitive ability testing, physical ability testing, and personality measures in employment selection. Data from cognitive and physical ability testing clearly support anticipated differences in terms of race and gender, respectively. Drawing on meta-analytical results, McDaniel and colleagues make the argument that most selection procedures (excluding personality) are likely to show racial differences. One only has to look at the Cooper Standards for physical abilities to conclude that the performance expectations for men and women on the same physical events are very different. To assume that we suspend our many decades of research and expect no differences on tests that have a long history of differences

https://doi.org/10.1111/j.1754-9434.2011.01392.x Published online by Cambridge University Press

is to deny reality. Yet that is what often passes as evidence of an unfair test in a court case. A test is shown to be problematic because it yields a difference greater than 0 that reaches statistical significance.

The role of sample size aside, the approach is problematic because it fails to recognize that a test that may actually help improve the selection of underrepresented groups by resulting in group differences that are smaller than expected will be held up as unfair. Rarely do we make large leaps in science. Often times we talk about standing on the shoulders of those who came before us. If tests that actually reduce group differences are rejected because they do not eliminate group differences completely, then we are imposing a requirement that will prevent testing from becoming increasingly better. It is time that we recognize that there is a large and valuable space between where we are in terms of group differences and 0 differences. We need to work toward improving what we do and accept that eliminating average group differences is not possible for many effective predictors and that it may take some time for the existing group differences to shrink or to find other solutions.

Central Role of Selection Ratio

In 1939, Taylor and Russell published a highly influential article arguing that although validity is important, the effectiveness of any selection tool is dependent upon far more than just the correlation between a test score or test composite and the criterion of interest. This marked the introduction of the base rate and selection ratio as important variables that influence the usefulness of a test in forecasting performance, above and beyond what we know about validity. They demonstrated that even with very high validity a test could be useless and that relatively low levels of validity could be quite helpful in predicting performance because of the complexity of any selection problem.

The definition of adverse impact is far too simplistic—it is not just about how a test performs. Like the contribution of Taylor and Russell we need to more completely define the space in which adverse impact occurs. There is no doubt that under certain circumstances even the worst offending test has no adverse impact. Should that lead us to conclude that the test is OK? There are situations when a test that shows minimal differences between two groups can have huge amounts of adverse impact only because the selection ratio is very low. Bobko and Roth (2004) showed that organizations with much lower selection ratios are more susceptible to adverse impact regardless of the selection procedures themselves. For example, in an organization with a selection ratio of .70, a difference in selection ratios in excess of 13% would still be within the acceptable limit of the 80% rule, whereas a selection ratio of .20 would violate the 80% rule with only a difference of 4%. The probability of finding adverse impact is highest for low base rates and low selection ratios and lowest for high base rates and high selection ratios. Clearly, adverse impact increases as selection ratio decreases, especially at low base rates, and as Cascio, Jacobs, and Silva (2010, p. 284) note, ''the 80% rule and other legal standards that focus solely on group differences do not reflect the intricacies of selection.''

It is clear that the current state of the *Uniform Guidelines* is not acceptable from a proscriptive perspective, and they certainly do not reflect the substantial knowledge we have gained over the time they have been in place. Right now a conclusion of adverse impact regarding a test (or selection procedure) is possible when in fact there is nothing wrong with the test. The way we evaluate adverse impact allows sample size, selection ratio, and expected outcomes regarding group differences to masquerade as a flaw in the test. Any changes to the *Uniform Guidelines* will require a massive effort. Any change that is coming should consider the three issues highlighted in this paper and provide more definitive direction in what truly reflects evidence of adverse impact.

https://doi.org/10.1111/j.1754-9434.2011.01392.x Published online by Cambridge University Press

References

Bobko, P., & Roth, P. (2004). The four-fifths rule for assessing adverse impact: An arithmatic, intuitive and logical analysis of the rule and implications for future research. In J. Martocchio (Ed.), *Reseach in personnel and human resources management* (Vol. 23, pp. 177–197). New York, NY: Elsevier.

Cascio, W. F., Jacobs, R., & Silva, J. (2010). Validity, utility, and adverse impact: Practical implications from 30 years of data. In J. L. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 271–288). New York, NY: Routledge.

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.

McDaniel, M. D., 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.

Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology*, *88*, 694–706.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, *23*, 565–578.

Vulcan Society of New York City Fire Department Inc. v. Civil Service. (1973). *Commission of City of New York 490 F2d387*.

https://doi.org/10.1111/j.1754-9434.2011.01392.x Published online by Cambridge University Press