

Corporate Office:

DCI Consulting Group, Inc.
1920 I Street, NW
Washington, DC 20006
Phone: (202) 828-6900
dciconsult.com

A Consideration of Practical Significance in Adverse Impact Analysis

Eric M. Dunleavy, Ph.D. - Senior Consultant
July 2010

One of the frequent statistical techniques used in EEO context is the adverse impact analysis, which compares the employment consequences of an organizational policy or procedure between two groups. This comparison often simplifies to a test of the difference between two rates or the ratio¹ of those rates. Perhaps most commonly considered in analyses of hiring data, adverse impact analyses often answer this basic question: Are the hiring rates between group 1 (e.g., males) and group 2 (e.g., females) ‘meaningfully’ different? It is important to note that, in this context, the notion of ‘meaningfully different’ can be interpreted in more than one way. For example, from a statistical significance perspective, ‘meaningfully different’ generally means ‘probably not due to chance’. In other words, what is the degree of uncertainty inherent in the conclusions of the analysis (i.e., that there is a meaningful difference between two groups)?

From the practical significance perspective, ‘meaningfully’ different could also mean ‘dissimilar enough for the EEO and/or scientific community to notice’. This perspective emphasizes the magnitude or size of the difference. As this notion suggests, practical significance measures include some inherent subjectivity, because EEO and scientific communities must determine how large a difference (or how much a ratio deviates from 1) must be to become a ‘red flag’ that may eventually be deemed unlawful discrimination. As described by the OFCCP statistical standards report (1979):

“First, any standard of practical significance is arbitrary. It is not rooted in mathematics or statistics, but in a practical judgment as to the size of the disparity from which it is reasonable to infer discrimination.”

¹ Please refer to Morris and Lobsenz (2000) for a review of tests that focus on the ratio of selection rates.

Second, no single mathematical measure of practical significance will be suitable to widely different factual settings.”

Practical significance is an important addition to statistical significance in the consideration of potential adverse impact. Because meaningless group differences will be “statistically significant” with large samples sizes, it is important to determine whether the size of the group difference represents potential discrimination. For example, Dunleavy, Clavette, & Morgan (2010) have demonstrated that a 1% difference in selection rates can become statistically significant when the sample size reaches 1,200: a difference in selection rates so small that discrimination cannot be reasonably inferred.

Although concrete practical significance standards are not available for all situations, a number of practical significance measures have been endorsed by EEO doctrine and accepted by U.S. courts dealing with EEO claims. Other practical significance measures, while not explicitly endorsed by EEO doctrine or courts, are generally accepted by the social scientific community. This paper reviews some practical significance measures that may be useful in the context of adverse impact analyses. These measures are particularly useful in combination with statistical significance² tests.

Practical significance measures appropriate for adverse impact analysis

Perhaps the most commonly used practical significance measure in the EEO context is the 4/5th or 80% rule, which uses an impact ratio (i.e., Group A pass rate divided by Group B pass rate) to measure magnitude. Codified in the Uniform Guidelines on Employee Selection Procedures (UGESP, section 4D), the rule is described as follows:

“A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant.”

Thus, the 4/5th rule is a measure of the magnitude of disparity. As the UGESP definition points out, the 4/5th rule is endorsed by Federal agencies, yet may need to be interpreted in light of particular context (e.g., sample size, in combination with statistical significance testing). However, case law suggests that the 4/5th rule can be interpreted as adequate stand alone

² Note that statistical significance tests like Z and Fisher's exact test are often useful, yet may be trivial in some situations.

evidence in some situations, although it is unclear exactly what circumstances warrant such interpretation.³ Note that the 4/5th rule is also explicitly endorsed in the Office of Federal

Contract Compliance Programs (OFCCP) Compliance Manual (1993; Section 7E06, titled “MEASUREMENT OF ADVERSE IMPACT”):

“80 Percent Rule: OFCCP has adopted an enforcement rule under which adverse impact will not ordinarily be inferred unless the members of a particular minority group or sex are selected at a rate that is less than 80 percent or four-fifths of the rate at which the group with the highest rate is selected (41 CFR 60-3.4D, Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures (Questions and Answers) (Nos. 10-27)). When a minority or female selection rate is less than 80 percent of that of White males a test of statistical significance should be conducted. (See SCRR Worksheets 17-6a, 6b, and accompanying instructions.) The 80 percent rule is a general rule, and other factors such as statistical significance, sample size, whether the employer's actions have discouraged applicants, etc., should be analyzed”.

Of course, it is important to note that the 4/5th rule analysis can be inaccurate in some situations. Shortly after the publication of the UGESP, the management science literature criticized the rule as stand-alone evidence of discrimination. These reasonable criticisms centered on (1) a series of inconsistencies regarding the interpretation of the rule (which are apparent in UGESP and the UGESP Question and Answers) and (2) some poor psychometric properties of 4/5th rule analyses. Most recently, in a study conducted by Roth, Bobko, and Switzer (2006), simulation research was used to identify some situations where the 4/5th rule provided erroneous conclusions. Specifically, the authors showed that false-positives (situations when the 4/5th rule was violated but selection rates were essentially equal in the population of applicants) occurred at an alarming rate, particularly when there were few hires, low minority representation, and small applicant pools. For these and other reasons⁴, most experts in the area of EEO view the 4/5th rule as a general rule of thumb that can be used in combination with other evidence, such as statistical significance testing (Meier, Sacks, Zabell, 1984). Having said that, little research has offered an alternative rule of thumb, so the 4/5th rule appears to be no worse conceptually than other social scientific rules of thumb for measures such as odds ratios, absolute differences in selection rates, or Cohen’s t transformations of the difference. These measures are described in more detail later in the paper.

Note that the rationale for combining practical and statistical significance results is an intuitive one. In situations where the measures come to identical conclusions, the EEO analyst can usually feel very confident in a finding of meaningful impact or no impact. In other

³ Also note that much of this case law is older, and many rulings were decided in the 10 years after UGESP were codified.

⁴ Please refer to Biddle (2005) for a description of how the 4/5th rule was developed, and the somewhat arbitrary nature of this rule of thumb.

situations, context may play an important role when statistical and practical significance measures produce different conclusions (i.e., when a standard deviation analysis is greater than 2.0 but the 4/5th rule is not violated).

Table 1 presents a framework for interpreting statistical and practical significance measures. As the table shows, statistically significant tests paired with meaningful practical measures point toward a disparity reasonable from which to infer discrimination. It is probably not reasonable to infer discrimination when a disparity is not statistically significant or practically meaningful. In other situations where the two perspectives disagree, context will play an important role. Note that it is difficult to conclude practical significance in the absence of statistical significance, because we are not confident that the difference is ‘real’.

Table 1: A Framework for Interpreting Statistical and Practical Significance Measures

	Practical Significance Measure (e.g., difference, impact ratio, etc.) Results		
		Meaningful	Trivial
Statistical Significance Test (e.g., Z, FET) Results	Significant	A disparity that is probably reasonable from which to infer discrimination	Somewhere in the Middle (but <u>chance is probably not</u> an explanation)
	Not Significant	Somewhere in the middle (but <u>chance is probably</u> an explanation)	A disparity that is probably not reasonable from which to infer discrimination

The issue of inconsistent results across disparity measurement perspectives was considered by the 2nd Circuit in **Waisome v. Port Authority (1991)** which ruled that practical significance evidence was required even in situations where a disparity was statistically significant at greater than two standard deviations:

“We believe Judge Duffy correctly held there was not a sufficiently substantial disparity in the rates at which black and white candidates passed the written examination. Plainly, evidence that the pass rate of black candidates was more than four-fifths that of white candidates is highly persuasive proof that there was not a significant disparity. See EEOC Guidelines, 29 C.F.R. § 1607.4D (1990); cf. Bushey, 733 F.2d at 225-26 (applying 80 percent rule). Additionally, though the disparity was found to be statistically significant, it was of limited magnitude, see Bilingual Bicultural Coalition on Mass Media, Inc. v. Federal Communications Comm'n, 595 F.2d 621, 642 n. 57 (D.C.Cir.1978) (Robinson, J., dissenting in part) (statistical significance tells nothing of the importance, magnitude, or practical significance of a disparity) (citing H. Blalock, Social Statistics 163 (2d ed. 1972)).....These factors, considered in light of the admonition that no minimum threshold of statistical significance mandates a finding of a Title VII

violation, persuade us that the district court was justified in ruling there was an insufficient showing of a disparity between the rates at which black and white candidates passed the written examination.”

Other practical significance measures have been used by courts as well. For example, numerous courts have evaluated practical significance using the actual percentage difference in selection rates. For example, in *Frazier v. Garrison I.S.D. (1993)*, a four and a half percent difference in selection rates was deemed trivial in a situation where 95% of applicants were selected. A similar practical significance measure was used in *Moore v. Southwestern Bell Telephone Co.*, where the court held that ‘*employment examinations having a 7.1 percentage point differential between black and white test takers do not, as a matter of law, make a prima facie case of disparate impact. Therefore, there was no meaningful discrepancy between minority and non-minority pass rates based on selection rate differences*’.⁵

‘Flip flop’ rules have also been endorsed by courts and the EEO community as measures of practical significance. Instead of measuring magnitude, these measures essentially impose some correction for sampling error on a practical significance measure, ensuring that a result wouldn’t drastically change if small changes to the hiring rates were made. This rationale is similar to statistical significance testing. For example, with the regard to the 4/5th rule, Question and Answer 21 from UGESP states:

“If the numbers of persons and the difference in selection rates are so small that it is likely that the difference could have occurred by chance, the Federal agencies will not assume the existence of adverse impact, in the absence of other evidence. In this example, the difference in selection rates is too small, given the small number of black applicants, to constitute adverse impact in the absence of other information (see Section 4D). If only one more black had been hired instead of a white the selection rate for blacks (20%) would be higher than that for whites (18.7%). Generally, it is inappropriate to require validity evidence or to take enforcement action where the number of persons and the difference in selection rates are so small that the selection of one different person for one job would shift the result from adverse impact against one group to a situation in which that group has a higher selection rate than the other group.”

A similar practical significance measure was articulated in *Contreras v. City of Los Angeles (1981)*. In this case practical significance was assessed via the number of additional ‘victim’ applicants that would need to be selected to eliminate a significant disparity. Practical significance was also assessed by determining the number of additional ‘victim’ applicants that would need to be selected to make rates very close between groups (i.e., around 2%).

⁵ It is important to note that in both of these cases overall selection rates and subgroup selection rates were very high, and that the 4/5th rule was not violated. It is unclear how differences in selection rates of this magnitude would be interpreted when selection rates are lower such that the 4/5th is violated (e.g., 4% vs. 8% and an impact ratio of .50 instead of 92% vs. 96% and an impact ratio of .96). Intuitively, such differences may be treated differently.

Another practical significance measure was used in *U.S. v. Commonwealth of Virginia* (1978) and in the *Waisome* case described above. This method required assessing the number of additional ‘victim’ applicants that would need to be selected to eliminate a statistically significant disparity (e.g., less than 2 standard deviations). In this context, if ‘one or two’ additional passes from the ‘victim’ group changed the statistical results, the difference would not be considered practically significant.⁶

Social scientific trends toward the use of practical significance measures

Practical significance is a general concept that has gained a great deal of support in the social scientific community in the last few decades.⁷ As advocated by Kirk (1996) in a special series on practical significance in *Educational and Psychological Measurement*, it is a concept whose time has come. This is because many in the social scientific community have identified an over-reliance of statistical significance testing in academic and applied research, and advocated a more balanced set of statistical standards that include practical significance measures in the form of effect sizes.⁸ For example, in the most recent Publication manual of the *American Psychological Association* (2010) a failure to report effect sizes (as practical significance measures) is considered a defect in the reporting of research:

“No approach to probability value directly reflects the magnitude of an effect or the strength of a relation. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relation in your Results section.”

Additionally, the *Journal of Applied Psychology*, generally considered a top-tier social scientific journal, now requires authors to:

“...indicate in the results section of the manuscript the complete outcome of statistical tests including significance levels, some index of effect size or strength of relation, and confidence intervals” (Zedeck, 2003, p. 4).

⁶ Note that this ‘statistical significance flip flop’ rule may be somewhat counter-intuitive, since the flip flop condition equates to the use of statistical significance testing with a lower alpha level (e.g., .04 instead of .05) or higher standard deviation criterion (e.g., 2.1 instead of 2.0). In this context alpha is being adjusted according to the number of additional selections that would be required to produce non-significant results.

⁷ Importantly, many statistically savvy researchers have advocated the use of both practical and statistical significance testing methods for many years (e.g., Cohen, 1988; Henkel, 1976; Reynolds, 1984; Tabachnick & Fidell, 2001). Meier, Sacks, & Zabell (1984) made this same recommendation specifically for analyses of disparity.

⁸ In this context, effect size refers to a measure capturing the magnitude of a relation between two variables, using an outcome and a predictor that influences that outcome. Note that this is not simply a probabilistic test as in statistical significance testing. For example, how strongly are gender and the likelihood of being hired correlated in the EEO context? In the discrimination context it is usually hypothesized that gender is a cause or explanation of being hired or rejected.

This paradigmatic shift is particularly noteworthy within the context of applied and present day EEO research because applicant pools are often very large and thus may produce trivial statistically significant results as a function of sample size alone. See Table 2 for an example of a disparity that appears practically meaningless, but as more and more data are collected (via multiplying sample sizes by a constant of 10), the statistical significance test eventually suggests meaningful disparity (at a sample size of 2,400). In other words, the impact ratio and difference in rates always suggests trivial disparity and are constant over sample size, yet the Z test changes simply as a function of sample size, and is eventually significant even though the difference in rates is only 1 percentage point. This phenomenon is likely when data are collected over time (e.g., 1 year, 2 years, 3 years, 10 years), across multiple locations, or across multiple jobs.

Table 2: A comparison of practical and statistical measures across sample sizes

# Applicants		# Selections		Selection Rates			Practical Measures		Statistical Test
Males	Females	Males	Females	Total	Males	Females	Impact Ratio	Diff in rates	SD (Z) test
100	100	99	98	0.985	0.99	0.98	0.99	0.01	0.58
1,000	1,000	990	980	0.985	0.99	0.98	0.99	0.01	1.84
1,200	1,200	1,188	1,176	0.985	0.99	0.98	0.99	0.01	2.01
10,000	10,000	9,900	9,800	0.985	0.99	0.98	0.99	0.01	5.82
100,000	100,000	99,000	98,000	0.985	0.99	0.98	0.99	0.01	18.40
1,000,000	1,000,000	990,000	980,000	0.985	0.99	0.98	0.99	0.01	58.17

Importantly, there are a variety of available practical significance measures capturing the magnitude of relation between protected group status and the likelihood of an employment decision. For example, the odds ratio, which captures the odds of experiencing a positive employment experience for one group relative to another, provides an intuitive metric of practical significance.⁹ This metric has been endorsed by statisticians with expertise in the EEO community (e.g., Gastwirth, 1988). For example, Gastwirth suggested that an odds ratio of 1.4 (or its reciprocal, .70) was a reasonable rule of thumb for moderate disparity, representing the case where members of one group are 1.4 times (or 40%) more likely to experience a positive employment decision than members of another group.

⁹ Note that the odds ratio is similar to the impact ratio used for 4/5th rule analysis. However, the odds ratio takes into consideration the rejection rate of each group in addition to the selection rate, while the impact ratio only considers selection rates. This difference can explain situations where the odds ratio and the impact ratio do not provide the same conclusion. For example, if one group is selected at 92% and another group is selected at 96%, the impact ratio suggests trivial significance (i.e., an impact ratio of 0.96), whereas the odds ratio would suggest meaningful disparity (i.e., an odds ratio of 0.458).

Measures of association (e.g., Phi) capturing the magnitude of relation between protected group and employment outcome can also be useful from a practical significance perspective. For readers familiar with the employment testing context, this notion is similar to ‘validity coefficients’ used to measure how well a test predicts job performance. Although there are some special considerations for assessing the relationship between two dichotomous variables, the same general 0 to 1 ‘validity coefficient scale’ applies, where 0 indicates no relationship between group and outcome and values closer to 1 indicate strong relationship. Unfortunately, there are no clear and obvious rules of thumb for these metrics,¹⁰ although a value close to zero can reasonably be interpreted as no relationship, and thus trivial disparity. Cohen’s h statistic, which is a transformation of the difference in two rates, may be another useful metric. Although no clear and universal rules of thumb are available for interpretation, Cohen (who later wrote that he regretted providing rules of thumb that were so easily misapplied) suggested the following starting points for interpretation:

- 0.2 = Small difference in rates
- 0.5 = Medium difference in rates
- 0.8 = Large difference in rates

It is important to reiterate that the various measures of practical significance may yield differing conclusions. The investigation of the validity of drug testing at the U.S. Postal Service by Normand, Salyards, and Mahoney (1990) provides an excellent example the potential for differing conclusions. Normand et al. found that applicants testing positive for drugs were 48.5% more likely to be heavy users of absenteeism leave than were applicants testing negative for drugs, a difference that is statistically significant. The 48.5% difference is of a magnitude that on the face appears to be of high practical significance and the odds ratio of 1.97 exceeds Gastwirth’s rule-of-thumb for moderate practical significance. However, if the same data are converted to a correlation coefficient, the resulting correlation of .10 would be considered a low level of practical significance.

Conclusion

Understanding the practical significance of a selection rate difference (or ratio of selection rates) is a critical issue in the high stakes situation where an employer faces charges of unlawful employment practices that discriminate against a protected group. A number of EEO-

¹⁰ The Department of Labor provided some general rules of thumb for interpreting the usefulness of correlations in the testing context, usually when a continuous test score predicts a continuous performance outcome. In this context, a correlation of .11 or less is ‘unlikely to be useful’, between .11 and .20 ‘depends on the circumstances’, between .21 and .35 is ‘likely to be useful’, and above .35 is ‘very beneficial’. However, given the special statistical case of two dichotomous variables, it is unclear how these DOL rules of thumb apply to adverse impact analyses. Future research should consider this issue.

endorsed and scientifically based practical significance measures are available for analysis of traditional employment decision data. These measures may be particularly useful in situations where sample sizes are very large, and statistical significance testing becomes a meaningless exercise. In fact, conducting and interpreting statistical significance tests alone when samples are very large is scientifically unsound and can be potentially misleading. We strongly recommend that EEO analysts consider both statistical significance tests and practical significance measures in adverse impact analyses.¹¹

¹¹ Again, it is important to note that UGESP may endorse a similar combination of tests, although the exact meaning of the following section (4D) is unclear: *‘Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group. Greater differences in selection rate may not constitute adverse impact where the differences are based on small numbers and are not statistically significant’.*

References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.
- Biddle, D.A. (2005). *Adverse impact and test validation: A practitioners guide to valid and defensible employment testing*. Burlington, VT: Gower.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Erlbaum Associates.
- Dunleavy, E. M., Morgan, D. M., & Clavette, M. (2010). Practical Significance: A concept whose time has come in adverse impact analyses. In Morrison, J., & Sinar, E. (Moderators). *The 4/5ths Is Just a Fraction: Alternative Adverse Impact Methodologies*. Symposium presented to the 25th annual SIOP conference in Atlanta, GA, April 2010.
- Gastwirth, J.L. (1988). *Statistical reasoning in law and public policy* (Vol. 1). San Diego, CA: Academic Press.
- Henkel, R. E. (1976). *Tests of Significance*. Sage University Series Quantitative applications in the social sciences. Newbury Park, CA: Sage Publications.
- Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56, 746-759.
- Meier, P., Sacks, J. & Zabell, S. (1984). What happened in Hazelwood: Statistics, employment discrimination, and the 80% Rule. *American Bar Foundation Research Journal*, 1, 139-186.
- Morris, S.B. & Lobsenz, R.E. (2000). Significance tests and confidence intervals for the adverse impact ratio. *Personnel Psychology*, 53, 89-111.
- Normand, J., Salyards, S. D., & Mahoney, J. J. (1990). An evaluation of preemployment drug testing. *Journal of Applied Psychology*, 75(6), 629–639.
- Office of Federal Contract Compliance Programs. (1993). *Federal contract compliance manual*. Washington, DC: U.S. Department of Labor.
- Reynolds, H. T. (1984). *Analysis of nominal data*. Sage University Series Quantitative applications in the social sciences. Newbury Park, CA: Sage Publications

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

Sobel, R., Michelson, S., Finklestein, M., Fienberg, S., Eisen, D., Davis, F. G., Paller, P. E. (1979). Statistical inferences of employment discrimination and the calculation of back pay. Part I: Inferences of discrimination. Unpublished OFCCP Statistical standards panel report.

Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn & Bacon.

Uniform guidelines on employee selection procedures. Fed. Reg., 43, 38,290-38,315 (1978).

Zedeck, S. (2003). Instructions for authors. *Journal of Applied Psychology*, 88, 35.

Cases Cited

Contreras v. City of Los Angeles (1981) 656 F.2d 1267

Frazier v. Garrison ISD (1993) 980 F.2d 1514

Moore v. Southwestern Bell Telephone Co. (5th Cir.1979) [593 F.2d 607](#), 608

U.S. v. Commonwealth of Virginia (1978) 620 F.2d 1018

Waisome v. Port Authority of New York & New Jersey (1991) 948 F.2d 1370