Applied HRM Research, 2001, Volume 6, Number 1, 13-32

Sample Size Required for Adverse Impact Analysis

Scott B. Morris
Illinois Institute of Technology

Practitioners who rely on adverse impact analysis for the evaluation and development of selection procedures need to carefully consider the adequacy of the samples used to compute adverse impact statistics. A method is presented for estimating the minimum sample size needed to test for adverse impact with adequate statistical power. The required sample size is often extremely large, particularly for small effects when the selection rate and the proportion of minority applicants are small. However, smaller samples will be adequate when the magnitude of the group difference is large.

The degree of adverse impact produced by a selection system is often evaluated using small sample statistics, despite warnings that such analyses are plagued by excessive sampling error and low statistical power (Cohen, 1994; Schmidt, 1996). When sample size is small, adverse impact statistics will vary considerably across test administrations. For example, Lawshe (1987) administered the same test for the same job in two consecutive years, and found dramatically different levels of adverse impact on the two administrations. In addition, when sample size is small, it is unlikely that a statistically significant difference in selection rates will be found (Morris & Lobsenz, 2000). On the other hand, when sample sizes are extremely large, even a trivial difference in selection rates will produce a statistically significant result. Therefore, sample sizes need to be carefully considered when evaluating adverse impact, as with any sample statistic.

The concerns about adequate sample sizes are particularly important in such areas as personnel selection, where decisions have substantial social and political implications for the individuals involved. The need for adequate sample size in test validation research has long been recognized (Schmidt, Hunter & Urry, 1976). However, little attention has been paid to the sample size requirements for adverse impact analysis.

As with test validity, adverse impact statistics are often used in the design and evaluation of selection systems. Adverse impact statistics play a central role in many employment discrimination claims (Gutman, 2000), and are often explicitly considered in the design of selection procedures. For example, when combining results from a test battery, more weight may be assigned to those subtests with less adverse impact (De Corte, 1999). In other cases, test developers may set lower cutoff scores for subtests with greater adverse impact (Hoffman & Thornton, 1997). Given the widespread use of adverse impact statistics in test development, it is important to understand the conditions under which these statistics have adequate statistical power.

Two general approaches are commonly used to evaluate adverse impact statistics. The first is the four-fifths rule, specified in the EEOC Uniform Guidelines on Employee Selection Procedures (1978). According to the four-fifths rule, adverse impact exists if the selection rate for the minority group is less than four-fifths of the selection rate for the majority group. An alternate approach is to test the difference in selection rates for statistical significance. Significance tests for adverse impact

are recommended by federal regulations (OFCCP, 1993), and have been utilized by the courts in a number of cases (Hazelwood School District v. United States, 1977; Rich v. Martin-Marietta, 1979).

Prior research has shown that neither approach to assessing adverse impact works well in small samples. The four-fifths rule will often indicate adverse impact when none exists in the population, and will often fail to identify true cases of adverse impact (Boardman, 1979; Greenberg, 1979). Statistical significance tests are less likely to falsely identify adverse impact. However, significance tests have extremely low power under many of the conditions where adverse impact is typically assessed (Morris & Lobsenz, 2000). That is, they will often fail to detect adverse impact when it does exist. Although past research identifies the potential for low power, it does not clearly indicate the conditions under which adverse impact tests would be appropriate. The current study builds upon this earlier work by examining the minimum sample size required to test for adverse impact with adequate statistical power.

Methods for determining sample size requirements are widely available (e.g., Cohen, 1988), and several previous studies have discussed the sample sizes needed for tests on proportions (Boardman, 1979; Cohen, 1988; Fleiss, 1981). However, each of these studies examined a single approach to testing proportions. By examining alternate analysis methods within the same study, the relative merits of the alternate methods can more readily be determined.

In addition, much of the work on power analysis is based on very general models that can be applied to a wide range of research problems (e.g., Cohen, 1988). This generality has many advantages, but may obscure the implications of the analysis for a particular application. For example, Cohen's (1988) work utilizes the harmonic mean of sample sizes and the difference between arcsin-transformed proportions. This makes it difficult to evaluate the implications of the results for adverse impact in selection settings.

The approach used in the current paper is entirely consistent with existing power analysis methods, and produces very similar results. However, because the current analysis is conducted in terms of the parameters relevant to personnel selection (e.g., the impact ratio, overall selection rate, and proportion of minority applicants), the implications of these results for adverse impact analysis are presented in a much clearer light.

Adverse Impact Statistics

The concept of adverse impact was first delineated by the Supreme Court decision in *Griggs v. Duke Power Company* (1971). Under the disparate impact theory, discrimination exists when there is evidence of a statistical disparity in selection or promotion rates, unless the practice meets a business necessity. The EEOC *Uniform Guidelines on Employee Selection Procedures* (1978) suggested the four-fifths rule,

> 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...(Section 4D, p. 38297).

The basic statistic used in the four-fifths rule is the Impact Ratio (IR), which is the ratio of the selection rate for the minority group (SR_{min}) to the selection rate for the majority group (SR_{maj}), or

$$IR = \frac{SR_{\min}}{SR_{maj}}. \quad \textbf{(1)}$$

A limitation of the four-fifths rule is that is that it does not consider the potential for sampling errors. When sample size is small, results are likely to vary dramatically from sample to sample, simply because a different random sample of applicants would show up for each test administration. For example, Lawshe (1987) compared the adverse impact of selection decisions for 11 jobs across two consecutive years. Although the same selection systems were used in both years, the impact ratios changed dramatically. As a result of sampling error, the rule would falsely indicate many cases of adverse impact where none exists in the population, and would often fail to identify adverse impact when it does exist in the population (Boardman, 1979; Greenberg, 1979).

To better account for the effects of sampling error, many practitioners test whether the difference in selection rates is statistically significant. A variety of tests are applied in different situations. The current paper will focus on the common situation where a selection system is administered to a pool of applicants, and adverse impact is assessed by comparing subgroup selection or passing rates. With such data, the two most common tests are the Z-test for the difference between two proportions, and the chi-square test for association in a 2x2 contingency table. Both tests are based on the normal approximation to the binomial distribution, and are mathematically equivalent (Fleiss, 1981).

Two alternative tests will be presented here: a *Z*-test on the difference in selection rates, and a *Z*-test on the ratio of selection rates. The computation of each test statistic will be illustrated using the following example. A employer administers a single test to a pool of 200 applicants, 50 (25%) of whom are from the minority group. Fifty of the majority applicants pass the test (SR_{maj} = 50/150 = .33), while only 10 of the minority applicants pass ($SR_{\min}$ = 10/50 = .20). The overall selection rate is 60/200 = .30. The resulting adverse impact ratio would be IR = .20/.33 = .60, which would indicate adverse impact according to the four-fifths rule.

The *Z*-test for the difference between proportions (Z_D) evaluates the null hypothesis of equal population selection rates for the two groups being compared. This test is described in the OFCCP Compliance Manual (1993), and can be written as,

$$Z_D = \frac{SR_{\min} - SR_{maj}}{\sqrt{\frac{SR_T(1-SR_T)}{(N)(P_{\min})(1-P_{\min})}}}, \quad \textbf{(2)}$$

where SR_{T} is the total selection rate, N is the number of applicants, and $P_{\min}$ is the proportion of applicants from the minority group. This approach is often labeled the 2-*SD* test, because the difference is considered significant if it is more than two standard deviations above or below zero (or more precisely, $|Z| > 1.96$, corresponding to a two-tailed α=.05). Using the example described above, the result would be:

$$Z_D = \frac{.20 - .33}{\sqrt{\frac{(.30)(1-.30)}{(200)(.25)(1-.25)}}} = -1.78. \quad \textbf{(3)}$$

This result would not be significant using a two-tailed α=.05.

An alternate statistical test (Z_{IR}) was suggested by Morris and Lobsenz (2000), who argued that the significance test should be based on the same effect size as the four-fifths rule. The null hypothesis of equal selection rates implies that the population impact ratio should be equal to one. This hypothesis can be tested using a Z-test on the natural log of the impact ratio (Z_{IR}),

$$Z_{IR} = \frac{\ln\left(\frac{SR_{\min}}{SR_{maj}}\right)}{\sqrt{\frac{1-SR_T}{(SR_T)(N)(P_{\min})(1-P_{\min})}}}. \quad \textbf{(4)}$$

Again, the null hypothesis can be rejected if $|Z| > 1.96$, corresponding to a two-tailed α=.05. Z_{IR} is quite similar to the Z_D test. In both tests, the numerator reflects an effect size and the denominator is the standard error of the effect size when the null hypothesis is true. The two tests differ only in how the effect size is defined (i.e., the difference or the ratio of selection rates), which leads to slightly different formulations of the standard error. Another difference is that, in the Z_{IR} test, the selection rates have been subjected to a natural log transformation, which results in a sampling distribution that is more closely normal (Fleiss, 1994).

Applying the Z_{IR} test to the example described above would result in

$$Z_{IR} = \frac{\ln\left(\frac{.20}{.33}\right)}{\sqrt{\frac{1-.30}{(200)(.30)(.25)(1-.25)}}} = -2.05. \quad \textbf{(5)}$$

In contrast to the Z_D test, this result is statistically significant at a two-tailed α=.05. This is consistent with past research, which has shown that Z_{IR} is generally more powerful than Z_D (Morris & Lobsenz, 2000).

Significance tests are never perfectly accurate. Even when the subgroup selection rates are equal in the population, some samples will produce statistically significant differences (i.e., Type I errors). In addition, a test may fail to reach statistical significance when there are true difference between groups in the population (i.e., Type II errors). The critical value of a test statistic (i.e., the score above which the test is considered significant), is set to maintain a desired Type I error rate. That is, in the long run, only a small proportion (typically 5%) of the samples will produce

significant results when there is no effect in the population.

The location of the critical value depends on whether a one-tailed or two-tailed test is used. In a two-tailed test, the researcher is willing to consider differences in either direction as statistically significant (e.g., either $SR_{min} > SR_{maj}$ or $SR_{min} < SR_{maj}$). In a one-tailed or directional test, the researcher wishes to test a specific hypothesis about the direction of the effect, and is willing to treat results on the opposite direction, no matter how large, as if they are non-significant. Because the two-tailed test considers results in either direction, it is twice as likely as a one-tailed test to produce a falsely significant result. In order to compensate, a higher critical value must be set for a two-tailed test in order to maintain the same Type I error rate as a one-tailed test. For example, when alpha=.05, a one-tailed Z-test is considered significant is $|Z| > 1.645$, whereas a 2-tailed test would be significant if $|Z| > 1.96$.

Use of a one-tailed versus a two-tailed test will also affect Type II errors. When groups differ in the predicted direction, a one-tailed test will be more likely than a two-tailed test to find a significant result. This can be seen in the example described above. The Z_D test produced a value of -1.78, which was not significant using the two-tailed test. Had a one-tailed significance test been specified, the critical value would have been 1.645, and the test would have been significant. Thus, a one-tailed test will produce fewer Type II errors than a two-tailed test, or put another way, the one-tailed test will have greater statistical power. Consequently, a smaller sample size will be required to achieve adequate power with a one-tailed test.

Two-tailed significance tests have been adopted for adverse impact analysis by federal agencies (OFCCP, 1993) and the courts (Hazelwood School District v. United States, 1977). However, adverse impact is generally only of interest if the focal group has a lower selection rate. As such, the test is directional, and a one-tailed test would be more appropriate. The use of the more liberal significance criterion ($|Z| > 1.645$) will increase the power of the test, while still maintaining an acceptable Type I error rate (α=.05).

Power and Minimum Sample Size for Adverse Impact Tests

A major limitation of significance tests is their sensitivity to sample size. When large samples are available, even small disparities will be statistically significant. In contrast, when samples are very small, only large differences will produce statistically significant results. With small samples, one cannot be certain that non-significant results were due to a lack of power or a true lack of adverse impact in the population.

Clearly, significance tests for adverse impact will be useful only to the extent that they have adequate statistical power. The next section describes methods for determining the power of adverse impact tests and the minimum sample size required to achieve a desired level of power.

Power is defined as the probability of rejecting the null hypothesis when a specific alternative hypothesis is true (Hays, 1994). Computationally, power is the probability of obtaining a result large enough to reach statistical significance, given a sampling distribution with known mean, standard error, and shape. For example, consider a pool of 500 minority and 500 majority applicants from a population with a 50% passing rate and an impact ratio of .8. Morris and Lobsenz (2000) showed that the power of the Z_D test in this sample would be .63. In other words, if a large number of random samples of 1000 applicants were drawn from this population, the Z_D test would be statistically significant in 63% of the samples. In contrast, if the sample size was 500, only 40% of

the samples would produce a significant Z_D test.

Clearly, high power is always desirable. Power refers to the probability of a statistically significant result when there are true differences in the population. Thus, a low-power test will often fail to identify real cases of adverse impact. Power near 100% would be ideal, but power around 80% is generally considered acceptable in psychological research (Cohen, 1988)

It should be noted that the four-fifths rule is not a significance test, and therefore the above definition of power does not apply. However, in practice the four-fifths rule is applied in samples in order to make inferences about a population. Therefore, the accuracy of the decision rule can be evaluated in much the same way as the power of a significance test. The sample size requirements for the four-fifths rule are based in the probability of the rule indicating adverse impact when adverse impact exists in the population. For simplicity, this will be referred to as the "power" of the four-fifths rule.

Before calculating power, it is first necessary to describe the relevant populations and the parameters reflected in each test. Independent random samples of applicants are assumed to be drawn from two independent populations, referred to as the minority and majority populations. A total of $\underline{N}$ cases are sampled from the two populations combined. The proportion of the sample from the minority group ($P_{\min}$) is assumed to be fixed across samples. The proportion of each population passing the selection criterion is indicated by $\pi_{\min}$ and π_{maj} for the minority and majority groups, respectively. For each group, π is estimated from the proportion of scores in the sample which pass the selection criterion, referred to as the selection rate ($SR_{\min}$ and SR_{maj}, for the minority and majority groups, respectively).

The overall selection rate for the sample is a function of the selection rates for each group, as well as the proportion of the sample from each group,

$$SR_T = (P_{\min})(SR_{\min}) + (1 - P_{\min})(SR_{maj}). \quad \textbf{(6)}$$

Given repeated samples of size $\underline{N}$ from the two populations with fixed $P_{\min}$, the expected value of $\underline{SR_T}$ (indicated by π_T) would be

$$\pi_T = E(SR_T) = (P_{\min})(\pi_{\min}) + (1 - P_{\min})(\pi_{maj}). \quad \textbf{(7)}$$

The four-fifths rule and the Z_{IR} test both evaluate the impact ratio. The population impact ratio is defined as $\psi=\pi_{\min}/\pi_{\text{maj}}$, and is estimated by the sample impact ratio given in Equation 1. Because the sampling distribution of the impact ratio is complex, analysis is generally performed on the natural log of the impact ratio, which is approximately normally distributed in large samples (Fleiss, 1994). The standard error of ln(IR) can be approximated by

$$SE_{IR} = \left(\frac{1}{\sqrt{N}}\right)\sqrt{\frac{1 - \pi_{\min}}{(\pi_{\min})(P_{\min})} + \frac{1 - \pi_{maj}}{(\pi_{maj})(1 - P_{\min})}}. \quad \textbf{(8)}$$

For the Z_D test, the key population parameter is the population difference in selection rates ($\pi_{min} - \pi_{maj}$). The sample estimate of the difference in selection rates ($SR_{min} - SR_{maj}$) is approximately normally distributed in large samples (Fleiss, 1994), with a standard error of

The second step in power analysis is to determine the minimum significant effect size. For each test, the minimum significant effect size can be written as a function of several factors: $\underline{Z}_{crit}$, the absolute value of the $\underline{Z}$-score corresponding to the significance criterion; π_T, the expected overall selection rate; $\underline{N}$, the number of applicants; and $\underline{P}_{min}$, the proportion of applicants from the minority group. Because we are interested in assessing adverse impact against the minority group, the effect

$$SE_D = \left(\frac{1}{\sqrt{N}}\right)\sqrt{\frac{(\pi_{min})(1 - \pi_{min})}{P_{min}} + \frac{(\pi_{maj})(1 - \pi_{maj})}{1 - P_{min}}}. \quad \textbf{(9)}$$

size will be significant if it is less than the negative critical value. $\underline{Z}_D$ will indicate a significantly smaller selection rate for the minority group when

$$SR_{min} - SR_{maj} < (-1)\, Z_{crit}\sqrt{\frac{(\pi_T)\,(1 - \pi_T)}{(N)(P_{min})(1 - P_{min})}}. \quad \textbf{(10)}$$

$\underline{Z}_{IR}$ will indicate a significantly lower selection for the minority group when

$$\ln\left(\frac{SR_{min}}{SR_{maj}}\right) < (-1)\, Z_{crit}\sqrt{\frac{(1 - \pi_T)}{(\pi_T)(N)(P_{min})(1 - P_{min})}}. \quad \textbf{(11)}$$

The four-fifths rule identifies adverse impact against the minority group when

$$\ln\left(\frac{SR_{min}}{SR_{maj}}\right) < \ln(.8). \quad \textbf{(12)}$$

The third step is to locate this minimum significant effect size (ES_{crit}) in the sampling distribution, relative to the population effect size (ES_{pop}). This can be represented as a standardized score,

$$Z_{power} = \frac{ES_{crit} - ES_{pop}}{SE_{ES}}, \quad \textbf{(13)}$$

where SE_{ES} is the standard error of the effect size, from Equations 6 and 7.

Power is defined as the probability of a score falling below Z_{power} in the sampling distribution. Because both sampling distributions are approximately normal (Fleiss, 1994), power can be determined as the proportion of scores in a standard normal distribution falling below Z_{power}.

The sample size required can be determined by reversing the process. A value of Z_{power} is found which corresponds to the desired level of power. For instance, for power of .8, the corresponding *Z*-score is .842 (i.e., 80% of the scores will fall below .842). Substituting this value into the above equations and solving for sample size produces the minimum N required for the specified level of power.

In order to determine the minimum sample size required, the researcher must first specify the expected values of the relevant parameters: the overall selection rate (π_T), the proportion of minority applicants ($P_{\min}$), and the degree of adverse impact in the population (ψ). To illustrate the process, consider the following example. In the population of applicants, it is believed that the selection rate for the minority group is only half the selection rate for the majority group (i.e., ψ=.5). The employer expects 30% of the applicants will be from the minority group (P_{min} = .3), and that 10% of the applicants will be selected (π_T=.1).

The equations also require specification of the population selection rates for both groups, but these can be computed from the other parameters. From the definition of ψ, $\pi_{\min}/\psi$ can be substituted for π_{maj} in Equation 5. Solving for $\pi_{\min}$ gives

$$\pi_{\min} = \frac{(\pi_T)(\psi)}{1 + P_{\min}(\psi - 1)}. \quad \textbf{(14)}$$

For the example,

$$\pi_{\min} = \frac{(.1)(.5)}{1 + (.3)(.5 - 1)} = 0558. \quad \textbf{(15)}$$

Similarly, the population selection rate for the majority group is

$$\pi_{maj} = \frac{\pi_T}{1 + P_{\min}(\psi - 1)}. \quad \textbf{(16)}$$

For the example,

$$\pi_{maj} = \frac{.1}{1 + (.3)(.5 - 1)} = .1177. \quad \textbf{(17)}$$

Formulas for computing the minimum sample size can be found by plugging the appropriate formulas for ES_{crit}, ES_{pop}, and SE_{ES} into Equation 13, and then solving for N. For the Z_{D} test, the minimum sample size can be computed from

$$N_D = \left[\frac{Z_{crit}\sqrt{\frac{(\pi_T)(1-\pi_T)}{(P_{\min})(1-P_{\min})}} + Z_{power}\sqrt{\frac{(\pi_{\min})(1-\pi_{\min})}{P_{\min}} + \frac{(\pi_{maj})(1-\pi_{maj})}{1-P_{\min}}}}{\pi_{\min} - \pi_{maj}} \right]^2. \quad \textbf{(18)}$$

Applying this to the example results in a minimum sample size of

$$N_D = \left[\frac{(1.96)\sqrt{\frac{(.1)(1-.1)}{(.3)(1-.3)}} + (.842)\sqrt{\frac{(.0588)(1-.0588)}{.3} + \frac{(.1177)(1-.1177)}{(1-.3)}}}{.0588-.1177} \right]^2 = 904.27. \quad \textbf{(19)}$$

The smallest sample size that would have 80% power would be the smallest integer greater than this value, or 905 applicants.

The minimum sample size for the Z_{IR} test is

$$N_{IR} = \left[\frac{Z_{crit}\sqrt{\frac{(1-\pi_T)}{(\pi_T)(P_{\min})(1-P_{\min})}} + Z_{power}\sqrt{\frac{(1-\pi_{\min})}{(\pi_{\min})(P_{\min})} + \frac{(1-\pi_{maj})}{(\pi_{maj})(1-P_{\min})}}}{\ln(\psi)} \right]. \quad \textbf{(20)}$$

Applying this to the example results in

$$N_{IR} = \left[\frac{(1.96)\sqrt{\frac{(1-.1)}{(.1)(.3)(1-.3)}} + (.842)\sqrt{\frac{(1-.0588)}{(.0588)(.3)} + \frac{(1-.1177)}{(.1177)(1-.3)}}}{\ln(.5)} \right]^2 = 796.2. \quad \textbf{(21)}$$

As expected, the minimum sample size for the Z_{IR} test (797) is smaller than the sample size required for the Z_D test (905).

The minimum sample size for the four-fifths rule is

$$N_{45} = \left\lceil \frac{Z_{power}\sqrt{\frac{(1-\pi_{\min})}{(\pi_{\min})(P_{\min})} + \frac{(1-\pi_{maj})}{(\pi_{maj})(1-P_{\min})}}}{\ln(.8) - \ln(\psi)} \right\rceil^2 . \quad \textbf{(22)}$$

Applying this to the example yields,

$$N_{45} = \left\lceil \frac{(.842)\sqrt{\frac{1-.0588}{(.0588)(.3)} + \frac{1-.1177}{(.1177)(1-.3)}}}{\ln(.8) - \ln(.5)} \right\rceil^2 = 204.9. \quad \textbf{(23)}$$

Because the four-fifths rule is not a significance test, it requires substantially smaller sample sizes than either of the other two approaches. However, the increased power comes at the expense of an inflated Type I error rate under many conditions (Boardman, 1979; Greenberg, 1979).

Computational Analysis of Minimum Sample Size

A computational analysis was conducted to examine the minimum sample size required given various levels of the relevant parameters. The analysis was conducted for π_T of .1, .3, and .5, P_{min} of .1, .3, and .5, and ψ ranging from .1 to .8. Two criteria for statistical significance were examined. Z_{crit} was set at 1.96 (two-tailed α=.05) and 1.645 (one-tailed α=.05).

The estimates of the minimum sample size are based on the assumption that the sampling distribution of each statistic is approximately normal. The normal approximation is fairly accurate as long as the expected frequency of each condition is greater than five (Fleiss, 1981). In this case, the smallest frequency refers to the expected number of minority hires (i.e., $\underline{N}*P_{min}*\pi_T$). As a result, the assumption of normality is justified when

$$N \geq \frac{5}{(P_{\min})(\pi_T)}. \quad \textbf{(24)}$$

Under some circumstances, the analysis produced estimates of the minimum $\underline{N}$ smaller than this value. However, because the sampling distribution would not be normal, these estimates may be inaccurate. In addition, the significance tests are not recommended under these conditions (Fleiss, 1981). Results below this level were marked in the tables by asterisks.

As shown in Table 1, extremely large samples were often required to achieve .8 power. In the worst condition, (10% selection, 10% minority), both significance tests required prohibitively large sample sizes, except for the extreme cases of adverse impact (ψ < .3). It is unlikely that any adverse impact study would have the 16,000 applicants required to detect ψ= .8. The four-fifths rule required substantially smaller samples then either significance tests, but the requisite $\underline{N}$ was still quite large unless ψ is less than .5.

The results were slightly better with 10% selection and 30% minorities. Studies with moderately large samples ($\underline{N}$=600) would have adequate power to detect large cases of adverse impact ($\psi < .4$) using the significance tests, although the sample sizes needed to detect low levels of adverse impact (.7-.8) are still prohibitively large (2800-7255). The four-fifths rule, on the other hand, required only moderate sample sizes ($\underline{N}$<485), except when $\psi > .7$.

With 30% selection and 30% minorities, the use of the significance tests appeared more reasonable. Although extremely large samples were required to detect small levels of adverse impact (ψ=.8), sample sizes around 400 were able to detect moderate levels of adverse impact (ψ=.6), and sample sizes less than 100 were sufficient to detect extreme cases of adverse impact (ψ<.3).

Of the two significance tests, Z_{IR} generally required smaller samples than Z_D. The difference was greatest when both the selection rate and proportion of minorities were small. With 10% selection and 10% minorities, the sample size required for the Z_{IR} test was between 7% and 32% smaller than that required for the Z_D test. The relative advantage of Z_{IR} decreased as either the selection rate or the proportion of minorities increased. With 50% selection and 50% minorities, there was little difference between the two statistics.

A similar pattern of results was found for the significance tests when the desired level of power was set at .6 (see Table 2). The significance tests still required sizable samples under many conditions, but the required sample size was reduced by 25-50%. In contrast, the four-fifths rule will have 60% power to detect adverse impact under almost all conditions, even with small sample sizes.

Given the inadequate power of the significance tests under many of the conditions, alternative procedures with greater power are needed. One possibility would be to adopt a more liberal significance criterion. Since the tests are often used to identify adverse impact against a particular group, the use of a two-tailed significance test is overly conservative. In such cases, a one-tailed test ($\underline{Z}_{crit}$= 1.645) would provide greater power while still controlling for Type I error. As shown in Tables 3 and 4, the use of a one-tailed test reduced the requisite sample size by about 20%. However, extremely large samples were still needed under many conditions.

Under some conditions, the sample size required for adequate statistical power was too small to meet the assumptions of the normal approximation (i.e., expected frequencies less than five). Under these conditions (indicated by asterisks in the tables), the estimated sample sizes may not be accurate, and the significance tests should not be used. In such cases, a more conservative estimate would be the minimum sample size required for the normal approximation, for which power would be equal to or greater than the level indicated in the table. These values are listed in parentheses for each condition.

A difficulty in the use of Tables 1 through 4 is that they require knowledge of the population adverse impact ratio (ψ). Although other population parameters, such as group mean differences in test scores, are well documented (Bobko et al, 1999; Neisser et al., 1996), population impact ratios are not often discussed. In order to better understand the practical implication of these results, a second analysis was conducted to assess the requisite sample size as a function of the standardized mean difference between groups on the predictor.

Table 1
Total Sample Size Required to Test for Adverse Impact with .8 Power (Two-tailed α=.05)

Impact Ratio	10% Minority			30% Minority			50% Minority		
	Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths
				Selection Rate = .1					
0.8	17982	16782	**	7256	6987	**	5720	5699	**
0.7	7620	6839	5492	2980	2801	2049	2267	2245	1482
0.6	4073	3503	1366	1543	1409	486	1128	1107	331
0.5	2467	2025	608	905	797	206	634	613	132
0.4	1613	1257	346*	572	483	111*	383	361	67*
0.3	1108	817	228*	381	305	70*	242	220	39*
0.2	785	547	170*	262	196	49*	157	136	26*
0.1	566	382*	149*	185	129*	41*	104	83*	20*
	(Normal approximation assumes N>500)			(Normal approximation assumes N>167)			(Normal approximation assumes N>100)		
				Selection Rate = .3					
0.8	4727	4423	**	1894	1829	**	1482	1480	**
0.7	2015	1820	1537	780	738	557	586	584	388
0.6	1082	942	392	405	374	135	291	289	88
0.5	658	551	179	238	214	59	163	161	36
0.4	431	347	105*	150	131	33*	98	96	19*
0.3	296	229	71*	100	84	21*	61	59	11*
0.2	210	157*	54*	68	55*	16*	39	37	8*
0.1	150*	113*	49*	48*	38*	14*	25*	24*	6*
	(Normal approximation assumes N>167)			(Normal approximation assumes N>56)			(Normal approximation assumes N>33)		
				Selection Rate = .5					
0.8	2075	1951	**	822	797	**	634	636	**
0.7	893	815	746	340	325	259	250	252	170
0.6	483	429	198	177	167	65	124	126	39
0.5	295	255	93*	104	97	29*	69	71	17*
0.4	194	164	56*	66	61	17*	41	43	9 *
0.3	134	111	39*	44	40	12*	25	27	6 *
0.2	94*	78*	31*	30*	27*	9*	16*	17 *	4 *
0.1	66*	59*	29*	20*	19*	8*	9*	12 *	4 *
	(Normal approximation assumes N>100)			(Normal approximation assumes N>33)			(Normal approximation assumes N>20)		

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .8 power for any sample size large enough to meet the assumptions of the normal approximation. **Power for the 4/5ths rule is always .5 when IR = .8.

Table 2
Total Sample Size Required to Test for Adverse Impact with .6 Power (Two-tailed α=.05)

Impact Ratio	10% Minority			30% Minority			50% Minority		
	Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths
				Selection Rate = .1					
0.8	11556	10079	**	4596	4272	**	3571	3548	**
0.7	4981	4008	498*	1905	1688	186	1416	1392	135
0.6	2714	1993	124*	996	833	44*	705	682	30*
0.5	1680	1110	56*	591	460	19*	397	373	12*
0.4	1126	657	32*	379	270	11*	240	216	7*
0.3	797	400*	21*	256	162*	7*	152	128	4*
0.2	586	242*	16*	180	96*	5*	99*	75*	3*
0.1	442	140*	14*	130*	54*	4*	66*	40*	2*
	(Normal approximation assumes N>500)			(Normal approximation assumes N>167)			(Normal approximation assumes N>100)		
				Selection Rate = .3					
0.8	3012	2630	**	1195	1112	**	926	921	**
0.7	1301	1050	140*	496	441	51*	367	362	36
0.6	710	525	36*	260	218	13*	183	178	8*
0.5	440	294	17*	154	121	6*	103	98	4*
0.4	295	175	10*	99	71	3*	62	57	2*
0.3	209	108*	7*	67	43*	2*	39	34	1*
0.2	154*	66*	5*	47*	26*	2*	26*	20*	1*
0.1	116*	39*	5*	34*	15*	2*	17*	11*	1*
	(Normal approximation assumes N>167)			(Normal approximation assumes N>56)			(Normal approximation assumes N>33)		
				Selection Rate = .5					
0.8	1303	1140**		515	480	**	397	395	**
0.7	565	458	68*	214	191	24*	157	156	16*
0.6	309	231	18*	112	95	6*	78	77	4*
0.5	192	130	9*	67	53	3*	44	42	2*
0.4	129	79*	6*	43	32*	2*	27	25	1*
0.3	91*	49*	4*	29*	20*	2*	17*	15*	1*
0.2	67*	31*	3*	20*	12*	1*	11*	9*	1*
0.1	50*	19*	3*	15*	7*	1*	7*	5*	1*
	(Normal approximation assumes N>100)			(Normal approximation assumes N>33)			(Normal approximation assumes N>20)		

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .6 power for any sample size large enough to meet the assumptions of the normal approximation. **Power for the 4/5ths rule is always .5 when IR = .8.

Table 3
Total Sample Size Required to Test for Adverse Impact with.8 Power (one-tailed α=.05)

Impact Ratio	10% Minority		30% Minority		50% Minority	
	Z_D	Z_{IR}	Z_D	Z_{IR}	Z_D	Z_{IR}
			Selection Rate = .1			
0.8	14081	13324	5699	5528	4506	4492
0.7	5945	5456	2336	2223	1785	1771
0.6	3165	2811	1207	1122	889	875
0.5	1908	1636	706	638	499	485
0.4	1241	1025	446	389	301	287
0.3	847	674	296	248	190	176
0.2	596	460*	203	162*	124	110
0.1	425*	330*	142*	110*	82*	68*
	(Normal approximation assumes N>500)		(Normal approximation assumes N>167)		(Normal approximation assumes N>100)	
			Selection Rate = .3			
0.8	3708	3519	1489	1448	1167	1166
0.7	1576	1456	613	586	462	461
0.6	844	759	317	299	229	229
0.5	511	448	186	172	128	128
0.4	333	285	117	106	77	76
0.3	228	191	78	69	48	48
0.2	160*	133	53*	46*	31*	30*
0.1	113*	99*	37*	32*	20*	20*
	(Normal approximation assumes N>167)		(Normal approximation assumes N>56)		(Normal approximation assumes N>33)	
			Selection Rate = .5			
0.8	1633	1557	647	632	499	501
0.7	702	656	268	259	197	199
0.6	379	348	139	134	97	99
0.5	231	209	82	78	54	56
0.4	151	136	52	49	32	34
0.3	103	93*	34	33	20	22
0.2	72*	67*	23*	23*	12*	14*
0.1	50*	52*	16*	17*	7*	10*
	(Normal approximation assumes N>100)		(Normal approximation assumes N>33)		(Normal approximation assumes N>20)	

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .8 power for any sample size large enough to meet the assumptions of the normal approximation.

Table 4
Total Sample Size Required to Test for Adverse Impact with .6 Power (one-tailed α=.05)

	10% Minority		30% Minority		50% Minority	
IR	Z_D	Z_{IR}	Z_D	Z_{IR}	Z_D	Z_{IR}
			Selection Rate = .1			
0.8	8476	7444	3376	3149	2627	2611
0.7	3647	2968	1398	1246	1041	1025
0.6	1983	1480	730	616	519	503
0.5	1225	828	433	341	292	275
0.4	819	493*	277	201	177	160
0.3	578	302*	187	121*	112	95*
0.2	423*	185*	131*	73*	73*	56*
0.1	318*	110*	94*	42*	48*	30*
	(Normal approximation assumes N>500)		(Normal approximation assumes N>167)		(Normal approximation assumes N>100)	
			Selection Rate = .3			
0.8	2211	1945	878	820	681	678
0.7	954	779	364	326	270	266
0.6	520	391	191	162	135	131
0.5	322	220	113	90	76	72
0.4	215	132*	73	53*	46	42
0.3	152*	82*	49*	33*	29*	25*
0.2	111*	51*	34*	20*	19*	15*
0.1	84*	31*	25*	12*	13*	9*
	(Normal approximation assumes N>167)		(Normal approximation assumes N>56)		(Normal approximation assumes N>33)	
			Selection Rate = .5			
0.8	958	845	379	354	292	291
0.7	415	341	158	142	116	115
0.6	227	172	83	71	58	57
0.5	141	98*	49	40	32	31
0.4	95*	60*	32*	24*	20	19*
0.3	67*	38*	21*	15*	12*	11*
0.2	49*	24*	15*	9*	8*	7*
0.1	37*	15*	11*	6*	5*	4*
	(Normal approximation assumes N>100)		(Normal approximation assumes N>33)		(Normal approximation assumes N>20)	

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .6 power for any sample size large enough to meet the assumptions of the normal approximation.

Table 5
Total Sample Size Required to Test for Adverse Impact with .8 Power as a Function of the Standardized Mean Difference Between Groups (Two-tailed α=.05)

Std. Mean Difference	Impact Ratio	10% Minority			30% Minority			50% Minority		
		Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths	Z_D	Z_{IR}	4/5ths
Majority Selection Rate = .1										
0.5	0.374	1564	1202	328*	644	536	121*	511	476	86*
1	0.113	653	439*	164*	270	188	57*	208	163	36*
		(Normal approximation assumes N>500)			(Normal approximation assumes N>167)			(Normal approximation assumes N>100)		
Majority Selection Rate = .3										
0.5	0.509	736	617	202	309	277	76	252	247	55
1	0.212	244	182	61*	103	81	22*	82	73	14*
		(Normal approximation assumes N>167)			(Normal approximation assumes N>56)			(Normal approximation assumes N>33)		
Majority Selection Rate = .5										
0.5	0.617	571	508	248	245	230	94	204	205	69
1	0.317	161	133	45*	70	61	16*	58	56	11*
		(Normal approximation assumes N>100)			(Normal approximation assumes N>33)			(Normal approximation assumes N>20)		

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .8 power for any sample size large enough to meet the assumptions of the normal approximation.

Table 6
Total Sample Size Required to Test for Adverse Impact with .8 Power as a Function of the Standardized Mean Difference Between Groups (One-tailed α=.05).

Std. Mean Difference	Impact Ratio	10% Minority			30% Minority			50% Minority		
		Z_D	Z_{IR}	45ths	Z_D	Z_{IR}	45ths	Z_D	Z_{IR}	45ths
Majority Selection Rate = .1										
0.5	0.374	1202	982	328*	501	432	121*	402	379	86*
1	0.113	491*	377*	164*	208	159*	57*	163	134	36*
		(Normal approximation assumes N>500)			(Normal approximation assumes N>167)			(Normal approximation assumes N>100)		
Majority Selection Rate = .3										
0.5	0.509	571	501	202	242	223	76	198	196	55
1	0.212	186	153*	61*	80	67	22*	64	59	14*
		(Normal approximation assumes N>167)			(Normal approximation assumes N>56)			(Normal approximation assumes N>33)		
Majority Selection Rate = .5										
0.5	0.617	447	411	248	192	184	94	161	162	69
1	0.317	125	111	45*	55	50	16*	45	45	11*
		(Normal approximation assumes N>100)			(Normal approximation assumes N>33)			(Normal approximation assumes N>20)		

*The requisite sample size was too small to be estimated precisely using the normal approximation. The tests would have greater than .8 power for any sample size large enough to meet the assumptions of the normal approximation.

Given certain assumptions about the distribution of scores in the population, it is possible to translate the standardized mean difference between groups into an adverse impact ratio (Sackett & Ellingson, 1997). The population selection rate for a particular group is the proportion of scores on a test above a specified cut score. If the shape and location of the population distribution are known, the selection rate can be expressed as an area under this distribution. In the following analysis, it was assumed that both minority and majority populations are normally distributed with equal variances. The standardized cut score is defined as the difference between the raw cut score and the group mean, divided by the standard deviation. The minority standardized cut score can be written as the majority standardized cut score (Z_{cut}) plus a constant representing the standardized mean difference between groups (Z_{Diff}). The selection rate for each group would be the area under the standard normal curve above the standardized cut score. Thus, the impact ratio would be

$$\psi = \frac{\int_{Z_{cut}+Z_{Diff}}^{\infty} f(x)\,dx}{\int_{Z_{cut}}^{\infty} f(x)\,dx}, \qquad \textbf{(25)}$$

where *f(x)* is the standard normal density function.

For the second analysis, the standardized mean difference (Z_D) was set at 0.5 and 1.0. The majority group selection rate was set at .1, .3 and .5, which were translated into standardized scores (Z_{cut}) of 1.28, .52, and 0, respectively. The proportion of minority applicants was set at .1, .3 and .5. The significance criterion was set at Z_{crit} = 1.96 and 1.645. For each combination of the parameters, the population impact ratio was computed using Equation 25. The integral of the normal distribution was approximated using the *cnorm* function in MathCad (MathSoft, 1997). This resulting impact ratio was then used in Equations 18, 20 and 22 to compute the minimum sample sizes required for each test.

Table 5 presents the sample sizes needed for .8 power using a two-tailed α=.05. The results for the one-tailed test are presented in Table 6. The pattern of results is similar to that found in the previous analysis. Larger sample sizes are needed when the proportion of minorities in the applicant pool is small, and the selection rate is low. In addition, the Z_{IR} test required slightly lower sample sizes than the Z_D test, and the four-fifths rule required substantially lower sample sizes than either of the significance tests.

In contrast to the first analysis, however, the sample sizes under many conditions are substantially more feasible than the worst cases from the previous analysis. This is because only moderate to large effect sizes were considered. Further, the impact ratios tended to be smaller, and therefore differ more from the null hypothesis, when the selection rate was small. Since larger effect sizes occurred under the conditions with the lowest power, extremely large sample sizes were not required.

When the difference between groups was large (one standard deviation), moderate samples generally had adequate power to detect adverse impact, even when using the more conservative two-tailed significance test (see Table 5). In the worst case, when the selection rate was 10% and

minorities comprised 10% of the applicant pool, a sample size of 655 was adequate to achieve 80% power using the Z_D test. The requisite sample size dropped to 270 or less if either the selection rate or the proportion of minorities is 30%. In the optimal case (50% selection, 50% minority applicants), only 58 applicants were required.

When the difference between groups was moderate (one-half standard deviation), considerable sample sizes were required under some conditions. For example, when the selection rate was 10% and 10% of the applicants were minorities, the two-tailed Z_D test required 1564 applicants in order to test for adverse impact with adequate power. The Z_{IR} test was slightly better, requiring only 1202 applicants. With 10% selection and 30% minority applicants, the required sample size dropped to 644 for the Z_D test and 536 for the Z_{IR} test. Under optimal conditions (50% selection, 50% minorities), the required sample size was only around 205 for either test.

Conclusion

When designing or evaluating a selection procedure, it is common to consider adverse impact along with test validity. Recent research suggests that many common selection practices will tend to result in adverse impact against minority groups (Bobko, ,Roth, & Potosky, 1999; Ryan, Ployhart, & Friedel, 1998; Sackett & Ellingson, 1997; Schmitt, Rogers, Chan, Sheppard & Jennings, 1997). However, for practitioners, it is more important to determine whether the combination of selection methods and decision rules applied in a particular organization will produce adverse impact in the relevant applicant pool. As a result, adverse impact analysis is typically based on sample statistics. Before making use of adverse impact statistics, it is important to consider the adequacy of the samples from which these statistics were derived.

In many selection situations, the organization only selects a small proportion of the applicant pool. In addition, the proportion of applicants who are minorities is often small. Under these common conditions, extremely large sample sizes are needed to detect adverse impact with adequate power. This raises questions about the reliance on sample-based adverse impact statistics in the evaluation and development of selection procedures. Given the instability of these statistics across samples, decisions based on adverse impact analysis will be equally unstable.

The requisite sample size is largely a function of the magnitude of the difference in selection rates. Given the wide acceptance of the four-fifths rule, the minimum level of adverse impact which is of practical importance would be .8. When the selection rate and the proportion of minority applicants are small, even large test administrations (N>1000) would not be sufficient to insure adequate power. On the other hand, when the selection rate and the proportion of minorities approach .5, moderate samples sizes should suffice. For example, a sample of 635 would allow .8 power for either significance test when the selection rate and proportion of minorities are both .5.

Commonly used selection tests often produce large group differences. For example, on general cognitive ability tests, Whites tend to score about one standard deviation above Blacks (Bobko et al., 1999; Neisser et al., 1996). Given the widespread use of cognitively loaded selection tests, the difference in selection rates may be substantially larger than the minimum value suggested by the four-fifths rule.

For the conditions examined in this study, a group difference of one standard deviation in test scores would result in adverse impact ratios ranging from .11 to .32. In addition, the smaller impact ratios tend to occur under conditions with lower power. As a result, the requisite sample size would

only be 653 in the worst case (10% selection, 10% minority applicants), and as low as 56 under the best conditions (50% selection, 50% minority applicants).

When the group difference is more moderate (one-half standard deviation), samples over 1000 would be required when both the selection rate and the proportion of minorities are small. However, sample sizes ranging from 200 to 750 would be adequate under other conditions.

Of the two significance tests, the Z_{IR} test consistently required smaller sample sizes; although the advantage was fairly small relative to the magnitude of the required sample size. The four-fifths rule was found to require substantially smaller samples sizes than either significance test. Unfortunately, the four-fifths rule is also likely to indicate adverse impact when none exists in the population (Boardman, 1979; Greenberg, 1979). Due to this trade-off, neither approach is ideal.

One way to increase statistical power is to conduct one-tailed significance tests. Since adverse impact analysis is generally used to evaluate discrimination against a particular group, a directional hypothesis is appropriate. The use of a one-tailed test was found to lower the required sample size by around 20%. However, large samples would still be required under many conditions.

An alternate approach would be to avoid null hypothesis testing altogether. Instead, researchers could report the impact ratio along with a confidence interval (Morris & Lobsenz, 2000). This approach does not avoid the problem of unstable estimates, but it does more clearly convey the ambiguous nature of the statistical results. In many cases, the impact ratio will be lower than .8, indicating adverse impact according to the four-fifths rule, but the confidence interval will include 1, suggesting that the difference in selection rates could be due to sampling error. While such results are more complex than a dichotomous decision based on significance tests, they are also more informative than simply concluding that the difference is non-significant.

A limitation of the current research was the use of the normal approximation of the sampling distribution, which is appropriate only for large samples large samples (Fleiss, 1994). In some cases, the results suggested a minimum sample size that was smaller than required for the normal approximation. In these cases, the estimates of the required sample sizes may not be accurate. However, the sample size required for the normal approximation, which would have more than sufficient power in these cases, was usually fairly small. Therefore, in most cases where the estimates are questionable, modest sample sizes would have adequate power. In addition, the significance tests, which are also based on the normal approximation, are not recommended under these conditions. As a result, the lack of precise estimates for these small sample sizes is not of great concern. Future research should explore power and sample size requirements under these conditions, and should incorporate alternate test procedures which are recommended for use in small samples, such as the Fisher Exact Test.

References

Boardman, A. E. (1979). Another analysis of the EEOC 'four fifths' rule. *Management Science, 25*, 770-776.

Bobko, P., Roth, P. L., & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1003.

De Corte, W. (1999). Weighing job performance predictors to both maximize the quality of the selected workforce and control the level of adverse impact. *Journal of Applied Psychology, 84*, 695-702.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). NY: Wiley.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 245-260). NY: Russell Sage Foundation.

Greenberg, I. (1979). An analysis of the EEOC "four-fifths rule," *Management Science, 25*, 762-769.

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

Gutman, A. (2000). *EEO law and personnel practice*. Newbury Park, CA: Sage.

Hazelwood School District v. United States, 433 U.S. 299 (1977).

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hoffman, C. C., & Thornton, G. C., III (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*, 455-470.

Lawshe, C. H. (1987). Adverse impact: Is it a viable concept? *Professional Psychology: Research and Practice, 18*, 492-497.

Mathsoft, Inc. (1997). *MathCad user's guide* (Version 7) [Computer software manual]. Cambridge, MA: Author.

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

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. Wade, Ceci, S. J., Halpern, Diane F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Office of Federal Contract Compliance Programs (1993). *Federal contract compliance manual*. Washington, D.C.: Department of Labor, Employment Standards Administration, Office of Federal Contract Compliance Programs (SUDOC# L 36.8: C 76/993).

Rich v. Martin-Marietta, 467 F. Supp. 587 (D. Col. 1979).

Ryan, A. M., Ployhart, R. E., & Friedel, L. A. (1998). Using personality testing to reduce adverse impact: A cautionary note. *Journal of Applied Psychology, 83*, 298-307.

Sackett, P. R., & Ellingson, J. E. (1997). The effects of forming multi-predictor composites on group differences and adverse impact. *Personnel Psychology, 50*, 707-721.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L, Hunter, J. E, &Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*, 473-485.

Schmitt, N., Rogers, W., Chan, D., Sheppard, L., & Jennings, D. (1997). Adverse impact and predictive efficiency of various predictor combinations. *Journal of Applied Psychology, 82*, 719-730.

U.S. Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43*, 38290-38315.

Author Notes

1. I would like to thank Nambury Raju for his helpful comments on this paper.

2. Portions of the manuscript were presented at the 14th annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA (April, 1999).

3. Correspondence regarding this article should be sent to Scott B. Morris, Institute of Psychology, Illinois Institute of Technology, 3101 S. Dearborn, Chicago, IL 60616. Electronic mail may be sent to scott.morris@iit.edu. Phone: 312-567-5932. Fax: 312-567-3932.