# How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing

By Richard E. Biddle*
(*Public Personnel Management*, Vol. 22 No.1 [Spring 1993])**

Why not set the cutoff score at 70 percent? That score has been used as passing throughout elementary school, junior high school, high school, and college. Most people are familiar with a 70 percent cutoff because of our experiences in the school system. If a score has been used for so long as a standard, why change it? Why challenge it? I have heard judges, administrators, attorneys, and analysts ask these questions.

The specific score that is used as the cutoff is what separates those who pass a test from those who do not. It is this score that determines the consequences of taking the test. Those who take a test and do not reach the cutoff score are not considered further for promotion. Or if the test is the conclusion to a training course, by missing the cutoff score, the person may have to take the training again. If the test is for certification, scoring below the cutoff means having to try again for certification. Failing the licensing test means having to wait another six months or a year to take the test again, and, perhaps, not getting a substantial pay increase.

The time that an unsuccessful test taker has to wait before taking a test again will vary. Two years is a common time period between test administrations when litigation is not involved. However, when litigation has been involved, hundreds of employees working for one employer waited 11 years between promotional tests. Hundreds of other employees have had to wait 6 years between promotional tests due to delays resulting from litigation.

A line must be drawn somewhere to distinguish between those who possess enough knowledge to pass a training course, to be considered further for promotion, to be certified, or to be licensed. It is the specific score called the cutoff score that creates the two classes of people: those who pass and those who fail. The group who passes rarely sues. Litigation comes from the group failing a test. But not for the litigants, 70 percent might still be a universally acceptable cutoff score for promotional tests, training tests, certification tests, and tests used for licensing.

Litigation, however, requires responses to penetrating questions directed to the person in charge of the test. These questions come in the form of interrogatories (written questions from the opposing party's attorney that must be answered under penalty of perjury), at depositions (a sit down session in a private office after receiving a subpoena where the opposing party's attorney verbally asks questions and a court reporter carefully takes down the reply), and court testimony where the opposing party's attorney cross examines the person testifying (asks questions of a party under oath in a court room after direct testimony has been given by that party). Questions asked will deal with the job analysis, test construction, validation procedures, and how the cutoff was determined.

If a minimum cutoff score is to be set, it makes sense to gather data needed to set the score in a defensible way and to consider the factors that incite litigation. See Cascio (1988) for a discussion of the factors. For a discussion of how the burdens are followed in court cases after the United States Supreme Court decision of *Wards Cove Packing v. Atonio,* see Biddle (1989).

The purpose of this paper is to suggest a process for setting a cutoff on knowledge tests used for promotion, training, certification, and licensing. The suggested cutoff setting process incorporates the advantages of the job related process reviewed by the United States Supreme Court, adds some job related features to it, then combines the modified job related process with a distribution-wide adverse impact analysis. The process described in this paper starts immediately *after* the job analysis, test specification, and test development work have been completed.

## Job Related Cutoff Setting Process

### Uniform Guidelines Requirement

The *Uniform Guidelines on Employee Selection Procedures* (1978) give us only vague guidance, stating that when setting cutoff scores, they should "normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force." See *Guidelines* (1978) Section 5H.

What process can be used to identify "normal expectations of acceptable proficiency within the work force"? Those of us working in the testing and selection field look to the profession and find different practices. Then we look to court cases to see what the courts have said about some of the profession's practices. We cannot work in a vacuum. We work in a hybrid field: half from the testing profession and half from what the courts say about our practices. If the profession thinks a method is great, but the courts have said it is unacceptable or has certain flaws, we need to rethink our method. Conversely, when the courts have reviewed a situation involving a practice of the profession and the employer won with the practice, then it makes sense to replicate that practice. If we are challenged about a process that has won before, our chances are substantially increased of winning again. After the practice has won in court, then we consider the level of the decision. A Federal District Court decision can be cited anywhere as precedent, but is not necessarily controlling on the next Federal Court. A Circuit Court decision is controlling to all the Federal District Courts within its boundary. However, taken outside the boundary of the Circuit, the decision can be cited as precedent, but is not controlling. When the United States Supreme Court selects one to three cases for review out of 100 sent to it, those few cases make up the precedence for all the Circuits and all the Federal District Courts as well as the state courts.

The case to be used as the foundation for the model presented in this paper and which answers the question as to how "normal expectations of acceptable proficiency"

are to be established has been reviewed by the United States Supreme Court. The practice used for establishing a minimum proficiency was one derived in the profession but the application was modified by a state board. It is *the application of the modified method* that received acceptance before the United States Supreme Court. The method was called the Angoff Method. It produces an average estimate of minimum competency using several Subject Matter Experts (incumbents, supervisors or trainers who can competently perform the duties for which the knowledge tested is needed). The modification lowered the Angoff average estimate by one, two, or three standard errors of measurement after consideration of several statistical and human factors. The standard error of measurement is designed for interpreting the reliability of test scores. It is a statistic expressed in test score units but derived from the reliability of the test. Differences from the average score and those who scored within the standard error of measurement can be attributed to chance.

## Unmodified Angoff Method

In a 93-page chapter in Thorndike's *Educational Measurement* titled "Scales, Norms, and Equivalent Scores," Angoff (1971) devotes one paragraph and one footnote to the process that has become the foundation for the method approved by the U.S. Supreme Court:

> "A systematic procedure for deciding on the minimum raw scores for passing and honors might be developed as follows: keeping the hypothetical 'minimally acceptable person' in mind, one could go through the test item by item and decide whether each such person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score earned by the 'minimally acceptable person.' A similar procedure could be followed for the hypothetical 'lowest honors person.'"

It is the footnote to this paragraph that describes the process followed frequently in the field:

> "A slight variation to this procedure is to ask each judge to state the *probability* that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons, instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. A parallel procedure of course, would be followed for the lowest honors score."

It has been reported that Angoff attributes the process described above to Ledyard Tucker (Smith, 1988). Regardless of who developed the method, on its face it addresses the *Uniform Guidelines'* requirement that cutoffs be set to be consistent with "normal expectations of acceptable proficiency within the work force." When the United States Supreme Court had the opportunity to review the process described above, modifications had been made to it. How this process was applied and accepted is critical to its successful replication.

The process of gathering opinions from Subject Matter Experts, adding up those opinions, and computing an average is called the *un*modified Angoff.

## Number of Subject Matter Experts Needed for the Angoff

How many Subject Matter Experts are required to serve as "judges"? Two court cases have answered this question for us. In *Contreras v. City of Los Angeles* (1981) seven Subject Matter Experts were used to give input on the job relatedness of a test and its items. In *U.S. v. South Carolina* (1978) ten Subject Matter Experts were used to review the test items and another ten Subject Matter Experts were used to give the Angoff estimates. Seven to ten Subject Matter Experts appears to be enough for the sampling process.

## Modifications Needed to the Angoff

*Agreement on Job Relatedness Modification.* It is important to note that not all the Subject Matter Experts need to agree on job relatedness judgments. In *South Carolina* only five of the ten (50%) Subject Matter Experts were needed. In *Contreras* five of seven (71%) Subject Matter Experts were needed for job relatedness decisions. Both employers were successful in their defenses with the different standards. However, the *South Carolina* case set the standard in 1978.

*Contreras* in 1981 used a higher percentage than *South Carolina* after *South Carolina* set the minimum in 1978. At least 50 percent need to agree on the job relatedness of a test item to include that item in the final pool of items. A preferred modification would be to reach this level of agreement (50% to 70% of the Subject Matter Experts) when identifying a duty from the job analysis for which knowledge measured by the test item is needed to competently perform the duty.

*Consequences Modification.* To defend test items as job related, two types of questions can be asked by methodology experts. One deals with the ease with which one could look up the answer to the item. If the answer to a test item can quickly be looked up in the normal flow of performing the job duties, the test item might be successfully challenged by plaintiffs. Another deals with the consequences of not knowing the answer

to the item. If there is no consequence when a person performing the duty does not know the answer to the item, again plaintiffs might successfully challenge the item. The more items successfully challenged on a test, the weaker the evidence of job relatedness.

Identifying what is likely to happen when information measured by the test item is not known will help document job relatedness of the item. A scale can be developed and used by Subject Matter Experts to identify levels of consequences for not knowing information measured by the test item.

*Differentiating Modification.* In order to establish a cutoff higher than even the *un*modified Angoff or to use a test for ranking, information on the differentiating nature of the test items is needed. A majority of the Subject Matter Experts should agree that a test item measures a knowledge, skill, or ability that differentiates in levels of duty performance. When a test is made up of test items that differentiate job performance, the test results can be used to rank candidates or to set higher than minimum competency cutoffs. See Section 14C(9) of the Uniform Guidelines.

*Standard Error of Measurement Modification.* An important modification to the Angoff method was recognized in the U.S. Supreme Court decision *U.S. v. South Carolina* (1978). It was not the Angoff method resulting in an average estimate of the Subject Matter Expert opinions on the test items that won. Nineteen tests with cutoffs were reviewed in the *South Carolina* decision. All nineteen won. *None of the nineteen* applied the Angoff *un*modified. Each Angoff-derived average score was lowered by one, two, or three standard errors of measurement after the board in South Carolina responsible for setting cutoff scores on the teacher licensing tests considered several statistical and human factors.

One standard error of measurement is the result of multiplying the square root of one minus the reliability of the test times the standard deviation of the test. This information can be obtained only after the test has been administered. The reliability of the test is the measure of consistency of the test that varies from zero for an inconsistent test to one for a perfectly consistent test. The standard deviation of the test is a measure of dispersion of the test scores around the mean test score. Therefore, a very *inconsistent* test would have the standard error of measurement of the test equal to the standard deviation of the test. To the extent the test is reliable, the formula will result in a reduction of the standard deviation. For example, a reliability of .75 will result in a standard error of measurement 50 percent the size of the standard deviation. A .91 reliability will result in a standard error of measurement 30 percent the size of the standard deviation. A .99 reliability (almost perfect) will result in a standard error of measurement 10 percent the size of the standard deviation.

While the KR-20 (Kuder Richardson 20) formula is the most well known method for measuring internal consistency or reliability of a test, it assumes the test items are of equal difficulty. The Horst modification of the KR-20 removes this assumption. See Gilford (1973). Although many times the differences between the two calculated

estimates of reliability are slight, state of the art item analysis software will include the Horst modification for accuracy.

The decision to use one, two, or three standard errors of measurement below the Angoff average should be based upon a variety of statistical and human factors: the size of the standard error of measurement, risk of error (risk of excluding a truly qualified candidate whose low score does not show the real level of knowledge compared to the risk of including an unqualified candidate whose low score does show an unacceptable level of knowledge), internal consistency of the Angoff panel (e.g., taken individually, the subject matter experts vary in their individual estimates of minimum competency), supply and demand for at-issue jobs, and the sex and race/ethnic composition of the at-issue jobs in the work force.

No formula was presented in *U.S. v. South Carolina* outlining how to apply human and statistical factors in the decision to reduce the *un*modified Angoff by one, two, or three standard errors of measurement to obtain the *modified Angoff*. The case simply states the board considered the human and statistical factors, then decided to lower the Angoff average by one, two, or three standard errors of measurement for each of the 19 tests. The Supreme Court appears to have given the employer or board the flexibility to consider the human and statistical factors as they deem appropriate before selecting one, two, or three standard errors of measurement to make the *modified Angoff* cutoff.

## Adverse Impact

The *modified Angoff* score is the lowest score that should be considered in the cutoff setting process. It is not necessarily the cutoff score that should be used. Other scores above the *modified Angoff* might better serve the employer's or board's purpose. A score may exist above even the *un*modified Angoff that does not have adverse impact against any group protected by Title VII of the Civil Rights Act and offer enough candidates for consideration. A score may exist above the *modified Angoff* and within the standard error of measurement that is substantially equally valid to the *modified Angoff* score with less adverse impact. By blindly taking the *modified Angoff* score, the employer or board may be ignoring other scores offering enough candidates, a more highly qualified pool of candidates, and which reduce or eliminate adverse impact. Adverse impact is the trigger that sets off class action Title VII discrimination suits. Taking this trigger away can save the employer or board hundreds of thousands of dollars in defense costs and save test takers years of time waiting for the next chance to take the test.

### Uniform Guidelines Procedures - Rate Comparisons

When setting cutoff scores, the *Uniform Guidelines* requires the consideration of several factors other than cutoffs should “normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force.” See

*Guidelines* (1978) Section 5H. Sections 3B and 4D specify requirements for alternate use and adverse impact considerations.

Section 3B requires consideration of alternative tests (i.e., practices, procedures, and tests) that are substantially equally valid but with less adverse impact. While in many situations this might be called an eternal search for truth, some research has been conducted in this area. Many employers and test publishers are not willing to advertise adverse impact results. However, in one situation involving entry-level firefighter ability tests, many different tests were compared from several different test publishers. The results of the study showed the test preparation manual concept reduced adverse impact while showing very good validity. Some test preparation manual tests had better results than others. See Campbell (1982).

Section 3B also requires consideration of alternative uses of tests that are substantially equally valid with less adverse impact. "Alternate uses" can include applying different cutoff scores or different weights than originally set. Substantially equally valid could mean correlations that are not significantly different with criterion-related validity. With content validity, Subject Matter Experts can be asked for a range of opinions regarding weights they consider substantially equal. Test scores that fall within one standard error of measurement could be considered for this purpose as substantially equally valid. In Section V of the Uniform Guidelines it states:

> "The concept of validation as used in personnel psychology involves the establishment of the relationship between a test instrument or other selection procedure and performance on the job. Federal equal employment opportunity law has added a requirement to the process of validation. In conducting a validation study, the employer should consider available alternatives which will achieve its legitimate business purpose with lesser adverse impact. The employer cannot concentrate solely on establishing the validity of the instrument or procedure which it has been using in the past.
>
> This same principle of using the alternative with lesser adverse impact is applicable to the manner in which an employer uses a valid selection procedure."

The key words for employers and boards to consider from Section 3B of the *Uniform Guidelines* are those which state that when scores are found which are "*substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact."* When conducting a content validation study, it is clear we need to consider alternative weights and cutoffs that are substantially equally valid with less adverse impact. We will not know the adverse impact until after test administration. Therefore, it is important not to set cutoff scores or weights prior to the administration of a test. See *Bouman v. Pitchess* (1988) and *San Francisco Police Officers Association v. City and County of San Francisco* (1987).

Section 4D specifies the consideration of statistical and practical significance to rate differences (e.g., passing rates, hiring rates, promotion rates, retention rates, etc.). See *Uniform Guidelines,* Section 4D (1978). A quick reading of Section 4D appears to require the application of a 4/5ths or 80 Percent Rule first. But a careful reading of the section shows that with statistical and practical significance, adverse impact can exist regardless of the 80 Percent Rule conclusion. Also, an earlier section of the *Guidelines* clarifies the role of the 80 Percent Rule as a Rule of Thumb and specifically states it is not a legal definition of adverse impact. See *Guidelines* Section II (1978).

*Rate differences* involve a comparison of two groups (e.g., men and women or whites and blacks). These two groups are each divided into two groups (e.g., those passing and those failing or those hired and those not hired). Rate comparisons compare the rate of one group (e.g., the rate of men passing) to the rate of another group (e.g., the rate of women passing).

*The 80 Percent Rule of Thumb* takes the rate of the group with the highest rate and puts that rate in the denominator of a fraction. The numerator is the rate of the comparison group. For example, if the group with the highest rate is Asians at .50 and the rate of the comparison group (e.g., Hispanics) is .30, then the .60 resulting (.30/.50=.60) is a violation of the 80 Percent Rule of Thumb. However, while this "Rule of Thumb" is easy to learn and apply, there are several reasons it cannot be a full definition of adverse impact. First, the *Guidelines* specifically say so. See *Guidelines* Section II (1978). The *Guidelines* also state in another section (4D) that although there may be an 80 Percent violation, this might not be adverse impact if the differences are not statistically and practically significantly different. See *Guidelines* Section 4D. That same section of the *Guidelines* states that even without an 80 Percent violation there might be adverse impact if the rate differences are statistically and practically different.

Rate differences made without considering the actual numbers can be very misleading. In our example above, if the Asians' 50 percent was derived from two Asians taking a test and one passing, you can see the rate is unstable. One person changing places from passing to failing changes the rate by 50 percent. The single biggest problem with the 80 Percent Rule of Thumb is that it has no probability distribution to it. When differences occur, we do not know from computing the 80 percent test the probability that the differences occurred by chance and chance alone. Ironically, with all of these problems the 80 Percent Rule of Thumb still has a major role in adverse impact determination- that role is in practical significance evaluation, discussed below.

*Statistical significance with rate differences* involves calculations with a hypergeometric approach, also called a two-sample approach. We call it a *Guidelines* approach because rate comparisons are called for in the *Guidelines.* See *Guidelines* Section 4D (1978). The fast way to make this calculation is with a chi-square formula. The square root of the chi-square result, when comparing two groups with a passing/failing type approach, results in a standard deviation. In the U.S. Supreme Court decision of *Hazelwood School District v. United States,* the level set for establishing statistical significance was between 2 or 3 standard deviations. See *Hazelwood* (1977)

and Technical Note 1 at the end of this paper. It is important to note that the U.S. Supreme Court has set the level of statistical significance in terms of a minimum number of standard deviations, and not a minimum probability level. See Technical Note 3 at the end of this paper.

*Practical significance with rate differences* involves at least three calculations. Each of these calculations involves the effects of small number changes on other statistics. How many more people need to be added to the disadvantaged group's passing number to (1) change the statistical significance conclusion, (2) change the 80 Percent Rule of Thumb conclusion, or (3) change the selection rates themselves from being different to being the same or very close to being the same. The court noted in *U.S. v. Commonwealth of Virginia* (1978) that by adding two more to the passing numbers in the plaintiff group, the statistical conclusion would be altered. In the *Contreras v. City of Los Angeles* decision, the court noted that with three more added to the plaintiff group, the 80 Percent Rule would be altered, and four more people added to the plaintiff group would bring the selection rates very close to one another. See *Contreras* (1981). Statistical differences that can be altered with very few number changes are not practically significant, and, therefore, do not create adverse impact.

*One way to plot rate differences for a distribution of scores* can be seen on the chart that follows. A chart showing the adverse impact graphically at each score in the distribution or showing the adverse impact in some easy to read way within a range of scores makes the final steps in setting a cutoff more manageable. At each score in the following example, the group with the highest rate is shown with an asterisk (*). The colons (:) show statistical significance between the rates with the hypergeometric probability. The exclamation point (!) shows 80 Percent Rule of Thumb differences in rates. The next several columns show the numbers of people who need to pass for that group to eliminate the statistically significant differences (VIR as the symbol for the *Virginia* reference), to eliminate the 80 Percent Rule of Thumb (80%) conclusion, and to bring the selection rate differences as close as they were in *Contreras* (SRD for selection rate differences). By spotting the colons (:) and reading the numbers in the practical significance columns, the employer or board can quickly find the zones with no adverse impact. Scores with no adverse impact above the *un*modified Angoff can be explored first to see if the score allows enough candidates to pass. Next, scores within the range from the *un*modified Angoff through the number of standard errors of measurement selected by the employer or board for the *modified Angoff* can be explored for no adverse impact. If all the scores have adverse impact, then the employer or board can see if there are scores that may minimize adverse impact between the *modified Angoff* and the *un*modified Angoff. Minimizing adverse impact could mean a higher proportion of an underutilized protected group who pass or one of the underutilized protected groups adversely impacted will no longer be adversely impacted at the alternative score.

***STATISTICAL CUTOFF ANALYSIS*** VERSION 6.0
PROGRAM BY BIDDLE & ASSOCIATES, INC.

TEST NAME: CUTOFF EXAMPLE
TEST DATE: 5/21/88
TEST ITEMS: 120
FILE NAME: COEXAMPL

*=GROUP WITH HIGHEST SELECTION RATE (5 OR MORE PASSING)
!=80% RULE OF THUMB VIOLATION FOUND
:=STATISTICAL SIGNIFICANCE VIOLATION FOUND
VIR=NUMBER NEEDED TO ELIMINATE STATISTICAL SIGNIFICANCE
80%=NUMBER NEEDED TO ELIMINATE 80% RULE OF THUMB VIOLATION
SRD=NUMBER NEEDED TO BRING SELECTION RATES CLOSE TO THE SAME

| | STATISTICAL SIGNIFICANCE | | | | | | | | PRACTICAL SIGNIFICANCE | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SEX | | | RACE/ETHNIC | | | | | SEX | | | | | | RACE/ETHNIC | | | | | | | | | | | | | | | | | |
| TEST | TOTAL | | | | | | | | MEN | | | | WOMEN | | | WHITE | | | BLACK | | | HISPAN | | | ASIAN | | | A.IND | | |
| SCORE | PASSED | MN | WN | WH | BL | HS | AS | AI | VIR | 80% | SRD | VIR | 80% | SRD | VIR | 80% | SRD | VIR | 80% | SRD | VIR | 80% | SRD | VIR | 80% | SRD | VIR | 80% | SRD |
| 90 | 6 | * | ! | * | ! | | | ! | | | | | 1 | 1 | | | | | 1 | 1 | | | | | | | | 1 | 1 |
| 89 | 14 | * | ! | * | ! | | | ! | | | | | 1 | 1 | | | | | 1 | 1 | | | | | | | | 1 | 1 |
| 88 | 15 | * | ! | * | ! | | | ! | | | | | 1 | 1 | | | | | 1 | 1 | | | | | | | | 1 | 1 |
| 87 | 20 | * | ! | | | * | | | | | | | 1 | 1 | | 1 | | | | | | | | | | | | | |
| 86 | 23 | * | ! | | | * | | | | | | | 1 | 1 | | 2 | | | | | | | | | | | | | 1 |
| 85 | 24 | * | ! | | | * | | | | | | | 1 | 1 | | 1 | | | | | | | | | | | | | 1 |
| 84 | 27 | * | | * | | | | ! | | | | | | 1 | | | | | | | 1 | | | | | | | 1 | 1 |
| 83 | 28 | * | | * | | | | ! | | | | | | 1 | | | | | | | 1 | | | | | | | 1 | 1 |
| 81 | 32 | * | ! | * | | | | ! | | | | | 1 | 1 | | | | | | 1 | | 1 | | | | | | 1 | 1 |
| 80 | 33 | * | ! | * | | | | ! | | | | | 1 | 2 | | | | | | 1 | | 1 | | | | | | 1 | 1 |
| 79 | 36 | * | ! | * | | | | | | | | | 1 | 2 | | | | | | 1 | | 1 | | | | | | | |
| 78 | 39 | * | ! | * | ! | | | | | | | | 2 | 2 | | | | | 1 | 1 | | 2 | | | | | | | |
| 77 | 44 | * | ! | * | ! | ! | | | | | | | 2 | 3 | | | | | 1 | 1 | 1 | 3 | | | | | | | |
| 76 | 46 | * | !: | * | ! | | | | | | | 1 | 2 | 3 | | | | | 1 | 1 | | 2 | | | | | | | |
| 75 | 48 | * | !: | * | ! | ! | | | | | | 1 | 3 | 4 | | | | | 1 | 1 | 1 | 3 | | | | | | | |
| 74 | 49 | * | !: | * | ! | ! | | | | | | 1 | 3 | 4 | | | | | 1 | 1 | 1 | 3 | | | | | | | |
| 73 | 51 | * | !: | * | ! | ! | | | | | | 2 | 3 | 4 | | | | | 1 | 1 | 1 | 3 | | | | | | | |
| 72 | 52 | * | !: | * | ! | ! | | | | | | 1 | 2 | 3 | | | | | 1 | 1 | 1 | 3 | | | | | | | |
| 69 | 53 | * | !: | * | ! | | | | | | | 1 | 2 | 3 | | | | | 1 | 1 | | 2 | | | | | | | |
| 67 | 54 | * | !: | * | ! | | | | | | | 1 | 2 | 3 | | | | | 1 | 1 | | 1 | | | | | | | |
| 65 | 56 | * | ! | * | ! | | | | | | | | 1 | 3 | | | | | 1 | 1 | | 2 | | | | | | | |
| 64 | 57 | * | !: | * | ! | | | | | | | 1 | 1 | 3 | | | | | 1 | 1 | | 1 | | | | | | | |
| 62 | 58 | * | !: | * | ! | | | | | | | 1 | 2 | 3 | | | | | 1 | 1 | | 1 | | | | | | | |
| 61 | 59 | * | ! | * | ! | | | | | | | | 1 | 2 | | | | | 1 | 1 | | 2 | | | | | | | |
| 60 | 61 | * | | * | !: | | | | | | | | | 1 | | | | 1 | 1 | 1 | | 2 | | | | | | | |
| 58 | 62 | | * | * | !: | | | | | 1 | | | | | | | | 1 | 1 | 1 | | 1 | | | | | | | |
| 52 | 63 | | * | * | | | | | | | | | | | | | | | | | | 1 | | | | | | | |
| 34 | 64 | * | | * | | | | | | | | | | | | | | | | | | | | | | | | | |

INTERPRETATION:

While there is statistical significance at the scores of 76-67, 64-62 and 60-58, the differences are not practically significant. The 80% rule of thumb appears somewhere for every score except 52-34. It is better to stay away from statistical significance, so look for a cutoff above 76 if 44 people can be processed to the next step, or the score of 65 with 56 people, or the score of 61 with 59 people. Use Tables 1-4 to set the final cutoff. For adverse impact to exist, both statistical and practical significance must be shown. Therefore, if you have to select a cutoff in the range of statistical significance (shown by the line :), then choose a cutoff that has no practical significance. When 2 or fewer added to a group at a score change statistical significance from a yes to a no, the numbers are too small to be practically significant (2 or less under the VIR column). Similarly, 3 or less under 80% and 4 or less under SRD are too few.

## U.S. Supreme Court Procedure - Pool Differences

While the *Uniform Guidelines* since 1978 have called for rate differences to be evaluated to determine adverse impact, the United States Supreme Court has expressed

the need to evaluate pool differences in cases decided in 1977 through 1989. Substantial differences occur very frequently when evaluating data with rate statistics versus pool statistics.

Rate differences compare the rate of one group's success to another group's success (e.g., passing a test, hired, retained after layoffs, getting raises, promoted, etc.). Pool differences compare the percentage a group makes up in the pool before an action starts to the pool after the action has occurred (e.g., the pool taking a test compared to the pool passing the test, pool applying compared to the pool hired, pool available for layoffs compared to the pool retained after the layoffs, pool available to get raises compared to the pool getting raises, pool available for promotion compared to the pool promoted).

As early as 1977, the United States Supreme Court in *Castaneda v. Partida* (1977) used pool differences to evaluate the pool of Mexican Americans on a jury compared to the pool of Mexican Americans in the population. In a selection case in the same year, the United States Supreme Court in *Hazelwood* (1977) compared the pool of Blacks selected as school teachers to the pool of Blacks with the skills to be teachers in the relevant labor force. Also, in the same year in *Teamsters v. U.S.* (1977), pool comparisons were described. As recently as June of 1989, the United States Supreme Court again called for the proper comparison to be pool comparisons in *Wards Cove Packing Co., Inc. v. Atonio* (1989). The Court quotes its decision in *Hazelwood:*

> "The 'proper comparison [is] between the racial composition of [the at-issue jobs] and the racial composition of the qualified . . . population in the relevant labor market.'"

The Court goes on to specify that: "It is such a comparison--between the racial composition of the qualified persons in the labor market and the persons holding at-issue jobs--that generally forms the proper basis for the initial inquiry in a disparate impact case. Alternately, in cases where such labor market statistics will be difficult if not impossible to ascertain, we have recognized that certain other statistics--such as measures indicating the racial composition of 'otherwise-qualified applicants' for at-issue jobs--are equally probative for this purpose."

The Court in *Wards Cove* uses the specific analysis of a pool comparison with: "...if the percentage of selected applicants who are nonwhite is not significantly less than the percentage of qualified applicants who are nonwhite, the employer's selection mechanism probably does not operate with a disparate impact on minorities." The Court footnotes this quote by stating that it used the word "probably" because in *Connecticut v. Teal* (1982) bottom line racial balance as a defense is not adequate if plaintiffs can show a particular hiring practice has a disparate impact on minorities, notwithstanding the bottom-line racial balance.

Comparing the beginning, starting, or available pool with the pool after the action has taken place appears to be the consistent pattern of U.S. Supreme Court mandates.

*Statistical significance with pool differences* is calculated with a binomial statistic, or one-sample statistic. We call this method the Hazelwood analysis in reference to the first U.S. Supreme Court decision which uses this statistic in an employment selection case. Statistical significance is found between 2 or 3 standard deviations. See *Hazelwood* (1977) and Technical Note 2 at the end of this paper.

*Practical significance with pool differences* can be calculated by determining the number of people that would have to be added to the plaintiff's group (e.g., selected) to increase its percentage in the pool to a high enough level to eliminate the statistical significance finding. If one or two people added to the plaintiff group eliminate the statistical significance finding, there is no adverse impact because there is no practical significance.

*Under what circumstances* would you apply the pool comparison versus the rate comparison statistics? When you apply the pool and rate statistics to applicant versus hire data, for example, pool and rate standard deviations most of the time will yield very different statistical conclusions. The rate statistics will show statistical significance before the pool statistics. When the sample sizes are infinitely large, the results will be the same. If the beginning pool is ten times larger than the comparison pool, the outcomes are more similar, but still often quite different. *The Supreme Court has identified the pools test as the threshold test.* The rate comparison test can be used as a conservative assessment for components of a selection process. But some time in the future we will have to see which way the courts go on this issue.

## Recommendations for Setting Cutoffs on Knowledge Tests Used for Promotion, Training, Certification, and Licensing

1. Use 7 to 10 Subject Matter Experts (incumbents, supervisors or trainers who can competently perform the duties for which the tested knowledge is needed).

2. Use the rule that at least 50 percent of the Subject Matter Experts need to agree on issues that determine inclusion of an item on a test. A higher standard would be 70 percent Subject Matter Expert agreement.

3. Have each Subject Matter Expert insure the key is accurate.

4. Ask each Subject Matter Expert to answer a question on the job relatedness of the item. A preferred option would be to have the Subject Matter Experts identify the duty(s) for which the knowledge measured by the test item is needed to competently perform the duty(s).

5. Ask each Subject Matter Expert to identify the level of consequence for what could likely happen in terms of duty performance if a person performing a duty needing the knowledge measured by the test item does not know the answer to the item.

6. Ask each Subject Matter Expert to identify the ease with which the item can be looked up in the normal performance of the duty.

7. Ask each Subject Matter Expert for an opinion regarding the differentiating aspect of the test item, if a cutoff is to be considered higher than minimum competency.

8. Ask each Subject Matter Expert to state the probability for each item that the "minimally acceptable person" would answer the item correctly. The Subject Matter Experts should first discuss the concept of the "minimally acceptable person" and for the exercise think of a number of minimally acceptable persons.

9. Sum the Subject Matter Expert probabilities, or proportions on each item, average the proportions for each item, then sum the averages. This average of averages represents the average minimally acceptable score, or average Angoff, also called the *un*modified Angoff.

10. Do not set a cutoff score or weights prior to administration of a non-repeating type of test (e.g., promotion test and licensing test). Flexibility may be needed to address the requirements of Section 3B of the *Uniform Guidelines*.

11. Calculate the reliability and standard deviation of the test after it is administered. Use an up-to-date item analysis program that calculates the Horst Modification of the KR-20 for the reliability estimate, as the Horst avoids the assumption of equal item difficulty.

12. Consider statistical and human factors such as the size of the standard error of measurement, risk of error (risk of excluding a truly qualified candidate whose low score does not show the real level of knowledge compared to the risk of including an unqualified candidate whose low score does show an unacceptable level of knowledge), internal consistency of the Angoff or Subject Matter Expert panel (e.g., taken individually, the Subject Matter Experts' range in their individual estimates of minimum competency), supply and demand for the at-issue jobs, and the sex and race/ethnic composition of the at-issue jobs in the work force. After consideration of the statistical and human factors, reduce the Angoff average score by one, two, or three standard errors of measurement. This score will be called the *modified Angoff* score.

13. Consider the adverse impact of the distribution of test scores. See if there is a score above the *un*modified Angoff that does not have adverse impact. If this score can produce enough candidates, consider using it. With this score, candidates not only are more than minimally qualified, according to the Subject Matter Experts, but the trigger to litigation (adverse impact) is removed. (If not enough candidates are identified with the no adverse impact method, see if it would be cost effective to select those you can with the no adverse impact score, and give another test. Cost

considerations include not only dollar costs but the effect on morale and lost time due to litigation delays.)

14. If no score can be found that identifies enough candidates without adverse impact above the *un*modified Angoff, evaluate the degree of adverse impact for scores *within* the first standard error of measurement below the *un*modified Angoff. The *Uniform Guidelines* requires consideration of alternative uses of a test that are substantially equally valid, if the alternative has less adverse impact. See *Uniform Guidelines,* Section 3B (1978). Less adverse impact could mean a higher proportion of an underutilized protected group who pass or one of the underutilized protected groups adversely impacted will no longer be adversely impacted at the alternative score.

15. If no score can be found without adverse impact that produces enough candidates above the *un*modified Angoff, and no score can be found *within* the first standard error of measurement that is considered substantially equally valid as the score of one standard error of measurement below the *un*modified Angoff with less adverse impact, then repeat this process until you reach your *modified Angoff* score. Repeat the process for two standard errors of measurement if your *modified Angoff* is two standard errors of measurement, or for up to three standard errors of measurement, if your *modified Angoff* score is three standard errors of measurement.

## Technical Notes

1. *Rate calculations.* One of the problems with the chi-square is that its answer does not give an accurate estimate of the hypergeometric probability. Several statisticians have attempted to develop statistical corrections to more closely estimate the probability. One of Cochran's corrections appears to be much better than no correction and much better than the traditionally used Yate's correction. See Haber (1980). (My experience applying no correction, Yate's correction, and Cochran's correction along with the exact probability on hundreds of adverse impact analyses supports Haber's findings.) Now that we have fast PCs we can use the technically most accurate way by calculating the hypergeometric probability directly and then inversely transforming it to a standard deviation. This produces the exact number of standard deviations without using the chi-square as an estimator. The disadvantage of the direct calculation of probability is the time it takes. Some calculations even on a fast PC take 15 minutes. The chi-square with the Cochran correction takes the slowest PC less than a second. It is my experience that for chi-square derived standard deviations from 1.80 to 2.10, the exact calculation is required in order to know if significance has been achieved. Fortunately, inexpensive software is available that will calculate the direct hypergeometric probability. A simple conversion table can be used to convert the probability to standard deviations. Some software will also do the automatic conversion from the calculated probability to the standard deviations.

2. *Pool calculations.* The binomial statistic can be calculated with an estimate process in a few seconds by the slowest computer or can be calculated precisely with the direct binomial probability calculation in a few minutes on fast PCs. With the direct calculation, the resulting probability can be inversely transformed back to a standard deviation that exactly equals the calculated probability. This inverse transformation of the calculated probability to the standard deviation is the most precise way to obtain the number of standard deviations from pool statistics.

3. *Statistical significance.* The 5% level of statistical significance can be shown with 1.645 standard deviations or 1.96 standard deviations, depending upon whether a one-tail or two-tail test is used. A one-tail test gives direction to the hypothesis. The two-tail test does not. For example, a one-tail test might be made to see if the passing rate for one group on a test is significantly less than the passing rate of another group. A two-tail test is used to see if one group's passing rate is different (higher or lower) than another group's rate. At least one circuit court addressed the problem. The Fourth Circuit rejected the plaintiff's attempt to use the 5% level of probability with a one-tail test for setting the level of statistical significance (i.e., using 1.645 standard deviations). See *EEOC v. Federal Reserve Bank* (1983). The EEO field must use standard deviations and not probability estimates to set statistical significance.

**References**


Angoff WH. (1971). Scales, Norms, and Equivalent Scores. In Thorndike RL, *Educational Measurement*, p. 508-600. Washington, DC: American Council on Education.

Biddle RE. (1989). Wards Cove Packing vs. Atonio Redefines EEO Analyses. *Personnel Journal*, June.

*Bouman v. Pitchess*, 47 EPD 38,212 (1988).

Cascio WF, Alexander RA, Barrett GV. (1988). Setting Cutoff Scores: Legal, Psychometric, and Professional Issues and Guidelines. *Personnel Psychology*, 41, 1-24.

*Castaneda v. Partida*, 97 S.Ct.1272 (1977).

Campbell T. (1982). Entry-Level Exam Examined in Court. *Western Fire Journal*, July.

*Connecticut v. Teal*, 102 S.Ct.2525 (1982).

*Contreras v. City of Los Angeles*, 656 F.2d 1267 (9th Cir. 1981).

*EEOC v. Federal Reserve Bank of Richmond*, 698 F.2d 633 (4th Cir. 1983).

Haber M. (1980). A Comparison of Some Continuity Corrections for the Chi-Square Test on 2x2 Tables. *Journal of the American Statistical Association*, 75, 371.

*Hazelwood School District v. United States*, 97 S.Ct.2736 (1977).

Guilford JP, Fruchter B. (1973). *Fundamental Statistics in Psychology and Education* (5th ed.). New York: McGraw-Hill.

*San Francisco Police Officers Association v. City and County of San Francisco*, 812 F.2d 1125 (9th Cir. 1987).

Smith RS, Smith JK. (1988). Differential Use of Item Information by Judges Using Angoff and Nedelsky Procedures. *Journal of Educational Measurement*, 25 259-274.

*Teamsters v. United States*, 97 S.Ct.1843 (1977).

Uniform Guidelines on Employee Selection Procedures. (1978). *Federal Register*, 43, 38290-38315.

*U.S. v. Commonwealth of Virginia*, 18 EPD 8779 (1978).

*U.S. v. South Carolina*, 434 US 1026 (1978).

*Wards Cove Packing Co., Inc. v. Atonio*, 109 S.Ct.2115 (1989).

*The information in this article is not intended to be legal advice.*

* Richard Biddle is President of Biddle & Associates, Inc., a Sacramento based EEO consulting, testing and software firm. He has served as a consultant or expert to attorneys on the plaintiff and defense sides in about 100 age, sex, race, or ethnic origin discrimination cases. His firm licenses software in the areas of tracking, test scoring, setting cutoff scores, affirmative action, and job analysis; licenses entry level police and fire tests; licenses occupational data for AAPs and court cases; and custom develops and validates tests. He can be contacted at richardbiddle@biddle.com .

** This article was reproduced with permission of *Public Personnel Management*, published by the International Personnel Management Association (IPMA), 1617 Duke St., Alexandria, VA 22314, 703-549-7100, www.ipma-hr.org.