Paper Presentation

Protective Service Physical Ability Tests: Establishing Pass/Fail, Ranking, and Banding Procedures

Stacy L. Bell, Managing Consultant
Firefighter Selection, Inc.

Evaluating Physical Ability Test Scores

- A recent survey of court-disputed police and fire physical ability tests showed a successful defense rate of less than 10% [1].
- Since job-related physical ability tests are likely to reflect such differences, setting pass/fail cutoffs that accurately reflect the physical ability levels required for successful job performance is a key consideration for any protective service agency involved in physical ability testing.
- This presentation will limit discussion to evaluating the use of physical ability test scores outside of other selection devices, although the principles herein may be used for combining physical ability test scores with other pre-employment tests.

---

1 Shepherd, M.R. (1997, May). *Court reviewed physical ability tests: The winning and losing characteristics of physical ability tests*. Paper presented at the May 16th Session of the Personnel Testing Council of Northern California, Sacramento, CA.

Shouldn't the cutoffs for physical ability tests be stringent?

- Agencies are often motivated in setting overly stringent cutoffs out of concern for public safety and the safety of police officers and firefighters.
- Take precautions to avoid overemphasizing the extent to which physical performance really contributes to the overall job performance.
- Setting cutoffs too low could unduly lower physical standards and endanger public safety. However, setting standards too high could also subject the agency to expensive and time-consuming litigation.

How do I decide on the type of cutoff to use?

- For purposes of this presentation, we will discuss three methods for establishing pass/fail cutoffs:
  - Modified Angoff
  - Norm-Referenced
  - Criterion-Referenced
- Using a combination of one or more of these methods is usually the appropriate approach for determining the cutoff that best represents the level required for successful job performance.

Sample Selection

- The first step in a cutoff procedure is the selection of the incumbent sample.
- Careful selection of a diverse subject-matter expert (SME) sample for the study is essential.
  - Courts are often skeptical of a physical ability test developed and validated without the input of women and minority subject-matter experts.
  - If women or minority groups are not adequately represented in the classification, they should be over-sampled in the validation study.
- SMEs should be full-duty, non-probationary incumbents who have at least one year experience in the relevant classification.
- Random selection of the sample and including performers from various age groups is also important.

Sample Selection

- If a criterion-referenced approach is used(either for pass/fail or for ranking), it is imperative to obtain a sample size that will yield sufficient statistical power.
  - Obtaining a .30 correlation is a court-established precedent for using a physical ability test as a ranking device.
  - Sample sizes of at least 20 are necessary for researchers to obtain validity coefficients of .30 or higher (a .306 correlation is required for significance at the .05 level using a 1-tail test).
- As with most criterion studies, the larger the sample size, the better the study.
  - With a sample of 30 subjects a researcher can only be 51% confident of finding a .30 correlation if it exists in the population.
  - With a sample of 60, one can be 78% confident of finding a .30 correlation is it exists.

Pass/Fail Cutoff Method 1: Modified Angoff

- What is the Angoff method?
  - The Angoff method has traditionally been used for setting pass/fail cutoffs on written exams.
    - SMEs provide judgments on the percentage of minimally-qualified applicants who would be expected to correctly answer each test item.
    - The judgments are then averaged and used as the pass/fail level of the test.
  - A similar procedure may be used for physical ability tests too.
    - SMEs would begin by taking the physical ability test and then complete surveys and provide their opinions on the test score that best represents where a minimally-qualified applicant would score.
    - The SME opinions are then averaged into a pass/fail cutoff.
    - SME opinions that are significantly lower or higher than their actual test scores should be carefully considered and the outliers removed from the study.

The Modified Angoff Method

- A modification of the Angoff method that received acceptance before the United States Supreme Court in *U.S. v. South Carolina* (for written tests) may also be used to effectively set pass/fail cutoffs for physical ability tests.
- The modification followed I U.S. v. South Carolina lowered the average Angoff estimate by one, two, or three standard errors of measurement.
- The approved modification was based on consideration of several statistical and human factors:
  - The size of the standard error of measurement
  - The risk of excluding a truly qualified candidate compared to the risk of including an unqualified candidate
  - The internal consistency of the Angoff panel
  - The supply and demand for at-issue jobs
  - The sex and race/ethnic composition of the jobs in the work force.
- Reducing the average Angoff by one, two or three standard errors of measurement would constitute the minimum passing level for the test.

Pass/Fail Cutoff Method 2: Norm-Referenced (on SMEs)

- Section 5H of the *Uniform Guidelines on Employee Selection Procedures* require:

  "Where cutoff scores are used, they should normally be set so as to be reasonable and consistent with *normal expectations of acceptable proficiency* within the work force…" (emphasis added).

- Evaluating SME performance on a physical ability test is an effective way to determine what constitutes "normal expectations of acceptable proficiency" providing that:
  - The SMEs provide reasonable exertion levels on the test
  - Measures of the SME job performance rating can be obtained
  - Range restriction didn't contribute to producing a SME sample that is *overqualified* for the job they perform.

Pass/Fail Cutoff Method 2: Norm-Referenced (on SMEs)

- *How do I determine a score that falls within the normal expectations of job performance?*
  - One possibility is to use the standard error of difference to determine the furthest score away from the mean (or other "normal" points of the distribution) that is not reliably different than the mean.

Standard Error of Difference (SED)

- SED = SEM * sqrt (2)
- In the context of banding, the SED provides a "confidence interval" answering:
  - How far away can I move from one score before I reach a score representing a different level of the KSAPC measured by the test?
  - How sure do I want to be that these two scores are really different? (1SED =68%, 2=95%; 3=99%)

Banding Using the SED

- Begin with top-most score on rank-ordered score list
- Deduct 1 or 2 SEDs from top score
- Treat all scores within band equally
- Pass entire group, or
- Make selections from within bands using other job related factors (professionalism, training, experience, etc.)

Example of Banding with SEDs

- If the average SME score on a continuously timed physical ability test is eight minutes and the standard error of difference is 45 seconds, setting a cutoff at 8:45 provides **68%** confidence that scores slower than 8:45 are reliably different than the eight minute average score of the SMEs.
- Using 2 standard errors of difference provides **95%** confidence that scores slower than 9:30 are reliably different than the eight minute average score of the SMEs

Precautions!

- The score that lies one (or more) standard error of difference below the average SME score represents a point in the distribution that is still within the range of the "normal," central score.
- Using this method to determine a passing point assumes that the mean of the SMEs represents normal workforce performance. While this assumption may be argued, it does avoid using the lowest performance level on the incumbent distribution as the cutoff point for the test.
- As with any cutoff procedure, individuals utilizing the test must decide if this is an assumption that they are willing to make.

Pass/Fail Cutoff Method 3: Criterion-Referenced (on SMEs)

- Another method that can be used for setting the pass/fail cutoff is a criterion-related validity approach.
  - Criteria usually include peer or supervisory ratings on incumbent performance on the physical aspects of the job, although other methods may be used.
- It is important to note that the scales used to obtain criterion ratings should not exceed the range of human judgment.
  - Scales ranging from 1-5, 1-7, or 1-9 are typically adequate to provide judgments on observable, physical performance.
  - Each rating on the criterion scale should be operationally defined in terms of observable aspects of job behavior that are pertinent to the criteria.

Pass/Fail Cutoff Method 3: Criterion-Referenced (on SMEs)

- It is important to include a wide range of job performers for a criterion-related validity study to reveal the minimum test performance necessary for satisfactory job performance.
- Given that typical entry-level fire and police recruitment involves rigorous selection, finding poor and marginal job performers to include in the study is not always possible.
  - This restriction in range creates a problem for setting minimum levels of competency due to the fact that the acquired data cannot extend to differentiate performance at minimum levels or lower.
  - Correcting for restriction in range by determining the variance of the unrestricted applicant population is one solution to remedy this problem.
- Once the criterion study is completed, the point at which the physical ability test data intersects with the marginal performance rating can be used to establish the pass/fail cutoff. Scores higher than the minimum competency level can be selected.

Setting Cutoffs Above Minimum Competency Levels, Ranking, and Banding

- **CAUTION!!!**
  - Setting cutoffs above the minimum-competency level, ranking, or banding all require similar support under the *Uniform Guidelines* and relevant court cases.
  - Generally speaking, the greater the adverse impact and the more stringent the test usage the greater the justification will be needed.
- Content validity methods are sufficient for providing support to use a test on ranking or banding basis; however, the courts have specifically endorsed using criterion-related validity to demonstrate that higher scores on a selection instrument equate to proportionately better job performance.
  - The courts require validity coefficients that often exceed the required .05 level of significance and have consistently required a validity coefficient of .30 or greater (regardless of the sample size in the study) as noted in the attached court cases.

Banding Methods for Physical Ability Tests

- Two banding methods are particularly useful for physical ability tests:
  - **Top-down bands** using the standard error of difference, or
  - **Expectancy bands** using the Lawshe model (or similar expectancy models).

Top-down Bands

- Using the top-down approach, the first band is created by subtracting the standard error of difference from the top score to arrive at the lowest score in the band.
  - All applicants in this band are considered reliably similar and are selected from within this band randomly or by using other job-related factors.
- To create a wider band, two or three standard errors of difference may be used (so long as the bandwidth does not reach below the minimum-competency level of the test).
  - For a more elaborate discussion of banding see Cascio et al (1991).

Expectancy Bands

- Using the results from a criterion-related validity study, expectancy bands rely on the principle that SMEs with high job performance ratings are expected to perform high on the test, while those with moderate performance ratings are expected to perform moderately, and those with low performance ratings are expected to perform poorly on the test.

Example of Using Expectancy Bands

- In a criterion-related validity study conducted with over 40 fire agencies in California, a correlation coefficient of -.44 was found (N = 62) between scores on a continuously-timed physical ability test for entry firefighters and performance ratings from peers on the physical aspects of performance on emergency scenes (fire suppression, emergency medical situations, rescue operations, and other emergency scenes).
- Expectancy bands were developed using the Lawshe model that demonstrate expected job performance increases in 10% increments based on test scores.

Example of Using Expectancy Bands

- Probability of Candidate Demonstrating Above Satisfactory Physical Job Performance:

Band	Test Score	Probability
D	574-522 seconds	56%
C	521-483 seconds	66%
B	482-419 seconds	75%
A	418 seconds and faster	85%

- A candidate who scores 521-483 seconds on this test has a 66% likelihood of performing at an above-satisfactory level on the physical aspects of the job at emergency scenes.
- Note: The scores within a band represent varying levels of expectancy. For example, expectancy scores within the 482-419 band range from 75% (482 seconds) to 84% (419 seconds).

Summary

- Setting a pass/fail cutoff should include a careful balance between selecting “the best of the best” physical performers and the selection of the level that represents the true physical ability required fro satisfactory job performance.
- Strict, top-down ranking is not necessarily the best use of physical ability test scores for a number of reasons and the authors would suggest against rank ordering oh physical ability tests.
- Banding is a more appropriate approach---especially when bands are created using a job performance expectancy model. Such bands can preserve much of the utility benefit of strict rank ordering while minimizing adverse impact.