Standard Setting for High Stakes Exams

Credentialing bodies put substantial resources into developing a valid and reliable exam, but may give less thought to how the resulting score should be used. However, a testing program is only as strong as its weakest component, and if the passing score is not clearly linked to a job-related criterion, it can undermine the entire credential. HumRRO has worked with dozens of associations and organizations on high-stakes testing programs and in the process has formulated the following best practices.

What is Standard Setting?

Standard setting is the process by which test programs establish a cut-score, or minimum score required to pass a test. Depending on whether the test is used in education, credentialing, or talent management, the cut-score can be used to determine if someone is eligible for certification or licensure, selected or promoted into a job, or placed in a course or training program.

How Does It Work?

Criterion-referencing compares people to an objective standard of performance or knowledge regardless of test form, time, and location by explicitly linking the passing standard to the purpose of the exam. Criterion-referenced standard setting is not strictly data driven. Rather, it is based on the sound professional judgment of subject matter experts (SMEs).

Before beginning the standard setting activity, HumRRO has SME participants take the test, so they can read the items in a context similar to test candidates. Next, SMEs

think about a hypothetical person who performs just well enough on the job or in the course to be considered successful. Then, SMEs describe the performance level required to be able to just pass the test (i.e., just good enough to be certified or move onto the next level). This is the minimum standard required to be certified, licensed, considered for selection/promotion, or placed into an academic or professional level or development program. Test candidates that meet that criterion are traditionally referred to as Just Sufficiently Qualified (JSQ) Candidates, Borderline Candidates, or Minimally Qualified Candidates.

Once the performance level is defined, SMEs review the test content and make multiple independent rounds of judgments about what type of test score constitutes a JSQ level.

Between rounds, SMEs share their first judgments with each other and facilitators provide impact data, such as the percentage of all candidates who answered a selected-response test item correctly, the score of an essay (or constructed response item), or the estimated percent of students classified into the different performance levels. When making judgments at the item-level, HumRRO consultants find that it is helpful to stop and share first judgments and impact data after every 10 to 15 items, to reduce the cognitive burden on SMEs who have to go back and consider why they judged each item the way they did.

The discussion and impact data are important to ensure that SMEs have a shared understanding of the JSQ level, which enhances their level of agreement. After the discussions are complete, the SMEs independently make a final judgment without further discussion. The analyst calculates a cut-score later and provides the recommendation to the policy making body.

Who Should Be Involved?

Since criterion-referencing is dependent on SME judgment, it is critical to identify very experienced job incumbents to participate in the process. The ultimate validity of a test is dependent on the accuracy of the cut-score. Given this, a credentialing program should recruit SMEs who are familiar with the constructs being measured (e.g., technical knowledge, task performance) and who are also familiar with job incumbents who perform at and near the minimum performance standard targeted by the test.

Additionally, SMEs should have a stake in the outcome. Senior employees who provide on-the-job training and immediate supervisors usually are appropriate standard setting SMEs. We also find that when SMEs represent the widest variety of professional settings, the richer the definition of the JSQ candidate.

Common Criterion-Referencing Methodologies

HumRRO regularly uses these criterion-referencing methodologies:

- **Modified Angoff** – Ideally suited for selected response items such as multiple-choice and multiple response items. This method is routinely used in the credentialing arena for programs with multiple-choice item formats.

- **Body of Work** – Ideally suited for tests that include open-ended items or where the candidate produces a product or work sample.

- **Bookmarking** – Ideally suited for test forms with multiple choice and constructed response items. This method is routinely used in the educational testing arena and for high volume credentialing programs with high volume testing that supports the use of item response theory (IRT)-based psychometric analysis.

For more information contact:

DAVID DORSEY, Ph.D.
Vice President,
Business Development Division
phone: 703.706.5670
e-mail: ddorsey@humrro.org

Visit **HumRRO.org** for additional information about all our services.

HumRRO was established in 1951 and is an independent, nonprofit corporation dedicated to the development and application of state-of-the-art scientific principles and technologies to solve the real-world challenges facing private and public sector organizations and educational institutions. Our professional staff is composed of psychologists with diverse expertise in: personnel selection, classification, and promotion; education assessment and accountability; strategic human capital management; program evaluation and policy analysis; employee development and training; credentialing; modeling and simulation; and survey research. Our client base includes the military, government agencies, private industry, and professional associations. HumRRO is an industry leader in working collaboratively with clients to create customized, best-practice solutions.