Criterion-Referenced Test Development Process: A Primer

Designing	Conferring
Developing	Accrediting
Implementing	Maintaining

Functional Manager	PMO
Component	Candidate

Description: Tests developed to certify individuals' mastery of a pre-defined set of knowledge and/or skills must meet a number of federal and professional guidelines (i.e., the Uniform Guidelines on Employee Selection Procedures published by the EEOC, the Standards for Educational and Psychological Testing jointly published by AERA, APA, and NCME). These guidelines attempt to ensure that the tests result in reliable information that stakeholders can use to make valid decisions regarding individuals who hold the certification. Although there are a number of approaches for developing certification tests that meet the established guidelines, one approach – the criterion-referenced test development (hereafter, CRTD) process – is considered "best-in-class." The CRTD process employs a number of evidence-based methodologies in stages to address three core questions: (1) are domains being addressed by the test critical to work performance?, (2) is the test assessing the content appropriately?, and (3) is the test classifying individuals appropriately? (see figure below). This job aid presents an overview of the CRTD process.

Process:

The first CRTD stage involves: (1) characterizing a practice area's (PA) *Essential Body of Work* (behaviorally-based descriptions of what a practitioner does in a given practice area), (2) codifying a PA's *Essential Body of Knowledge* (knowledge/skills that practitioners need to perform the work specified in the EBW, and (3) generating a *Certification Blueprint* that list terminal and enabling objective statements that explicitly describe what a practitioner must know and be able to do to perform the work specified in the EBW). The goal of this CRTD stage is to ensure that: (1) there is a shared and agreed-upon understanding of the knowledge and skills that the test will need to address, and (2) those knowledge and skills are ones that are critical to work performance.

The second CRTD stage involves: (1) using documents to develop test items that assess the content specified in the certification blueprint, (2) SMEs reviewing those items to ensure that they accurately depict the content specified in the certification blueprint, (3) piloting the items to determine if they are "good" or "bad" items (from a psychometric perspective), and (4) compiling a test with the right characteristics (i.e., it is covering the content appropriately per the certification blueprint, and is psychometrically sound). Test developers rely on data obtained from piloting the items to gauge whether or not the items (hence, the resulting test) are psychometrically "good" or "bad." This includes analyzing the following statistics:

- Item Difficulty: The proportion of individuals answering a question correctly (a "good" item is one that is not too hard nor too easy)
- Item Discrimination and Item-Total Correlation: How well an item is differentiating between strong and weak performers (a "good" item is one that strong performers get correct, and weak performers get incorrect)
- Test Reliability: The degree to which a test will produce stable and consistent results (a "good" test is one with a reliability of .80 or above).
- CSEM: Variation between the scores of an individual if that individual were to take the same test repeatedly (a "good" test is one with a CSEM less than or equal to 4).

The final CRTD stage involves: (1) collecting judgments from SMEs regarding how "minimally competent individuals" would do in each item, (2) calculating a provisional cut-score based on the SMEs' judgments, and (3) agreeing upon an operational cut-score based on the provisional cut score and pilot data.

Success of this stage depends on how well SMEs agree on the definition of "minimally competent individuals," and SMEs judgments regarding how well such individuals would do on each item on the test. Appropriateness of the SME judgments are then analyzed using data from the pilot. Finally, the final arbiter of the cut-score can choose to make adjustments to the provisional cut-score. However, the adjustments are limited based on acceptable limits (i.e., using the CSEM statistic).