Clause 9.3.5

- Appropriate methodology and procedures (e.g. collecting and maintaining statistical data) shall be documented and implemented in order to affirm, at justified defined intervals, the fairness, validity reliability and general performance of each examination, and that all identified deficiencies are corrected.

Survey Question

- What are appropriate methods to reaffirm the fairness, validity reliability and general performance of examinations?

Terms

- **Fairness** – equal opportunity for success provided to each candidate in the certification process
  - Note 1 to entry: Fairness includes freedom from bias in examinations
- **Validity** – evidence that the assessment measures what it is intended to measure, as defined by the certification scheme
- **Reliability** – indicator of the extent to which examinations scores are consistent across different examination times and locations, different examination forms, and different examiners.
- **General Performance**

Fairness

- What kind of data exists to show that the examination process is fair?
- Does the CAB look at data associated with various language and cultural groups to ensure that some particular group is not performing worse on the examination form
- Bias review of test questions and test process
- Differential Item Functioning (DIF)

Validity

- Many types of validity
  - ***Content validity***
  - ***Concurrent validity***
  - ***Predictive validity***
  - ***Face validity***

Validity

- ***Content validity***
  - the most important type of validity for most certification and licensure programs is probably that of content validity.
  - is a logical process where connections between the test items and the job-related tasks are established.
  - is typically estimated by gathering a group of subject matter experts (SMEs) together to review the test items.
  - The SMEs are asked to indicate whether or not they agree that each item is appropriately matched to the content area indicated. Any items that the SMEs identify as being inadequately matched to the test blueprint, or flawed in any other way, are either revised or dropped from the test.

Validity

- ***Concurrent validity***
  - a statistical method using correlation. Examinees who are known to be either masters or non-masters on the content measured by the test are identified, and the test is administered to them under realistic exam conditions. Once the tests have been scored, the relationship is estimated between the examinees' known status as either masters or non-masters and their classification as masters or non-masters (i.e., pass or fail) based on the test.
  - This type of validity provides evidence that the test is classifying examinees correctly. The stronger the correlation is, the greater the concurrent validity of the test is.
  - the degree to which the examination results correlates with other measures of the same construct measured at the same time. If an examination results is compared to another similar examination result, with the results correlate?

Validity

- ***Predictive validity***
  - the relationship of test scores to an examinee's future performance as a master or non-master that is estimated.
  - the degree to which the examination can predict (correlate with) the performance or knowledge (the construct) measured at some time in the future.
  - the degree to which the test predict examinees' future status as masters or non-masters?"
  - correlation computed is between the examinees' classifications as master or non-master based on the test and their later performance, perhaps on the job.

Validity

- ***Face validity***
  - determined by a review of the items and not through the use of statistical analyses.
  - anyone who looks over the test, including examinees and other stakeholders, may develop an informal opinion as to whether or not the test is measuring what it is supposed to measure.
  - Does it appear to candidates to “look like” it measures what it is supposed to measure?
  - If we asked parents if a spelling test for school children is a good test, this is an example of face validity.

Validity

- Validity is on a continuum.
- More or less evidence for validity of an assessment process.

VALIDITY OF EXAMINATIONS AND EXAMINATION SCHEMES

Validity

Less Evidence

More Evidence

Threats to Validity

- Security Breach
- Wrong Content (doesn't relate to the job)
- Candidate unfamiliar with exam question format
- Poor exam questions
- Subjectivity in scoring
- Wrong passing score
- Non-Standardized Examination Administration

Evidence of Validity

- Job Task Analysis
- Evidence of the involvement of interested parties in identifying tasks and competencies
- Examination content, weight and format related to the tasks and KSAs
- Passing Score Study
- Item Analysis and Test Level Analysis
- Regular review of assessments and corrective actions where required
- Standardized Exam Administrations
- Rater Calibration and Standardization
- Security policies and procedures
- Elimination of subjectivity

Reliability

- has to do with the consistency, or reproducibility, of an examinee's performance on the test.
- For example, if you were to administer a test with high reliability to an examinee on two occasions, you would be very likely to reach the same conclusions about the examinee's performance both times.
- A test with poor reliability, on the other hand, might result in very different scores for the examinee across the two test administrations.

Reliability

- Many types of Reliability
  - ***Test-retest reliability***
  - ***Parallel forms reliability***
  - ***Decision consistency***
  - ***Internal consistency***
  - ***Interrater reliability***

Reliability

- ***Test-retest reliability***
  - administer a test form on two separate occasions only a few days or a few weeks apart
  - the time should be short enough so that the examinees' skills in the area being assessed have not changed through additional learning.
  - relationship between the examinees' scores from the two different administrations is estimated, through statistical correlation, to determine how similar the scores are.
  - This type of reliability demonstrates the extent to which a test is able to produce stable, consistent scores across time.

Reliability

- ***Parallel forms reliability***
  - parallel forms are all constructed to match the test blueprint and to be similar in average item difficulty.
  - Parallel forms reliability is estimated by administering both forms of the exam to the same group of examinees.
  - While the time between the two test administrations should be short, it does need to be long enough so that examinees' scores are not affected by fatigue.
  - The examinees' scores on the two test forms are correlated in order to determine how similarly the two test forms function.
  - This reliability estimate is a measure of how consistent examinees' scores can be expected to be across test forms.

Reliability

- ***Decision consistency***
  - For many criterion referenced tests (CRTs) a more useful way to think about reliability may be in terms of examinees' classifications.
  - For example, a typical CRT will result in an examinee being classified as either a master or non-master; the examinee will either pass or fail the test.
  - It is the reliability of this classification decision that is estimated in decision consistency reliability.
  - If an examinee is classified as a master on both test administrations, or as a non-master on both occasions, the test is producing consistent decisions.
  - This approach can be used either with parallel forms or with a single form administered twice in test-retest fashion.

Reliability

- ***Internal consistency***
  - The internal consistency method estimates how well the set of items on a test correlate with one another; that is, how similar the items on a test form are to one another.
  - Many test analysis software programs produce this reliability estimate automatically.
  - Quantifiable. Kuder-Richardson or coefficient alpha measure of test reliability

Reliability

- ***Interrater reliability***
  - When a test includes performance tasks, or other items that need to be scored by human raters, then the reliability of those raters must be estimated.
  - This reliability method asks the question, "If multiple raters scored a single examinee's performance, would the examinee receive the same score.
  - Interrater reliability provides a measure of the dependability or consistency of scores that might be expected across raters.

WHAT????

Testing of People is a Science

- Psychometrics
  - technique of mental measurement (Merriam-Webster)
  - The measurement of mental traits, abilities, and processes. (Dictionary.com)
  - Field of study concerned with the theory and technique of measurement. . . of skills, knowledge, abilities, attitudes. . . (Wikipedia)

Testing of People is a Science

- You wouldn’t try to fly a plane on your own or do brain surgery on your own

Testing of People is a Science

- If you are struggling doing this on your own, find a psychometric or measurement expert in your country to help you (universities, etc. have experts).

Questions?