TEST DEVELOPMENT PROCESS REVIEW
Examination Development Process Overview

1. Need for Examination Identified
2. Job Analysis Performed
3. Examination Specification Developed
4. Item Writing & Review
5. Examination Assembly & Review
6. Examination Scoring and Analysis
7. Standard Setting Study
8. Equating of Examination

After Cut Score Established
APMLE Structure

+ Structure of each part:
  + Part I – LOFT
  + Part II – LOFT
  + Part III – Fixed forms

+ Test Development Activities:
  + Similarities and differences depending upon the model.
  + Will discuss general components, and then aspects related to fixed forms and those related to LOFT
Validity

+ **Validity Evidence**
  + Anything that supports the meaning of test score inferences about the construct being assessed
  + Anything that supports the use of test scores
  + A test is valid in that it is used only for its intended purpose

+ **Job / Practice Analysis**
  + Systematic study of a profession, job, occupation
  + Other names – professional activities and knowledge study (PAKS), task analysis, practice analysis, role delineation
What are Test Specifications?

+ A description of the test that includes:
  + **test content**
    - With enough detail to be clear what is on the test,
  + the **number** of items,
    - With breakdown according to various content domains
  + **item types,**
    - Such as multiple choice, constructed response, use of graphics, etc
  + **format** of the examination to be followed for every form of the exam
+ AKA: test blueprint, test map, content map, test grid, test matrix
Test Development: Steps after Job Analysis (Form Based Exams)

- **Content:** found to be important for job as determined by job analysis
- **Sampling of content:** How many items are needed in the test form necessary to assess minimal competency?
- **Importance of content domains:** What is the emphasis on specific content domains?

- Based on identified test specifications, select items that match content domains
- Evaluate total item bank
- Pretest new items
- Evaluate statistical parameters: verify appropriate performance of items

- Review and edit items to ensure correct grammatical structure and adherence to fairness and sensitivity guidelines
- Forms assembled and equated to ensure comparability of test scores for different test forms
- Prepare test forms for administration: computer delivery

- **Outcome:** Valid, reliable test that is psychometrically sound and legally defensible
STATISTICS
Item Analysis

+ Items do not always perform the way they appear
  + Estimates about difficulty of items prior to testing are more often wrong than right.
+ To be valid, items need to behave statistically in an expected manner
+ Conduct item analysis
  + Item Discrimination
    • Point-biserial or biserial correlation (rbiserial)
  + Item Difficulty
    • Proportion correct (P+)
  + Distracter Analysis
    • Shows how options behave relative to the keyed answer
Why Are These Statistics Important?

+ Impact on Reliability
  + If items do not perform well across test, then overall reliability of the measurement is impacted
  + A test with poor reliability is also not a valid exam.

+ Impact on Fairness
  + Candidates expect to receive an exam that fairly assesses their knowledge

+ Impact on Validity of Results
Setting the Standard

CUT SCORE STUDIES
Minimum Competence

- Statement of minimal competences
  - What level of knowledge does the candidate need to have to pass the exam
    - Basic, Essential, Working
- Developed by panel members
- Applied to test items by panel members
- Reviewed by decision-makers as part of final determination of passing score
Cut Score Study Goal

+ Recommend to the Decision-making Body (NBPME):
  + A Standard
    • What a candidate needs to know and be able to do at a minimum --
      o For competence
      o License
  + A Passing Score
    • Reflective of that Standard
A Primer on Equating

+ Once a standard has been determined on a form, each subsequent form needs to be equated to the original “reference” form.

+ Score results need to be equitable:
  + Are the scores derived in a standardized manner?
  + Are the test forms equivalent?
LOFT – What is the benefit?

Linear on the Fly: Fixed length linear exam assembled dynamically at test center

Item Exposure Control
- Large number of candidates
- Continuous or window testing

Efficient Use of Item Bank
- Item Bank represents a significant investment to an organization
- Utilize all items with less exposure

Equivalent Forms
- Pre-equated
- Meet same content requirements
- Allows for on-site, immediate scoring
Goals for a LOFT Exam

**Content**
- Equal content distribution across exam
- Enemy items controlled
- Pretest items delivered

**Psychometrics**
- Each exam matches the psychometric targets
- Raw cut score within tightly controlled range

**Exposure**
- Item Exposure is controlled
- Overexposure can be set for 20%

**Result**
- Unique valid and reliable exams to all candidates
## Conversion to LOFT

<table>
<thead>
<tr>
<th>LOFT procedures</th>
<th>Combined all items with acceptable statistics administered 2008-2014</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Items reviewed by panel of subject matter experts in for accuracy and overlap</td>
</tr>
<tr>
<td></td>
<td>List of duplicate items identified by subject matter experts, with a subsequent computer scan for duplicates</td>
</tr>
<tr>
<td></td>
<td>Calibrated all approved items using 1PL Rasch Item Response Theory model</td>
</tr>
<tr>
<td></td>
<td>Conducted simulations on a final item bank</td>
</tr>
</tbody>
</table>
Benefits of IRT and LOFT Testing Model

+ **IRT statistics:**
  + Relates item difficulty to candidate ability
  + Estimates difficulty levels for each item and for the test overall
  + Estimates the accuracy of candidate scores across all levels of the ability distribution

+ **LOFT model**
  + Allows more efficient use of an item bank
  + Permits more precise psychometric targeting of cut score
  + Controls item exposure well