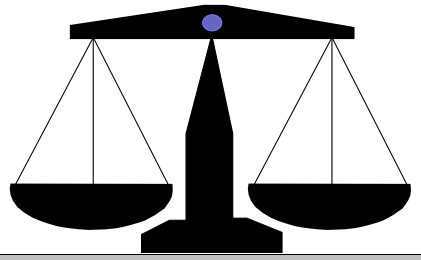
Test Development Process Flow-Chart for Licensing Examinations
*** Components of Validity, Fairness, and Reliability ***

Validity | **Fairness** | **Reliability**






Some Processes & Key Terms Associated with Each Step

Validity ("What they s*hould* know")

I—Job Analysis
- 1—Subject Matter Expert (SME) committee delineates broad tasks and pertinent knowledge, skills, and abilities (KSA's) for job performance
- 2—Practitioner survey *may* be developed for field sampling to rank & rate SME determinations

II—Test Specifications, or 'Blueprint'
- 1—Define and Label 'Content Domains'
  - a—Develop broad headings
  - b—Identify subtopics and KSA key-term clusters
- 2—Weighting of Topics and Subtopics
  - a—Determine relative importance and emphasis among broad headings and within subtopics
  - b—Establish relationship, compensating vs. noncompensating, among exam parts *("Compensating" means a strong performance in one area "compensates" for a weak showing in another rather than having to pass each.)*
- 3—Sufficiency of Sampling in Content Domains
  - a—Determine adequate coverage per area
  - b—Establish overall exam length

III—Content Outline
- 1—Published: concise presentation of broad specs
- 2—Internal: comprehensive breakdown of specs

IV—Item Conception, Creation, Review
(See next column: "Fairness – Section II, esp. #2")

Fairness ("What they w*ill* know")

I—Setting the Passing Standard, or "Cut-Score," Using SME Committee Members as 'Judges'
- 1—Establish a score-performance threshold for demonstration of ***minimal competence*** by:
  - a—Discussing concepts of minimal competence for entry-level candidates
  - b—Discussing consequences of licensing ill-prepared candidates
  - c—Orienting 'judges' to knowledge level and mindset of a representative sample of minimally competent candidates *in the testing environment*
  - d—Discussing and conducting the cut-score study
- 2—Cut-score studies: Two approaches
  - a—Angoff ('yes/no') and Modified Angoff (percent probability) studies, generally performed on items with no statistics
  - b—Adaptive approach, a.k.a. "item-mapping," performed using items with administration statistics; this is an alternative use of Computer Adaptive Testing (CAT)

II—Periodic, Routine Review to Revalidate and/or Update Test Specifications and All Items
- 1—Currency: Are topic areas and items still important? Any new ones? Any change in emphases or number of questions to ask?
- 2—Three critical questions for ***each*** item in the available-for-use pool:
  - a—Is it ***appropriate***? Does it fit under an identified outline topic?
  - b—Is it ***relevant***? Is it a big "Who cares?" Or just 'nice to know'?
  - c—Is it ***accurate***? Is the key correct, and the only one present?
- 3—Statistical properties: Is the item too hard, too easy, or confusing?

Reliability (*Trusting* scores)

I—Issues with Scores
- 1—Repetition: Same score on retake?
- 2—Comparability: Do similar scores actually indicate similar ability?

II—Exam Equating
- 1—Adjusting overall difficulty differences
  - a—internal equating: strict stat specs
  - b—score equating: using scaled scores

III—Scoring Considerations
- 1—Scoring of items
  - a—Classical statistics
    - 1—p-values: % answering correctly
    - 2—discrimination index: who gets it?
  - b—Item Response Theory (IRT)
  - c—Weighting, or adjusting 'worth'
- 2—Scoring of exams
  - a—Criterion-referenced, e.g. pass/fail
  - b—Norm-referenced & rank-ordered

IV—Score Reporting Options
- 1—Pass/Fail: referenced to a cut-score
- 2—Raw score: tally of correct answers
- 3—Percent correct: #correct / #total
- 4—Scaled scores: 'scaling' matches raw score to its counterpart on a 'scale'
- 5—Percentiles: used in 'norm-referencing'
- 6—Alternate scales: 200-800; stanines; etc.

Prepared for Summary/Overview Purposes by Morgan Testing Services ©2002 –2008. All Rights Reserved.
300 State Street, Suite 413A – Box 1982 – New London, Connecticut 06320 – (860) 857-4936 / www.MorganTestingServices.com.