A Comparison of CAT with LOFT Methods for Certification Examinations
Speakers

G. Gage Kingsbury, Ph.D., NWEA
Brian Bontempo, Ph.D., Mountain Measurement
Anthony R. Zara, Ph.D., Pearson VUE
Purposes Today

• Background: Business Problem
• Describe Linear On-The-Fly Testing (LOFT)
• Describe Computerized Adaptive Testing (CAT)
• Compare LOFT to CAT
• Situations that might benefit from LOFT or CAT
Context

• Certification exam
• Offered year round (Continuous Testing)
• Hundreds of locations around the world
• Thousands take the exam each year
• There is a thriving network of candidates that regularly share information amongst each other
Business Problem

How do we provide a fresh test?

• To avoid excessive item repeating, controlling exposure
• To assure fair and valid assessment
• To assure sufficient coverage of content domain
Solution #1: Multiple Fixed Forms

- 150 Item Pool
- 60 Item Exam
  - 10 Content Areas
  - 6 Items per Content Area
- 3 fixed forms
  - 40 unique items per form
  - 10 overlapping items
  - Spiral Design for Linking
Problems with Multiple Fixed Forms

• Repeat test takers can receive the exact same test
• Easiest to compromise the test content
• Need to equate the test forms to ensure test fairness
Solution #2: Randomly Select Items from a Pool

- 150 Item Pool
- 60 Item Exam
  - 10 Content Areas
  - 6 Items per Content Area
- Select Items at Random
- NOTE: This is very different than
  - Randomized Item Ordering
  - Randomized Response Options
Problems with Random Selection

• Some test takers receive easy items while others receive hard items
• Need to calibrate the item pool
• Need to score the exam using Item Response Theory (IRT)
Solution #3: LOFT
What is LOFT?

- LOFT is a testing process that creates a test with known psychometric and content characteristics from an item pool.
- LOFT allows test takers to see different tests, while maintaining the psychometric properties of each test.
- LOFT creates a new fixed-test form for each test taker.
- LOFT allows more control than sampling items randomly from an item pool.
Varieties of LOFT

• Shadow testing
• Target information function
• Content constraints
LOFT Considerations

• The items within a pond need to be written to be homogeneous with respect to
  – Difficulty
  – Content Area
• Item Blocking Rule
  – Difficulty Strata
  – Content Area
Example: LOFT Design

• 150 Item Pool
• 60 Item Exam
  – 10 Content areas
  – 6 Items per content area per candidate
• 30 buckets
  – 10 Content Areas X 3 Difficulty Strata
• 5 Items per bucket
Solution #4: Computerized Adaptive Testing
Computerized Adaptive Testing

- Computerized adaptive testing (CAT) uses a computer to dynamically create a unique test for each test taker.
- CAT adjusts the difficulty of the test questions for each person, creating a test that is challenging.
- CAT selects questions from a large, calibrated item pool, which makes scores comparable and reliable.
The scores from an adaptive test are as reliable and valid as those from a traditional test of twice the length.

The scores from adaptive tests allow valid interpretations that are criterion-referenced, norm-referenced and standards-referenced.
CAT Components

- IRT Model
- Pretested and Calibrated Item Pool
- Ability Estimation Algorithm
- Content Control Mechanism
- Item Selection Algorithm
- Item Exposure Control Mechanism
- Stopping Rule
Test Information Functions for Grade 4 Mathematics

Student Mean = 211.7  s.d = 11.11  Proficiency = 205  Basic = 192

RIT
0.00 0.02 0.04 0.06 0.08 0.10 0.12
165 175 185 195 205 215 225 235 245

Information
Comparability of CAT and P&P

A quick study using 1,200 grade 3 and 4 students

Spring – all students took P&P (ALT)

Fall – half CAT and half P&P
Relationship Between Spring and Fall Reading Scores

Spring RIT  vs  Fall RIT

- PP to CAT
- PP to PP

2009 NOCA Annual Educational Conference
Shine on: Exploring the Spectrum of Credentialing

Education. Innovation. Advocacy
Varieties of Adaptive Testing

- Adaptive Mastery Testing (AMT)
- Computerized Classification Testing (CCT)
- Branching Tests
- Shadow Testing
CAT Considerations

• All items need to be calibrated using Item Response Theory (IRT) before being used as operational items
• The psychometric properties of the items must be stable
• Item exposure needs to be controlled to avoid item overuse
Review

• Problem: How do we provide a fresh test?
• Solution #1: Multiple Fixed Forms (1940s)
• Solution #2: Random Item Selection (2000s)
• Solution #3: LOFT (Today)
• Solution #4: CAT (Today)
Comparing LOFT and CAT
LOFT is appropriate when...

- Small to Medium Testing Volume
- Small Item Pools
- Defined Content Structure
- Organizations that produce sufficient items to build multiple parallel forms
  - Quantity
  - Difficulty
  - Content
CAT is appropriate when…

• Medium to Large Testing Volume
• Medium to Large Item Pools
• Stable Content Domain
• Item Security is a Concern
• Interested in reducing testing time
• Organizations with ongoing item development processes
<table>
<thead>
<tr>
<th>LOFT</th>
<th>CAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smaller Item Pool</td>
<td>Larger Item Pool</td>
</tr>
<tr>
<td>Smaller Testing Volume</td>
<td>Larger Testing Volume</td>
</tr>
<tr>
<td>Longer Test</td>
<td>Shorter Test</td>
</tr>
<tr>
<td>Less Expensive to Develop</td>
<td>More Expensive to Develop</td>
</tr>
<tr>
<td>Less Precise Measurement</td>
<td>More Precise Measurement</td>
</tr>
</tbody>
</table>
Appropriate Conditions
When should you consider LOFT?

- Notice a test compromise problem
- Transitioning from event-based testing to continuous testing
- Ramping up item development efforts
When should you consider CAT?

• Need precise measurement for all test takers
• Item and security needs
• Testing time is too long
• Continuous testing desired
LOFT and CAT both provide unique tests for each test taker, within the limits of the item pool.

LOFT gives a test of similar difficulty to each test taker; CAT adjusts difficulty.

LOFT has test information similar to a single, fixed-form test; CAT can deliver equiprecise measurement or equally confident decisions by administering fewer items.
Questions and Discussion
Contact Information

• Anthony R. Zara – tony.zara@pearson.com
• Brian Bontempo – brian@mountainmeasurement.com
• G. Gage Kingsbury – gage.kingsbury@nwea.org