Improving Educational Assessment Through Computer-Based Testing

Stephen G. Sireci
University of Massachusetts Amherst
Keynote Presentation delivered at the Connecticut Assessment Forum
Rocky Hill, CT, August 13, 2010

(c) Stephen G. Sireci, 2010
Currently,

- Educational tests are used for
  - Accountability
  - Evaluating students’ strengths and weaknesses
  - Instructional planning
  - Monitoring students’ progress
  - Certification of competency
  - Many other uses

Thus, tests are being called on for more and broader purposes.
Why?

- Tests are seen as providing objective information
- in a fair (“Standardized”) manner
However,

- We are limited in what we can measure in a standardized testing format.
- Testing takes time (takes time away)
- The students we test are diverse with respect to language, culture, financial resources, parental education level, disability status, etc.
What have psychometricians done to help tests meet these demands?

- **Statistical contributions**
  - Item response theory
  - Differential item functioning
  - Test validation/evaluation techniques
  - Test equating

- **Test development contributions**
  - Universal test design
  - Computer-based testing (CBT)
  - Computerized-adaptive testing (CAT)
In this presentation, I will concentrate on CBT and how it can help us achieve the laudable and multifaceted goals driving contemporary educational assessments.
Topics to be addressed:

- Promises of CBT
- Limitations of CBT
- Introduction to computerized-adaptive testing (CAT)
  - Item-level adaptive
  - Multistage testing
- Problems to be addressed
- Future directions
CBT: The future is now

- Most, if not all, states are using or moving toward CBT
- Many large-scale testing programs use CBT:
  - Admissions Tests: GRE, GMAT, ASVAB, ACCUPLACER, TOEFL
  - Licensure tests: CPA Exam, PT Exam, NYSE, NBME, NCARB, Driver’s license, PRAXIS, MICROSOFT, NCLEX
- Formative Assessments (NWEA’s MAP)
- Massachusetts Adult Proficiency Tests
Promises of CBT

- Measure skills not measurable in paper-based formats
- Improved test security
- Reduce testing time w/o reducing measurement accuracy (CAT)
- On-demand testing
- Automatic scoring
- Immediate score reporting
- Reduced test anxiety (CAT)
Measuring New Skills

Idea is to address criticism of “selecting” correct response to “generating” a response.

- Research
- Writing
- Higher-order skills

And reduce possibility of getting item correct by guessing
Some examples of innovative computer-based items follow.

- Many were put together by my colleague April Zenisky
“Examinee ability estimates are not group-dependent. Ability estimates obtained from different sets of items will be the same (except for measurement errors), and item parameter estimates obtained in different groups of examinees will be the same (except for sampling errors). In item response theory, item and ability parameters are said to be invariant. The property of invariance of item and ability parameters is obtained by incorporating information about the items into the ability-estimation process and by incorporating information about the examinees’ abilities into the item-parameter estimation process.”

Marco has $7.00 to spend on his lunch. Click on a drink, a sandwich, and a side that Marco could have for lunch so that the price of the three items adds up to $7.00 or less.

<table>
<thead>
<tr>
<th>Drinks</th>
<th>Sandwiches</th>
<th>Sides</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1.25</td>
<td>Milk $3.75</td>
<td>$1.50 Chips</td>
</tr>
<tr>
<td>$.95</td>
<td>Tea $4.50</td>
<td>$1.00 Cookies</td>
</tr>
<tr>
<td>$1.50</td>
<td>Soda $4.75</td>
<td>$1.75 Rice</td>
</tr>
<tr>
<td>$1.25</td>
<td>Juice $5.00</td>
<td>$1.50 Green Salad</td>
</tr>
<tr>
<td>$.95</td>
<td>Coffee $4.75</td>
<td>$1.50 Fruit Salad</td>
</tr>
</tbody>
</table>
Highlighting Text (Carey, 2001; Walker & Crandall, 1999)

Directions: Click on the one value in the number sentence below that must be changed in order for the number sentence to be correct.

\[ 3 + 2 + 5 = 5 + 2 + 9 = 10 \]

This is one way to allow students to demonstrate understanding of the commutative property of addition.
The equation for the two-parameter logistic IRT model is:

\[ P_i(\theta) = \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \]

Using the equation tool below, change the equation given above to correctly represent the mathematical expression for the three-parameter logistic IRT model. (Your answer will appear in the ‘My Answer’ box after closing Equation Master.)

\[ P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta-b_i)}}{1 + e^{Da_i(\theta-b_i)}} \]
Scores on two items for five examinees are given below. Complete the table by computing the classical item difficulty for each of the items and typing it into the highlighted boxes provided.

(You may use the calculator tool as needed.)

<table>
<thead>
<tr>
<th>Examinees</th>
<th>Item 1</th>
<th>Item 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Difficulty: 0.60

Difficulty: 0.80
In the space provided, please write an essay addressing the topic given below.

“The earliest known evidence of testing dates back to Chinese employment testing thousands of years ago. Today, we are talking about computer-based testing and the Internet. Briefly (yet completely) describe three of the major developments in measurement practice that have taken place in history.”

As educational and psychological tests have found new uses in aspects of everyday life, such tests have evolved in many interesting and exciting ways from their earliest beginnings. Three critical advances that have particularly changed measurement practices include (1) the test development efforts related to the Army Alpha tests during the World War I era, (2) the theoretical innovations associated with the conceptualization of item response theory, and (3), the introduction of the computer as a mechanism for delivering tests and recording examinee responses. In this essay, I will detail how such changes have impacted…
Graphical Modeling
(Bennett, Morley, & Quardt, 2000)

1. A car starts out slowly and then goes faster and faster until a tire blows out. Graph the distance the car might have traveled as a function of time.
Drag-and-Drop Example 1:

How would you balance the scale pictured below? Drag the weights from below the scale onto the scale to make it balance.
Site Planning Vignettes (Architecture Exam Bejar, ‘91)
Dynamic Problem Solving Simulation
(Clauser, et al., 1997)

**Initial history**

**Reason(s) for Visit:**
Chest pain; respiratory distress

**History of Present Illness:**
The patient, a 65-year-old accountant, was brought to the emergency department by ambulance from the trucking company where he works. Oxygen was administered during transport. About 10 minutes before arrival, he developed excruciating, sharp pain in the right side of his chest and marked respiratory distress. The pain increases with respiration. He is unable to answer questions. A coworker who accompanied the patient to the hospital says that this never happened before, but the patient has had emphysema and asthma for years.

All other history unobtainable.
In addition to new item formats designed to measure new skills, computer-based testing also offers:

- Automated scoring of constructed responses (e.g., D. Williamson, this conference),
- Adaptive testing
What is Computerized-Adaptive Testing (CAT)?

- “Tailored Testing”
- Each examinee takes a unique test targeted to their current achievement level (“proficiency”).
- The test “adapts” itself to the examinee.
- If the examinee answers a question incorrectly, an easier question is administered.
  - and vice versa
Example of (Item-Level) CAT (Sireci, 2004)

Answering an item correctly moves the test-taker to the right.
Answering an item incorrectly moves the test-taker to the left.
General Issues for CAT

- Should you allow item review and revision (changing answers)?
- Should you allow examinees to omit items?
- How do you select items for examinees?
- When does the test end?
  - Norm-referenced/criterion-referenced algorithms
What are the advantages of CAT?

- Psychometrically more “efficient”
  - Shorter tests with same reliability
- Tailored testing is thought to reduce test anxiety and boredom, and increase interest.
- Estimate of measurement precision for each examinee as test progresses
- Less overall item “exposure.”
Can achieve = measurement precision with about $\frac{1}{2}$ # items

From Luecht & Sireci (2002)
What are some disadvantages of CAT?

- Less control over test “forms”
- Need lots of items and large sample sizes to calibrate them

For these reasons “multistage-adaptive” testing designs are becoming more popular.

One example of an MST is the MAPT: Massachusetts Adult Proficiency Tests
The Massachusetts Adult Proficiency Tests (MAPT)

• Developed to fulfill requirements of the National Reporting System in Adult Education
  – The NCLB of adult ed
  – States must develop curriculum frameworks in Math, Reading, & ESOL
  – Must measure educational gain across “educational functioning levels”
  – Tests in Math and Reading (40 items)
## Figure 2

**Multi-Panel, MST Test Structure for Current MAPT for Reading Fiscal 2009**

<table>
<thead>
<tr>
<th>Stage</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Beginning Basic Education</td>
<td>Low Intermediate Basic Education</td>
<td>High Intermediate Basic Education</td>
<td>Low/High Adult Secondary Education</td>
</tr>
<tr>
<td>Stage 1</td>
<td>15 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 2</td>
<td>5 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 3</td>
<td>5 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 4</td>
<td>5 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 5</td>
<td>5 items</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage 6</td>
<td>5 items</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- Panel A
- Panel B
More on the MAPT

- Approved by the US DOE (Office of Vocational and Adult Ed)
- Administered via the Internet through UMASS’ OWL system
- MAPT for Mathematics & Numeracy
  - Tied to the MA Adult Ed Curriculum Frameworks for Mathematics and Numeracy
- MAPT for Reading
  - Tied to the MA Adult Ed Curriculum Frameworks for ELA
## MA ABE Math Tests: Level 1

<table>
<thead>
<tr>
<th>Content Strands</th>
<th>Knowledge &amp; Comprehension</th>
<th>Application</th>
<th>Analysis, Synthesis, Evaluation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number Sense</td>
<td>15</td>
<td>15</td>
<td>5</td>
<td>35</td>
</tr>
<tr>
<td>Patterns, Functions, Algebra</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Statistics and Probability</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>Geometry and Measurement</td>
<td>10</td>
<td>10</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>41%</strong></td>
<td><strong>41%</strong></td>
<td><strong>18%</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
MAPT Test Development Process

Test Specification Meetings

Item Writing Training → Item Development → Item Review and Revision → Pilot Testing → Item Analysis

Started for Fall 2005 pilot tests and repeated for Spring 2006 pilot tests

Sensitivity Review → Content Validity Panels → Standard Setting Panels

System Testing → Operational MAPT
Development of MAIT Panels

- Systematically alter difficulty of “paths”
- Control content specifications
- Control cognitive specifications
- Control item “enemies”

Seamlessly
MAPT for Reading Test Information Functions (Panel A)
Some MAPT Statistics

- **MAPT for Math:**
  - About 7,600 tests administered last year

- **MAPT for Reading**
  - About 8,700 tests administered last year
MAPT Challenges

- Developing items: To target curriculum, over 200 teachers trained to write items
- Piloting items: Over 3,000 written; 800 operational
- Students who test more than twice may see some of the same items
- Security issues with Internet-based testing
- How do we get items out of system?
CBT, CAT, and MST

- Are they in your future?
- Probably, but let’s review some current problems.
DISadvantages of CBT

- More expensive to develop (or buy)
- Can be more expensive to administer.
  - In some cases, more expensive for test takers/districts
- Fewer people can take the test at the same time
- Entire item bank could be stolen!
- Computer proficiency may interfere with what we want to measure
- Scrolling
DISadvantages of CAT

- More complex to:
  - develop (IRT models needed)
  - administer (requires item selection)
- New (and scary) ways to “cheat”
- Harder to identify problem items
- Harder to evaluate “tests” (content validity)
- Need large item pools
- More difficult score reports
Massachusetts Adult Proficiency Test (MAPT) for Mathematics and Numeracy: Student Score Report

Site: Northampton Adult Resource Center  Class: Higgins

Student Name: David Thewlis  Student’s Score: 483  [Score Range: 459 to 507]  Test Date: 9/24/09

<table>
<thead>
<tr>
<th>Geometry and Measurement</th>
<th>Item Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>4G-1.3 Use visual methods to describe and compare shape,</td>
<td>549</td>
</tr>
<tr>
<td>4G-4.2 Fahrenheit/Celsius,</td>
<td>453</td>
</tr>
<tr>
<td>3G-3.1 Use direction, distance, maps</td>
<td>444</td>
</tr>
<tr>
<td>2G-4.8 Perimeter of rectangle</td>
<td>407</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number Sense</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4N-2.1 Correct operation for multi-step narrative problem</td>
<td>475</td>
</tr>
<tr>
<td>3N-2.3 Addition/subtraction relations up to 1,000,000</td>
<td>450</td>
</tr>
<tr>
<td>4N-3.7 Add/subtract integers</td>
<td>447</td>
</tr>
<tr>
<td>2N-1.3 Halves and quarters of quantities</td>
<td>420</td>
</tr>
<tr>
<td>3N-3.5 Rounding numbers up to 1,000,000</td>
<td>399</td>
</tr>
<tr>
<td>3N-1.5 Fraction, decimal, % equivalents ½, ¼</td>
<td>395</td>
</tr>
<tr>
<td>2N-2.3 Addition/subtraction relations up to 1,000</td>
<td>388</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patterns, Functions, and Algebra</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4P-2.1 Simple expression or representing verbalual relations</td>
<td>491</td>
</tr>
<tr>
<td>3P-3.3 Substitute the value for the variable in one-step</td>
<td>483</td>
</tr>
<tr>
<td>3P-4.1 Change in one variable relates to second variable</td>
<td>456</td>
</tr>
<tr>
<td>2P-1.1 Simple repeating number patterns up to 1,000</td>
<td>443</td>
</tr>
<tr>
<td>3P-3.4 Value of the variable in one-step equations</td>
<td>419</td>
</tr>
<tr>
<td>3P-1.1 Complete number sequences whole numbers two-step progressions</td>
<td>388</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics and Probability</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3S-2.5 Values on a bar graph</td>
<td>490</td>
</tr>
<tr>
<td>3S-1.4 Find a total from subtotaled categories</td>
<td>469</td>
</tr>
<tr>
<td>2S-2.4 Values on a bar graph</td>
<td>433</td>
</tr>
<tr>
<td>3S-5.3 Probability as a ratio in multiple forms</td>
<td>420</td>
</tr>
<tr>
<td>4S-4.6 Know when percent figures don’t add up to 100%</td>
<td>412</td>
</tr>
<tr>
<td>3S-3.3 Average (mean), range</td>
<td>397</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Geometry and Measurement</th>
<th>Item Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>3G-4.1 Add, subtract, multiply $$</td>
<td>422</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number Sense</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>3N-3.6 Common parts of whole number quantities</td>
<td>496</td>
</tr>
<tr>
<td>2N-3.2 Estimate 10, 100</td>
<td>486</td>
</tr>
<tr>
<td>2N-1.3 Halves and quarters of quantities</td>
<td>469</td>
</tr>
<tr>
<td>5N-3.2 Ratio, proportion</td>
<td>469</td>
</tr>
<tr>
<td>4N-2.1 Operation for multi-step word problem</td>
<td>459</td>
</tr>
<tr>
<td>3N-3.11 Calculations with multiplication &amp; division</td>
<td>447</td>
</tr>
<tr>
<td>3N-3.4 Basic $ calculations</td>
<td>447</td>
</tr>
<tr>
<td>3N-3.2 Calculations w/ 3-digit whole numbers</td>
<td>419</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patterns, Functions, and Algebra</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4S-4.1 graph/table connects to an argument</td>
<td>476</td>
</tr>
<tr>
<td>2P-4.2 Describe quantitative changes</td>
<td>439</td>
</tr>
<tr>
<td>2P-3.6 Expression or equation representing verbal expression</td>
<td>439</td>
</tr>
<tr>
<td>3P-1.1 Number sequences w/ whole numbers 2-step</td>
<td>439</td>
</tr>
<tr>
<td>4P-2.2 Simple formulas from tables</td>
<td>386</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics and Probability</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2S-3.3 Support simple statements with data</td>
<td>476</td>
</tr>
<tr>
<td>4S-3.3 Find the mean</td>
<td>409</td>
</tr>
</tbody>
</table>
Questions to ask when considering CBT

- Do we need to test knowledge and skills we cannot test using paper?
- Will computer literacy be an issue?
- Is computer availability a problem?
- How soon/how often can students re-test?
Summary: How can CBT Help?

1. Engage students
2. Target students’ proficiencies
3. Measure gain
4. Provide diagnostic information
5. Provide immediate results
6. Measure new skills
Summary: BUT

1. Resources are a problem
2. Aligning CBTs with curricula and instruction takes a lot of work
3. Scrolling is a problem
4. Need to take better advantage of “adaptive” capability
   - Test accommodations
   - Instruction vs. assessment
Future Research Directions

- Automated scoring of CR items
- Validating we are really testing new skills
- Ensuring construct-irrelevant variance is not present
- Web-based test delivery
- Game industry interface and sequencing
- What is best CBT design (CAT, MST, randomly equivalent, etc.) for your situation?
Future

- Internet-based testing
- Internet-based proctoring
- Adapting tests based on students’ needs
  - English language learners
  - Students with disabilities
  - Disengaged students
Closing remarks

Thanks for your attention, and thanks to Gil Andrada and CT Department of Education for the invitation.

Questions, Comments:
Sireci@acad.umass.edu