Evaluating the Psychometric Properties of Generated Test Items

Mark J. Gierl
Centre for Research in Applied Measurement and Evaluation
University of Alberta

Hollis Lai
Faculty of Medicine and Dentistry
University of Alberta

André-Philippe Boulais, André De Champlain, Claire Touchie, Debra Pugh
Medical Council of Canada

FISHBOWL SESSION

Association of Test Publishers Annual Conference
Palm Springs, CA March 1-3, 2015
Automatic item generation (AIG) is the process of using item models to generate test items with the aid of computer technology.

AIG requires three steps:

1. Content specialists begin by identifying the content required to produce new items by specifying cognitive models.

2. Then, the cognitive model content is placed in an item model—an item model is similar to a template that highlights the variables in an item that must be manipulated to produce new items.

3. Finally, computer algorithms are used to place the cognitive model content into the item models producing new test items.
A 54-year-old woman has a laparoscopic cholecystectomy. On post-operative day 3 she has a temperature of 38.5c. Physical examination reveal a red and tender wound and calf tenderness. Which one of the following is the best next step?

a. Mobilize  
b. Antibiotics  
c. Anti coagulation  
d. Reopen the wound
A structure of problem solving knowledge for issues related to post-operation fevers

**STEP #1**

**Problem and Scenarios**

- Urinary Tract Infection (UTI)
- Atelectasis (A)
- Wound Infection (WI)
- Pneumonia (P)
- Deep Vein Thrombosis (DVT)
- Deep Space Infection (DSI)

**Sources of Information**

- Timing of Fever
- Physical Examination
- Type of Surgery

**Features**

- **1-2 Days**
  - Guarding and Rebound
    - Element 1-2 days
      - Constraint
        - UTI: Very unlikely
    - No guarding or rebound
      - Constraint
        - UTI: Present

- **2-3 Days**
  - Fever
    - Element 2-3 days
      - Constraint
        - UTI: Very likely
        - WI, P, DVT: Average
    - Normal range of body temp
      - No Constraint

- **4-5 Days**
  - Element 4-5 days
    - Constraint
      - UTI: Unlikely
      - DVT: Very likely

**Abdominal Examination**

- Element Normal
  - Constraint
    - UTI, A, WI, P, DVT: Very likely
    - DSI: Unlikely
  - DSI: Tenderness

**Gastrectomy**

- Element Gastrectomy
  - Constraint
    - UTI, A, WI, P, DVT: Very likely
    - DSI: Unlikely

**Left Hemicolectomy**

- Element Left Hemicolectomy
  - Constraint
    - WI, DVT: Very likely
    - UTI, A, P, DVT: Unlikely

**Appendectomy**

- Element Appendectomy
  - Constraint
    - WI, DVT: Very likely
    - UTI, A, P, DSI: Unlikely
STEP #2: Item models are created using the cognitive model content, where an item model is like a template of the assessment task.

A 54-year-old woman has a <TYPE OF SURGERY>. On post-operative day <TIMING OF FEVER> the patient has a temperature of 38.5c. Physical examination reveal <PHYSICAL EXAMINATION>. Which one of the following is the best next step?
AIG METHODOLOGY

• **STEP #3**: After the item model is specified, we combine this information systematically to produce new items.

• To accomplish this complex combinatoric task, we created software for item generation called IGOR (Item GeneratOR)
When we used our three-stage process with 5 different surgery item models, more than 20,000 items were generated.

Item Model 1: Gallstones—288
Item Model 2: Hernias—256
Item Model 3: Aneurism—5,184
Item Model 4: Post Operation Management—7,488
Item Model 5: Post Operation Fever—7,680

TOTAL NUMBER OF GENERATED ITEMS: 20,896
### AIG Item Quality

<table>
<thead>
<tr>
<th>Item Quality Indicator</th>
<th>Traditional</th>
<th>AIG</th>
</tr>
</thead>
<tbody>
<tr>
<td>The question measures specific content, as outlined in the test specifications.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>The question is based on important topics in the curriculum and is designed to measure key thinking and problem-solving skills.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>The question is carefully edited, formatted, and presented using correct grammar, punctuation, capitalization, and spelling.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>The central idea is included in the stem, not the options.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>The stem of the question is worded positively, and avoids negatives such as NOT or EXCEPT.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>Only one of the options is clearly correct.</td>
<td></td>
<td>≈≈</td>
</tr>
<tr>
<td>The correct option is not cued due to item writing errors such as presenting a conspicuous correct options or blatantly incorrect options.</td>
<td></td>
<td>≈≈</td>
</tr>
</tbody>
</table>
AIG ITEM QUALITY

SUBSTANTIVE RESULTS

• Predictive accuracy was computed using a *relative criterion*, meaning that a perfectly accurate outcome is determined relative to the total number of AIG items correctly identified by the panelist.

• In other words, if a panelist identifies 5 AIG items correctly and 0 AIG items incorrectly, then predictive accuracy is 100%.

• The overall predictive accuracy for the panelists was 42% indicating that they were able to correctly identify the AIG items on the evaluation test about 40% of the time.
• Medical test items were generated in 4 different content areas—generated items that each had a different correct option and satisfied an important blueprint requirement were included in the field test administration

• The candidates included 4,101 medical graduates

• Our results focus on the psychometric outcomes for 9 pediatric test items generated from a single item model (neonatal jaundice)
### AIG ITEM QUALITY

#### RESULTS—CORRECT OPTION

<table>
<thead>
<tr>
<th>Generated Item</th>
<th>Classical Item Analysis</th>
<th>IRT Item Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion Correct</td>
<td>Biserial Correlation</td>
</tr>
<tr>
<td>1 EASY</td>
<td>0.95</td>
<td>0.37</td>
</tr>
<tr>
<td>2 LOW DISCRIM</td>
<td>0.80</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>4 HIGH DISCRIM</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>5</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>0.16</td>
</tr>
<tr>
<td>7</td>
<td>0.93</td>
<td>0.14</td>
</tr>
<tr>
<td>8</td>
<td>0.52</td>
<td>0.36</td>
</tr>
<tr>
<td>9 HARD</td>
<td>0.19</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Mean (Std. Dev.)

- Classical Item Analysis: Proportion Correct 0.50 (0.29)  
- Classical Item Analysis: Biserial Correlation 0.25 (0.10)  
- IRT Item Analysis: \( b \) 0.20 (3.28)  
- IRT Item Analysis: \( a \) 0.46 (0.22)
## RESULTS—DISTRACTORS

<table>
<thead>
<tr>
<th>Item</th>
<th>Difference</th>
<th>Biserial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 EASY</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>-0.01</td>
<td>-0.56</td>
</tr>
<tr>
<td>b</td>
<td>-0.01</td>
<td>-0.56</td>
</tr>
<tr>
<td>c*</td>
<td>0.10</td>
<td>0.37</td>
</tr>
<tr>
<td>d</td>
<td>-0.01</td>
<td>-0.27</td>
</tr>
<tr>
<td>e</td>
<td>-0.06</td>
<td>-0.26</td>
</tr>
<tr>
<td><strong>9 HARD</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>-0.09</td>
<td>-0.16</td>
</tr>
<tr>
<td>b</td>
<td>-0.02</td>
<td>-0.24</td>
</tr>
<tr>
<td>c</td>
<td>0.00</td>
<td>-0.20</td>
</tr>
<tr>
<td>d*</td>
<td>0.14</td>
<td>0.29</td>
</tr>
<tr>
<td>e</td>
<td>-0.03</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

*correct option*
## RESULTS—DISTRACTORS

<table>
<thead>
<tr>
<th>Item</th>
<th>Difference</th>
<th>Biserial</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 LOW DISCRIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>-0.02</td>
<td>-0.31</td>
</tr>
<tr>
<td>b</td>
<td>-0.03</td>
<td>-0.66</td>
</tr>
<tr>
<td>c*</td>
<td><strong>0.03</strong></td>
<td><strong>0.10</strong></td>
</tr>
<tr>
<td>d</td>
<td>-0.02</td>
<td>-0.23</td>
</tr>
<tr>
<td>e</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>4 HIGH DISCRIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>-0.06</td>
<td>-0.46</td>
</tr>
<tr>
<td>b</td>
<td>-0.05</td>
<td>-0.23</td>
</tr>
<tr>
<td>c</td>
<td>-0.14</td>
<td>-0.16</td>
</tr>
<tr>
<td>d*</td>
<td><strong>0.28</strong></td>
<td><strong>0.38</strong></td>
</tr>
<tr>
<td>e</td>
<td>-0.02</td>
<td>-0.04</td>
</tr>
</tbody>
</table>

*correct option*
1. What are the current challenges we face in developing test items?

2. We began our research in 2007 because we could not get enough items for our computer adaptive tests. Can the current demand for test items be satisfied using the traditional item development approach? (Do we need AIG?)

3. To demonstrate that AIG has some potential for producing operational items, we gathered evidence on both substantive and statistical outcomes. The results look promising. But is this evidence enough? In other words, what motivates organization to change?
THANK YOU

Dr. Mark J. Gierl (mark.gierl@ualberta.ca)
Automatic item generation is the process of using item models to generate test items with the aid of computer technology—it requires a merger between the “art” and “science” of item development.

The **test developer** identifies the knowledge and skills required to solve problems—we associate these activities with the “art” of test development because it requires judgement, expertise, and experience.

**Computer technology** is required for the generative task of systematically combining large amounts of information in each item model—we associate this with the “science” of modern computing.

By merging the outcomes from the content-based creative task with the technology-based generative task, automated processes can be used to promote a **new approach to item development**.