Moving beyond Efficiency to Allow CAT to Provide Better Diagnostic Information

Brian D. Bontempo, Ph.D.
*Mountain Measurement, Inc.*
www.mountainmeasurement.com
Background

• As CAT researchers and practitioners, our focus is typically on how to build more efficient CATs.

• For a moment, let’s focus on what do we do with the CAT data.
Common Test Score Report

John Doe  Oct 1, 2011
PASS
The Future

http://scores.espn.go.com/mlb/boxscore?gameldId=310928130

Final in 12

Yankees 7  Rays 8
(97-65, 45-36 away)  (91-71, 47-34 home)

7:00 PM ET, September 28, 2011
Tropicana Field, St. Petersburg, Florida

W: J. McGee (5-2)
L: S. Proctor (0-3)

New York Yankees

<table>
<thead>
<tr>
<th>Hitters</th>
<th>AB</th>
<th>R</th>
<th>H</th>
<th>RBI</th>
<th>BB</th>
<th>SO</th>
<th>#P</th>
<th>AVG</th>
<th>OBP</th>
<th>SLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Jeter SS</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>23</td>
<td>.297</td>
<td>.355</td>
<td>.388</td>
</tr>
<tr>
<td>R. Pena 2B</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>13</td>
<td>.100</td>
<td>.159</td>
<td>.175</td>
</tr>
<tr>
<td>C. Granderson CF</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
<td>7</td>
<td>.262</td>
<td>.364</td>
<td>.552</td>
</tr>
<tr>
<td>a-G. Golson PH-CF-LF</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>15</td>
<td>.182</td>
<td>.250</td>
<td>.182</td>
</tr>
<tr>
<td>M. Teixeira 1B</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td></td>
<td>17</td>
<td>.248</td>
<td>.341</td>
<td>.494</td>
</tr>
<tr>
<td>E. Chavez 3B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>6</td>
<td>.263</td>
<td>.320</td>
<td>.356</td>
</tr>
<tr>
<td>R. Cano DH</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>21</td>
<td>.302</td>
<td>.349</td>
<td>.523</td>
</tr>
<tr>
<td>b-J. Posada PH-DH</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>7</td>
<td>.233</td>
<td>.315</td>
<td>.398</td>
</tr>
<tr>
<td>N. Swisher RF</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>17</td>
<td>.260</td>
<td>.374</td>
<td>.449</td>
</tr>
</tbody>
</table>

Tampa Bay Rays

<table>
<thead>
<tr>
<th>Hitters</th>
<th>AB</th>
<th>R</th>
<th>H</th>
<th>RBI</th>
<th>BB</th>
<th>SO</th>
<th>#P</th>
<th>AVG</th>
<th>OBP</th>
<th>SLG</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Jennings LF</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>27</td>
<td>.259</td>
<td>.356</td>
<td>.449</td>
</tr>
<tr>
<td>B. Upton CF</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
<td>27</td>
<td>.243</td>
<td>.331</td>
<td>.429</td>
</tr>
<tr>
<td>E. Longoria 3B</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td></td>
<td>31</td>
<td>.244</td>
<td>.355</td>
<td>.495</td>
</tr>
<tr>
<td>M. Joyce RF</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
<td>9</td>
<td>.277</td>
<td>.347</td>
<td>.478</td>
</tr>
<tr>
<td>b-R. Canzler PH</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>5</td>
<td>.333</td>
<td>.400</td>
<td>.333</td>
</tr>
<tr>
<td>J. Ruggiano RF</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>.248</td>
<td>.273</td>
<td>.400</td>
</tr>
<tr>
<td>c-J. Jaso PH</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
<td>.324</td>
<td>.299</td>
<td>.354</td>
</tr>
<tr>
<td>E. Johnson PR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td>0</td>
<td>.194</td>
<td>.257</td>
<td>.338</td>
</tr>
<tr>
<td>J. Lobaton C</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td>3</td>
<td>.116</td>
<td>.231</td>
<td>.147</td>
</tr>
</tbody>
</table>
Types of Reports

- Examinee Reports
- Group Reports (School, Company, Jurisdiction)
- Psychometric Reports
Evolution of Reporting - Examinee Content

- Pass/Fail Decision -> Scores
- Achievement -> Growth
- Raw/Percent Scores -> Universal Ability Estimates
- Overall Score -> Content Subscores
Issues in CAT Examinee Reports

- When modeled unidimensionally, subscores have limited utility.
- Error associated with subscore ability estimates is high because there are not enough items in each content area.
- Item selection is not optimal because it is based on efficiency of overall score.
- Person fit is useless.
Solutions to CAT Examinee Reports

- When modeled unidimensionally, content area scores have limited utility
  - Increase Randomization in Item Selection
  - MIRT for subscores
  - Multiple unidimensional scales for subscores

- Not enough items in each content area
  - Non-mutually exclusive content categorization
  - Implement Multiple Content “Schema”
Solutions to CAT Examinee Reports

• Item selection is not optimal because it is based on efficiency of overall score
• Person fit is useless
  - Smart CATs which add some items that are intentionally selected to
    • ISV
    • Person Fit
    • Cheating
    • More information in smaller test areas
Research Questions for CAT Examinee Reports

• What are the strengths and weaknesses associated with each model for adding more items?
  - Use all items in overall score (But this violates the test blueprint)
  - Select item responses at random from each content area to meet blueprint
  - Select item responses intentionally (highest ISV) from each content area to meet blueprint
  - Select the first set of items administered that meet blueprint
Evolution of Reporting - Group Content

- Passing Rates -> Mean Scores
- Descriptive Statistics -> Distribution
- Normative -> Criterion Referenced
- Percentile Ranks -> Distance to Goal (e.g., cutscore)
- Comparisons -> Profiling
Issues in CAT Group Reports

- Individual subscore ability estimates typically have large amounts of error
- Group sizes may be small
Solutions to CAT Group Reports

- Aggregate the data over the group and calculate subscore ability estimates for the group directly
Research Questions for CAT Group Reports

- What are the parameters around “group ability estimates” being more accurate than the median of individual ability estimates when the sample sizes are small?
  - How many items per person or SEM per person
  - How many examinees per group
Evolution of Reporting - Psychometric Content

- Error
- Fit
- Cheating
- Speed/Time
- Motivation
- Validity
Issues in CAT Psychometric Reports

- Traditional item quality statistics have limited usefulness
Solutions to CAT Psychometric Reports

- Implement item selection algorithms that increase the dispersion of the ability of the examinees taking each item
Research Questions for CAT Psychometric Reports

• How can the progressive item selection Algorithm be used to improve dispersion?
  - Bontempo, Kingsbury & Zara (2010)

• Are there other item selection algorithm approaches that might improve dispersion?
The Future of Test Reporting

- Data Visualization
- Interactivity
- Data Driven Interpretation
- MULTImedia
- Integration
Conclusion

- Future CATs will provide better diagnostic information to
  - Examinees
  - Regulators, Educators, Employers
  - Test Developers
- This goal will be accomplished by
  - Smart CATs which collect additional information during the test
  - Psychomagic
- The time is now for Reporting
Thank you for your kind attention

brian@mountainmeasurement.com