Item Response Theory (IRT) Models for Questionnaire Evaluation: Response to Reeve

Ron D. Hays

October 22, 2009, ~3:45-4:05pm
http://twitter.com/RonDHays
http://gim.med.ucla.edu/FacultyPages/Hays/
Features of IRT with diagnostic utility

- Category response curves
- Information/reliability
- Differential item functioning
- Person fit
- Computer-adaptive testing
Category Response Curves (CRCs)

- Reeve’s Figure 7 showed that 2 of 6 response options are never most likely to be chosen
  - No, very small, small, moderate, great, very great change

- He suggests 1 or both of the response categories could be dropped or reworded to improve the response scale
Figure 7

Appreciating each day.

Probability of Response

Posttraumatic Growth

No change

Small change

Very small change

Moderate change

Great change

Very great change

Posttraumatic Growth

θ
Drop response options?

• No, very small, small, moderate, great, very great change

→

• No, moderate, great, very great change
Reword?

• Might be challenging to determine what alternative wording to use so that the replacements are more likely to be endorsed.
Keep as is?

- CAHPS global rating items
  - 0 = worst possible
  - 10 = best possible
- 11 response categories capture about 3 levels of information.
  - 10/9/8-0 or 10-9/8/7-0
- Scale is administered as is and then collapsed in analysis
Information/Reliability

• For z-scores (mean = 0 and SD = 1):
  – Reliability = 1 – SE² = 0.90 (when SE = 0.32)
  – Information = 1/SE² = 10  (when SE = 0.32)
  – Reliability = 1 – 1/information

• Lowering the SE requires adding or replacing existing items with more informative items at the target range of the continuum.
  – But this is …
Easier said than done

• Limit on the number of ways to ask about a targeted range of the construct
• One needs to avoid asking the same item multiple times.
  – “I’m generally said about my life.”
  – “My life is generally sad.”
• Local independence assumption
  – Significant residual correlations
### Item parameters (graded response model) for global physical health items in Patient-Reported Outcomes Measurement Information System

<table>
<thead>
<tr>
<th>Item</th>
<th>A</th>
<th>b1</th>
<th>b2</th>
<th>b3</th>
<th>b4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global01</td>
<td>7.37 (na)</td>
<td>-1.98 (na)</td>
<td>-0.97 (na)</td>
<td>0.03 (na)</td>
<td>1.13 (na)</td>
</tr>
<tr>
<td>Global03</td>
<td>7.65 (2.31)</td>
<td>-1.89 (-2.11)</td>
<td>-0.86 (-0.89)</td>
<td>0.15 (0.29)</td>
<td>1.20 (1.54)</td>
</tr>
<tr>
<td>Global06</td>
<td>1.86 (2.99)</td>
<td>-3.57 (-2.80)</td>
<td>-2.24 (-1.78)</td>
<td>-1.35 (-1.04)</td>
<td>-0.58 (-0.40)</td>
</tr>
<tr>
<td>Global07</td>
<td>1.13 (1.74)</td>
<td>-5.39 (-3.87)</td>
<td>-2.45 (-1.81)</td>
<td>-0.98 (-0.67)</td>
<td>1.18 (1.00)</td>
</tr>
<tr>
<td>Global08</td>
<td>1.35 (1.90)</td>
<td>-4.16 (-3.24)</td>
<td>-2.39 (-1.88)</td>
<td>-0.54 (-0.36)</td>
<td>1.31 (1.17)</td>
</tr>
</tbody>
</table>

Note: Parameter estimates for 5-item scale are shown first, followed by estimates for 4-item scale (in parentheses). na = not applicable

Global01: In general, would you say your health is …? Global03: In general, how would you rate your physical health? Global06: To what extent are you able to carry out your everyday physical activities? Global07: How would you rate your pain on average? Global08: How would you rate your fatigue on average?

a = discrimination parameter; b1 = 1st threshold; b2 = 2nd threshold; b3 = 3rd threshold; b4 = 4th threshold
Differential Item Functioning (DIF)

• Probability of choosing each response category should be the same for those who have the same estimated scale score, regardless of their other characteristics

• Evaluation of DIF
  – Different subgroups
  – Mode differences
  – Different response options
Person Fit

• Large negative $Z_L$ values indicate misfit.

• Person responded to 14 items in physical functioning bank ($Z_L = -3.13$)
  – For 13 items the person could do the activity (including running 5 miles) without any difficulty.
  – However, this person reported a little difficulty being out of bed for most of the day.
Unique predictors of person misfit

- Less than high school education
- Non-white
- More chronic conditions
Computer Adaptive Testing (CAT)
http://www.nihpromis.org/

- Patient-reported outcomes measurement information system (PROMIS) project
  - Item banks measuring patient-reported outcomes
  - Computer-adaptive testing (CAT) system
PROMIS Banks (454 items)
http://www.assessmentcenter.net/ac1/

- Emotional Distress
  - Depression (28)
  - Anxiety (29)
  - Anger (29)
- Physical Function (124)
- Pain
  - Behavior (39)
  - Impact (41)
- Fatigue (95)
- Satisfaction with Participation in Discretionary Social Activities (12)
- Satisfaction with Participation in Social Roles (14)
- Sleep Disturbance (27)
- Wake Disturbance (16)
Time to complete item

- Polimetrix panel sample
- 12-13 items per minute (automatic advance)
- 8-9 items per minute (next button)
  - 6 items per minute among UCLA Scleroderma patients
CAT

• Context effects (Lee & Grant, 2009)
  – 1,191 English and 824 Spanish respondents to 2007 California Health Interview Survey
  – Spanish respondents self-rated health was worse when asked before compared to after questions about chronic conditions.
### Differential Item Functioning (DIF) Information

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PROMIS Male - Female</td>
<td>1.22</td>
<td>0.2700</td>
<td>0.32</td>
<td>0.5695</td>
</tr>
</tbody>
</table>

### Calibration Sample

- **Description:** PROMIS Promis Wave 1
- **IRT Model:** Graded Response Model
- **Category Response Function**
- **Sample Population:** PROMIS Wave 1 Full Bank
- **Item Information Function**

### Model Fit & Scalability Indices

<table>
<thead>
<tr>
<th>Value</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S - X^2$</td>
<td>73.87</td>
</tr>
<tr>
<td>$S - G^2$</td>
<td>79.41</td>
</tr>
<tr>
<td>$X^2$</td>
<td>33.00</td>
</tr>
<tr>
<td>$G^2$</td>
<td>49.90</td>
</tr>
<tr>
<td>Loewinger H Scalability</td>
<td>0.00</td>
</tr>
<tr>
<td>Location At Max Information</td>
<td>-0.10</td>
</tr>
</tbody>
</table>

Threshold 1 is between the lowest scored response and the next highest response(s).

### IRT Parameters

<table>
<thead>
<tr>
<th>Slope ($a$)</th>
<th>Guessing ($c$)</th>
<th>Threshold (CB) 1</th>
<th>Threshold (CB) 2</th>
<th>Threshold (CB) 3</th>
<th>Threshold (CB) 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.56467</td>
<td>0</td>
<td>-1.1758</td>
<td>0.1564</td>
<td>1.8111</td>
<td>3.2708</td>
</tr>
</tbody>
</table>
Assessment Center/Q-Bank
Thank you!