AN INTRODUCTION TO ITEM RESPONSE THEORY AND ITS APPLICATIONS TO HEALTH ASSESSMENTS

Qiong (Joan) Wu
Harvard Center for Population and Development Studies

INDEPTH-SAGE WORKSHOP
April 20, 2010
ITEM RESPONSE THEORY

- IRT vs Classical test theory (CTT)
  + CTT: focuses test scores
    observed score = true score + error (O=T+E)
  + IRT: focuses on individual item characteristics

- IRT is a scaling method
  + Assigns numerical scores based on a set of item responses

- IRT is an item analysis tool
  + Evaluates quality of individual items based on estimated item parameters
ITEM RESPONSE THEORY

- Also called latent trait models
  - Many variables in health assessments cannot be measured directly, such as physical functioning ability, fatigue, depression etc.

Physical functioning ability
- Walking more than 1 mile
- Climbing flights of stairs
- Extending arms above shoulder
- Stooping/kneeling
- Picking up a coin
ALTERNATIVE WAYS OF DERIVING SUCH SCORES

- Raw total/Raw percentage
  - Items weighted equally
  - Item dependent

- Factor scores
  - Usually assume continuous observed variables
CHARACTERISTICS OF IRT SCALED SCORES

- Designed for dichotomous or polytomous items
- Pattern scoring instead of number scoring
- When the assumptions are met
  - Scores are item invariant
- Considered equal-interval, so preferred scores for longitudinal analysis
PARAMETERS IN DICHTOMOUS IRT

- Person parameter (\(\theta\))
  + Latent scores, theta scores (mean=0, SD=1)
- Item parameters
  + Item difficulty/location (b), usually -3<b<3
  + Item discrimination (a), >=0
  + Pseudo-guessing (c), 0<=c<=1
- Item difficulty/location and theta scores on the same scale
**MATHEMATICAL MODELS FOR BINARY DATA**

- **One-parameter IRT (Rasch model)**
  \[ P_i(\theta) = \frac{e^{\theta-b_i}}{1 + e^{\theta-b_i}} \]

- **Two-parameter IRT**
  \[ P_i(\theta) = \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}} \]

- **Three-parameter IRT**
  \[ P_i(\theta) = c_i + (1-c_i) \frac{e^{a_i(\theta-b_i)}}{1 + e^{a_i(\theta-b_i)}} \]
During the past week, did you ever feel sad?

Theta scores

Probability of a “yes” response

Item1 (b = 0)

P=0.5

a = 2

b = 0
ITEMS WITH DIFFERENT DIFFICULTY/LOCATION
ITEMS WITH DIFFERENT DISCRIMINATION
POLYTOMOUS IRT

- Deals with items with more than 2 response options
- Model probability of each response option conditional on latent trait

<table>
<thead>
<tr>
<th>Q1025</th>
<th>In the last 30 days, how much difficulty did you have ...</th>
<th>None</th>
<th>Mild</th>
<th>Moderate</th>
<th>Severe</th>
<th>Extreme/cannot do</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>... in standing for long periods (such as 30 minutes)?</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>98</td>
</tr>
</tbody>
</table>
POLYTOMOUS IRT (GRADED RESPONSE MODEL)

- Parameters
  - Latent trait: theta
  - Discrimination parameters
  - Threshold parameters

- Dichotomous IRT
  - Theta
  - Discrimination par
  - Difficulty/location par

- Measurement questions
  - How many categories are needed?
  - Are response categories really different one another?

\[ P(X_i = k \mid \theta) = \frac{1}{1 + \exp(\alpha_i(\theta - \delta_{ix}))} - \frac{1}{1 + \exp(\alpha_i(\theta - \delta_{ix-1}))} \]
AN IDEAL ICC FOR A 3-OPTION ITEM

How often do you feel sad?

1: rarely
2: sometimes
3: most or all of the time
A REAL ITEM

1: rarely or none of the time;
2: some or a little of the time
3: occasionally or a moderate amount;
4: most or all of the time

In the last week, I felt everything I did was an effort.
ASSUMPTIONS OF IRT

- Unidimensionality
- Local independence (residual covariance = 0)

Unidimensionality implies LI, not vice versa.

Physical functioning ability:
- Walking more than 1 mile
- Walking more than 1 block
- Bathing/dressing oneself
- Doing laundry
- Feeding self
APPLICATIONS OF IRT

- Test construction and item analysis
- Computerized adaptive testing (CAT) and item banking
- Test equating/linking
- Differential item functioning/item bias
TEST CONSTRUCTION AND ITEM ANALYSIS

- Basic rule
  - Select items with appropriate $b$ & high $a$ for dichotomous items
  - Check ICCs for polytomous items
CAT AND ITEM BANKING

- Basic idea
  + Items tailored to individuals’ trait levels

- Why do we need it?
  + Reduce test length, minimize fatigue
  + Minimize floor/ceiling effects

- Challenges
  + Large item pool
  + Large sample of subjects for initial item calibration
CAT AND ITEM BANKING (2)

- Different types of adaptive testing
  - Two-stage testing

*Figure 1. A Model of Two-Stage Testing*
C.A.T AND I.T.E.M B.A.N.K.I.N.G (2)

- Different types of adaptive testing
  - Multi-stage testing
TEST EQUATING/LINKING

- Definition
  - Setting a common metric for scores from tests composed of different sets of items

- When do we need it?
  - When two groups of subjects take different test forms

- How do we do it?
  - Needs a valid link, either through common-item design, or common subject sample design
TEST EQUATING/LINKING

+ Linking test scores from two test forms (common item design)
  - Group A

  \[ \text{common items} \rightarrow \text{unique items} \]

  \[ \text{unique items} \rightarrow \text{common items} \]

  - Group B

  \[ \text{common items} \rightarrow \text{unique items} \]

  \[ \text{unique items} \rightarrow \text{common items} \]

  - Fixing item parameters to be equal across two forms
Differential Item Functioning

- Definition
  - different groups of subjects display different probabilities of endorsing a response option conditional on latent trait

- Why do we care?
  - Test/survey bias when those items are included

- Anchor items
  - Use anchor items to equate the groups on the latent trait. IRT often use rest items
DIF ACROSS GENDER GROUPS

- “Do you have any difficulties with making phone calls?”

![Functional ability graph showing comparison between country A and country B.](image-url)
DEMYSTIFYING IRT

- High correlation between number right scores and IRT scaled scores

- IRT estimates can be sample dependent (DIF)

- Large-sample technique
  - >200 for 1PL models; >400 for 2PL models; >600 for 3PL models
WHEN IS IRT RECOMMENDED

- Computerized adaptive testing and test construction
  - A large item pool AND
  - A large number of subjects
- Test scoring when item discrimination vary a lot
- Cross-cultural comparison
- Longitudinal analysis
OTHER ISSUES

- Multidimensional IRT

- Specialized software program
  + BILOG, MULTILOG, Mplus
  + R module