Item Response Theory (IRT): Enhancing Health Outcomes Measurement

Bryce B. Reeve, Ph.D.

e-mail: bbreeve@email.unc.edu
Presentation Overview

• IRT Models
  – Theory of IRT (Reeve)
  – IRT item, scale, and person properties (Reeve)
  – Comparison with Classical Test Theory (Reeve)
  – IRT Assumptions and Model Fit (Orlando Edelen)
  – IRT Scoring (Orlando Edelen)

• Applying IRT to enhancing health outcomes measurement
  – Designing and evaluating scales (Siemons; Krishnan)
  – Assessing Differential Item Functioning (DIF) (Orlando Edelen)
  – Linking scales (Glas; Oude Voshaar)
  – Item Banking and Computerized Adaptive Testing (Bjorner; Nikolaus)
Please Note:

• The **quality** of a health outcomes measure is related to the attention the developer(s) took to use **qualitative** and **quantitative** methods integrating **multiple perspectives** throughout the process.

• IRT Methods do not replace the **classical/traditional test theory** methods for item/scale analysis.

• IRT analysis is **not a magic wand**!
  – It cannot fix bad data or poorly defined constructs
  – By itself, it does not address all forms of validity and other attributes that evaluate the quality of a questionnaire.
# The Need for Better Outcome Measures

<table>
<thead>
<tr>
<th>Needs</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop measures that are valid, reliable, and sensitive to detect clinically meaningful change</td>
<td>Have a minimum set of questions to reduce respondent burden.</td>
</tr>
<tr>
<td>Different forms of an instrument to measure different health levels.</td>
<td>Different forms to be linked on the same metric for group comparisons</td>
</tr>
<tr>
<td>Non-biased measurement across groups</td>
<td>Detect differences in group perceptions</td>
</tr>
</tbody>
</table>
What is Item Response Theory (IRT)?

- IRT is designed for:
  - Modeling latent “unobservable” variables (traits, domains, \( \theta \))
  - Multi-item Scales/Questionnaires
IRT Model: Item Characteristic Curves

I am unhappy some of the time?

Depression

Probability of Response

False

True

θ
IRT Model

I am unhappy some of the time?

Probability of Response

b (item) parameter:
- Threshold
- Location
- Difficulty
- Severity

b = .25
IRT Models

I am unhappy some of the time.
I don’t care what happens to me.

$b$ parameter:
- Threshold
- Location
- Difficulty
- Severity

$b = .25$
$b = 1.33$
IRT Model

I am unhappy some of the time?

Probability of Response

Depression

True

$a = 2.83$

$b = .25$

$a$ (item) parameter:
- slope
- discrimination
- relationship with trait

None

Depression

Severe

\( \theta \)
IRT Models

I am unhappy some of the time.
I don’t care what happens to me.
I cry easily.

\[ a = 2.20 \]
\[ a = 2.83 \]
\[ a = 1.11 \]

\[ b = -0.23 \]
\[ b = 0.25 \]
\[ b = 1.33 \]
IRT Model: Item Characteristic Curves

I am unhappy some of the time?

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}$$

Probability of Response

Depression

None

Severe

$$\theta$$
IRT: Item Information Curves
(The range of the latent construct over which an item is most useful for distinguishing among respondents)

- I am unhappy some of the time
- I don’t care what happens to me
- I cry easily

\[ a = 2.83 \]
\[ a = 2.20 \]
\[ a = 1.11 \]

Depression

None

Severe

\[ b = -0.23 \]
\[ b = 0.25 \]
\[ b = 1.33 \]
Building reliable and efficient measures...

I am unhappy some of the time.

I don’t seem to care what happens to me.

I cry easily.

10 Items from the MMPI-2 Depression Scale
Scale (Test) Information Curve 
(The range of the latent construct over which a scale is most useful for distinguishing among respondents)

\[ TI(\theta) = \sum_{i=1}^{m} I(\theta) \]

Reliability \( y = 1 - \text{Error Variance} \)
\[ = 1 - \left( \frac{1}{\text{Information}} \right)^2 \]

\[ = 1 - \frac{1}{\text{Information}} \]
Questions on the MMPI-2 depression scales were chosen because they maximally discriminate a clinically depressed group from a non-clinical group.
Standard Error of Measurement Curve
(The range of the latent construct over which a scale
is most useful for measuring respondent trait levels)

\[
\text{SEM} = \frac{1}{\sqrt{I}}
\]
What is the reduction in information going from a 22 to 12 item scale?

$r = .80$
$r = .90$
$r = .93$
$r = .95$

$\theta$

* $r =$ approximate reliability
What about IRT models for questions with more than two response categories?

Data from responses to the PROMIS Depression Item Bank.
In the past 7 days, I felt unhappy.

In the past 7 days, I felt I had no reason for living.
Item Response Theory (IRT)

In the past 7 days, I felt unhappy.

Category Response Curves

Never

Rarely

Some times

Often

Always

Depressive Symptoms

very mild

severe

Information Function

very mild

severe

very mild

severe
In the past 7 days, I felt I had no reason for living.
I felt unhappy.
I felt depressed.
I withdrew from other people.
I felt worthless.
I felt I had no reason for living.

1. Never
2. Rarely
3. Sometimes
4. Often
5. Always
IRT Family of Models

IRT models come in many varieties (over a 100) to handle:

- Unidimensional and multidimensional data
- Binary, polytomous, and continuous response data
- Ordered as well as unordered response data
<table>
<thead>
<tr>
<th>Model</th>
<th>Item Response Format</th>
<th>Model Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch / 1-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination power equal across all items. Threshold varies across items.</td>
</tr>
<tr>
<td>2-Parameter Logistic</td>
<td>Dichotomous</td>
<td>Discrimination and threshold parameters vary across items.</td>
</tr>
<tr>
<td>Graded Response</td>
<td>Polytomous</td>
<td>Ordered responses. Discrimination varies across items.</td>
</tr>
<tr>
<td>Nominal</td>
<td>Polytomous</td>
<td>No pre-specified item order. Discrimination varies across items.</td>
</tr>
<tr>
<td>Partial Credit (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination power constrained to be equal across items.</td>
</tr>
<tr>
<td>Rating Scale (Rasch Model)</td>
<td>Polytomous</td>
<td>Discrimination equal across items. Item threshold steps equal across items.</td>
</tr>
<tr>
<td>Generalized Partial Credit</td>
<td>Polytomous</td>
<td>Variation of Partial Credit Model with discrimination varying among items.</td>
</tr>
</tbody>
</table>
Applications of IRT models for Health Outcomes Measurement
1. Design and Evaluation

I’m able to do better things with my life.
1. Design and Evaluation
1. Design and Evaluation
2. Testing for Differential Item Functioning (DIF)

In the past 7 days, I cried

In the past 7 days, I felt blue

<table>
<thead>
<tr>
<th>None of the time</th>
<th>A little of the time</th>
<th>Some of the time</th>
<th>Most of the time</th>
<th>All of the time</th>
</tr>
</thead>
<tbody>
<tr>
<td>None of the time</td>
<td>A little of the time</td>
<td>Some of the time</td>
<td>Most of the time</td>
<td>All of the time</td>
</tr>
</tbody>
</table>

Depression
3. Linking Health Outcome Measures

PROMIS Depression Measure  ↔  CES Depression Scale
3. Linking Health Outcome Measures

- PROMIS Depression Measure
- Becks Depression Inventory
- CES Depression Scale

Depression Scale:
- 20
- 30
- 40
- 50
- 60
- 70
- 80
4. Item Banking and Computerized Adaptive Testing (CAT)

In the past 7 days, I felt unhappy:

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Always</th>
</tr>
</thead>
</table>

no depression  |  |  |  |  |
mild depression |  |  |  |  |
moderate depression |  |  |  |  |
severe depression |  |  |  |  |
extreme depression |  |  |  |  |

Depression Item Bank

Item 1  | Item 2  | Item 3  | Item 4  | Item 5  | Item 6  | Item 7  | Item 8  | Item 9  | Item n
Traditional Measurement Theory (Classical Test Theory, CTT) versus Modern Measurement Theory
<table>
<thead>
<tr>
<th>Classical Test Theory</th>
<th>Item Response Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures of precision fixed for all scores</td>
<td>Precision measures vary across scores</td>
</tr>
<tr>
<td>Longer scales increase reliability</td>
<td>Shorter, targeted scales can be equally reliable</td>
</tr>
<tr>
<td>Scale properties are sample dependent</td>
<td>Item &amp; scale properties are invariant within a linear transformation</td>
</tr>
<tr>
<td>Comparing person scores dependent on item set</td>
<td>Person scores comparable across different item sets</td>
</tr>
<tr>
<td>Comparing respondents requires parallel scales</td>
<td>Different scales can be placed on a common metric</td>
</tr>
<tr>
<td>Mixed item formats leads to unbalanced impact on total scale scores</td>
<td>Easily handles mixed item formats</td>
</tr>
<tr>
<td>Summed scores are on an ordinal scale</td>
<td>Scores on interval scale</td>
</tr>
<tr>
<td></td>
<td>Graphical tools for item and scale analysis</td>
</tr>
</tbody>
</table>
Questions on the MMPI-2 depression scales were chosen because they maximally discriminate a clinically depressed group from a non-clinical group.
Conclusions

• IRT serves as a powerful analytic tool to help design health outcomes measures.

• Limitations
  – Lack of user-friendliness of software
  – Required knowledge of measurement theory.
  – Needs large sample sizes
Important Deadlines

**April 12:**
Oral and Poster Presentation Abstract Submissions Due

**May 31:**
Scholarship Applications and Award Nominations Due

**July 1:**
Presenters Confirm Participation (Oral and Poster Presentations)

**August 12:**
Early Registration Deadline

**September 16:**
Advanced Registration Deadline

**September 16:**
ISOQOL Hotel Room Block Closes

ISOQOL 20th Annual Conference
October 9-12, 2013
**Miami, Florida, USA**

ISOQOL Conference in Berlin, Germany
October 15-18, 2014

Energizing the Science of Quality of Life Research:
Where have we been and where can we go?

isoqol.org/2013conference
Sample Size Issues
Sample Size Issues

- The IRT model to be estimated
  - Parameters ↑, Sample Size ↑ - Rasch models need less data.
- The number of items or questions.
  - Number of items ↑, Sample Size ↑
- The number of response options.
  - Number of response categories ↑, Sample Size ↑
- Unidimensionality of construct
  - Better the data meet assumption of unidimensionality, sample size ↓
- The item properties
  - Items at the extremes need more data
- Population distribution
  - Distributed across theta continuum, Sample Size ↓
- Purpose of Study
  - Evaluation of an instrument, smaller sample sizes needed
  - Estimate accurate respondent scores, larger sample sizes needed.
  - Calibrating items for an item bank, larger sample sizes
**Rasch / 1-Parameter Logistic IRT Model**

\[
P(X_i = 1|\theta) = \frac{1}{1 + e^{-(\theta - b_i)}}
\]

- For severe depression, \(b = 1.33\)
- For moderate depression, \(b = 0.25\)
- For none, \(b = -0.23\)

**Graph**

- X-axis: Depression
- Y-axis: Probability of Response

**Items**
- I am unhappy some of the time
- I don’t care what happens to me
- I cry easily

**Equations**

\[
P(X_i = 1|\theta) = \frac{1}{1 + e^{-a(\theta - b_i)}}
\]
2-Parameter Logistic IRT Model

\[ P(X_i = 1 | \theta) = \frac{1}{1 + e^{-1.7a_i(\theta - b_i)}} \]

\[ a_i(\theta - b_i) \]

\[ a = 2.20 \]
\[ a = 2.83 \]
\[ a = 1.11 \]

\[ b = -0.23 \]
\[ b = 0.25 \]
\[ b = 1.33 \]