Multi-item Scales and Tests: Development and Validation Methods

Elizabeth A. Hahn
Associate Professor
Department of Medical Social Sciences
Feinberg School of Medicine, Northwestern University
e-hahn@northwestern.edu

Biostatistics in Medical Research
Biostatistics Collaboration Center (BCC) & Outcomes Measurement and Survey Core (OMSC)
November 8, 2011
Learning Objectives

1. Describe General Measurement Concepts and Methods
2. Learn about Classical and Modern Test Theory
3. Define Reliability and Validity
Creating Multi-item Scales
“Objective”

Exercise test versus physical functioning, $r = 0.40$

“Subjective”
<table>
<thead>
<tr>
<th>Item</th>
<th>Question</th>
<th>Not at all</th>
<th>Very little</th>
<th>Somewhat</th>
<th>Quite a lot</th>
<th>Cannot do</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFA01</td>
<td>Does your health now limit you in doing vigorous activities, such as running, lifting heavy objects, participating in strenuous sports?</td>
<td>☐ 5</td>
<td>☐ 4</td>
<td>☐ 3</td>
<td>☐ 2</td>
<td>☐ 1</td>
</tr>
<tr>
<td>PFC36</td>
<td>Does your health now limit you in walking more than a mile?</td>
<td>☐ 5</td>
<td>☐ 4</td>
<td>☐ 3</td>
<td>☐ 2</td>
<td>☐ 1</td>
</tr>
<tr>
<td>PFC37</td>
<td>Does your health now limit you in climbing one flight of stairs?</td>
<td>☐ 5</td>
<td>☐ 4</td>
<td>☐ 3</td>
<td>☐ 2</td>
<td>☐ 1</td>
</tr>
<tr>
<td>PFA05</td>
<td>Does your health now limit you in lifting or carrying groceries?</td>
<td>☐ 5</td>
<td>☐ 4</td>
<td>☐ 3</td>
<td>☐ 2</td>
<td>☐ 1</td>
</tr>
<tr>
<td>PFA03</td>
<td>Does your health now limit you in bending, kneeling, or stooping?</td>
<td>☐ 5</td>
<td>☐ 4</td>
<td>☐ 3</td>
<td>☐ 2</td>
<td>☐ 1</td>
</tr>
</tbody>
</table>
Advantages of Multi-item Scales

- *Latent variables* are usually complex and not easily measured with a single item.
- Usually more reliable and less prone to random measurement errors than single-item measures.
- A single item often cannot discriminate between fine degrees of an attribute.
Creating Multi-item Scales

Latent construct vs. Index
Latent Construct

- Estimation of a *unidimensional latent trait*
  - abstract concept
  - cannot be measured directly
  - examples: attitudes, satisfaction, patient-reported outcomes (PRO)

- However, it is possible to measure *indicators of the latent trait*
  - use observed responses to questionnaire items
Index

- Summary of individual components
  - symptoms
  - comorbid conditions
Comorbidity

Heart Attack
Diabetes
Asthma
Stroke
Hypertension
8 attributes recommended by the Medical Outcomes Trust for health status and quality of life instruments (Scientific Adv Comm, Qual Life Res 2002)

1. a conceptual and measurement model
2. reliability
3. validity
4. responsiveness
5. interpretability
6. low respondent and administrative burden
7. alternative forms
8. cultural and language adaptations
The Life Story of a PROMIS Item
Patient-Reported Outcomes Measurement Information System
www.nihpromis.org
Classical and Modern Test Theory

Classical Test Theory assumptions:
- “parallel tests”: each item is a “test” that reflects the underlying level of the trait
- item responses differ only due to random error
- a scale score is computed by simple summation

Modern Test Theory assumptions:
- each item reflects a different level of the trait
- respondents with a particular trait level have a probability of responding positively to different items
Example: Measuring “Liking for Science” in School Children

Less liking for science

More liking for science
A “Liking for Science” Variable
Writing Questions

3 Elements of a question

1. Context
2. Stem
3. Response

How much do you like each activity?

Going to the zoo.
## Results of ordering by 9 judges

<table>
<thead>
<tr>
<th>Activity</th>
<th>easy-to-like</th>
<th>hard-to-like</th>
<th>median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>1 2 3 4 5 6 7 8 9 10 11</td>
<td>2 2 5 11</td>
<td></td>
</tr>
<tr>
<td>Hard</td>
<td>1 1 2 1 4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>learn names of weeds</td>
<td>1 1 2 1 4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>watch the grass change over seasons</td>
<td>2 2 2 2 1</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>watch bird make nest</td>
<td>4 1 2 1 1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>going to the zoo</td>
<td>2 1 1 1 1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>making a map</td>
<td>1 1 1 1 1</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Administered 25 science activity items to children ($n=75$)

<table>
<thead>
<tr>
<th>Activity</th>
<th>Judges</th>
<th>Children</th>
</tr>
</thead>
<tbody>
<tr>
<td>learn names of weeds</td>
<td>hard</td>
<td>hard</td>
</tr>
<tr>
<td>watch grass change</td>
<td>somewhat hard</td>
<td>somewhat hard</td>
</tr>
<tr>
<td>watch bird make nest</td>
<td>somewhat easy</td>
<td>somewhat easy</td>
</tr>
<tr>
<td>going to the zoo</td>
<td>easy</td>
<td>easy</td>
</tr>
<tr>
<td>making a map</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
## Interpretation

<table>
<thead>
<tr>
<th>Items</th>
<th>Child 1</th>
<th>Child 2</th>
<th>Child 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. learn names of weeds</td>
<td>😊</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>b. watch grass change</td>
<td>😊</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>c. watch bird make nest</td>
<td>😊</td>
<td>😊</td>
<td>😞</td>
</tr>
<tr>
<td>d. going to the zoo</td>
<td>😊</td>
<td>😊</td>
<td>😊</td>
</tr>
</tbody>
</table>

Less Liking for Science ("easy")

More Liking for Science ("hard")
Types of Respondent Data and Methods/Modes of Survey Administration

- Self-report vs. proxy/observer

- Self-administration
  - paper-and-pencil
  - telephone
  - computer

- Interviewer-administration
  - paper-and-pencil
  - telephone
  - computer
Considerations for Question Wording in Surveys of Culturally Diverse Populations
Instrument dimensions of equivalence across language and cultural groups

- **Content**: Content is *relevant*
- **Semantic**: *Meaning* is the same
- **Technical**: *Method* of assessment is comparable
- **Criterion**: *Interpretation* remains the same
- **Conceptual**: Instrument measures the same *theoretical construct*

Flaherty et al, 1988
Scoring Multi-item Scales
Scoring Multi-item Scales

- Determine how a high scale score will be interpreted
- Reverse scores and/or recalibrate scores (if warranted)
- Address missing item responses
- Compute scale scores (may involve transforming)
- Perform scoring checks
Positive and negative scoring examples

How much of the time during the past 4 weeks...

1. Have you been a very nervous person?
   1. All of the time
   2. Most of the time
   3. A good bit of the time
   4. Some of the time
   5. A little of the time
   6. None of the time

2. Have you felt calm and peaceful?
   1. All of the time
   2. Most of the time
   3. A good bit of the time
   4. Some of the time
   5. A little of the time
   6. None of the time
Scoring: Missing Data

1. Treat the scale score as missing
   - ignores other scale items with valid data
   - missing items may be related to outcome

2. Simple mean imputation
   - most common strategy; > 50% scale items completed
   - assumes missing item’s value = average of non-missing items

3. General imputation methods
   - may reduce non-response bias if done appropriately
   - can be mathematically and computationally difficult

4. Use Item Response Theory measurement models
The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

(Circle One Number on Each Line)

<table>
<thead>
<tr>
<th>Activity</th>
<th>Yes, Limited a Lot</th>
<th>Yes, Limited a Little</th>
<th>No, Limited at All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Lifting or carrying groceries</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Climbing several flights of stairs</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Climbing one flight of stairs</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Bending, kneeling, or stooping</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Walking more than a mile</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Walking several blocks</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Walking one block</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Bathing or dressing yourself</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Scoring

Sum
- Prorate for missing items
  \[(\text{sum of items}) \times (\# \text{ of items in scale}) / (\# \text{ of items answered})\]
  \[(13 \times 10) / 9 = 14.44\]

Sum and Average
- Result is on the same scale as the original items
- Example \((13 / 9 = 1.4)\):
  1. Yes, Limited a Lot
  2. Yes, Limited a Little
  3. No, Not Limited at All
  \[\{\text{Average} = 1.4\}\]

Transform
- Most common transformation is to a 0-100 scale

\[
P_{F\text{tot}} = \left[ \frac{(PF_{\text{pro}}) - \min}{\text{range}} \right] \times 100 = \left[ \frac{(14.44) - 10}{20} \right] \times 100 = 22.2
\]
Reliability and Validity
Distinction between Reliability and Validity

- a measure may be reliable (always yields the same score for the same respondent), but it may be consistently measuring the wrong thing (not measuring what it is supposed to measure)

- reliability is necessary, but not sufficient for valid measurement
Reliability

- the extent to which a measure yields the same number or score each time it is administered, all other things being equal (i.e., true change has not occurred)
Reliability

- How you measure reliability depends on the type of measurement scale

  - Nominal: categories
  - Ordinal: ordered categories
  - Interval: differences have meaning
  - Ratio: interval with true zero
Reliability

- A reliable measure is free from random error.

- Two different reliability characteristics of a measure:
  
  - Repeatability/reproducibility
  
  - Internal consistency
Reliability: Repeatability/Reproducibility

- over time (test-retest reliability)
- over observers (inter-rater or intra-rater reliability)
- over different variants of an instrument (equivalent forms reliability)

example: measurement of blood pressure
reliability of measures over a 24-hour period or by different health care providers or using different cuffs
Reliability for Nominal and Ordinal Scales

- relevant statistic for estimating repeatability/reproducibility reliability is Kappa or Weighted Kappa

- Kappa ($\kappa$) quantifies the amount of agreement between measurements that is greater than the amount expected by chance alone

  - if $\kappa=0$, chance agreement
  - if $\kappa<0$, less than chance agreement (rare)
  - if $\kappa=1$, perfect agreement
Reliability for Interval and Ratio Scales

- relevant statistic for estimating repeatability/reproducibility reliability is an Intraclass Correlation Coefficient ($r_{ICC}$)
- numerous versions of ICCs
- if $r_{ICC}$ near 0, almost all variation is due to measurement error and the measure is unreliable
- if $r_{ICC}$ near 1, minimal measurement error and the measure is very reliable
Reliability: Internal Consistency

- the extent to which a set of questions measures a single underlying dimension
  - e.g., fatigue, depression, physical function
Reliability: Internal Consistency

- as the number of items is increased, the reliability will increase
- diminishing returns with increasing items
- reliability can be increased by deleting an item with poor item-total correlations
Reliability: Internal Consistency

For multi-item scales comprised of items with interval response choices, reliability is most commonly assessed using Cronbach’s coefficient alpha ($r_a$).

- values $\geq 0.90$ are considered the standard for individual-level applications
- values $\geq 0.70$ are considered the standard for group-level applications
## Guidelines for instrument reliability/precision

<table>
<thead>
<tr>
<th>Data type</th>
<th>Relevant statistic</th>
<th>High/excellent reliability (minimal/no error)</th>
<th>Moderate/good reliability (acceptable error)</th>
<th>Low reliability (high error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal (dichotomous)</td>
<td>Kuder-Richardson 20 (KR-20)</td>
<td>≥ 0.90</td>
<td>0.70 – 0.89</td>
<td>&lt; 0.70</td>
</tr>
<tr>
<td>Nominal</td>
<td>Kappa ($\kappa$)</td>
<td>&gt; 0.74</td>
<td>0.40 – 0.74</td>
<td>&lt; 0.40</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Weighted Kappa</td>
<td>&gt; 0.74</td>
<td>0.40 – 0.74</td>
<td>&lt; 0.40</td>
</tr>
<tr>
<td>Interval/Ratio</td>
<td>Intraclass Correlation Coefficient ($r_{ICC}$)</td>
<td>≥ 0.90</td>
<td>0.70 – 0.89</td>
<td>&lt; 0.70</td>
</tr>
<tr>
<td></td>
<td>Internal Consistency Reliability ($r_{\alpha}$)</td>
<td>≥ 0.90</td>
<td>0.70 – 0.89</td>
<td>&lt; 0.70</td>
</tr>
</tbody>
</table>
Validity

degree to which the measure reflects what it is supposed to measure (rather than something else)
Types of Validity

- content validity
- construct validity (including criterion validity)
- responsiveness
Content Validity

- the extent to which a measure samples a representative range of the content
- need a clear idea of what is to be measured
- fairly subjective (compare to existing standards, well-accepted theoretical definitions, expert opinions, interviews with the target population)
Construct Validity

- hypothesize how the measure should “behave”
  - the direction of relationships
  - the strength of relationships
- an iterative process

Diagram:
- Testing ➔ Empirical results ➔ Revisions ➔ Testing
Construct Validity

- **convergent validity**
  - extent to which different ways of measuring the same trait are interrelated

- **discriminant (divergent) validity**
  - measures of different traits should be relatively unrelated

- **criterion validity**
  - use of a “gold standard” measure
FACT-B Convergent Validity (ECOG Performance Status Rating)

Mean TOI Rasch Measure

- No symptoms (n=100): 57.1
- Some symptoms (n=67): 54.5
- Some bedrest (n=16): 50.3

$p<0.001$
FACT-C
Convergent and Divergent Validity
(Pearson correlations)

FACT-C and FLIC: $r=0.74$

FACT-C and Social Desirability Scale:
$r=0.02$

Responsiveness Validity

- measure should be able to detect small, but meaningful, changes over time
FACT-B Trial Outcome Index (TOI)
Sensitivity to Change in Patient-rated PSR

Mean FACT-B TOI Change

PSR worse  
n=8 (d=.65)  

PSR same   
n=29 (d=.10)  

PSR better 
n=10 (d=.55)
Construct and Responsiveness Validity

Conceptual equivalence: association between hemoglobin response and improvement in fatigue

Mean Change in FACT Fatigue Subscale Scores

- **Hemoglobin Non-Responder**
- **Hemoglobin Responder**

<table>
<thead>
<tr>
<th>Condition</th>
<th>N</th>
<th>Mean Change</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solid Tumor with Chemotherapy</td>
<td>436</td>
<td>1.1</td>
<td>0.43</td>
</tr>
<tr>
<td>Lymphoprolif. Malignancy with chemotherapy</td>
<td>214</td>
<td>0.5</td>
<td>0.39</td>
</tr>
<tr>
<td>Non-Myeloid Malignancy without Chemotherapy</td>
<td>76</td>
<td>1.9</td>
<td>0.85</td>
</tr>
<tr>
<td>All Study Groups Combined</td>
<td>726</td>
<td>1.0</td>
<td>0.50</td>
</tr>
</tbody>
</table>

\[ P = 0.0001 \quad P = 0.0030 \quad P < 0.0001 \quad P < 0.0001 \]

\[ ES^b = 0.43 \quad ES = 0.39 \quad ES = 0.85 \quad ES = 0.50 \]

---

\(^a\)Hemoglobin response is defined as an improvement of at least 2 g/dL in hemoglobin.

\(^b\)Effect size (ES) = mean change in responder group / standard deviation of responder group at baseline.

Cella et al. Annals of Oncology 2004
Reliability and Validity are not static characteristics

- demonstrating reliability is essentially accumulating evidence about the stability of the measure

- demonstrating validity involves accumulating evidence of many different types which indicate the degree to which the measure denotes what it was intended to represent
Item Response Theory (IRT) 
Item Banks

- comprised of a large collection of items measuring a single concept
- enables test instruments of various lengths and even computerized adaptive tests (CATs)
TOFHLA Numeracy: Item Response Theory Analysis (1-p model) 
\((n=1,891\) English-speaking patients) 

High Literacy 

Low Literacy 

# represents 18 people 

Need items for higher literacy people
Item Response Theory Analysis Results (1-p model)  
\(n=616\) English-speaking primary care patients

**High Literacy**

- EIZ3Q5
- ECA4Q1

**Hard Items**

- EHT3Q1
- EIZ3Q1
- EMD2Q3
- ESI2Q1
- EIC1Q3
- ECA1Q2
- EOB1P2
- EOB3D2
- EIC1P1
- EOB3D1
- EOB2Q2
- EIC1Q8
- EOB4P4
- EHT3D3
- EHT1P2
- EHT3D4
- EOB5P5
- EIC1P3
- EOB1P4
- EOB5P2

**Low Literacy**

- EAS5D2
- EHT1P3
- EIN1Q10
- EIN1Q5
- EIZ3Q2
- EMD2Q3
- EDB5P2
- EIC3P8
- EOB3D2
- EHT2P1
- EIN1Q9

**Easy Items**

- EAS5D1
- EIC1P9
- EIN2D3
- EIC3P8
- EOB3D2
- EHT2P1
- EIN1Q9
- EIC1P6
- EAS1P6

Mean Patient Score →

Mean Item Difficulty ←

# represents 3 patients
The Advantage of IRT-based PRO Measures Over Traditional PRO Measures

<table>
<thead>
<tr>
<th>Traditional PRO Questionnaires</th>
<th>IRT-Based Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed # of items presented serially or in branched designs (skip patterns)</td>
<td>Variable # of items, tailored for each person using Computerized Adaptive Testing (CAT) technology</td>
</tr>
<tr>
<td>Floor and/or ceiling effects</td>
<td>Reduced floor and ceiling effects</td>
</tr>
</tbody>
</table>
### The Advantage of IRT-based PRO Measures Over Traditional PRO Measures

<table>
<thead>
<tr>
<th>Traditional PRO Questionnaires</th>
<th>IRT-Based Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability higher as number of items increases</td>
<td>Reliability can be equal to or higher than in other instruments</td>
</tr>
<tr>
<td>Larger number of items increases respondent burden</td>
<td>Fewer, targeted items (with CAT) reduces respondent burden</td>
</tr>
</tbody>
</table>
The Advantage of IRT-based PRO Measures Over Traditional PRO Measures

<table>
<thead>
<tr>
<th>Traditional PRO Questionnaires</th>
<th>IRT-Based Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very few instruments can cross-walk scores to other instruments for combining or comparing scores</td>
<td>Can create multiple instruments from psychometrically-linked item banks</td>
</tr>
<tr>
<td></td>
<td>Can maintain cross-walks with several leading PRO scales</td>
</tr>
</tbody>
</table>
Reference Material


www.nihpromis.org

www.rasch.org
Thank You!

What questions do you have?