Rasch Measurement in Medical Rehabilitation

Allen Heinemann, Ph.D.
Northwestern University Medical School
Rehabilitation Institute of Chicago

Webinar Objectives

• Provide an introduction to contemporary measurement approaches applicable to medical rehabilitation
• Describe rating scale (Rasch) analysis
• Illustrate applications of rating scale analysis in medical rehabilitation
## Rehabilitation Measurement Problems

- Does a total FIM™ score of 70 obtained from two different patients (e.g., SCI, CVA) imply the same level of disability?
- Is Mary’s FIM™ improvement from 13 to 23 equivalent to Sam’s improvement from 116 to 126?
- How do we know these items ‘add up’ to mean anything?
- Can we apply arithmetic to numbers like these?

### More Rehabilitation Measurement Problems

- Is the improvement between a FIM™ rating scale point of 1 & 2 equivalent to the distance between 5 & 6?
- Does the Barthel Index item weighting affect conclusions about disability severity?
- Rating scale definitions are arbitrary
- Item weights reflect instrument developers’ biases
Misunderstandings about Measurement

- Limit our ability to construct measures of patients’ impairment, disability, handicap
- Create opportunities for “misinference”
- Make discussions of improvement difficult
- Limit our ability to compare patients’ with different underlying conditions

Defining Measurement

- Determination of the magnitude of a specified attribute of an object, organism, event
- Acquired purposefully
- A fixed unit
- Expressed by a numeral
Example: Height

- Attribute: Height
- Measuring instrument: Ruler
- Magnitude of attribute: Distance from toe to head
- Fixed unit: Centimeter
- Expressed as a numeral: Height equals some number of centimeters

The Ruler as a Measurement System

- A ruler is independent of its construction
- A ruler implements the idea of a unit
- The unit is the same everywhere along a ruler
- The centimeter is the concrete realization of the unit
- Concrete units operationalize additivity
- Rulers employ a natural number system
- Rulers implement a correspondence between units and numbers
Measurement within a Larger Perspective of Scientific Inquiry

- Experience
- Observation
- Measurement
- Analysis
- Theory

Where Measurement Belongs in Scientific Inquiry

- Experience
- Observation
- Measurement
- Analysis
- Theory
How Constructs Achieve Definition

• Personal knowledge, intuition, subjectivity
  – too hot to touch — as cold as snow
  – safe radiation exposure — unsafe exposure
• Data-based instrument calibrations
  – test-specific functional status
• Theory-based instrument calibrations
  – manufactured thermometers
  – Lexile framework for reading comprehension

How Do Numbers Occur?

• As labels
  – Zip codes, house addresses, product codes
• As counts, scores, ranks
  – FIM™ scores, Olympic placements
• As measures: Reliable linear units
  – Hourly pay rate, Celsius degrees
Levels of Measurement

- Nominal (categorical)
- Ordinal (counts, ordered categories—20s, 30s)
- Equal-Interval (linear)
- Ratio (local or absolute zero defined)

Stevens, 1946, 1951

Kinston’s Levels of Representation

I. Entity (idea)
   making a distinction (identifying, distinguishing)
II. Observable (thingness)
   making an intersection (pointing to, counting)
III. Comparable (quantity)
   making a comparison (ordering, ranking, valuing)
IV. Measurable (standard unit)
   making a measurement (measuring, using arithmetic)
V. Relatable
   making a relation (connecting, formulating)

W Kinston: Structure of Scientific Analysis, Systems Research, 2, 1985
Kinston’s Levels of Representation: A Rehabilitation Example

I. Entity (*an idea or concept*)
   - Disability

II. Observable (*thingness*)
   - Wheelchair mobility

III. Comparable (*a quantity, possessed by someone*)
   - More or less assistance, skill level

IV. Measurable (*standard unit*)
   - Rehab

V. Relatable
   - Disability is reduced by environmental modification

Subjectivity and Objectivity

<table>
<thead>
<tr>
<th>Level</th>
<th>Site of Experience</th>
<th>Relevant Notion</th>
<th>Error Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Entity</td>
<td>Subjective</td>
<td>Plausibility</td>
<td>Losing contact with external reality</td>
</tr>
<tr>
<td>II. Observable</td>
<td>Objective</td>
<td>Probability</td>
<td>Making wrong distinctions, using unsuitable categories</td>
</tr>
<tr>
<td>III. Comparable</td>
<td>Subjective</td>
<td>Plausibility</td>
<td>Emphasizing irrelevant attributes</td>
</tr>
<tr>
<td>IV. Measurable</td>
<td>Objective</td>
<td>Probability</td>
<td>Examining unpromising relations, forgetting to control for crucial factors</td>
</tr>
<tr>
<td>V. Relatable</td>
<td>Subjective</td>
<td>Plausibility</td>
<td>Producing findings of limited applicability</td>
</tr>
</tbody>
</table>
We Use Numbers to Compare and Evaluate

- 10 rehab days @ $1,000/day > 7 rehab days @ $1,200
- Home discharge after 25 rehab days is a better outcome than nursing home discharge after 15 days
- A 50 FIM™ point gain over 10 days is an average 5 point gain/day

This Kind of Number Use Requires Arithmetic

- Add, subtract, multiply, average
- Arithmetic done on labels, raw scores, counts and ranks is nonsense
- Conclusions can be misleading
- Equal-interval, constant unit, linear measures are required
Our Assumptions

- We judge correctly,
- According to standard criteria,
- With ratings recorded accurately, and that
- 1,2,3,4,5,6,7 are real measures for task performance that add up to scores as good as measures

Are Unsound!

The Reality

- Ratings are merely guesses
- Standards vary by time and person
- Rating can be recorded incorrectly
- 1,2,3,4,5,6,7 are merely ranks that may not add up
- Raw scores are never measures
Transcending Scores: Constructing Measures

• All we can observe are counts and scores
• They may provide the basis for constructing measures
• We need a mature understanding of score meaning
• What matters is not what a score is but what it implies
• We need a method for using counts and scores to estimate the measures we want them to imply

This Method is called

Rasch Measurement
or
Rating Scale Analysis
Essential Conditions Causing a Response

How Differences between Person Ability and Item Difficulty Ought to Affect the Probability of a Correct Response
The Logistic Response Curve

- Log-odds units
- log [probability of success/probability of failure] = Person ability - Item Difficulty
- Can be estimated independently of the items included in a particular test
- More complex models can be specified
A More Complex Example: Rating Functional Status

<table>
<thead>
<tr>
<th>Observable Categories</th>
<th>Can’t do</th>
<th>Needs help</th>
<th>Does OK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numeric Labels</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Interpreted to imply category probabilities</td>
<td>$P_0$</td>
<td>$P_1$</td>
<td>$P_2$</td>
</tr>
<tr>
<td>With transition odds</td>
<td>$P_1/P_0$</td>
<td>$P_2/P_1$</td>
<td></td>
</tr>
</tbody>
</table>

This interpretation gives the observed raw ratings inferential meaning.

Conjoint Additivity

When these transition odds are modeled as the consequence of conjoint additivity parameters like

$$\log[P_{ni/x}/P_{ni/x-1}] = B_n - D_i$$

then statistical analyses of raw categorical ratings can *construct* and *control* inferentially stable linear measures, where
The Model Parameters are defined as

- the Functional Measure $B_n$ of patient $n$, when rated
- on Task $i$ of Difficulty $D_i$

Additional Model Parameters are defined as

- the Functional Measure $B_n$ of patient $n$, when rated
- on Task $i$ of Difficulty $D_i$

- by Clinician $j$ of Rigor $C_j$
- at Step $x$ with Gradient $F_x$
Rating Scale Information

- Difficulty of each Task
- Measure for each Person
- Rigor of each Rater
- Gradient of each rating Step, and
- Fit of each task to an underlying construct
- Fit of each person to this construct
- Separation of sample into distinct strata

How Good are the Measures Produced by the Data?

- Do data fit the measurement model?
- Quality control method is needed
- Fit statistics ensure that all
  - items measure the same construct
  - people possess the same quality
  - judges understand the same rating scheme
Fit Statistics

• Infit
  – Dominated by unexpected inlying patterns among informative, on-target observations
  – Inlier sensitive ($B_v \sim D_i$)

• Outfit
  – Dominated by unexpected outlying, off-target, low information responses
  – Outlier sensitive ($B_v < D_i$ or $B_v > D_i$)

Fit Problems

• Muted Patterns (small fit statistics)
  – Unmodeled dependence
  – Redundancy
  – Error trends

• Noisy Patterns (large fit statistics)
  – Unexpected, unrelated irregularities
Reliability and Separation

- Observed variance = “True” variance + error variance
- Reliability = “True” variance/Error variance
- Separation = $\sqrt{\text{reliability} / (1-\text{reliability})}$
  = “True” SD / Error SD
- Separation is the number of statistically different performance strata the test can identify in the sample

Person Separation

- The ability of the items to resolve the sample into distinct strata
- Describes the spread of the persons in this sample
- “Real” separation adjusts for misfitting persons
Relationship between Reliability and Separation

<table>
<thead>
<tr>
<th>Separation Ratio G</th>
<th>KR-20 Alpha G^2/(1+G^2)</th>
<th>% Variance Not Due to Error/Due to Error</th>
<th>Distinct Strata (4G+1)/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.00</td>
<td>0/100</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>.50</td>
<td>50/50</td>
<td>1</td>
</tr>
<tr>
<td>1.5</td>
<td>.70</td>
<td>70/30</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>.80</td>
<td>80/20</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>.90</td>
<td>90/10</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>.94</td>
<td>94/6</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>.96</td>
<td>96/4</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>.97</td>
<td>97/3</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>.98</td>
<td>98/2</td>
<td>9</td>
</tr>
</tbody>
</table>

Common Problems with Rating Structures

• More categories provided than respondents differentiate
• Some categories vaguely defined, e.g.,
  – FIM™ assistance required
  – “often” vs. “frequently” vs. “sometimes”
• Item difficulties become inverted
• Item fit worsens
Sample Size and Item Calibration Stability: Dichotomous Items

<table>
<thead>
<tr>
<th>Item Calibrations stable within</th>
<th>Confidence</th>
<th>Minimum sample size range (best to poor targeting)</th>
<th>Size for most purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>± 1 logit</td>
<td>95%</td>
<td>16-36</td>
<td>30</td>
</tr>
<tr>
<td>± 1 logit</td>
<td>99%</td>
<td>27-61</td>
<td>50</td>
</tr>
<tr>
<td>± ½ logit</td>
<td>95%</td>
<td>64-144</td>
<td>100</td>
</tr>
<tr>
<td>± ½ logit</td>
<td>99%</td>
<td>108-243</td>
<td>150</td>
</tr>
</tbody>
</table>

Rating scales need smaller samples than do dichotomous items.

How is Factor Analysis Different from Rasch Analysis?

- Same kind of data, but
  - different interpretations of numerical status
- Same estimation methods, but
  - different measurement models
- Solve the same problems, but
  - substantially different utility
Problems with Factor Analysis

- Factor scores are not linear measures
- Factor structure is sample dependent
- Factor structure reflects rating scale used
- Ordinal variables and highly correlated factors create confusion
- Complete data are required
- Identity of smaller factors is overwhelmed by residuals from preceding larger factors
- Guidelines for stopping are vague
- Standard errors for factor scores hard to obtain

In summary:
Rating scale analysis provides

- Person and Item Separation
- Person and Item Fit Statistics
- Person Ability and Item Difficulty Estimates
- Rating Scale Adequacy Indices
Suggested Procedure for Improving Rating Scales

- Maximize person separation (2.0 minimum)
- Assure that item step structures are clear
- Examine item fit statistics
- Identify misfitting persons
- Examine item targeting on sample
- Review person and item map to establish construct

Strategies for Improving Rating Scales

- Consider single vs. multiple rating scale model
- Verify rating scale behavior
- Survey the range of item difficulties, fit statistics
- Consider range of person ability estimates, fit statistics, subgroup possibilities
Answers to Rehabilitation Measurement Problems

- Does a total FIM™ score of 70 obtained from two different patients (e.g., SCI, CVA) imply the same level of disability?
- Is Mary’s FIM™ improvement from 13 to 23 equivalent to Sam’s improvement from 116 to 126?
- Approximately the same; disability measure can be fine-tuned for specific impairments
- No - raw scores aren’t measures

Answers to More Rehabilitation Measurement Problems

- Is the improvement between a FIM™ rating scale point of 1 & 2 equivalent to the distance between 5 & 6?
- Does the Barthel Index item weighting affect conclusions about disability severity?
- No. The improvement can be calibrated
- Yes, it confounds measurement
Examples of Rasch-Developed Measures

Rasch articles 1986-August 2013
(Rasch + Rehabilitation, Web of Knowledge)
Where were these 525 articles published?

- Archives of Physical Medicine and Rehabilitation
- Journal of Rehabilitation Medicine
- Disability and Rehabilitation
- Investigative Ophthalmology Visual Science

November 2013 APMR articles with “Rasch” keyword

- Hays: Upper-Extremity and Mobility Subdomains From the Patient-Reported Outcomes Measurement Information System (PROMIS) Adult Physical Functioning Item Bank
- Heinemann: Measuring Enfranchisement: Importance of and Control Over Participation by People With Disabilities
- Angst: Effects of Inpatient Rehabilitation in Hip and Knee Osteoarthritis: A Naturalistic Prospective Cohort Study With Intraindividual Control of Effects
- Major: Validity and Reliability of the Berg Balance Scale for Community-Dwelling Persons With Lower-Limb Amputation
## MetaMetrics – Lexile Framework

**TEXT LEXILE RANGES TO GUIDE READING FOR COLLEGE AND CAREER READINESS**

<table>
<thead>
<tr>
<th>GRADES</th>
<th>CCSS LEXILE TEXT RANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 - 12</td>
<td>1185L - 1385L</td>
</tr>
<tr>
<td>9 - 10</td>
<td>1050L - 1335L</td>
</tr>
<tr>
<td>6 - 8</td>
<td>925L - 1185L</td>
</tr>
<tr>
<td>4 - 5</td>
<td>740L - 1010L</td>
</tr>
<tr>
<td>2 - 3</td>
<td>420L - 820L</td>
</tr>
<tr>
<td>1</td>
<td>190L - 530L</td>
</tr>
</tbody>
</table>

---

## Matching Readers with Text

### SAMPLE TITLES

**LITERATURE TITLES**

140L: The Plot Against America (Roth)
146L: Sons of the Storm (Scott)
152L: The Good Earth (Buck)
152L: A Fable (Allende)
190L: The Decameron (Boccaccio)

### INFORMATIONAL TITLES

160L: Sustaining Life: How Human Health Depends on Biodiversity (Grau & Bernstein)
1550L: The Art of War (Sun Tzu)
1520L: Fair Play: The Ethics of Sport (Simon)
1500L: Critiques of Pure Reason (Kant)

### MATCHING READERS WITH TEXT

- **146L**: The Legend of Sleepy Hollow ( Irving)
- **146L**: Billy Budd*** (Melville)
- **142L**: Life of King Arthur and His Knights (Pyle)
- **142L**: The Scarlet Letter*** ( Hawthorne)
- **135L**: Robinson Crusoe (Defoe)
- **138L**: The Secret Sharer (Joyce)
- **134L**: The Metamorphosis*** ( Kafka)
- **134L**: Fever Pitch (Crichton)
- **128L**: The House of the Spirits (Allende)
- **127L**: Tarzan of the Apes (Burroughs)
- **128L**: A Brief History of Time (Hawking)

---

(c) Allen Heinemann 2014
Example
Measuring Environmental Factors for Persons with Stroke, Traumatic Brain Injury and Spinal Cord Injury

Conceptual Framework
### Comparison of ICF and Developed Conceptual Frameworks

<table>
<thead>
<tr>
<th>ICF Taxonomy</th>
<th>Initial Framework</th>
<th>Binned Items</th>
<th>Winnowed Items</th>
<th>Revised Framework</th>
<th>Final Item Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Products and Technology</td>
<td>Assistive technology</td>
<td>178</td>
<td>104</td>
<td>Assistive technology</td>
<td>15</td>
</tr>
<tr>
<td>Natural Environment and human made changes</td>
<td>Natural environment</td>
<td>79</td>
<td>37</td>
<td>Built and natural environment</td>
<td>36</td>
</tr>
<tr>
<td>Support and Relationships Attitudes</td>
<td>Social supports and attitudes</td>
<td>710</td>
<td>91</td>
<td>Social environment</td>
<td>82</td>
</tr>
<tr>
<td>Services, Systems and Policies</td>
<td>Services, systems, policies</td>
<td>411</td>
<td>35</td>
<td>Services, systems, policies</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Transportation</td>
<td>136</td>
<td>28</td>
<td>Access to information and technology</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Access to information and technology</td>
<td>112</td>
<td>37</td>
<td>Access to information and technology</td>
<td>25</td>
</tr>
<tr>
<td>Economic-financial</td>
<td>42</td>
<td>35</td>
<td>Economic Quality of Life</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,273</strong></td>
<td><strong>392</strong></td>
<td><strong>265</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Economic quality of life: Rating scale analysis results

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Items Categories People</th>
<th>Person Item Reliability</th>
<th>Rating Scale</th>
<th>Misfitting Items (mn sq infit &gt;1.4)</th>
<th>Item-Measure r’s &lt;.4</th>
<th>PCA Residual Variance</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24i 5c 1812p</td>
<td>.91</td>
<td>1.00</td>
<td>Monotonic</td>
<td>None</td>
<td>5.6%</td>
<td>Combine response categories 1,2</td>
</tr>
<tr>
<td>2</td>
<td>24i 4c 1812p</td>
<td>.92</td>
<td>1.00</td>
<td>Monotonic</td>
<td>None</td>
<td>5.2%</td>
<td>Delete 3 misfitting items and 3 over fitting items (6, 7, 10)</td>
</tr>
<tr>
<td>3</td>
<td>18i 4c 1812p</td>
<td>.90</td>
<td>1.00</td>
<td>Monotonic</td>
<td>None</td>
<td>6.0%</td>
<td></td>
</tr>
</tbody>
</table>
Map of Economic QOL Items and Respondents

Economic QOL Rating scale – measure map
### Economic QOL item statistics in measure order

<table>
<thead>
<tr>
<th>Measure S.E.</th>
<th>Infit</th>
<th>I-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.37 .04</td>
<td>.96  -1.2</td>
<td>.77</td>
</tr>
<tr>
<td>1.37 .04</td>
<td>1.23  5.9</td>
<td>.73</td>
</tr>
<tr>
<td>1.29 .03</td>
<td>.98  -.4</td>
<td>.75</td>
</tr>
<tr>
<td>1.24 .03</td>
<td>1.16  4.4</td>
<td>.71</td>
</tr>
<tr>
<td>20 .03</td>
<td>1.82  9.2</td>
<td>.70</td>
</tr>
<tr>
<td>24 .03</td>
<td>1.10  9.2</td>
<td>.70</td>
</tr>
<tr>
<td>55 .03</td>
<td>1.35  9.2</td>
<td>.70</td>
</tr>
<tr>
<td>59 .03</td>
<td>1.82  9.2</td>
<td>.70</td>
</tr>
<tr>
<td>57 .03</td>
<td>1.10  9.2</td>
<td>.70</td>
</tr>
<tr>
<td>56 .03</td>
<td>1.35  9.2</td>
<td>.70</td>
</tr>
</tbody>
</table>

### Rasch Measurement Resources

![Rasch Measurement Resources](image.png)
Rasch Measurement Resources

Rasch Measurement in Health Sciences

Edited by
Nikolaus Benczcko

JAM Press is pleased to announce the new book, Rasch Measurement in Health Sciences, is now available. The book is available in soft cover ($44, ISBN 0-9755351-3-7) and hard cover ($54, ISBN 0-9755351-2-9). Postage and handling are additional. Information on ordering the book is located on the reverse of this announcement. Information is also available at the Journal of Applied Measurement web site (www.jampress.com). Please go to the JAM Press Books page on the web site and scroll down to the new books section.

21 August 2014

www.rasch.org

21 August 2014
They may all fly (the construct)…
but, we won’t know if birds of different feather can be calibrated together without verifying they “add up” to the same construct

Discussion
Acknowledgments

RRTC on Improving Measurement of Medical Rehabilitation Outcomes

Midwest Regional Spinal Cord Injury Care System

This presentation does not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.