Assessing Treatment Outcomes: Questioning Measurement Precision

Measurement Precision: Using the Rasch Model

Ann Doucette, Ph.D.

8 March 2005

VANDERBILT UNIVERSITY

18th Annual Research Conference – System of Care for Children’s Mental Health, Tampa, FL

Understanding Disappointing Findings: Issues To Consider

- Implementation
- Measurement

Presentation Overview

- Measurement models
  - Issues and advantages
  - Rasch Measurement Model
  - Longitudinal data
    - Measurement issues and changes in sample characteristics
    - Raw Scores versus Measure Scores
    - Estimating change
      - Regression to the mean

Measurement Precision: The Need For A New Approach

Measures developed using Classical Test Theory (CTT) assume:

- All items contribute equally to the the overall scale score
- Response options (e.g. Likert scales) are equal interval scales
- Error applies equally to all scores across the population

Model-based Measurement: Contrasting IRT and Classical Test Theory Approaches

<table>
<thead>
<tr>
<th>Item Response Theory (IRT)</th>
<th>Classical Test Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error of measurement differs across scores/response patterns, generalizes across populations</td>
<td>Standard error of measurement applies to all score in a specific population</td>
</tr>
<tr>
<td>shorter measures can be more reliable than longer measures</td>
<td>longer measures are more reliable than shorter measures</td>
</tr>
<tr>
<td>comparable scores across multiple measures are optimized — “difficulty” varies across persons — IRT control for item differences between test forms</td>
<td>test equating is needed to compare scores across multiple measures — equating error can be problematic</td>
</tr>
</tbody>
</table>

Model-based Measurement: Contrasting IRT and Classical Test Theory Approaches

<table>
<thead>
<tr>
<th>Item Response Theory (IRT)</th>
<th>Classical Test Theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>unbiased estimates of item characteristics can be obtained from non-representative samples</td>
<td>unbiased assessment of item characteristics is dependent on representative samples from target populations</td>
</tr>
<tr>
<td>meaningful scores are provided from IRT trait score estimates</td>
<td>meaningful scores are provided by standard scores (norm referenced)</td>
</tr>
<tr>
<td>interval scale properties are achieved by justifiable measurement models essentially the log odds that individual endorsies item is the difference between trait level and item difficulty</td>
<td>interval scale properties are achieved by identifying items to obtain normal raw score distributions — relative distances between interval levels are not the same across multiple measures</td>
</tr>
</tbody>
</table>
Summarizing The Advantages Of The Rasch Measurement Model

- Ability to perform item level analysis
- Error estimates and item fit indices
- Reliability (both person and scale reliability)
- Item independence
- Category (scale) analysis
  - Identification of response scale categories that offer little or no information
  - Identification of idiosyncratic use of scale categories
- Items are calibrated in terms of difficulty, and contribute differentially to the construct being measured
- Differential item function (DIF)
  - Group bias (age, gender, racial/ethnic, cultural, language groups)

The Rasch Measurement Model

- The Rasch model, as opposed to 2- and 3-parameter models, questions how well empirical data (measure scores/responses) fit in terms of the measurement model constraints.
- The additional parameters in 2PL (item difficulty) and 3PL (respondents guessing) models are used to explain variance in the measurement model.
- The Rasch model provides “sample free” (sample independent) item calibrations, item difficulties (δ), from easy to hard — no impairment to severe impairment.
- Rasch also yields fit statistics that provide information regarding a respondent’s expected response in comparison to his/her actual response.

Hypothetical Example

Scale items represent constructs along a continuum from low to high, minimal to maximal, etc. Every scale item is calibrated along this continuum.

CONSTRUCT

Low probability of endorsing items beyond low end of the continuum.

Low → Measured Construct → High

High probability of answering items at the high end of the continuum.

Measuring the Construct Calibrating Items

Person1 has no opportunity to demonstrate improvement. Scores can only indicate stability or deterioration.
- Person1 has no opportunity to demonstrate deterioration or minimal improvement.

Baseline

Person Mean

6-Month Follow-up

Person Mean

All items are above the person mean. No items are below the mean. Maximum item fit indices indicate depressive concerns.
Segmenting The Respondent Sample

Separation Index*

Baseline = 2.65
6-Months = 1.90
12-Months = 1.85

* Separation Index=the number of statistically distinct strata of “trait difficulty” (anxious-depressed) that can be represented in the sample using this measure.

Interpreting The Data

- 45% of youth at baseline assessment had scores below the clinical cutoff indicating mild to moderate impairment
- 76% of those youth with mild/low moderate scores maintained that status between baseline and follow-up (6 and 12 months) assessments
- Scores indicated that these youth made no progress during the 12 month period

Interpreting The Data Within The Context Of Measure-Sample Fit

Longitudinal Intervention Studies
- Individuals exposed to treatment are hypothesized to improve
- Measures sensitive to the initial, more severe levels of impairment may not be sensitive to later moderate and/or mild impairment levels
  - A lack of items at the mild end of the continuum provides no opportunity to demonstrate improvement for individuals with low baseline scores
- Measures used to screen and identify clinician and non-clinical groups may not be sensitive beyond cutoff scores
  - A lack of items at either end of the continuum provides restricted opportunity to demonstrate deterioration and/or improvement

Raw Scores Versus Measure Scores

- All items contribute equally to scale score
- Error generalizes across all scale items
- Raw scores are essentially counts
- Items differentially contribute to scale score
- Error differs across scale score
- Measure scores satisfy the requirements of interval scaling and additivity
Sample Distribution: Raw and Rasch Measure Scores

Estimating Change

Measurement Precision
Necessary But Insufficient To Estimate Meaningful Change

Reliable Change: Assumptions
- Pre and posttest scores are parallel measurements.
- Change that cannot be attributed to measurement error and related regression effects.
- Change is attributed as evidence of the effectiveness of treatment services.

Regression To The Mean
- Statistical phenomenon that occurs when
  - Repeated measures are taken on the same participant over time
  - Repeated measures are taken on groups of participants that have been categorized based on baseline measures
  - Natural variation appears as real change
- Extreme high or low scores are likely to be followed by lower or higher scores that are closer to the mean

Identifying Meaningful Change: Regression to the Mean
- Because of imperfect correlation, the predicted score on a variable (posttest) tends not to be as extreme as the predictor variable (pretest)
- The more extreme the score the greater the regression toward the mean – extreme scores fan in toward the mean
- Regression toward the mean = 1 - correlation between pretest/posttest
- Regression toward the mean should be considered in interpreting results across population samples, and appropriate adjustments should be made if needed
  - Adjustment – estimated RTM subtracted from observed change score
  - ANCOVA – adjusts individual follow-up scores according to baseline assessments

Regression To The Mean
- Individual score = true score + error
  - Scores above the mean tend to have positive errors of measurement
  - Scores below the mean tend to have negative errors of measurement
  - High scores in either direction have high error of measurement estimates
  - Errors of measurement are assumed to be uncorrelated
  - Obtained scores underestimate true scores for those below the mean and overestimate for those above the mean
Addressing Regression To The Mean

- Random assignment to comparison groups
  - All participants would be assumed to be equally affected by regression to the mean
  - Mean change for control/placebo group that takes into account regression to the mean, which then can be used to adjust the treatment effect
- Use of multiple baseline measures
  - Regression to the mean increases with larger measurement variability (error)
  - Multiple measures provide more precise estimates of the “true” mean and within participant variability
- Use measure scores (Rasch logit scores)

Identifying Change – Differences Versus Meaningful Difference

- Simple Difference Score: Follow-up score minus baseline score.
- Reliable Change Index: follow-up score minus baseline score divided by standard error of differences.
- Edwards-Nunnally Confidence Interval: two standard errors of measurement (pre/post) centered on baseline true score — follow-up score located relative to interval (accounts for regression to the mean).
- Nunnally-Kotsch: Pooling of variances from baseline and follow-up scores in calculations of standard error estimates.
- Growth Curve: HLM, makes use of all data available (baseline, concurrent, follow-up and post-treatment.
- Recovery: movement from clinical range to non-clinical range (CBCL), or from severe/marked to moderate/mild range (CAFAS)

Edwards-Nunnally Confidence Interval

- Reliable change: ± 2 standard errors of measurement = confidence interval
- Standard error of measurement = $SEM = SD \times \sqrt{(1-\gamma_p)}$
- Confidence interval is centered on pretest true score.
  - Estimated True Score Change: posttest score is regressed toward the mean using the reliability estimates of the pretest score
    \[\text{mean of pretest} + \text{reliability of pretest} \times (\text{posttest score} - \text{mean of pretest})\]
- Not subject to the effects of regression to the mean

Reliable Change

- Failed to Improve
- Reliable Improvement
- Stable, No Reliable Change
Contact Information

Ann Doucette, PhD
Vanderbilt Institute for Public Policy Studies
Vanderbilt University
1207 18th Avenue South
Nashville, TN 37212

Tele: 615.343.1655
eFax: 615.250.0518

Email: ann.doucette@vanderbilt.edu
adoucette@aol.com