Assessing Treatment Outcomes: Questioning Measurement Precision

Measurement Precision

Using the Rasch Model

Ann Doucette, Ph.D.

VANDERBILT UNIVERSITY

8 March 2005

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

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

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

Item Response Theory (IRT)	Classical Test Theory
unbiased estimates of item characteristics can be obtained from non-representative samples	unbiased assessment of item characteristics is dependent on representative samples from target populations
meaningful scores are provided from IRT trait score estimates	meaningful scores are provided by standard scores (norm referenced)
interval scale properties are achieved by justifiable measurement models essentially the log odds that individual endorses item is the difference between trait level and item difficulty	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

Summarizing The Advantages Of The Rasch Measurement Model

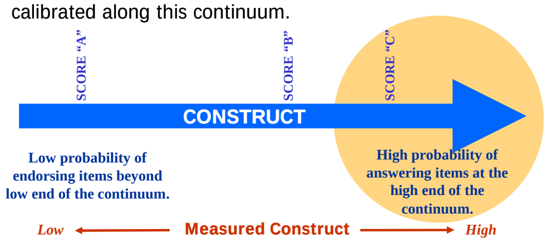
- Ability to perform item level analysis
  - Error estimates and item fit indices
  - Reliability (both person and scale reliability)
  - Item independence
  - Category (scale) analysis
    - Identification of response scale categories that offer little or no information
    - Identification of idiosyncratic 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.

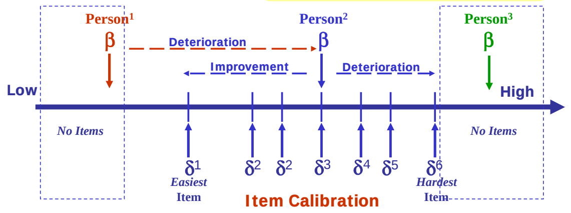
Hypothetical Example

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

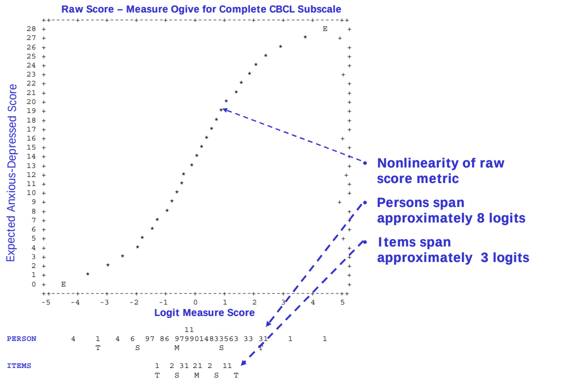

SCORE "A"
SCORE "B"
SCORE "C"
CONSTRUCT
Low probability of endorsing items beyond low end of the continuum.
High probability of answering items at the high end of the continuum.
Low
Measured Construct
High

Measuring the Construct Calibrating Items

- Person¹ has no opportunity to demonstrate improvement. Scores can only indicate stability or deterioration.
- Person³ has no opportunity to demonstrate deterioration or minimal improvement.



Raw Scores – Measure Scores

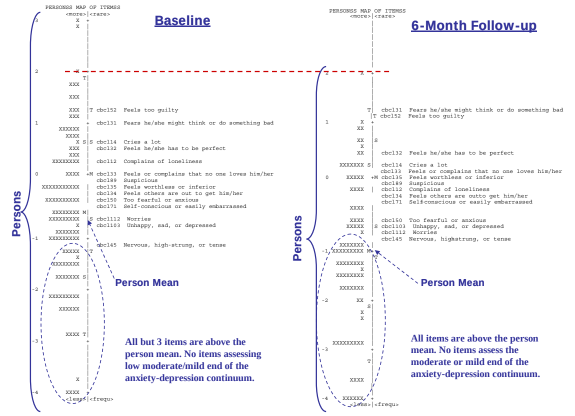


```
  ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
28 +                                                        E    +
27 +                                                     *     +
26 +                                                *            +
25 +                                             *               +
24 +                                           *                 +
23 +                                          *                +
22 +                                        *                    +
21 +                                       *                     +
20 +                                     *                       +
19 +                                    *                        +
18 +                                   *                         +
17 +                                  *                          +
16 +                                 *                         +
15 +                                *                            +
14 +                               *                             +
13 +                              *                              +
12 +                            *                              +
11 +                            *                                +
10 +                           *                                 +
 9 +                          *                              +
 8 +                         *                                   +
 7 +                       *                                     +
 6 +                      *                                      +
 5 +                    *                                      +
 4 +                   *                                         +
 3 +                *                                            +
 2 +             *                                               +
 1 +         *                                                   +
 0 +   E                                                         +
  ++-----+-----+-----+-----+-----+-----+-----+-----+-----+-----++
  -5    -4    -3    -2    -1     0     1     2     3     4     5

                                  11
PERSON       4    1    4   6   97 86 9799014833563 33 31    1      1
                  T        S         M        S       T

ITEMS                          1  2 31 21 2   11
                               T   S   M   S   T
```

- Nonlinearity of raw score metric
- Persons span approximately 8 logits
- Items span approximately 3 logits

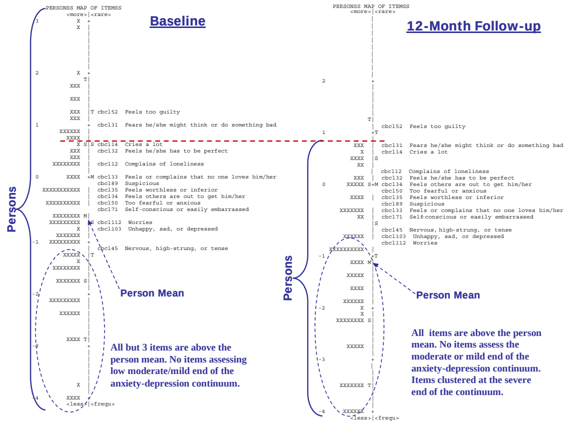


```
PERSONSS MAP OF ITEMSS
        <more>|<rare>
 3          X +
            X |
              |
              |
              |
              |
 2          X +
             T|
          XXX |
          XXX |
          XXX |T cbc152  Feels too guilty
          XXX |
 1            +  cbc131  Fears he/she might think or do something bad
       XXXXXX |
         XXXX |
          X S|S cbc114  Cries a lot
          XXX |  cbc132  Feels he/she has to be perfect
          XXX |
     XXXXXXXX |  cbc112  Complains of loneliness
              |
 0       XXXX +M cbc133  Feels or complains that no one loves him/her
              |  cbc189  Suspicious
  XXXXXXXXXXX |  cbc135  Feels worthless or inferior
              |  cbc134  Feels others are out to get him/her
   XXXXXXXXXX |  cbc150  Too fearful or anxious
              |  cbc171  Self-conscious or easily embarrassed
     XXXXXXXX M|
    XXXXXXXXX |S cbc112  Worries
            X |  cbc1103  Unhappy, sad, or depressed
      XXXXXXX |
-1  XXXXXXXXX +
              |  cbc145  Nervous, high-strung, or tense
        XXXXX |T
            X |
     XXXXXXXX |
      XXXXXXX S|
-2            +
    XXXXXXXXX |
       XXXXXX |
              |
         XXXX T|
-3            +
              |
              |
            X |
-4       XXXX +
       <less>|<frequ>
```

All but 3 items are above the person mean. No items assessing low moderate/mild end of the anxiety-depression continuum.

```
PERSONSS MAP OF ITEMSS
        <more>|<rare>
              |
              |
 2          X +
              |
              |
             T|  cbc131  Fears he/she might think or do something bad
              |T cbc152  Feels too guilty
 1          X +
           XX |
           XX |S
            X |
           XX |  cbc132  Feels he/she has to be perfect
      XXXXXXX S|  cbc114  Cries a lot
              |  cbc133  Feels or complains that no one loves him/her
 0      XXXXX +M cbc135  Feels worthless or inferior
              |  cbc189  Suspicious
         XXXX |  cbc112  Complains of loneliness
              |  cbc134  Feels others are outto get him/her
              |  cbc171  Selfconscious or easily embarrassed
         XXXX |
         XXXX |  cbc150  Too fearful or anxious
        XXXXX |S cbc1103  Unhappy, sad, or depressed
            X |  cbc112  Worries
              |  cbc145  Nervous, highstrung, or tense
      XXXXXXX |
-1  XXXXXXXXX M+
              |T
     XXXXXXXX |
            X |
     XXXXXXXX |
      XXXXXXX |
-2         XX +
             S|
            X |
            X |
    XXXXXXXXX |
-3            +
             T|
         XXXX |
-4     XXXXXX +
       <less>|<frequ>
```

All items are above the person mean. No items assess the moderate or mild end of the anxiety-depression continuum.



Segmenting The Respondent Sample

Separation Index*

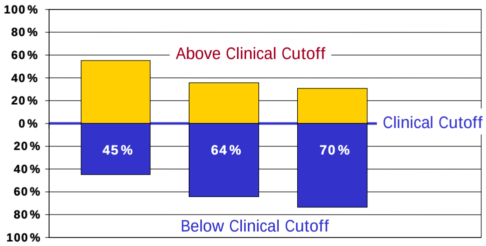
Baseline = 2.65

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.

Percent Sample Above and Below Clinical Anxious-Depressed Cutoff



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

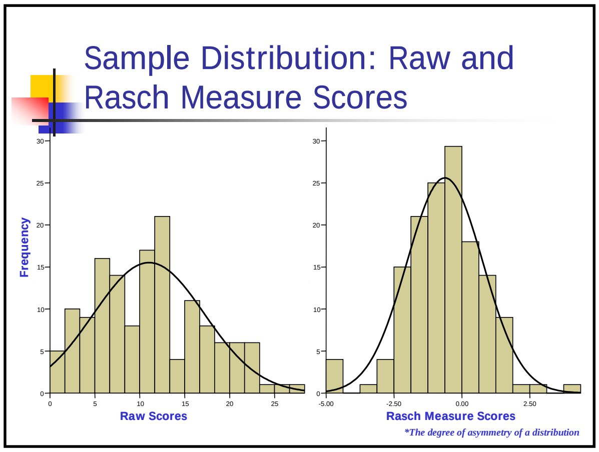
Longitudinal Intervention Studies

- 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 by 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

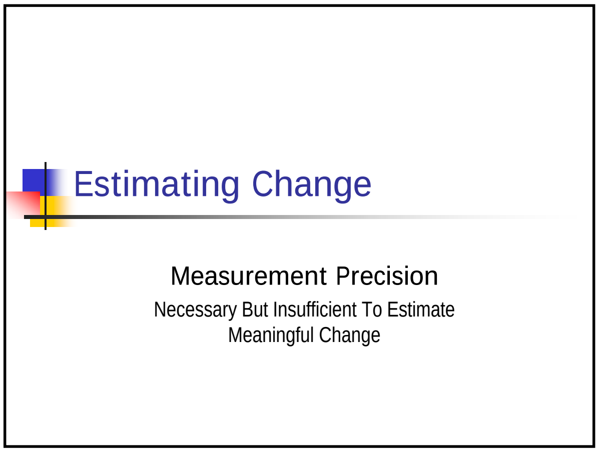
- All items contribute equally to scale score
- Error generalizes across all scale items
- Raw scores are essentially counts

- 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

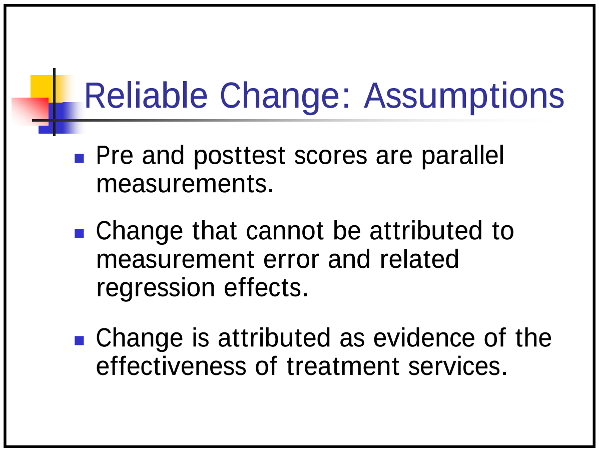
**The degree of asymmetry of a distribution*



Estimating Change

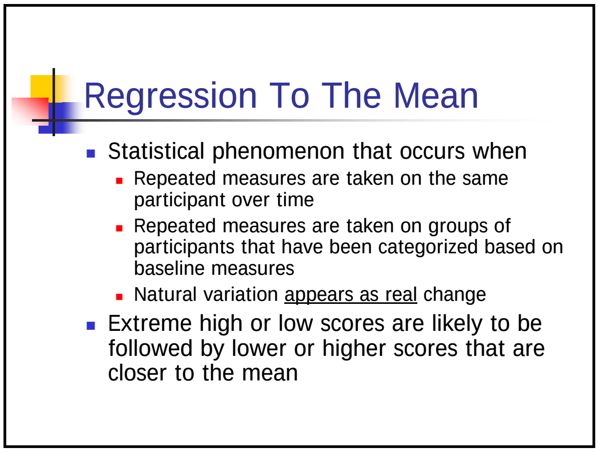
Measurement Precision

Necessary But Insufficient To Estimate
Meaningful Change



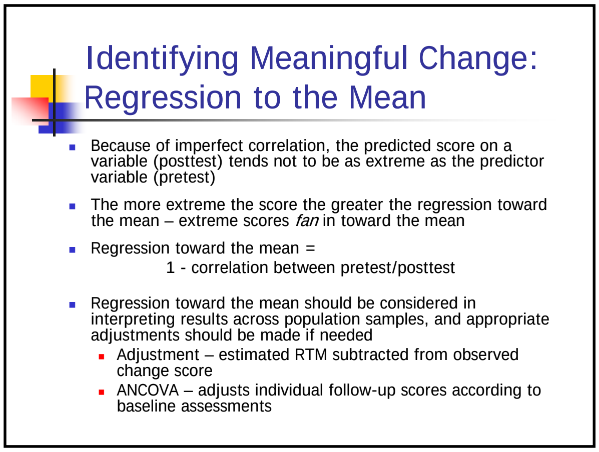
Reliable Change: Assumptions

- 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.



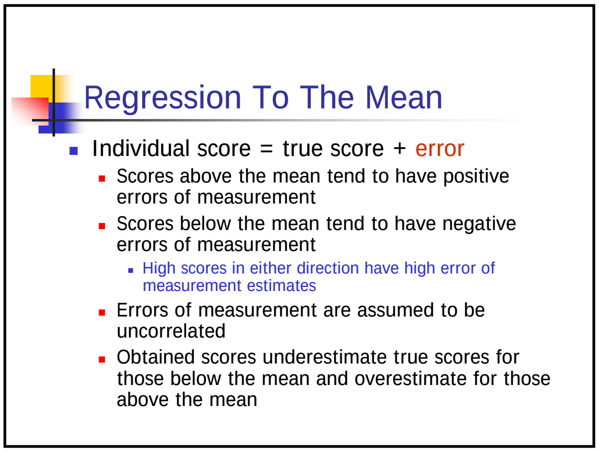
Regression To The Mean

- 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

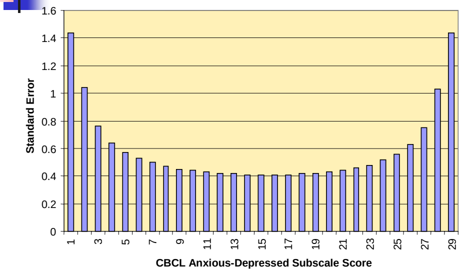
- Because of imperfect correlation, the predicted score on a variable (posttest) tends not to be as extreme as the predictor variable (pretest)
- The more extreme the score the greater the regression toward the mean – extreme scores *fan* in toward the mean
- Regression toward the mean =
  1 - correlation 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

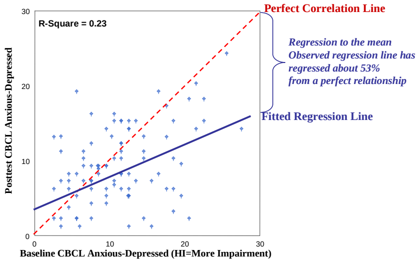
- Individual score = true score + error
  - Scores above the mean tend to have positive errors of measurement
  - Scores below the mean tend to have negative errors of measurement
    - High scores in either direction have high error of measurement estimates
  - Errors of measurement are assumed to be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

Standard Error: CBCL Anxious-Depressed Subscale Scores



Extreme scores (high and low) contain less information.

Regression To 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 (plus/minus) 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

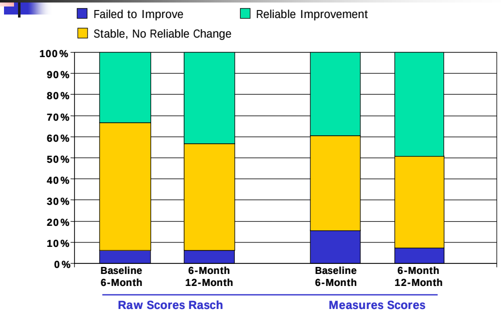
- Reliable change: ± 2 standard errors of measurement = confidence interval
- Standard error of measurement =

$$SEM = SD_t\sqrt{(1-\gamma_{tt})}$$

- 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

  [mean of pretest + reliability of pretest x (pretest score -mean of pretest)]
- Not subject to the effects of regression to the mean

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