Examining How Rasch/IRT Can Increase Measurement Precision

Ann M. Doucette, Ph.D.

The Evaluators’ Institute

American Evaluation Association Annual Meeting
Chicago, Illinois
12 November 2015
Learning Objectives

• Recognize the role of measurement precision in obtaining reliable and generalizable psychotherapy research outcomes
• Understand the basic differences between Classical Test Theory (CTT) and Item Response theory (IRT)
• Discuss the advantages IRT offers over CTT in terms of measurement precision and testing the assumption often made in psychotherapy research of scaled interval-level data
• Describe the basic characteristics and assumptions (e.g., unidimensionality and local independence) of one-parameter logistic IRT models (Rasch) and multiple-parameter models (2PL, 3PL, and 4PL models)
• Recognize the pivotal role of measurement as a foundation of evaluation practice
Performance measurements:
balanced scorecard model for, 106;
converting data into information, 111–116fig; cost effectiveness of, 102; criteria for good, 107–108;
current challenges to, 119–122; customer satisfaction, 103;
disability adjudicators’ claims processing, 110t; EA (evaluability assessment),
90–91; efficiency of, 102–103;
improving performance through, 120–121; logic modeling used in, 72–75fig; networked environments
implementation of, 121–122; outcomes of, 101–102; outputs, 102;
reliability of, 13–14b, 107, 558b, 559;
RFE (rapid feedback evaluation), 95–96; service quality, 103; STEM
instruction program, 75fig; strategy maps for, 106; tables of performance
indicators, 86; validity of, 12–17, 107, 558b, 559. See also
Data; Data analysis; Logic models; Program evaluation; Statistics

Measurement validity: description of, 12, 558b; internal, 14–15, 558b; process for achieving, 12–13b; statistical conclusion, 16–17
Measurement

A scientific method – a way of finding out (more or less reliably) what level of an attribute is possessed by the object or objects under investigation . . . the magnitude of a level of an attribute via its numerical relation (ratio) to another level of the same attribute (Michell, 2001, p. 212).
Measurement Requirements

Measures must be*

• Linear – ordered/monotonic, **conjointly additive** → permitting arithmetic computations can be conducted
  o Assumes unidimensionality
• Item calibrations must not be dependent on whose responses are used to estimate the measurement model (**sample invariant**)
• Person estimates/measures must not be dependent on which items are answered (**test free**)
• Missing data should not matter

* Thurstone, 1925, 1926)
Basic Assumptions: Measurement

- **Reliability**: the consistency across observed scores for a particular attribute (depression, achievement, vocational interest, etc.). Respondents would obtain the same scale score placement on different occasions.

- **Validity**: the accuracy of the measurement in terms of expected scores for a universe of observations defining a specific attribute for a given population and a given purpose.

- **Dimensionality**: measure scores are interpreted as assessing a single trait (depression, anxiety, etc.).
  - Multidimensionality confounds, making it impossible to distinguish the attribute assessed by the measure.
Measurement Function

- Assigning a **numerical value** across a set of item that represents **person placement** on a define attribute
If the student gets all items correct we can assume the student likely gets the algebra, geometry, addition, subtraction, etc., correct; and has exceptional math skill.

If the student gets all items correct we cannot assume the student has exceptional math skill.
“Everything gets much more complicated after this.”
Approaches To Measurement

• Classical Test Theory (CTT)
  o Currently the approach most frequently used

• Latent Trait Models
  o Used frequently in educational testing and healthcare
    • Standardized tests and high stakes tests
    • Increasing acceptance in behavioral healthcare
    • NIH Roadmap for Medical Research Initiative – Patient-Reported Outcomes Measurement Information System (PROMIS)
Measurement Approaches

Classical Test Theory
(True Score Theory)

• Classical Test Theory (CTT)
  o Currently the approach most frequently used
• Latent Trait Models
  o Used frequently in educational testing and physical healthcare
  • Increasing acceptance in behavioral healthcare
  • NIH Roadmap for Medical Research Initiative – Patient-Reported Outcomes Measurement Information System (PROMIS)

Latent Trait Models
(Strong True Score Theory)

Rasch

Item Response Theory

1PL  2PL  3PL  4PL
Classical Test Theory (CTT)

- Classical Test Theory is sometimes referred to as **true score theory**
- Measure score = true score + random error
  - Measurement error is normally distributed
  - Measurement error is uncorrelated with the true score
- Item characteristics are measure and sample dependent
- All items contribute equally to the overall scale score (unless separately weighted using a scoring protocol)
- Response options (e.g. Likert scales) are assumed and frequently analyzed as equal interval scales
Classical Test Theory: Limitations

- Observed score is dependent on choice of items
- Item statistics – item difficulty and item discrimination are sample dependent
  - No modeling of data at item level
- Assumed common measurement error across all items
- Measure reliability is sample dependent
- Limits
  - Ability to equate across tests
  - Develop computer adaptive applications
Item Response Theory (IRT)

- Item Response Theory (IRT): **modern latent trait theory/Strong True Score Theory** – an alternative scaling methodology that enables the examination of the hierarchical structure, unidimensionality and additivity of items and measure scores
- Item – ability relationships: probability of correct response is a function of responder ability, an estimate of his/her level of the measured latent trait
  - Higher ability $\rightarrow$ increased probability of favorable score
- Estimates are provided for **each** individual item and **each** response scale option
- Items provide differential weight/contribution to the overall measure scale score
  - Answering more difficult items yields a higher score
Assumptions: IRT Models

- **Unidimensionality**: items measure the same underlying construct – a dominant first factor
  - Unidimensionality is seldom perfectly achieved – it is a matter of the degree of the departure from unidimensionality
- **Local Independence**: response to an item is conditionally independent of responses to other items in a scale after controlling for the latent trait measured by the scale.
  - Latent variable explains how items are related – e.g., person with high depression would have a specific response profile across items
- **Monotonicity**: consistent increasing relationship between theta ($\theta$ – proficiency/ability) and item responses
  - Expectation that respondent will use consistent response options for items of like difficulty
Item Response Theory Models

- Data modeling, statistical estimation of parameters that represent the location of persons and items on a latent continuum, the measurement construct of interest

- Statistical models specifying the probability of a discrete outcome (e.g., a correct response to an item, the selection of a response option on a Likert scale), in terms of person and item parameters.
  - Person parameters, ability, indicate
    - The level of impairment along the measured construct (e.g., depression, distress, etc.)
    - The strength of an individual’s attitude
    - Competence in an academic area (e.g., mathematics, etc.)
  - Item parameters refer to
    - Difficulty
    - Discrimination (slope or correlation)
    - Guessing (lower asymptote)
    - Carelessness (upper asymptote)
Item Response Theory Models

- The Rasch model/1PL IRT model assumes that item discrimination parameters are equal, implying that two items of the same difficulty \((b)\) on a test are equally precise at measuring ability.

- 2PL models allow discrimination parameters to vary – item difficulty \((b)\) and discrimination \((a)\) are both modeled.
  - Does the item discriminate across individuals with the same ability?

- 3PL models estimate the contribution of guessing \((c)\). Guessing is of questionable importance in behavioral healthcare, although is being used in the area of personality research.

- 4PL models estimate the contribution of carelessness.
IRT Models: Response Scales

- Dichotomous (yes-no) –

- Polytomous (Likert scales, rating scales, etc.)
  - Partial Credit Model (PCM): thresholds for each item response are independently estimated – ordinal scale properties are tested
    - Example: multiple choice items where some responses are more correct than others
    - Additional parameters are estimated → increasing model fit
  - Rating Scale Model (RSM): assumes all items share the same rating scale (e.g., Likert scaling)

Commonly used in evaluations
Rasch versus 1PL

Rasch

- Local fit of the data to the model, one parameter at a time
- Parameterizes each member of the respondent sample individually
- Item characteristic curves (ICCs) modeled to be parallel with a slope of 1 (the natural logistic ogive)
- Item data violating models assumptions (not supporting parameter separability) in a linear framework is examined for possible deletion from the model – *data fit the model*

1PL

- Global fit of the model to the data
- Summarizes the respondent sample as a normal distribution
- ICCs modeled to be parallel with a slope of 1.7 (approximating the slope of the cumulative normal ogive)
- Model reconsidered in the event of data misfit in terms of adding additional parameters, such as discrimination and guessing – *fit of the model to the data*
2PL Model: Item Discrimination ($a$)

- The $a$ parameter is the steepness of the curve at its steepest point (slope of the line tangent to the ICC at $b$)
  - Its closest relative in classical test theory is the item total correlation
- The steeper the curve, the more discriminating the item is, and the greater its item total correlation.
- As the $a$ parameter decreases, the curve gets flatter until there is virtually no change in probability across the ability continuum.
- Items with very low $a$ values provide little useful information for distinguishing among respondents
  - In classical test theory, these items would have low item total correlations.
- The 2-parameter model allows both $a$ and $b$ parameters to vary to describe the items. This model is used where there is no suspected guessing.
Example: Item Discrimination

Logit Distance $a_i = 2.0$

Formula for $a_i$ is obtained from the Newton-Raphson estimation equation:

$$\hat{a}_i = 1 + \left[ \frac{\sum_n (x_{ni} - p_{ni})(\theta_n - b_i)}{\sum_n p_{ni}(1 - p_{ni})(\theta_n - b_i)^2} \right]$$
3PL/4PL Model: Guessing – Carelessness

- Individuals responding to certain measures, such as tests of knowledge may guess in selecting a response.
  - Example: standardized achievement tests where it is suggested that if you do not know the correct response, guess
- In the 3PL IRT model, guessing is parameterized as a lower asymptote to the item's logistic ogive of the probability of a correct answer – *guessability*
- In the 4PL IRT model (4-PL), carelessness is parameterized as an upper asymptote – *mistake-ability*
- The literature suggests that when the lower or upper asymptote is .10 or greater, it is significant
- The lower and upper asymptote estimates are used to modify the item difficulty and person ability estimates.
- The Rasch measurement model handles *guessability* and *mistake-ability* as misfit, and does not make adjustments for item difficulties and person abilities.
Example: Guessing – Carelessness (3PL/4PL)

Difference between Logit and Probit models: logistic approach yields flatter tails – normal/probit curves approach axes more quickly.

Quantitatively Logit and Probit models yield similar results, however estimates are not directly comparable.

MPlus uses a Probit IRT model and provides a conversion factor that yields comparable Probit and Logit estimates.
Advantages of IRT Models

- Sample invariant
- Modeling and estimation of measurement properties at the item level
- Estimation of measurement precision at the score (response option) level
- Estimation of item contribution to the overall test precision for score ranges
- Identification of individuals (groups) responding to items/response option in inconsistent manner
- Calibration across measures – equating
- Foundation for computer adaptive testing (CAT)
IRT Models: Limitations

• Separate software is needed to conduct IRT analyses (e.g., Winsteps, Bilog, Bilog-MG, Multilog, Parscale, Conquest, Rumm, Testfact, etc.)
  o Each software package has unique functions – some are idiosyncratic to specific software packages
  o IRT software has varying ability in terms of its integration with other statistical programs (SPSS, SAS, etc.)

• Expertise is needed in selecting models and interpreting parameter estimates

• Large samples are needed for multi-parameter models (the more parameters the larger the sample needed)

• Less emphasis in assessment, measurement, statistical academic courses
Application of the Rasch Measurement Model
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 assume items are distinguishable in terms of item difficulty ($\theta$).
• The Rasch model provides “sample invariant” (sample independent) item calibrations, item difficulties ($\theta$), from easy to hard — no impairment to severe impairment.
The Rasch Model . . .
Selecting a Few Good Items

The Rasch model asserts that . . .

- Variation in discrimination (2PL) is symptomatic of multi-dimensionality and item bias
- Guessing (3PI) and carelessness (4PL) is an unreliable person characteristic and a measurement liability
  - Guessing and carelessness are not item characteristics

The Rasch model eliminates items violating model assumptions
Scale items represent a construct along a continuum from low to high, minimal to maximal, etc. Every scale item is calibrated along this continuum.

Persons with no/little need for intervention

Moderate

Persons with high need for intervention

Items about adversity, severe problems

Items about day-to-day issues, small problems

Low Need

Measured Construct

High Need
Calibrating Items

- Person\(^1\) has no opportunity to demonstrate improvement. Scores can only indicate stability or deterioration.
- Person\(^3\) has no opportunity to demonstrate deterioration or minimal improvement.

Person\(^1\) has no opportunity to demonstrate improvement. Scores can only indicate stability or deterioration.
Person\(^3\) has no opportunity to demonstrate deterioration or minimal improvement.

Person\(^2\)

Item Calibration

Mild
No Items

δ\(^1\) Easiest Item

δ\(^2\) δ\(^2\) δ\(^3\) δ\(^4\) δ\(^5\) δ\(^6\) Hardest Item

Severe
No Items

β

Improvement

Deterioration
Examples: Data Sources
Data Source: Examples

• Large Commercial Health Plan: Assessment of Global Distress (Life Status Questionnaire 30 items)
  o Behavioral healthcare (N=258,393)
    • 32% male
    • Adults (18-24: 9%, 25-35: 26%, 36-45: 31%, 45-60: 30%)
    • Diagnosis (Depression 42%, Adjustment 20%, Anxiety 9%, Bipolar 7%, Alcohol/Drugs 6%)

• Project Match: URICA – Stage of Change
  o Funded by NIAAA, 6-year study ending in 8/2005
  o ~ 1700 volunteers with alcohol abuse/dependence
    • N=1726 (952 outpatients, 744 after care; 1307 males)
Questionnaires

**URICA**
- 32 item, self-report scale
- Respondents asked to think about entry into treatment or about problems they experience in their lives
- Four eight-item subscales
  - Pre-contemplation
  - Contemplation
  - Action
  - Maintenance
- Scoring: 5-point Likert scale
  - Scale: 1 = strong disagreement → 5 = strong agreement (3 = undecided)
  - Total score: mean of C + A + M – PC = second order continuous Readiness to Change score

**Life Status Outcome Questionnaire (OQ 30)**
- 30 item, self-report scale
- Respondents asked to think back about how they felt over the past
- 30 items – no subscales
  - Items ask about psychological, emotional function, physical function, and substance abuse
- Scoring: 5-point Likert scale
  - Scale: 0 = Never, 5 = almost always (3 = sometimes)
  - Total score: sum across all 30 items – scale score: 0 - 120
Basic Measurement Question

Are Items Sufficiently Co-located with Respondents?
Persons and items are well matched at baseline. Items differentiate a significant proportion of the sample.

No items are available to differentiate individuals who are mildly or severely distressed, except for alcohol/drug and suicide focused items.

Items clustering together, although added to the overall score, reflect little in terms of increased or decreased distress, as they assess the same level of the construct.
41.3% respondents.

Global Distress

Item Cluster

Decreasing person ability
Decreasing item difficulty

Increasing person ability
Increasing item difficulty
No items to differentiate individuals in terms of higher degrees of readiness to change.

Items clustering together, although added to the overall score, reflect little in terms of increased or decreased distress, as they assess the same level of the construct.
URICA: Readiness to Change

56.3% of the sample
No items to accurately locate readiness to change status
Measurement Model Fit: Global Distress

| ENTRY | RAW    | SCORE  | COUNT | MEASURE | MODEL | S.E. | INFIT | OUTFIT | PTBIS | EXACT MATCH | ESTIM | P | DISCR | VALUE | ITEMS
|-------|--------|--------|-------|---------|-------|------|-------|--------|-------|--------------|-------|---|-------|--------|--------|
|       | ENTRY  | RAW    | COUNT | MEASURE | MODEL | S.E. | INFIT | OUTFIT | PTBIS | EXACT MATCH | ESTIM | P | DISCR | VALUE | ITEMS
| 24    | 30095 | 258004 | 2.20  | .00 | 1.29 | 9.9 | .21 | 9.9 | .17 | 89.7 | 90.9 | .91 | .12 | 24 TROUBLE BECAUSE OF ALCOHOL
| 11    | 40340 | 257941 | 1.84  | .00 | 1.47 | 9.9 | .46 | 9.9 | .19 | 86.8 | 88.5 | .87 | .16 | 11 USE ALCOHOL/DRUGS TO GET GOING
| 20    | 65489 | 257906 | 1.82  | .00 | 1.53 | 9.9 | .50 | 9.9 | .15 | 76.0 | 80.5 | .77 | .25 | 20 PEOPLE CRITICIZE MY DRINKING
| 7     | 15827 | 257603 | 1.47  | .00 | 1.02 | 6.9 | .99 | 6.7 | .50 | 57.1 | 56.7 | 1.00 | .61 | 7 SUICIDAL THOUGHTS
| 23    | 30679 | 255346 | 0.69  | .00 | 1.04 | 9.9 | 1.02 | 9.9 | .53 | 49.3 | 48.3 | .96 | 1.20 | 23 TROUBLE GETTING ALONG WITH FRIENDS
| 12    | 37898 | 257605 | 0.32  | .00 | 0.74 | 9.9 | .73 | 9.9 | .72 | 49.5 | 41.8 | 1.35 | 1.47 | 12 FEEL WORTHLESS
| 19    | 37872 | 256349 | 0.30  | .00 | 1.06 | 9.9 | 1.08 | 9.9 | .57 | 41.8 | 41.5 | .92 | 1.48 | 19 DISTURBING THOUGHTS
| 28    | 39407 | 257615 | 0.18  | .00 | 0.92 | 9.9 | .92 | 9.9 | .65 | 42.3 | 39.5 | 1.12 | 1.53 | 28 SOMETHING WRONG WITH MIND
| 21    | 41766 | 258025 | 0.15  | .00 | 1.27 | 9.9 | 1.30 | 9.9 | .45 | 39.6 | 43.2 | .63 | 1.62 | 21 UPSET STOMACH
| 25    | 41766 | 257807 | 0.12  | .00 | 0.90 | 9.9 | .90 | 9.9 | .64 | 45.7 | 42.2 | 1.14 | 1.62 | 25 FEEL SOMETHING BAD GOING TO HAPPEN
| 30    | 46369 | 257896 | 0.06  | .00 | 1.05 | 9.9 | 1.07 | 9.9 | .53 | 51.5 | 49.0 | .95 | 1.80 | 30 SATISFIED WITH RELATIONSHIPS
| 18    | 46843 | 255779 | 0.01  | .00 | 0.84 | 9.9 | 8.9 | 9.9 | .65 | 56.7 | 49.8 | 1.17 | 1.83 | 18 HAPPY PERSON
| 15    | 43891 | 257407 | -0.04 | .00 | 1.38 | 9.9 | 1.41 | 9.9 | .36 | 39.9 | 45.9 | .55 | 1.70 | 15 FREQUENT ARGUMENTS
| 9     | 47307 | 257474 | -0.06 | .00 | 0.95 | 9.9 | 9.6 | 9.9 | .58 | 52.9 | 48.3 | 1.05 | 1.84 | 9 SCHOOL/WORK SATISFYING
| 17    | 46092 | 256786 | -0.09 | .00 | 0.73 | 9.9 | 0.74 | 9.9 | .73 | 51.7 | 43.1 | 1.36 | 1.79 | 17 HOPELESS ABOUT FUTURE
| 8     | 47129 | 255913 | -0.13 | .00 | 0.89 | 9.9 | 0.90 | 9.9 | .64 | 49.2 | 45.0 | 1.13 | 1.84 | 8 FEEL WEAK
| 2     | 47586 | 257112 | -0.13 | .00 | 0.81 | 9.9 | 0.82 | 9.9 | .68 | 51.9 | 45.7 | 1.23 | 1.85 | 2 NO INTEREST
| 10    | 47936 | 256375 | -0.16 | .00 | 0.98 | 9.9 | 0.96 | 9.9 | .60 | 45.4 | 43.9 | 1.02 | 1.87 | 10 FEARFUL
| 5     | 51039 | 256460 | -0.24 | .00 | 0.93 | 9.9 | 0.95 | 9.9 | .60 | 53.8 | 48.6 | 1.07 | 1.99 | 5 SATISFIED WITH LIFE
| 22    | 51013 | 255327 | -0.34 | .00 | 0.95 | 9.9 | 0.96 | 9.9 | .63 | 44.5 | 41.0 | 1.08 | 2.00 | 22 NOT WORKING/STUDYING AS WELL AS NEED TO
| 27    | 51485 | 257073 | -0.37 | .00 | 0.84 | 9.9 | 0.86 | 9.9 | .67 | 49.2 | 42.8 | 1.21 | 2.00 | 27 NOT DOING WELL SCHOOL/WORK
| 14    | 55825 | 257386 | -0.57 | .00 | 0.96 | 9.9 | 0.96 | 9.9 | .62 | 45.2 | 43.2 | 1.06 | 2.17 | 14 FEEL LONELY
| 29    | 57020 | 256874 | -0.65 | .00 | 0.71 | 9.9 | 0.71 | 9.9 | .73 | 55.0 | 46.2 | 1.34 | 2.22 | 29 FEEL BLUE
| 26    | 57399 | 257728 | -0.68 | .00 | 0.83 | 9.9 | 0.84 | 9.9 | .67 | 50.2 | 44.9 | 1.21 | 2.23 | 26 FEEL NERVOUS
| 16    | 56262 | 257369 | -0.70 | .00 | 0.84 | 9.9 | 0.84 | 9.9 | .66 | 51.3 | 46.4 | 1.20 | 2.19 | 16 DIFFICULTY CONCENTRATING
| 1     | 58430 | 257786 | -0.70 | .00 | 1.26 | 9.9 | 1.29 | 9.9 | .47 | 40.3 | 42.5 | .65 | 2.27 | 1 TROUBLE FALLING ASLEEP
| 4     | 60845 | 256446 | -0.96 | .00 | 0.98 | 7.4 | 0.98 | 6.4 | .57 | 48.8 | 47.1 | 1.03 | 2.37 | 4 BLAME SELF
| 6     | 60982 | 256324 | -1.00 | .00 | 0.98 | 6.7 | 0.99 | 4.7 | .54 | 56.4 | 52.0 | 1.02 | 2.38 | 6 IRRITATED
| 13    | 65891 | 257368 | -1.05 | .00 | 1.56 | 9.9 | 1.67 | 9.9 | .34 | 35.9 | 42.5 | .24 | 2.56 | 13 CONCERN ABOUT FAMILY TROUBLES
| 3     | 66576 | 256301 | -1.28 | .00 | 1.97 | 9.9 | 9.6 | 9.9 | .58 | 49.7 | 47.1 | 1.05 | 2.60 | 3 STRESSED

Estimates identified as having misfit are reviewed for potential deletion from the measurement model. These estimates are indicative of multidimensionality, guessability and mistake-ability (carelessness) – the need for a multi-parameter model to explain variance from the measurement model.
<table>
<thead>
<tr>
<th>ENTRY</th>
<th>TOTAL</th>
<th>SCORE</th>
<th>COUNT</th>
<th>MEASURE</th>
<th>S.E.</th>
<th>MNSQ</th>
<th>ZSTD</th>
<th>ZSTD</th>
<th>CORR.</th>
<th>OBS%</th>
<th>EXP%</th>
<th>DISCR.</th>
<th>VALUE</th>
<th>ITEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUMBER</td>
<td>SCORE</td>
<td>COUNT</td>
<td>MEASURE</td>
<td>S.E.</td>
<td>MNSQ</td>
<td>ZSTD</td>
<td>ZSTD</td>
<td>CORR.</td>
<td>OBS%</td>
<td>EXP%</td>
<td>DISCR.</td>
<td>VALUE</td>
<td>ITEM</td>
<td></td>
</tr>
<tr>
<td>ENTRY</td>
<td>TOTAL</td>
<td>SCORE</td>
<td>COUNT</td>
<td>MEASURE</td>
<td>S.E.</td>
<td>MNSQ</td>
<td>ZSTD</td>
<td>ZSTD</td>
<td>CORR.</td>
<td>OBS%</td>
<td>EXP%</td>
<td>DISCR.</td>
<td>VALUE</td>
<td>ITEM</td>
</tr>
<tr>
<td>MEAN</td>
<td>7155.8</td>
<td>1710.2</td>
<td>0.00</td>
<td>0.03</td>
<td>1.25</td>
<td>2.5</td>
<td>62.8</td>
<td>58.0</td>
<td>12.2</td>
<td>8.5</td>
<td>5.2</td>
<td>21.7</td>
<td>19.2</td>
<td>7.6</td>
</tr>
<tr>
<td>S.D.</td>
<td>581.8</td>
<td>1.8</td>
<td>0.19</td>
<td>0.9</td>
<td>3.7</td>
<td>5.5</td>
<td>6.9</td>
<td>6.9</td>
<td>6.9</td>
<td>6.9</td>
<td>6.9</td>
<td>6.9</td>
<td>6.9</td>
<td></td>
</tr>
</tbody>
</table>
Rasch Model
Examining Assumptions

- Dimensionality
- Local Independence
- Monotonicity
- Additivity
The Rasch model specifies that item discrimination (item slope) is uniform across items
- This supports additivity and construct stability

Item discrimination is NOT parameterized
- The Rasch slope is the average discrimination of all the items. It is not the mean of the individual slopes because discrimination parameters are non-linear
- Mathematically, the average slope is set at 1.0 when the Rasch model is formulated in logits, or 1.70 when it is formulated in probits (2-PL and 3-PL – 0.59 is the conversion from logits to probits)

Issue to consider
- Are the misfitting items distinct?
  - Could they be considered a separate subscale?
- Does removal of these item appreciable improve the measurement model?
Item Discrimination

Is The One Parameter Model Sufficient?

Item differences (difficulty)/person differences (ability) contribute additively to the probability (log-odds) of responding positively.
Challenges – Adding the 2nd Parameter

Interpretation: Which Item Is More Difficult?

**Lord’s Paradox**

Probability of scoring higher on **green item** is .9. At this level of impairment, the **red item** is more difficult.

Probability of scoring higher on **red item** is .6. At this level of impairment.

- **Mild**
- **Moderate**
- **Severe**
2PL – Is More Better?

• Measurement development
  o Parsimonious model
  o Good data fit to the measurement model
  o Objective is to minimize measurement variance – development of measures that can be generalized to diverse population groups, conditions, etc.

• Performance on a developed measure
  o Additional parameters may shed light on interpretation of measurement variance.
  o Example: items may function differently for individuals at different positions on the measured construct.
Questioning The Assumption Of Additivity, Interval Level Data and Uncorrelated error
Approaching Uniform Response
Global Distress

An individual has a 50% probability of responding with “0” or a “1” and so forth where the probability curve intersects (Rasch-Step Calibration).

Satisfied with relationships

Happy

0 = Never
1 = Rarely
2 = Sometimes
3 = Frequently
4 = Almost Always
Non-uniform Response
Global Distress

Having trouble with alcohol
Using alcohol/drugs to get going

0 = Never       3 = Frequently
1 = Rarely      4 = Almost Always
2 = Sometimes   4 = Almost Always

7.7 SUICIDAL THOUGHTS

Category Probability

Measure relative to item difficulty

Category Probability

Measure relative to item difficulty

.18 logits
.42 logits
7.11 logits
5.99 logits

0 = Never
1 = Rarely
2 = Sometimes
3 = Frequently
4 = Almost Always
Assumptions: Non-uniform And Non-interval Response Patterns

Assumptions:
- Response options are equal intervals – a score of “4” is twice as much as a score of “2”
- All respondents systematically interpret and use response options across items in the same manner

People criticize my drinking
I have thoughts of ending my life
I am concerned about family troubles
I feel stressed at work/school

Assumed Interval Responses

<table>
<thead>
<tr>
<th>OBSERVED AVERAGE MEASURES BY AN OBSERVED CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>-6</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>012 3 4</td>
</tr>
<tr>
<td>7 SUICIDAL THOUGHTS</td>
</tr>
<tr>
<td>13 CONCERN ABOUT FAMILY TROUBLES</td>
</tr>
<tr>
<td>3 STRESSED</td>
</tr>
</tbody>
</table>
Stress | Alcohol/Drugs | Suicidality

-5 | -5 | -5
-4 | -4 | -4
-3 | -3 | -3
-2 | -2 | -2
-1 | -1 | -1
0  | 0  | 0
1  | 1  | 1
2  | 2  | 2
3  | 3  | 3
4  | 4  | 4
5  | 5  | 5

- Never (0)
- Rarely (1)
- Sometimes (2)
- Frequently (3)
- Almost Always (4)

Rasch Ruler – Logit Measure Scores
Questioning the Assumption of Additivity – URICA Readiness to Change

STAGES
1. Pre-contemplation
2. Contemplation
3. Action
4. Maintenance

• Scales have some overlap
• Estimates for the Contemplation scale indicate respondents have an easier time responding positively to these items – little distinction between the two scales
Error – Assumed To Be Uncorrelated

Standard Error as a Function of Ability

\[ \text{SE} = 0.25 \]
\[ \theta = -2.34 \text{ to } 2.47 \]

\[ \text{SE} = 1.0 \]
\[ \theta = -5.21 \text{ / } \theta = 5.04 \]
Interpreting Score Change

Do All Items Provide the Same Level Of Information?
### Total Scores – Equivalent Meaning

**Does It Matter Where Change Occurs?**

<table>
<thead>
<tr>
<th>Item</th>
<th>Baseline</th>
<th>Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>I feel stressed at work/school</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>I am concerned about family troubles</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>I feel irritated</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>I blame myself for things</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total Score:</strong></td>
<td>16</td>
<td>6</td>
</tr>
<tr>
<td>I am annoyed by people who criticize drinking</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>I need a drink the next morning to get going</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>I have thoughts of ending my life</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>I have trouble at work because of drinking drug use</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total Score:</strong></td>
<td>16</td>
<td>6</td>
</tr>
</tbody>
</table>
Response Option Approaches

• Dominance model – Likert scaling
  o The more strongly an opinion or perspective is held the higher the probability of endorsing a favorable response
    ▪ All items of less difficulty would also be endorsed favorably - additive

• Ideal point model
  o Probability of endorsement is associated with the respondent’s position on the latent trait ($\theta$)
    • Probability decreases as the item location moves away (either below or above) from his/her position on the latent trait
    • Person with moderately strong opinion would likely not endorse (disagree with) items reflecting stronger and weaker positions
    • Person can disagree for two reasons, the trait level ($\theta$) is below item level, and conversely the trait level ($\theta$) is above the item location

• Multi-dimensional pairwise preference (MUPP)
  o Forced-choice statements
    • Statements representing trait levels ($\theta$) that are nearly proximate – balanced item sets
    • Increased measurement precision

• Unfolding

Strongly Disagree  Mid-point  Strongly Agree
─ ← + ─
The estimated Reliable Change Index for the General Distress Measure is reported to be 10. An individual’s score must change by at least 10 points on the measure to be considered clinically significantly changed.

10 points has a differential effect depending where on the continuum the change occurs.
The complexity of assessment includes dimensions not directly measured:
- Literacy level
- Conscientiousness
- Emotional stability
- Interpretation of construct meaning

MIRT:
- *between-item* (assessment is multidimensional, (multiple latent traits), however each item assess only one dimension)
- *within item* (multiple skills/competencies are needed to respond to an item) dimensionality
- Models test information as a response surface representing the relationship between correct response, person ability (θ) and discrimination (α)

MIRT scores – complexity in interpretation:
- Multidimensional item difficulty and discrimination over a range of θ vectors
MIRT Item Surface

- Item surface increases more quickly across $\theta_1$
  - Consequence of the $a$ parameter (discrimination)
Reliable Change*  
Evaluating and Characterizing Effective Treatment Providers

<table>
<thead>
<tr>
<th>Scores</th>
<th>Raw Score(^1)</th>
<th>Raw Score(^2)</th>
<th>Rasch (^a) Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement</td>
<td>22%</td>
<td>24%</td>
<td>20%</td>
</tr>
<tr>
<td>Stability</td>
<td>58%</td>
<td>56%</td>
<td>72%</td>
</tr>
<tr>
<td>Deterioration</td>
<td>20%</td>
<td>20%</td>
<td>8%</td>
</tr>
</tbody>
</table>

\[^*\]Reliable Change (Jacobson & Truax, 1991)

\[ RCI = \frac{X_2 - X_1}{SE_{DIFF}} \]

Note: Suggested reliable raw score change for the General Distress Measure is 10 points (Raw Score\(^1\)).

Sample specific RCI is estimated at a change of 12 points (Raw Score\(^2\)).

\(^a\) Multidimensional random coefficient multinomial logit model – alcohol/drug items treated as a separate subscale.
Assessing Outcomes

DILBERT by Scott Adams

DOGGBERT CONSULTS
YOU NEED A DASHBOARD APPLICATION TO TRACK YOUR KEY METRICS.

THAT WAY YOU'LL HAVE MORE DATA TO IGNORE WHEN YOU MAKE YOUR DECISIONS BASED ON COMPANY POLITICS.

WILL THE DATA BE ACCURATE?

OKAY, LET'S PRETEND THAT MATTERS.
Thank You!

Ann M. Doucette, Ph.D.

Email: ann.doucette9@gmail.com