Panel Discussion 2

Mixed Methods Approach to Assuring Content Validity

THIRD ANNUAL
PATIENT-REPORTED OUTCOME (PRO) CONSORTIUM WORKSHOP

April 4, 2012 ■ Silver Spring, MD

Co-sponsored by

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Mixed Methods Approach to Assuring Content Validity

J. Jason Lundy, PhD

Asst. Director, PRO Consortium

Critical Path Institute

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Mixed Methods Panelists

Joseph C. Cappelleri, PhD, MPH

- Senior Director, Biostatistics, Pfizer Inc.

Jeremy Hobart MD, PhD, FRCP

- Professor of Clinical Neurology and Health Measurement, Peninsula College of Medicine and Dentistry

Ron D. Hays, PhD

- Professor, Department of Medicine, David Geffen School of Medicine, UCLA

Panel Objectives

Provide quantitative recommendations for:

- Content validity stage
  - Exploratory analyses to refine scales
  - Methods appropriate for small samples
- Psychometric analysis stage
  - Confirmatory analyses of the measurement model
  - Utilizing larger samples, in the clinical trial context

Instrument Development Process

- Previously, development was conducted in two linear stages
  - Qualitative Analysis Stage & Quantitative Analysis Stage
- Subsequently, the two stages of research were redefined
  - Content Validity Stage & Psychometric Testing Stage

Instrument Development Process

Previous Process		Current Process
Scoping Stage	→	Scoping Stage
Qualitative Research Stage • Qualitative Interviews, no quantitative testing	→	Content Validity Stage • Mixed Methods – Qualitative Interviews & Quantitative Assessments
Quantitative Research Stage • Confirmatory Psychometric Analyses	→	Psychometric Analysis Stage • Confirmatory Psychometric Analyses

Mixed Methods

- Blends qualitative and quantitative methodologies into the assessment of content validity
- The approach is cyclical, iterative, and hypothesis-driven
- Anomalies that are detected should be explained, modifications to the instrument should be made, and further testing conducted

Presentation Overview

- Classical Test Theory and Item Response Theory: A Brief Overview
  - *Joseph C. Cappelleri, PhD, MPH*
- Rasch Measurement Theory and the achievement of content validity
  - *Jeremy Hobart MD, PhD, FRCP*
- Multiple Methods are Needed to Develop Survey Instruments
  - *Ron D. Hays, PhD*
- FDA Response
  - *James P. Stansbury, PhD, MPH*

Classical Test Theory and Item Response Theory: A Brief Overview

Joseph C. Cappelleri, PhD, MPH

Senior Director, Biostatistics

Pfizer Inc.

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Assumptions

- Assumes each person has true score on a concept of interest
  - Observed score = True score + Error
  - Obtained if there were no errors in measurement
  - Expected over an infinite number of independent administrations
  - True score not observed but estimated by observed score
- Key assumptions
  - Random errors are normally distributed
  - Random errors are uncorrelated with true score
  - Expected value of error is zero

Item Difficulty

- Consider a set of binary items (can be extended to ordinal items)
- Item difficulty is measured by the proportion of respondents who “endorse” an item (here “endorsing” implies a favorable response)
- Items with high proportions of endorsement are easy items while items with low proportions of endorsement are difficult items
- Total score for an individual is based on how many items endorsed
- Items with proportions of 0 or 1 are useless because they do not differentiate among individuals on the concept of interest
- Best to create items with varying difficulty with an average proportion of endorsement across items of 0.50

Item Discrimination

- Proportion of endorsement (item difficulty) and the “extreme group method” can be used to calculate an *item discrimination index*

- The more the item discriminates among subjects with different attributes, the higher its discrimination index

- The opportunity of an item to have the highest discrimination index occurs when its proportion of endorsement is 0.50

Item Discrimination Index – Extreme Group Method

- Step 1: Partition subjects who have the highest and lowest overall scores into upper and lower groups
  - For example, upper group: top 25%, lower group: bottom 25%
- Step 2: Determine the proportion who endorsed each item in the upper and lower groups
- Step 3: Subtract this pair of proportions from the two groups to arrive at a discrimination index for each item

Item Discrimination Index - Illustration

Item	Proportion Endorsed for Upper Group	Proportion Endorsed for Lower Group	Item Discrimination Index
1	0.90	0.10	0.80
2	0.85	0.20	0.65
3	0.70	0.65	0.05
4	0.10	0.70	-0.60

Item 1 is the best at discriminating

Item 2 is the second best

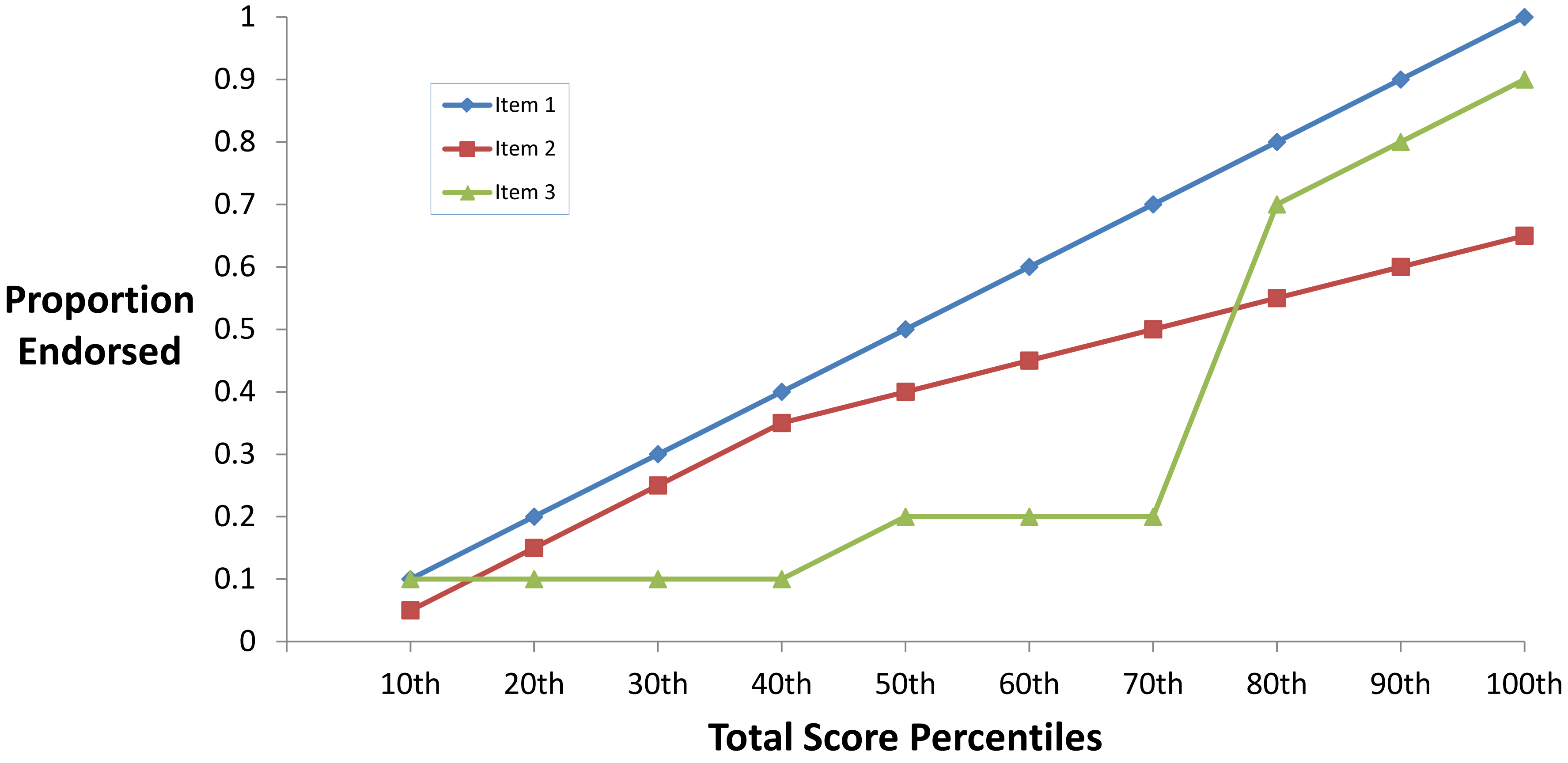
Item 3 is poor at discriminating

Item 4 seems very poor (or not, depending on the nature of the item)

Item Curves

- Provides more fine-grained information on an item than the overall proportion endorsed or discrimination index
- Produced by plotting the percentage of subjects choosing each response option on the vertical axis by the total score on the horizontal axis (expressed as a percentile)
- A good item has its probability of endorsing increasing monotonically with increasing total score (e.g., by showing an S-shaped curve)

Item Curves - Illustration



Item 1: Equally good at discriminating across the continuum of the attribute
Item 2: Discriminates better at the lower end than at the upper end of the attribute
Item 3: Discriminates better at upper end, especially between 70th and 80th percentiles

Corrected Item-to-Total Correlation

- Another assessment of item discrimination

- Measures how well an item correlates with the sum of the remaining items

- Best to have moderate-to-high correlations

- Items with low correlation indicate that they do not go with the rest of the items

Reliability

- Internal consistency – Cronbach’s alpha
  - If items are measuring the same concept, they should elicited similar response
  - Function of average inter-item correlation and number of items

- Test-retest
  - Captures the stability or reproducibility of the measure
  - Correlation of measure on two occasions between which there is no change

Sample Size Considerations

- Samples as small as 30 individuals can provide useful descriptive information about the psychometric performance of measures
  - Based on empirical evidence and experience as well as knowledge of statistical theory
- Multivariate methods, such as exploratory factor analysis and confirmatory factor analysis, can be considered but require larger samples

Item Weighting

- Differential when item are given more weight or less weight when being combined into a total score
  - Three ways to assign differential weights: item reliability, factor loadings, corrected item-to-total coefficients

- This is in contrast to giving each item equal weight
  - Each item contributes equally
  - Generally preferred strategy when items are substantially inter-correlated in measuring a single concept

- Items can be averaged or summed to produce total (raw) scores
  - Scores can be linearly transformed to a Z-score so that the mean is 0 and standard deviation of 1, analogous to the ability parameter (“theta”) metric in item response theory

Classical Test Theory (CTT) vs. Item Response Theory (IRT)

- CTT: Focus most often is on the total score
- IRT: Focus is on entire pattern of responses to all test items by an individual
- CTT and IRT provide different yet useful and complementary ways to examine responses to a series of items

Some IRT Considerations

- Non-linear monotonic function describing the association between a subject’s level on a latent trait and the probability of a particular response to an item

- Major assumptions
  - Unidimensionality (scale is measuring only one attribute)
  - Local independence (for people with the same latent trait, there is no correlation among the items)
  - Monotonicity (probability of endorsing an item should increase monotonically with higher scores on the scale)

- Item characteristic curves (ICCs) depict the correspondence between the item responses and a latent trait
  - Characterized by one, two or three parameters

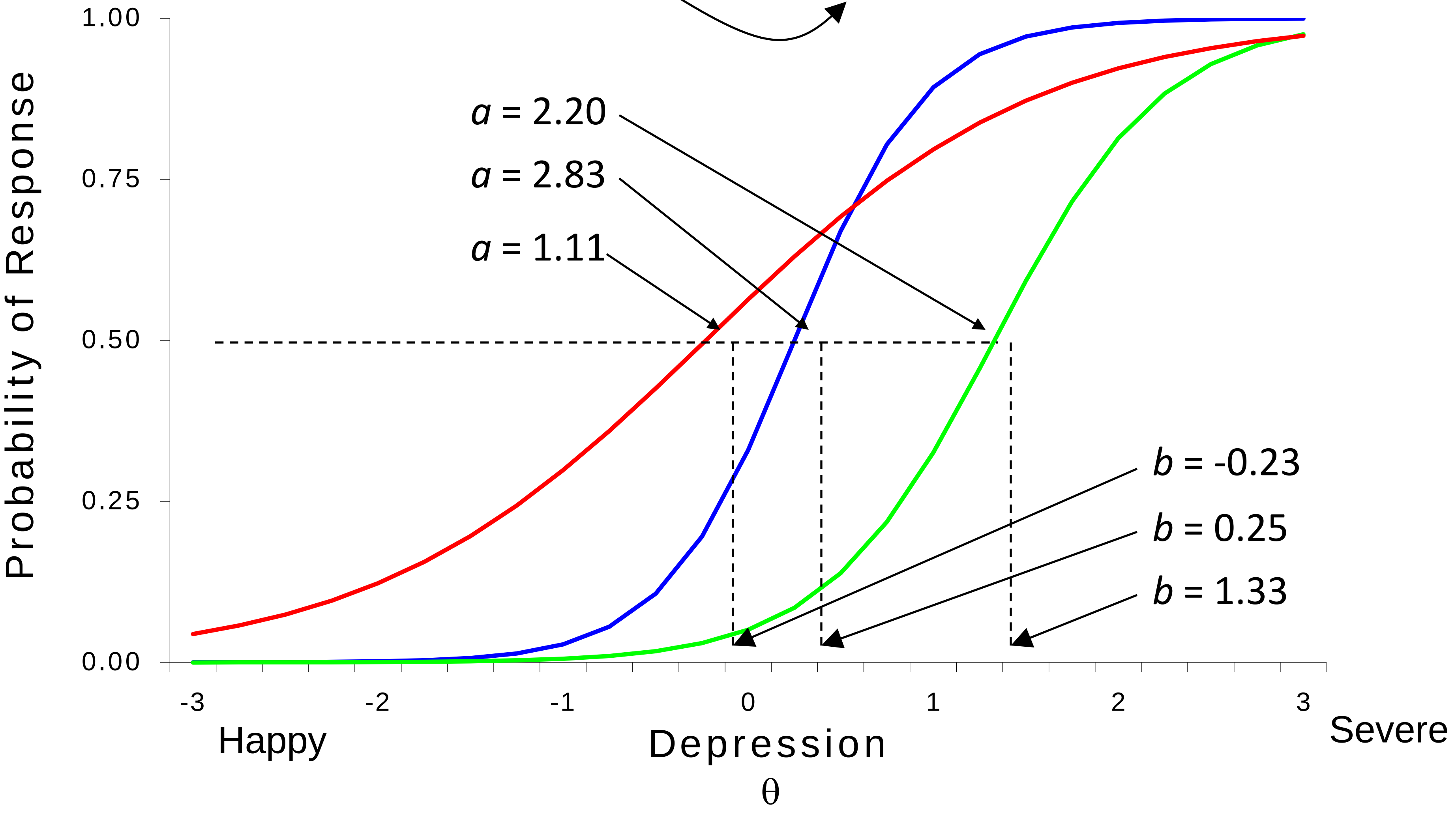
Common IRT Models

Model	Item Response Format	Model Characteristics
Rasch / 1-Parameter Logistic	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
2-Parameter Logistic	Dichotomous	Discrimination and threshold parameters vary across items.
Graded Response	Polytomous	Ordered responses. Discrimination varies across items.
Nominal	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit (Rasch Model)	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale (Rasch Model)	Polytomous	Discrimination equal across items. Item threshold steps equal across items.
Generalized Partial Credit	Polytomous	Variation of Partial Credit Model with discrimination varying among items.

2-Parameter Logistic IRT Model: ICCs for 3 Items (b_i is difficulty, a_i is discrimination, θ is trait)

$$P\left(X_i = 1 \middle| \theta\right) = \frac{1}{1 + e^{-1.7 a_i (\theta - b_i)}}$$

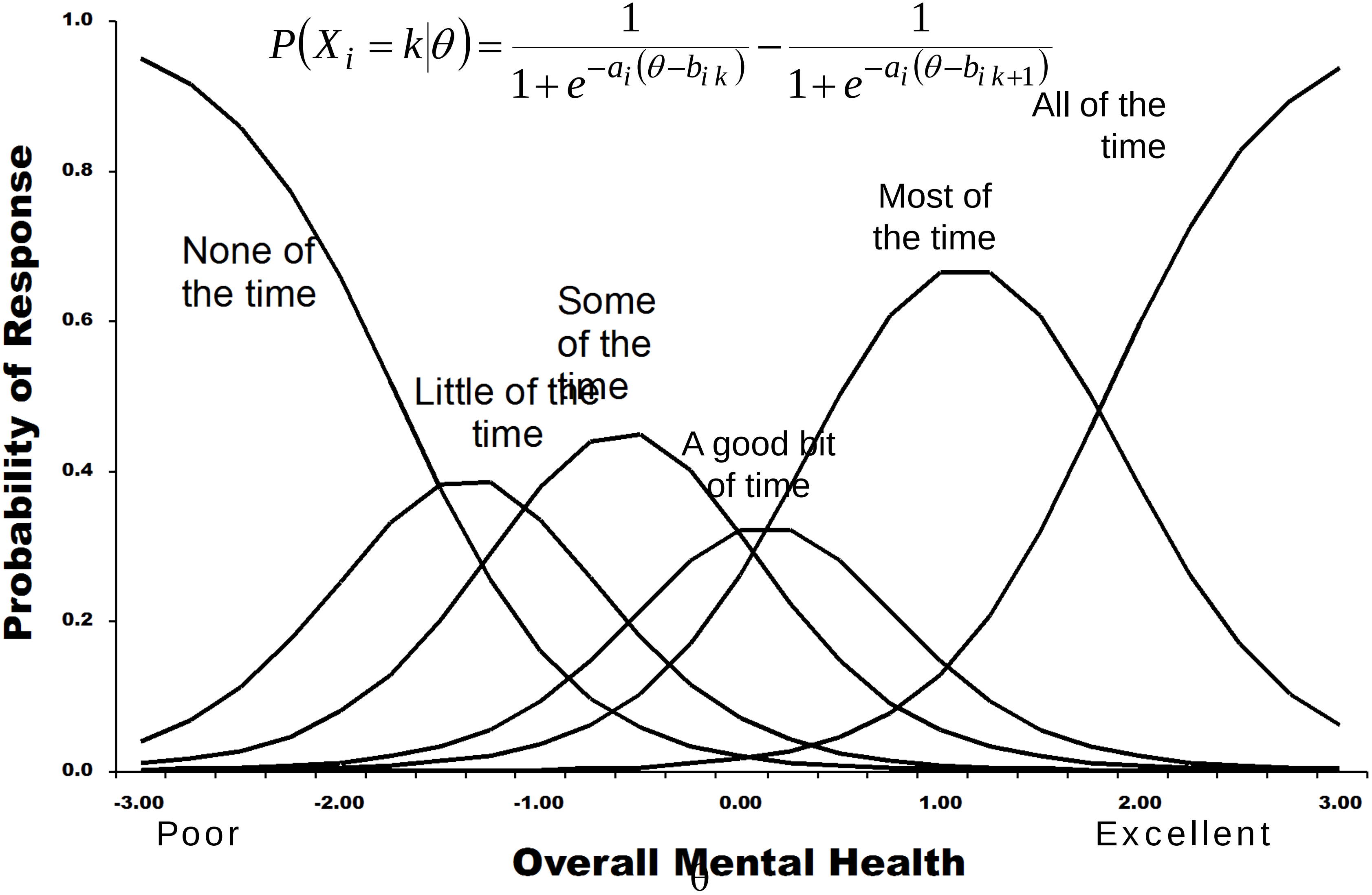
$$a_i(\theta - b_i)$$



IRT Graded Response Model: ICC for One Item

Did you have a lot of energy?

$$P(X_i = k|\theta) = \frac{1}{1+e^{-a_i(\theta-b_{i\,k})}} - \frac{1}{1+e^{-a_i(\theta-b_{i\,k+1})}}$$



Sample Size Considerations

- Depends on IRT model to be estimated
  - Parameters ↑, Sample Size ↑ - Rasch models need less data.
- Depends on number of items or questions
  - Number of items ↑, Sample Size ↑
- Depends on number of response options
  - Number of response categories ↑, Sample Size ↑

Rasch Measurement Theory and the achievement of content validity

Jeremy Hobart MD, PhD, FRCP

Professor of Clinical Neurology and Health Measurement, Peninsula College of Medicine and Dentistry

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

Overview

- A few words on content validity
- Rasch Measurement Theory (RMT):
  - What is it?
  - Why does it foster a mixed methods approach
- RMT approach to scale develpment
- 3 brief examples

A few words on content validity

- More than domain coverage
- Item & response category wording & working
- Variable mapping
- Scale-to-sample targeting
- Scale performance
- …the extent to which a scale measures…..

What is Rasch Measurement Theory-RMT ?

- An experimental measurement paradigm for scale development & evaluation

- A theory-driven approach with hypothesis generation and testing

- Identification, explanation and investigation (diagnosis) of anomalies

(Anomaly =departure of hypothesis from hypothesis test)

- Hypothesis revision and re-testing

What is the hypothesis test in RMT ?

$$Pr\{x_{ni}|\beta_n,\delta_i\} = \frac{e^{x_{ni}(\beta_n-\delta_i)}}{1+e^{(\beta_n-\delta_i)}}$$

Why is the Rasch model an hypothesis test?

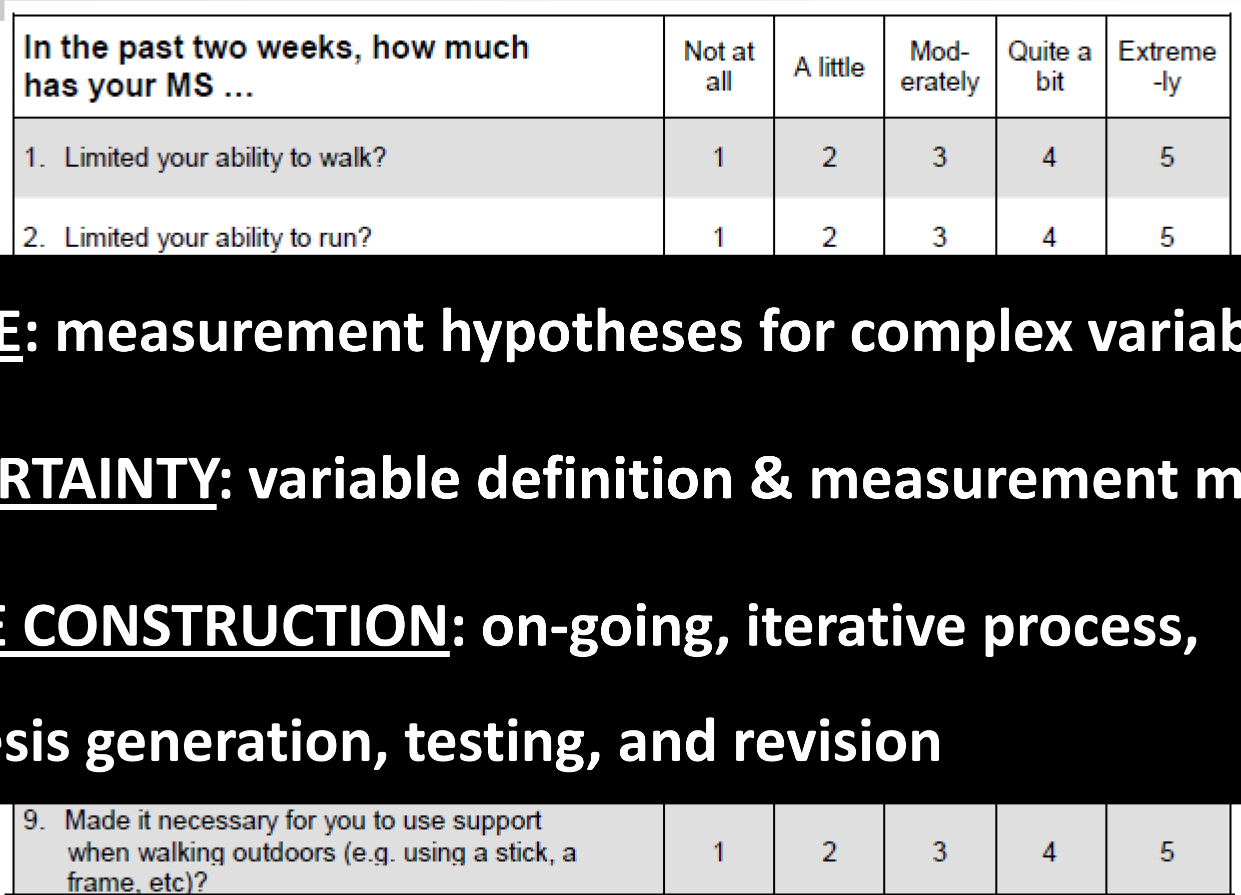
- derived to articulate the requirements of scales for them to enable measurement

- Model derived from 1st principals to enable invariant comparisons

- Model is independent of data

- Therefore, model provides a test against which the data can be compared

Why is RMT suited to scale development ?



In the past two weeks, how much has your MS ...	Not at all	A little	Mod-erately	Quite a bit	Extreme-ly
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors (e.g. using a stick, a frame, etc)?	1	2	3	4	5

1. **SCALE**: measurement hypotheses for complex variables

2. **UNCERTAINTY**: variable definition & measurement method

3. **SCALE CONSTRUCTION**: on-going, iterative process, hypothesis generation, testing, and revision

Summary of an RMT-based approach

- Conceptual clarity. Invest ++

- <u>Round 1</u>
- “qualitative” work to generate the hypotheses
- “quantitative” work tests hypotheses (bespoke, RA, n=small)
- Review, reflect, explain, inform, investigate, revise

- <u>Round 2</u>
- “qualitative” work to develop the hypothesis further
- “quantitative” work tests hypotheses (bespoke, RA, n=small
- Review, reflect, explain, inform, investigate, revise

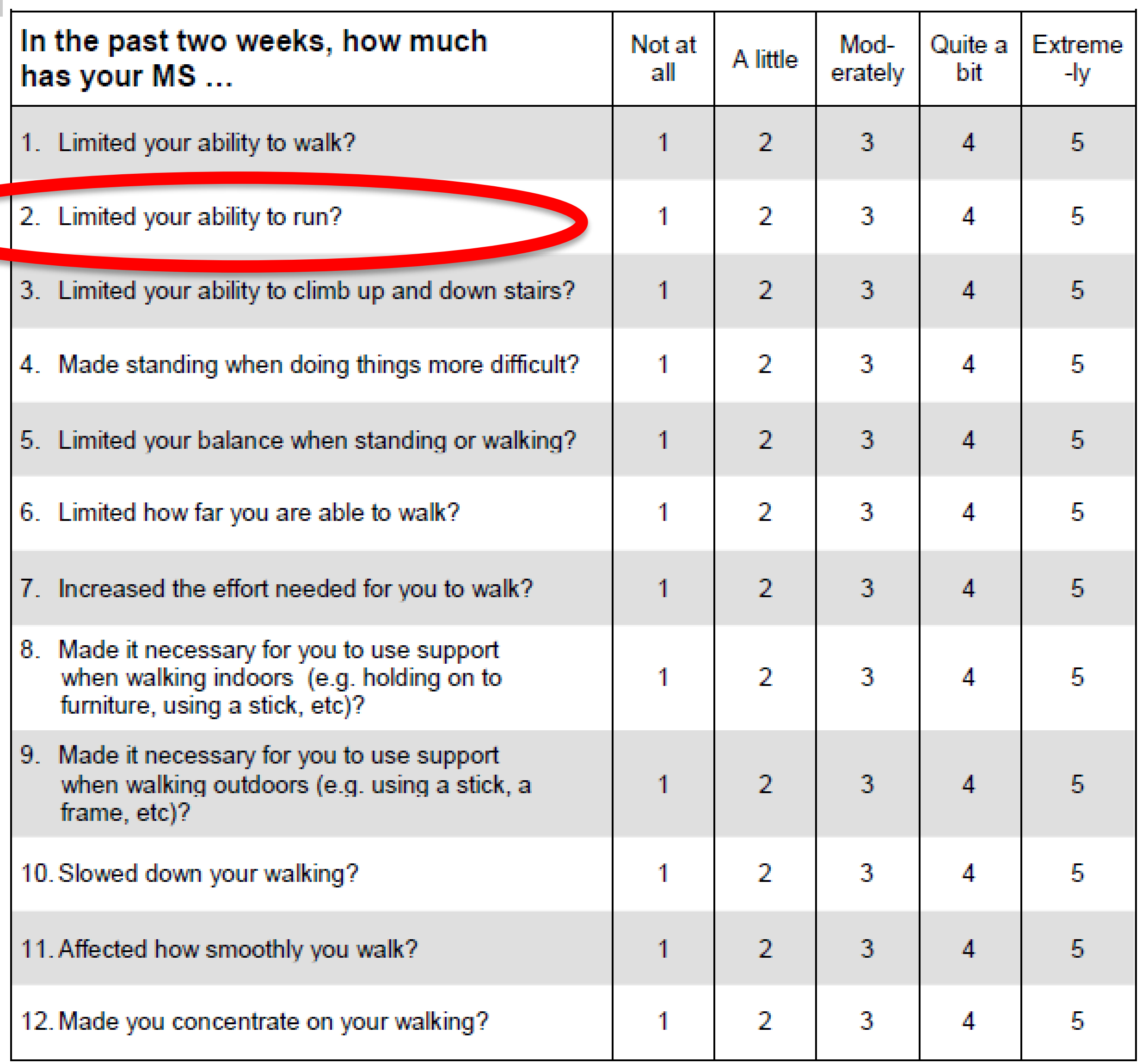
- Additional rounds as required: iterative,

- Ultimately larger sample quantitative evaluations

Examples of how RMT can assist scale content development

- Item fit
- Response category ordering
- Targeting & precision

Anomaly = item misfit

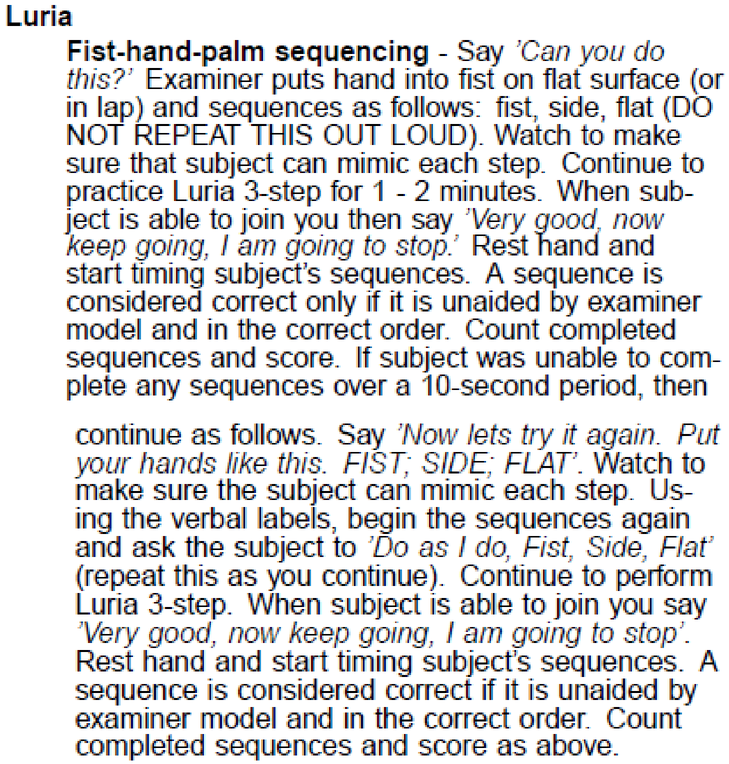


In the past two weeks, how much has your MS ...	Not at all	A little	Mod-erately	Quite a bit	Extreme-ly
1. Limited your ability to walk?	1	2	3	4	5
2. Limited your ability to run?	1	2	3	4	5
3. Limited your ability to climb up and down stairs?	1	2	3	4	5
4. Made standing when doing things more difficult?	1	2	3	4	5
5. Limited your balance when standing or walking?	1	2	3	4	5
6. Limited how far you are able to walk?	1	2	3	4	5
7. Increased the effort needed for you to walk?	1	2	3	4	5
8. Made it necessary for you to use support when walking indoors (e.g. holding on to furniture, using a stick, etc)?	1	2	3	4	5
9. Made it necessary for you to use support when walking outdoors (e.g. using a stick, a frame, etc)?	1	2	3	4	5
10. Slowed down your walking?	1	2	3	4	5
11. Affected how smoothly you walk?	1	2	3	4	5
12. Made you concentrate on your walking?	1	2	3	4	5

Anomaly = item response categories don’t work empirically as intended conceptually

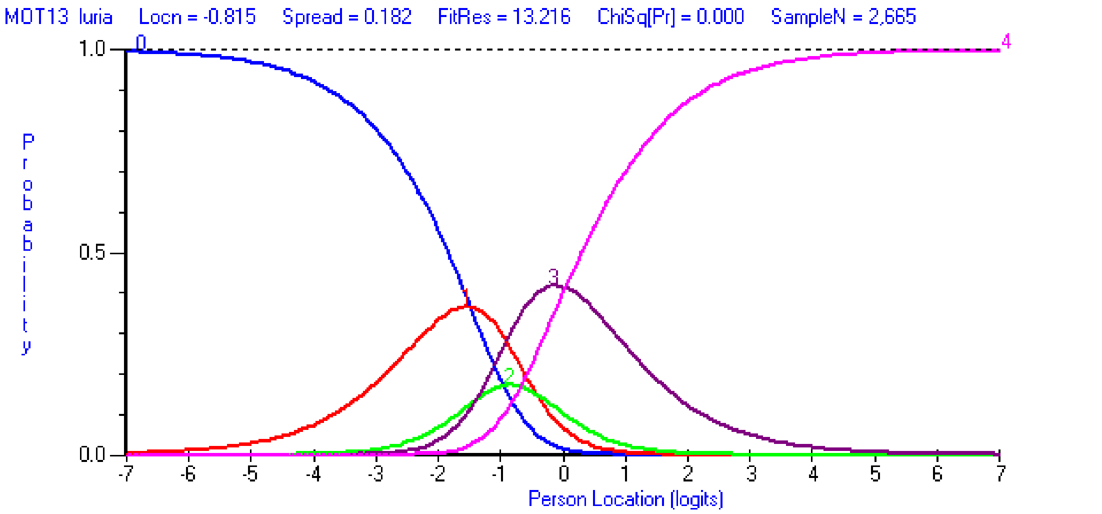
Luria:

0 = ≥4 in 10 sec, no cue
1 = <4 in 10 sec, no cue
2 = ≥4 in 10 sec with cues
3 = <4 in 10 sec with cues
4 = cannot perform

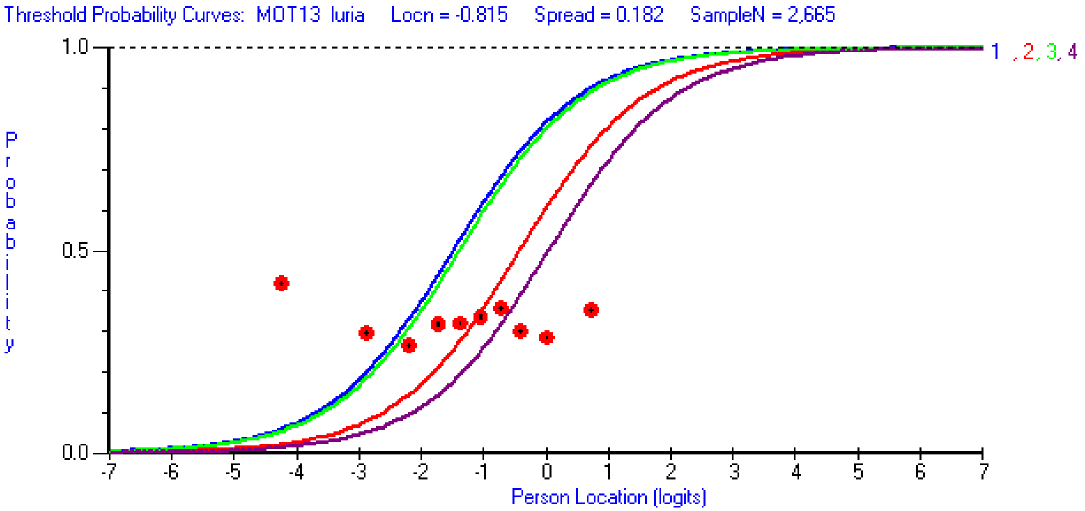


Luria

Fist-hand-palm sequencing - Say *'Can you do this?'* Examiner puts hand into fist on flat surface (or in lap) and sequences as follows: fist, side, flat (DO NOT REPEAT THIS OUT LOUD). Watch to make sure that subject can mimic each step. Continue to practice Luria 3-step for 1 - 2 minutes. When subject is able to join you then say *'Very good, now keep going, I am going to stop.'* Rest hand and start timing subject's sequences. A sequence is considered correct only if it is unaided by examiner model and in the correct order. Count completed sequences and score. If subject was unable to complete any sequences over a 10-second period, then continue as follows. Say *'Now lets try it again. Put your hands like this. FIST; SIDE; FLAT'*. Watch to make sure the subject can mimic each step. Using the verbal labels, begin the sequences again and ask the subject to *'Do as I do, Fist, Side, Flat'* (repeat this as you continue). Continue to perform Luria 3-step. When subject is able to join you say *'Very good, now keep going, I am going to stop'*. Rest hand and start timing subject's sequences. A sequence is considered correct if it is unaided by examiner model and in the correct order. Count completed sequences and score as above.

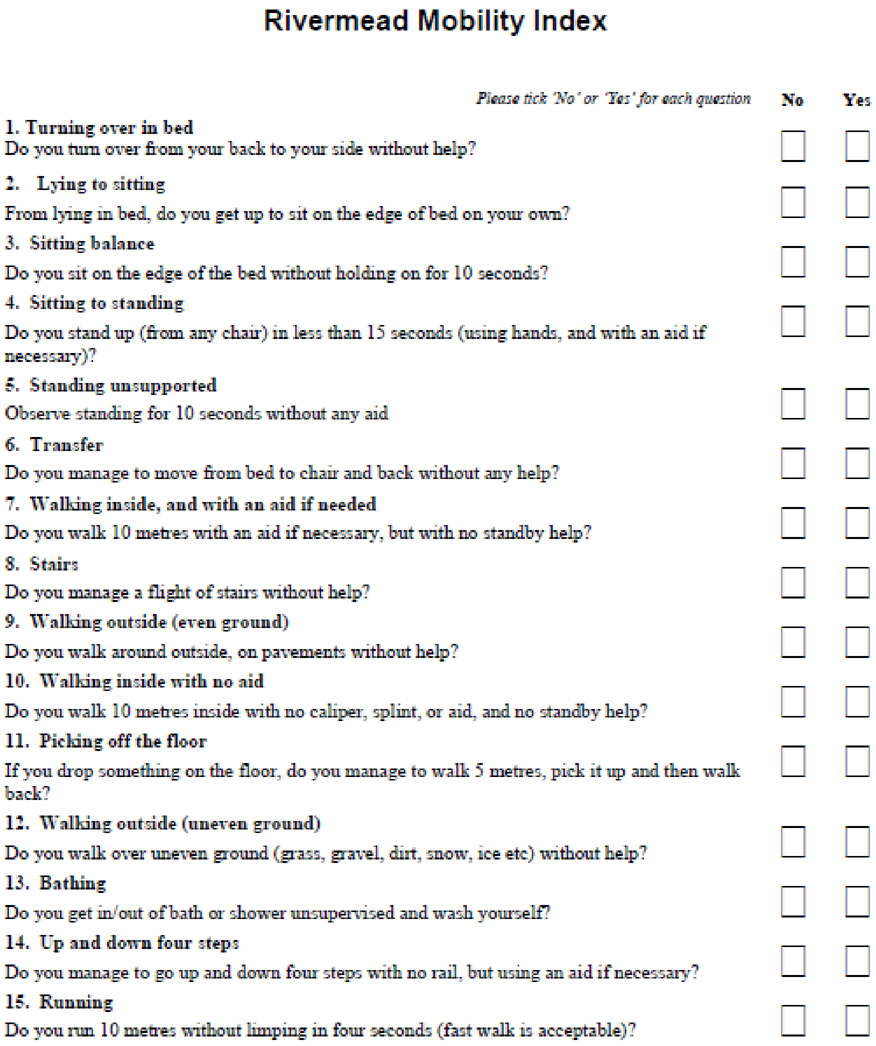


MOT13 luria Locn = -0.815 Spread = 0.182 FitRes = 13.216 ChiSq[Pr] = 0.000 SampleN = 2,665



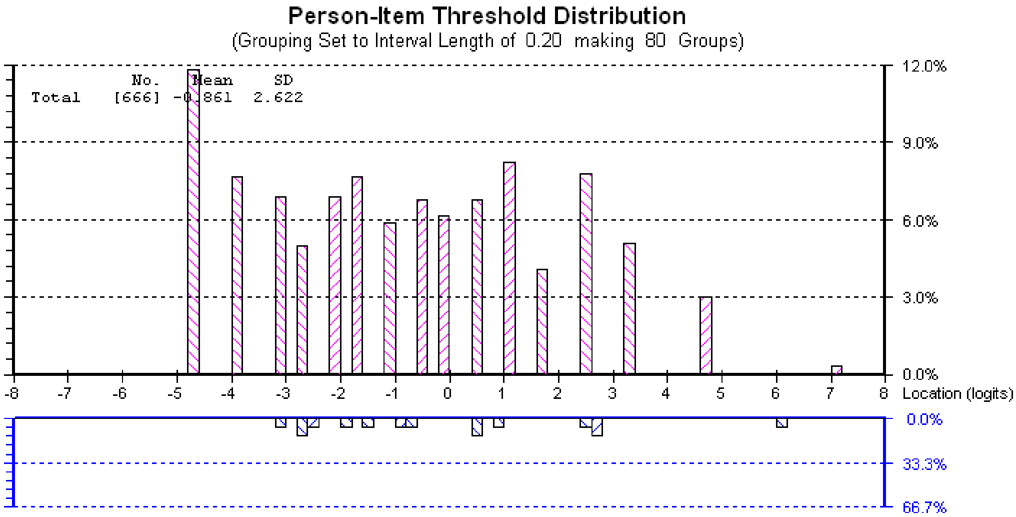
Threshold Probability Curves: MOT13 luria Locn = -0.815 Spread = 0.182 SampleN = 2,665

Anomaly = poor targeting and precision for a clinical trial



Rivermead Mobility Index

Please tick 'No' or 'Yes' for each question	No	Yes
1. Turning over in bed Do you turn over from your back to your side without help?	□	□
2. Lying to sitting From lying in bed, do you get up to sit on the edge of bed on your own?	□	□
3. Sitting balance Do you sit on the edge of the bed without holding on for 10 seconds?	□	□
4. Sitting to standing Do you stand up (from any chair) in less than 15 seconds (using hands, and with an aid if necessary)?	□	□
5. Standing unsupported Observe standing for 10 seconds without any aid	□	□
6. Transfer Do you manage to move from bed to chair and back without any help?	□	□
7. Walking inside, and with an aid if needed Do you walk 10 metres with an aid if necessary, but with no standby help?	□	□
8. Stairs Do you manage a flight of stairs without help?	□	□
9. Walking outside (even ground) Do you walk around outside, on pavements without help?	□	□
10. Walking inside with no aid Do you walk 10 metres inside with no caliper, splint, or aid, and no standby help?	□	□
11. Picking off the floor If you drop something on the floor, do you manage to walk 5 metres, pick it up and then walk back?	□	□
12. Walking outside (uneven ground) Do you walk over uneven ground (grass, gravel, dirt, snow, ice etc) without help?	□	□
13. Bathing Do you get in/out of bath or shower unsupervised and wash yourself?	□	□
14. Up and down four steps Do you manage to go up and down four steps with no rail, but using an aid if necessary?	□	□
15. Running Do you run 10 metres without limping in four seconds (fast walk is acceptable)?	□	□



Person-Item Threshold Distribution

(Grouping Set to Interval Length of 0.20 making 80 Groups)

Two Philosophies of IRT Measurement

- Develop a well-fitting model to reflect the item response data
- The model should reflect the properties of the data sufficiently and accurately, so that the behavior of the item is summarized by the model parameters
- Philosophy: Items are assumed to measure as they do, not as they should.

Non-Rasch Modelers

- Obtain specific measurement properties defined by the model to which the item response data must fit.
- If an item or a person does not fit within the measurement properties of the model, the item or person is discarded.
- Philosophy: Model the data as it should behave using models that yield strong mathematical properties.

Rasch Modelers

Source: D. Thissen & H. Wainer (Eds.). (2001). Test Scoring. LEA

Two Paradigms of scale development

- Develop a well-fitting model to reflect the item response data
- The model should reflect the properties of the data sufficiently and accurately, so that the behavior of the item is summarized by the model parameters
- Philosophy: Items are assumed to measure as they do, not as they should.

IRT

- Posits a theory-driven hypothesis testing approach
- Views an item set (scale) as a hypothesis of how a complex variable might be measured
- Uses the Rasch model as the hypothesis test (it articulates measurement requirements)
- Treats “misfit” as anomalies in measurement hypothesis (scale) requiring explanation and investigation.
- Findings advance scales to achieve better measurement.

RMT

Multiple Methods are Needed to Develop Survey Instruments

Ron D. Hays, PhD

Professor of Medicine and Professor of Health Services, UCLA

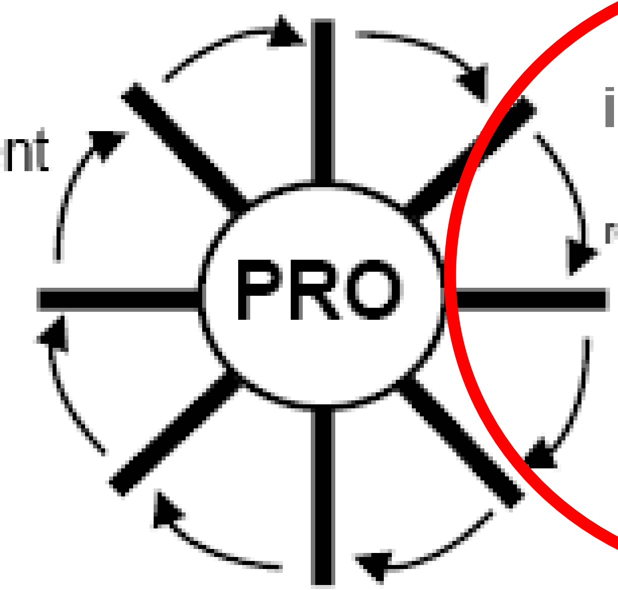
CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA

FDA PRO Process
(released February 2006)

i. Identify Concepts and Develop Conceptual Framework

Identify concepts and domains that are important to patients.
Determine intended population and research application.
Hypothesize expected relationships among concepts.



ii. Create Instrument

Generate Items.
Choose administration method,
recall period, and response scales.
Draft Instructions.
Format Instrument.
Draft procedures for scoring and administration. Pilot test draft instrument. Refine instrument and procedures.

iii. Assess Measurement Properties

Assess score reliability, validity, and ability to detect change.
Evaluate administrative and respondent burden. Add, delete, or revise items.
Identify meaningful differences in scores. Finalize instrument formats, scoring, procedures, and training materials.

iv. Modify Instrument

Change concepts measured,
populations studied,
research application,
instrumentation,
or method of administration.

Documentation

- Chronology of all item development activities
- Protocols for qualitative interviews, focus groups, cognitive interviews and other research used to identify concepts, generate items, or revise an existing instrument, including training of interviewers
- Development of response options, modes of administration and scoring
- Size, characteristics, location, and (if requested) transcripts of each qualitative interview and focus group
- Documentation on how saturation was achieved (i.e. no new information was obtained from additional qualitative interviews or focus groups)

Documentation (Cont.)

- Description of any pilot test, including cognitive interviewing, cognitive interview transcripts (if requested)
- Versions of the instrument at various milestones of development
- Item tracking table that list the source of each item in the final instrument, and how it changed during development
- A summary statement of qualitative research in support of content validity of the PRO instrument

D. Patrick et al, Value in Health 2007, 10, S125-37

Iterative Process

- Literature review, existing items, focus groups, cognitive interviews
- Traditional “classical test theory” analyses
  - Item frequencies, means, variances, correlations, internal consistency reliability, factor analysis, etc.
- Rasch and item response theory analyses
  - Item fit, ordering of response categories, item location, item discrimination, precision

Literature Review

Focus Groups

Expert Input and Consensus

Existing Items

Newly Written Items

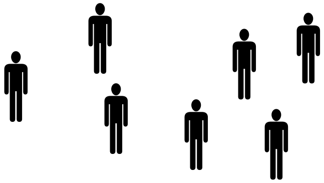
Initial Item Pool

Expert Review

Translation

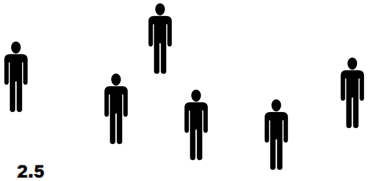
Cognitive Interviews

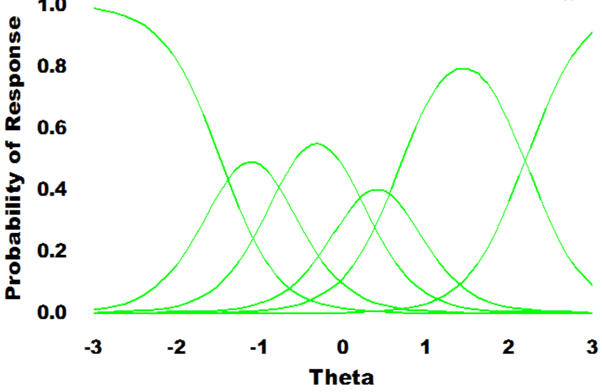
Data Analysis



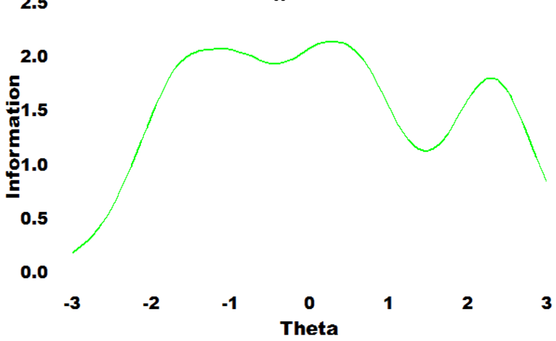
Questionnaire

administered to large representative sample



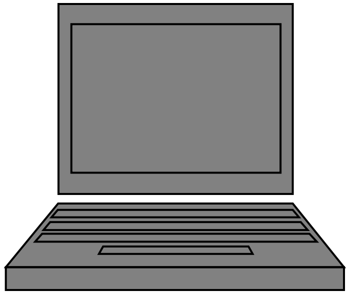


Psychometric Testing



Final Items

Short Form Instruments



Process yields preliminary (but solid) ideas about

- Clarity of item instructions, stems, and response categories
- Item
  - Fit
  - Location of response categories
  - Information (precision)
- Scale information

Mixed Methods Approach to Assuring Content Validity

FDA Response

James P. Stansbury, PhD, MPH
Consumer Safety Officer, SEALD
Study Endpoints, FDA

CRITICAL PATH INSTITUTE
collaborate · innovate · accelerate

FDA