Evaluating Multi-Item Scales

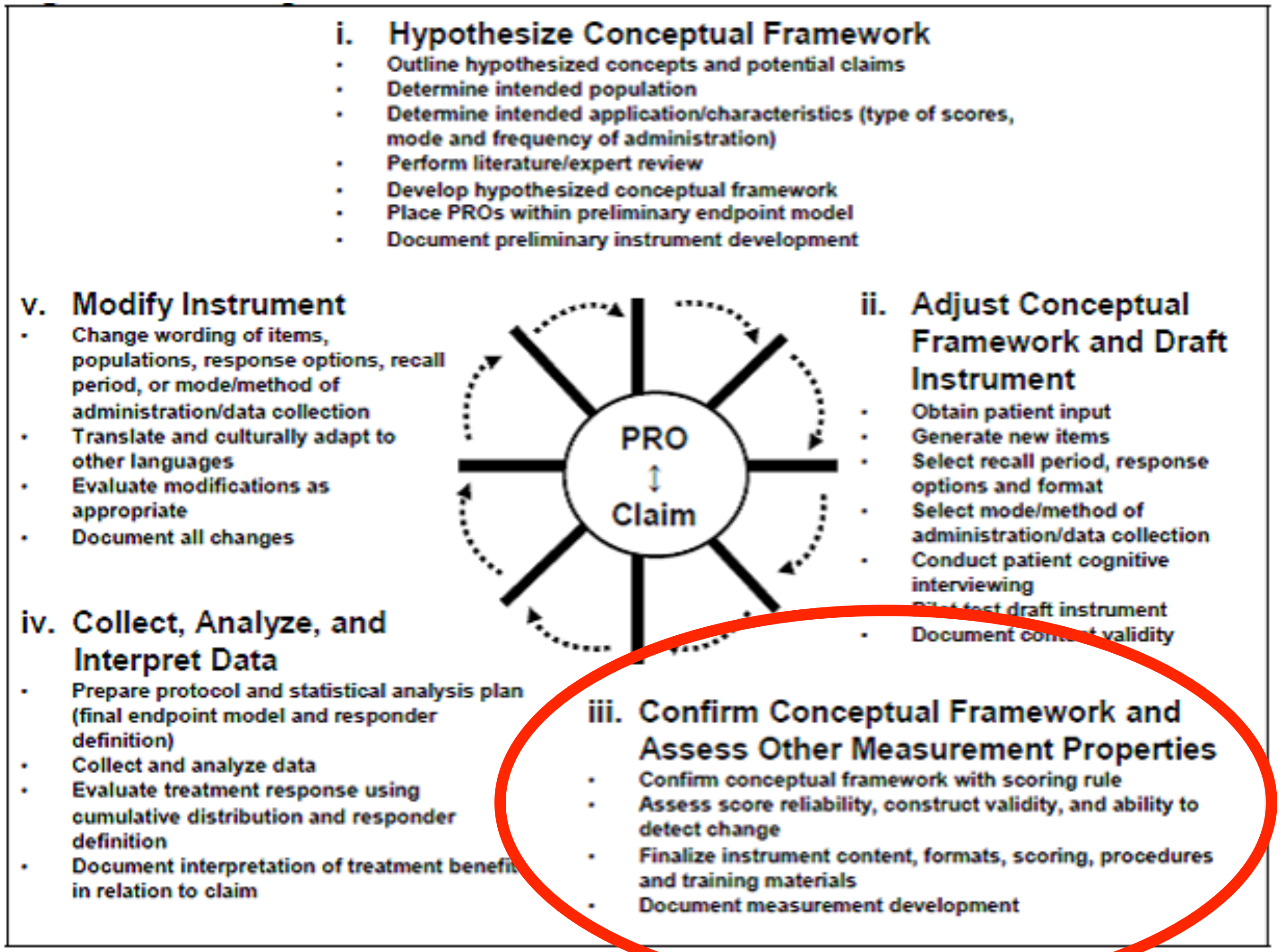
Ron Hays

Health Services Research Design (HPM 225B)

January 25, 2016, 1:00-3:00pm

CHS 61-269

Iterative Development



i. Hypothesize Conceptual Framework

- Outline hypothesized concepts and potential claims
- Determine intended population
- Determine intended application/characteristics (type of scores, mode and frequency of administration)
- Perform literature/expert review
- Develop hypothesized conceptual framework
- Place PROs within preliminary endpoint model
- Document preliminary instrument development

ii. Adjust Conceptual Framework and Draft Instrument

- Obtain patient input
- Generate new items
- Select recall period, response options and format
- Select mode/method of administration/data collection
- Conduct patient cognitive interviewing
- Pilot test draft instrument
- Document content validity

iii. Confirm Conceptual Framework and Assess Other Measurement Properties

- Confirm conceptual framework with scoring rule
- Assess score reliability, construct validity, and ability to detect change
- Finalize instrument content, formats, scoring, procedures and training materials
- Document measurement development

iv. Collect, Analyze, and Interpret Data

- Prepare protocol and statistical analysis plan (final endpoint model and responder definition)
- Collect and analyze data
- Evaluate treatment response using cumulative distribution and responder definition
- Document interpretation of treatment benefi in relation to claim

v. Modify Instrument

- Change wording of items, populations, response options, recall period, or mode/method of administration/data collection
- Translate and culturally adapt to other languages
- Evaluate modifications as appropriate
- Document all changes

http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM205269.pdf

Physical Functioning

- Ability to conduct a variety of activities ranging from self-care to running

- Predictor of
  - Hospitalizations, institutionalization, and mortality

- Six physical functioning items included in 2010 Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Medicare Survey

Because of a health or physical problem are you unable to do or have any difficulty doing the following activities?

- Walking?
- Getting in or out of chairs?
- Bathing?
- Dressing?
- Using the toilet?
- Eating?

  - I am unable to do this activity (0)
  - Yes, I have difficulty (1)
  - No, I do not have difficulty (2)

Medicare beneficiary sample (n = 366,701)

- 58% female
- 57% high school education or less
- 14% 18-64; 48% 65-74, 29% 75-84, 9% 85+

© Ron Leishman * www.ClipartOf.com/439705

% of Medicare beneficiaries (n = 366,701) selecting each response option

Item	Unable to do	Have difficulty	No difficulty
Walking	4	27	69
Chairs	3	19	78
Bathing	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating	3	3	94

% of Medicare beneficiaries (n = 366,701) selecting each response option

Item	Unable to do	Have difficulty	No difficulty
Walking	4	27	69
Chairs	3	19	78
Bathing r = 84	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating r = .51	3	3	94

Item-Scale Correlations

Item	Item-Scale Correlations
Walking (0, 1, 2)	0.71
Chairs (0, 1, 2)	0.80
Bathing (0, 1, 2)	0.83
Dressing (0, 1, 2)	0.86
Toileting (0, 1, 2)	0.84
Eating (0, 1, 2)	0.75

Possible 6-item scale range: 0-12 (2% floor, 65% ceiling)

Reliability

Degree to which the same score is obtained when the *target* or thing being measured (person, plant or whatever) has not changed.

- ✓Internal consistency (items)
  - ✓Need 2 or more items
- ✓Test-retest (administrations) correlations
  - ✓Need 2 or more time points

Reliability

Model	Reliability	Intraclass Correlation
Two-way random	$\frac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS} + k(MS_{JMS} - MS_{EMS})/N}$
Two-way mixed	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS}}$
One-way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (k-1)MS_{WMS}}$

BMS = Between Ratee Mean Square
WMS = Within Mean Square
JMS = Item or Rater Mean Square
EMS = Ratee x Item (Rater) Mean Square

N = n of ratees
k = n of items or raters

Alpha

Reliability Formulas

Model	Reliability	Intraclass Correlation
Two-way random	$\frac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS} + k(MS_{JMS} - MS_{EMS})/N}$
Two-way mixed	$\boxed{\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS}}$
One-way	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (k-1)MS_{WMS}}$

BMS = Between Ratee Mean Square N = n of ratees
WMS = Within Mean Square k = n of items or raters
JMS = Item or Rater Mean Square
EMS = Ratee x Item (Rater) Mean Square

Internal Consistency Reliability (Coefficient Alpha)

- Coefficient alpha = 0.92

  $(MS_{bms} – MS_{ems})/MS_{bms}$

- Ordinal alpha = 0.98

http://support.sas.com/resources/papers/proceedings14/2042-2014.pdf

http://gim.med.ucla.edu/FacultyPages/Hays/utils/

Item-scale correlation matrix ("Multi-trait Scaling")

	Mobility	Basic
Walk	0.80*	0.20
Chairs	0.80*	0.20
Toilet	0.80*	0.20
Bathing	0.20	0.80*
Dress	0.20	0.80*
Eating	0.20	0.80*

*Item-scale correlation, corrected for overlap.

Item-scale correlation matrix (“Multi-trait Scaling”)

	Mobility	Basic
Walk	0.80*	0.80
Chairs	0.80*	0.80
Toilet	0.80*	0.80
Bathing	0.80	0.80*
Dress	0.80	0.80*
Eating	0.80	0.80*

*Item-scale correlation, corrected for overlap.

Item-scale correlation matrix (“Multi-trait Scaling“)

	Mobility	Basic
Walk	0.74*	0.66
Chairs	0.81*	0.74
Toilet	0.69*	0.85
Bathing	0.78	0.82*
Dress	0.79	0.87*
Eating	0.70	0.74*

*Item-scale correlation, corrected for overlap.

Item Response Theory (IRT)

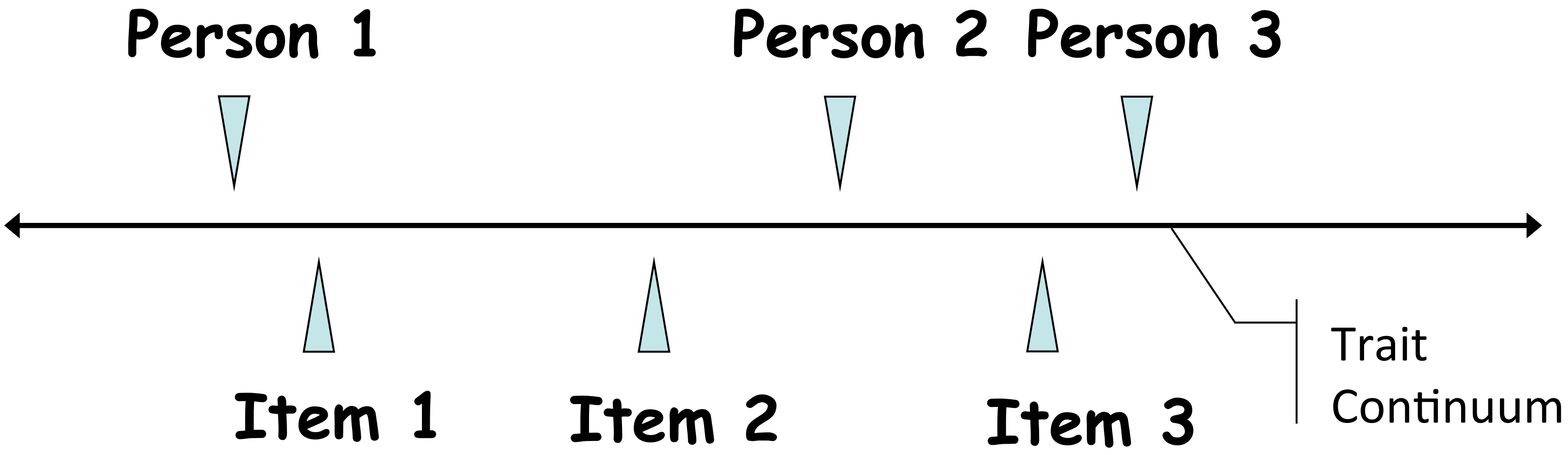
IRT models the relationship between a person's response Y_i to the question (i) and his or her level of the latent construct (θ) being measured by positing

$$\Pr(Y_i \geq k) = \frac{1}{1 + \exp(-a_i\theta + b_{ik})}$$

b_{ik} estimates how difficult it is to have a score of k or more on item (i).

a_i estimates the discriminatory power of the item.

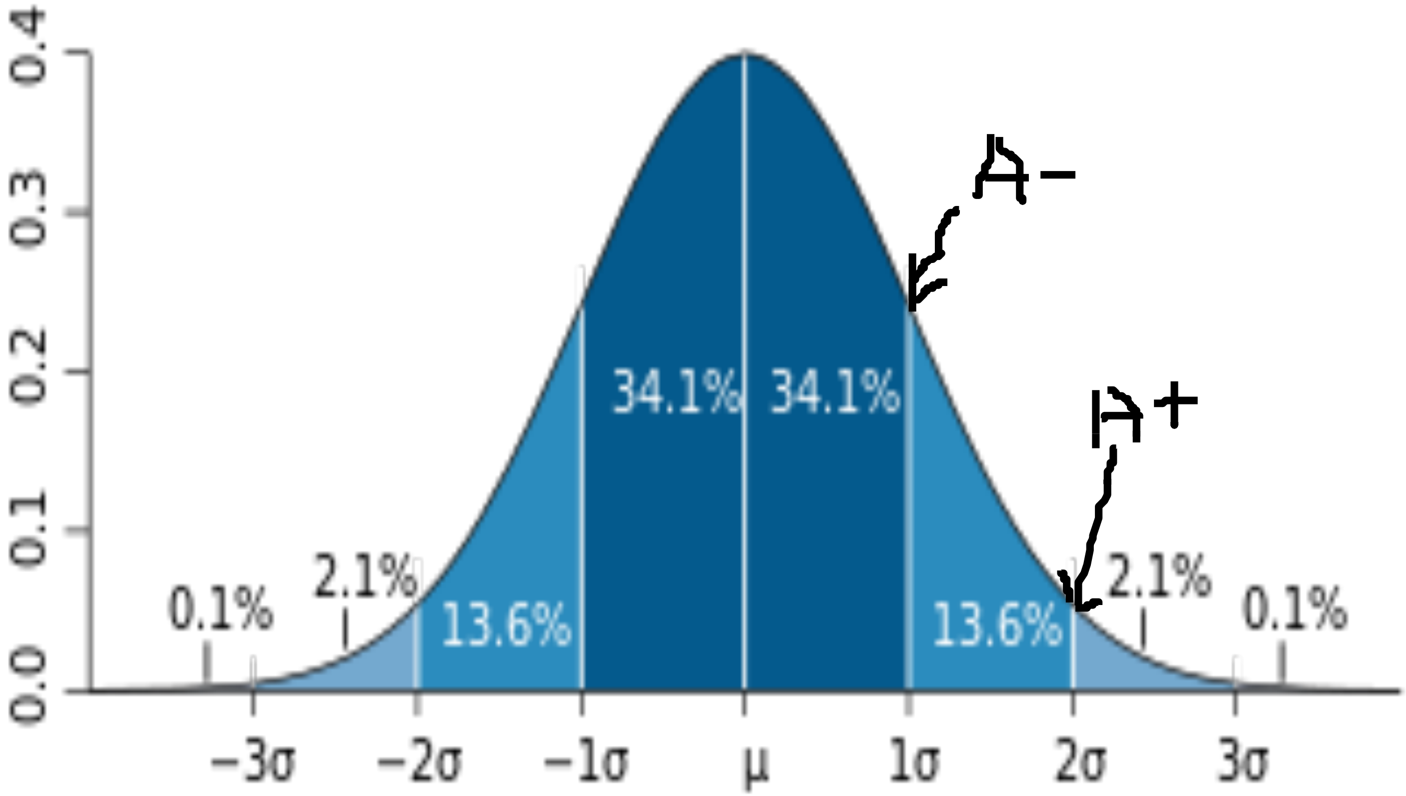
Item Responses and Trait Levels



www.nihpromis.org

Normal Curve (bell-shaped)

z = -1 to 1 (68.2%); z = -2 to 2 (95.4%); z = -3 to 3 (99.6%)



% of Medicare beneficiaries (n = 366,701) selecting each response option

Item	Unable to do 0	Have difficulty 1	No difficulty 2
Walking	4	27	69
Chairs	3	19	78
Bathing	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating	3	3	94

First Threshold

Item	Unable to do	Have difficulty	No difficulty
Walking	4	27	69
Chairs	3	19	78
Bathing	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating	3	3	94

Second Threshold

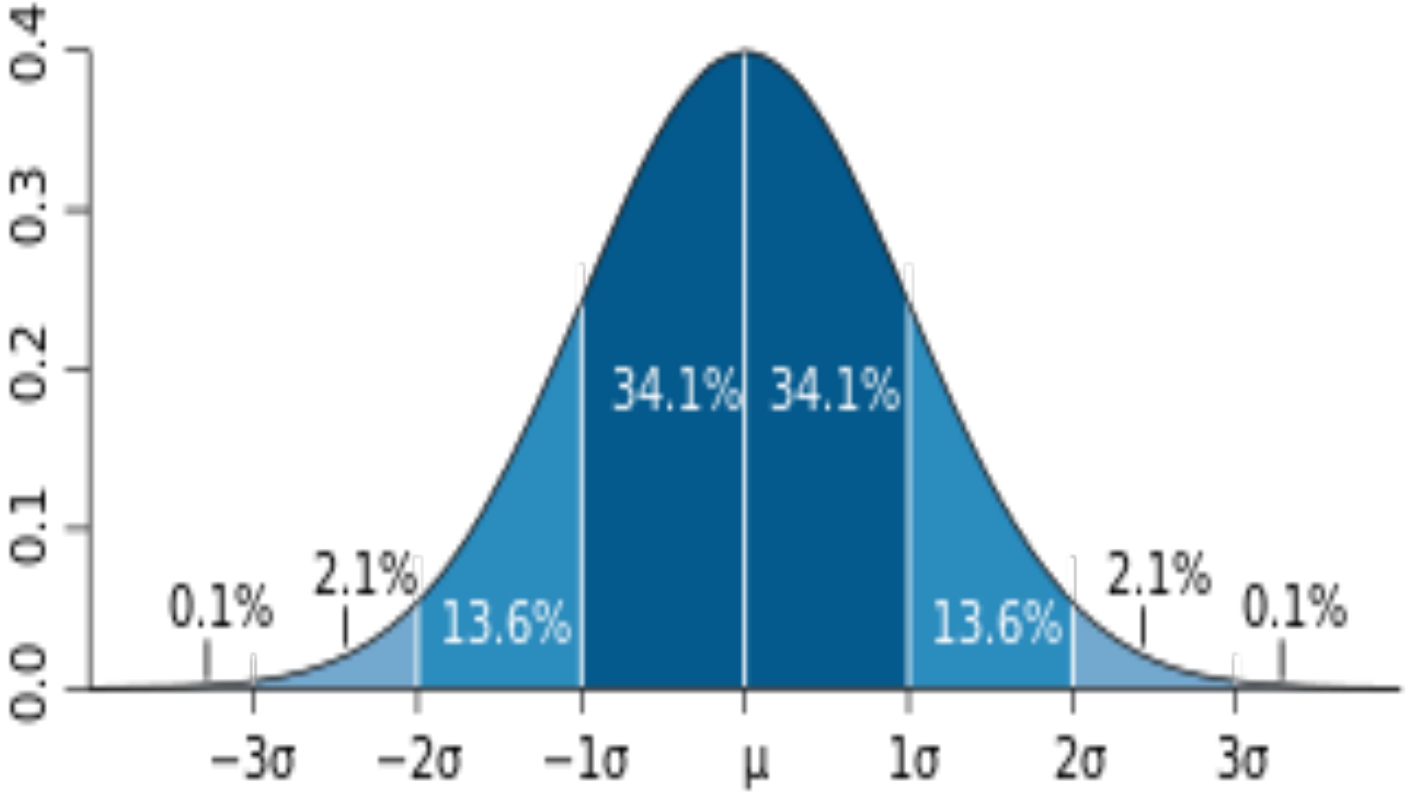
Item	Unable to do	Have difficulty	No difficulty
Walking	4	27	69
Chairs	3	19	78
Bathing	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating	3	3	94

Threshold #1 Parameter (Graded Response Model)

Physical Functioning	1st Threshold *Unable to do*
Walking	-1.86
Chairs	-1.91
Bathing	-1.72
Dressing	-1.78
Toileting	-1.87
Eating	-1.98

Normal Curve (bell-shaped)

z = -1 to 1 (68.2%); z = -2 to 2 (95.4%); z = -3 to 3 (99.6%)

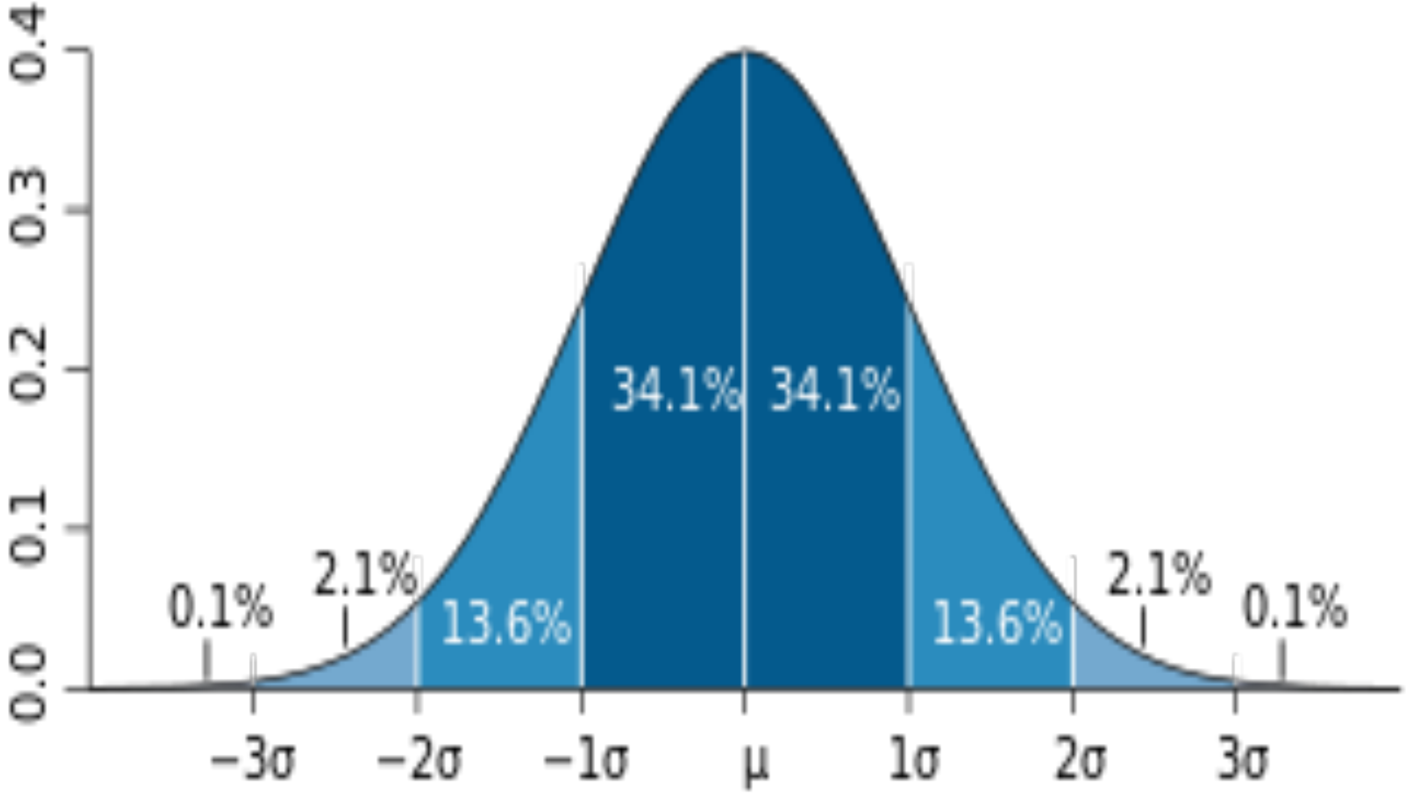


Threshold #2 Parameter (Graded Response Model)

Physical Functioning		**2nd Threshold** ***Unable to do or have difficulty***
Walking		-0.55
Chairs		-0.81
Bathing		-1.02
Dressing		-1.10
Toileting		-1.27
Eating		-1.53

Normal Curve (bell-shaped)

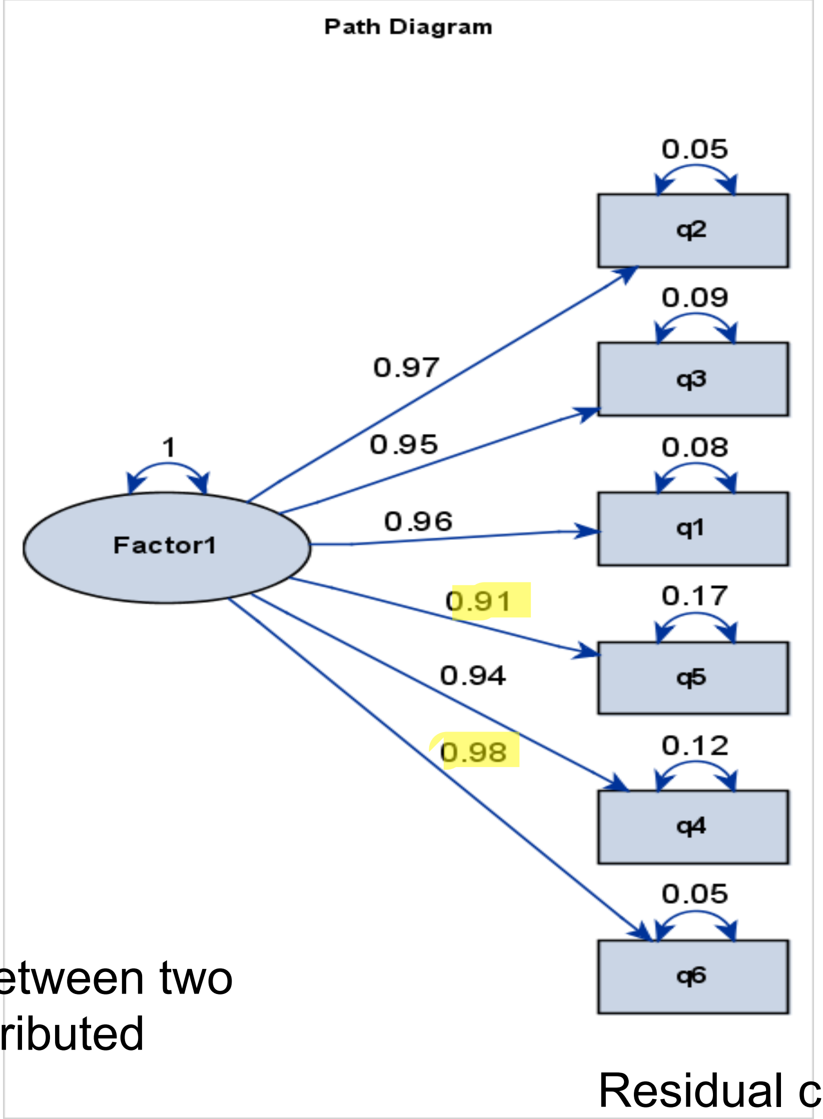
z = -1 to 1 (68.2%); z = -2 to 2 (95.4%); z = -3 to 3 (99.6%)



Item Parameters (Graded Response Model)

Physical Functioning	1st Threshold *Unable to do*	2nd Threshold *Have difficulty*	Slope (Discrimination)
Walking	-1.86	-0.55	4.63
Chairs	-1.91	-0.81	5.65
Bathing	-1.72	-1.02	6.34
Dressing	-1.78	-1.10	8.23
Toileting	-1.87	-1.27	7.23
Eating	-1.98	-1.53	4.87

Confirmatory Factor Analysis (Polychoric* Correlations)



Dressing

Eating

Bathing

Walking

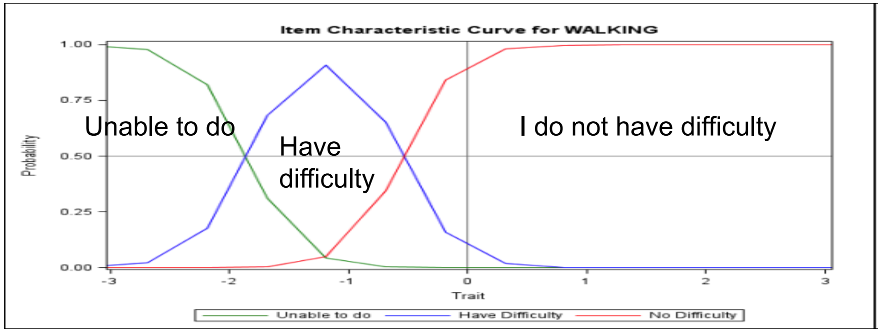
Chairs

Toileting

Residual correlations <= 0.04

* Estimated correlation between two underlying normally distributed continuous variables

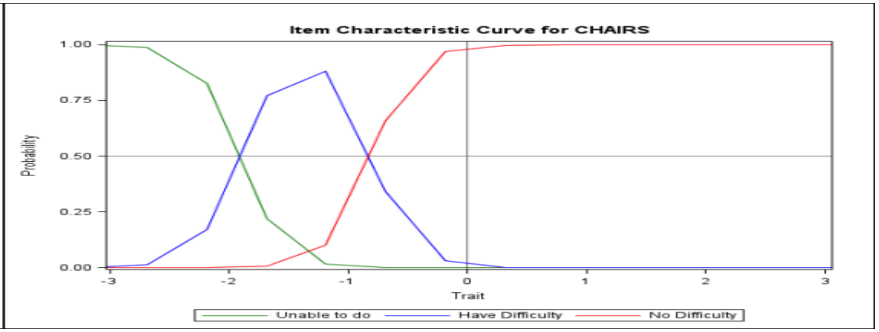
Item Characteristic Curves

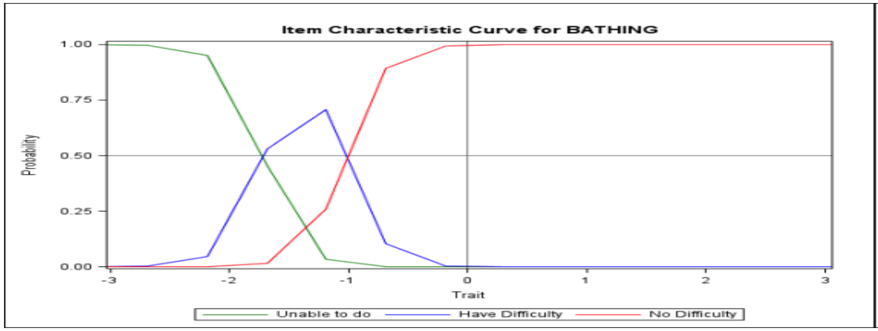


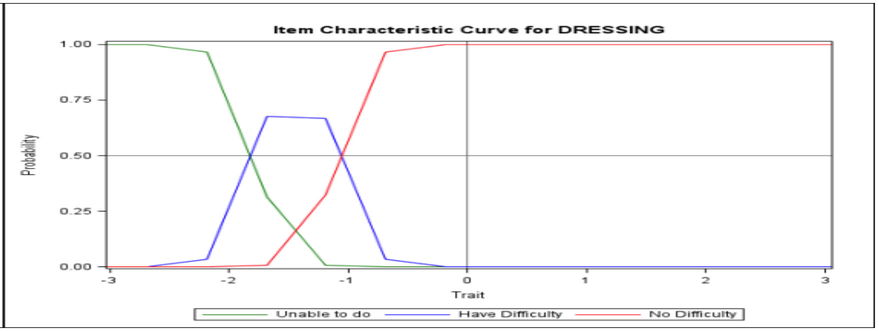
Unable to do

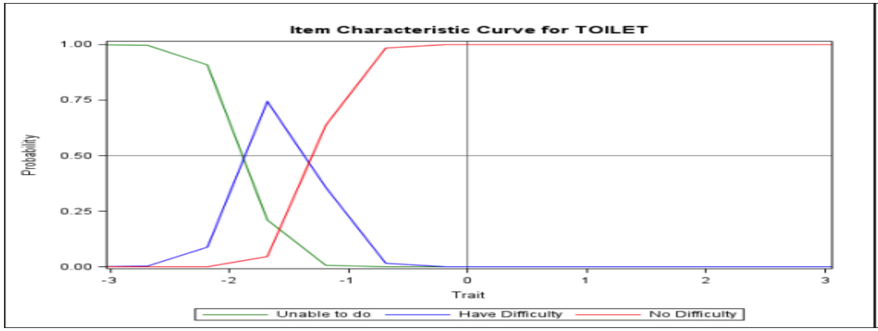
Have difficulty

I do not have difficulty









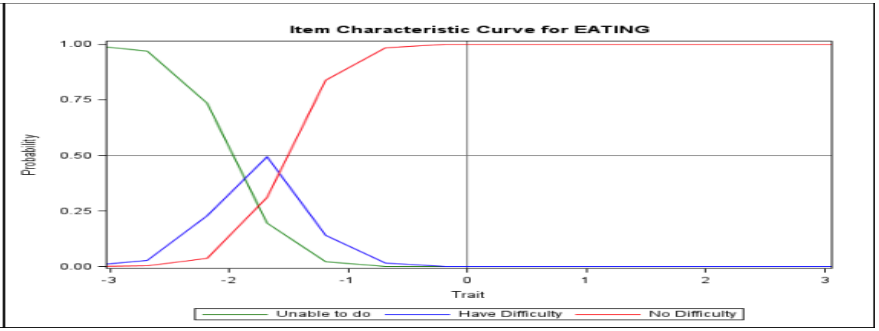
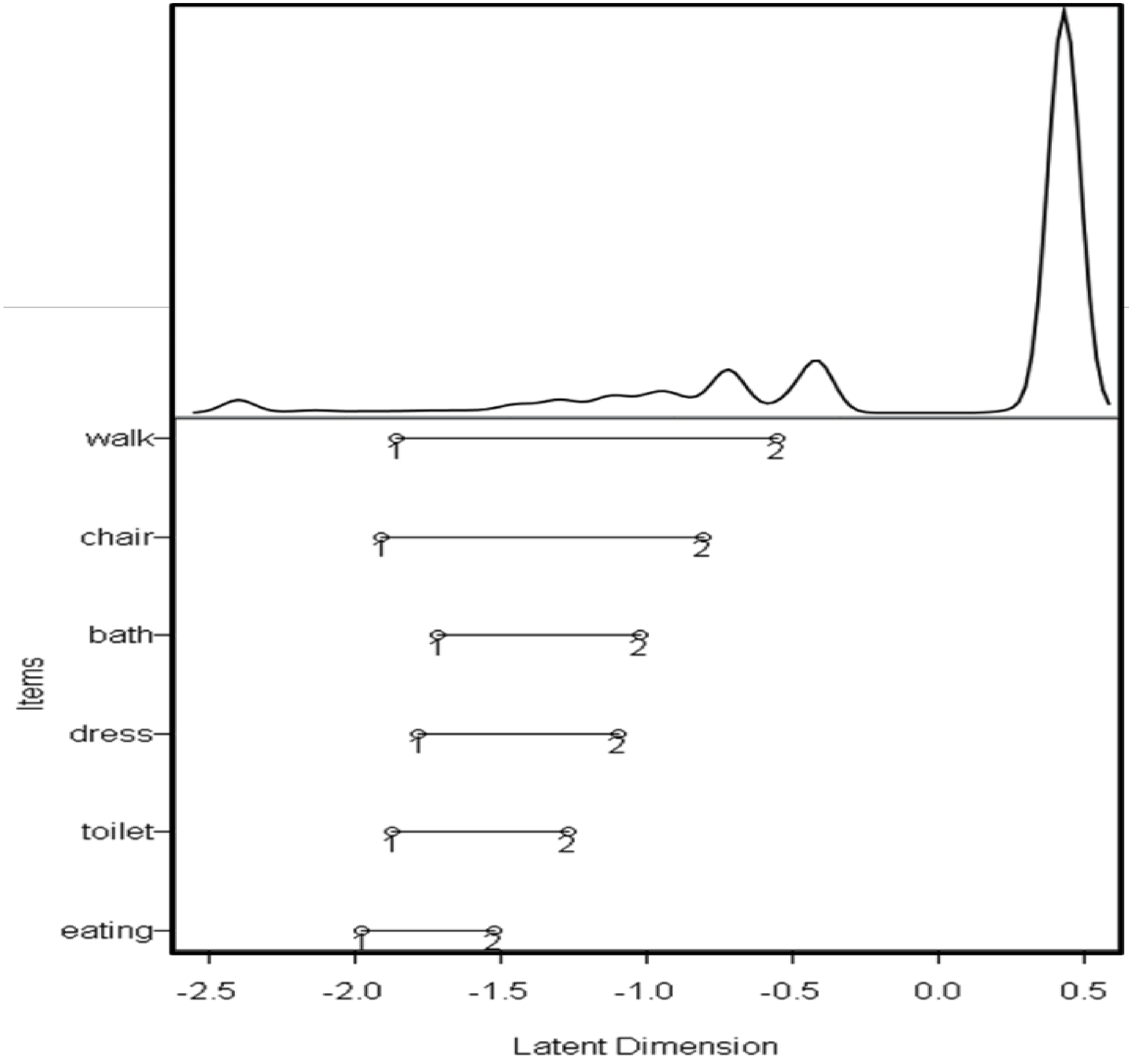
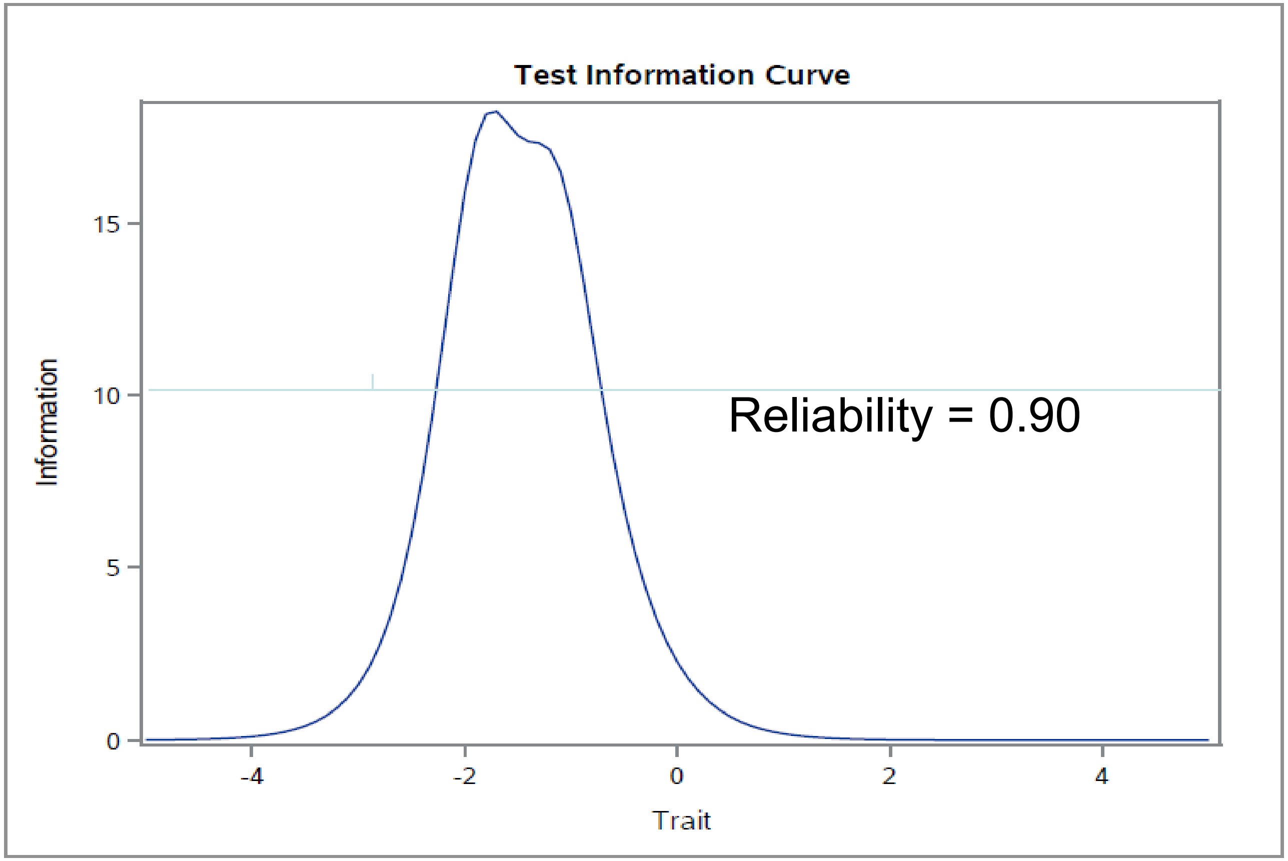


Figure 2. Person-Item Map



Reliability = (Info - 1) / Info

The IRT Procedure



Validity

- Content validity: Does measure “appear” to reflect what it is intended to (expert judges or patient judgments)?
  - Do items operationalize concept?
  - Do items cover all aspects of concept?
  - Does scale name represent item content?
- Construct validity
  - Are the associations of the measure with other variables consistent with hypotheses?

Physical Function Scale Correlations

r = 0.39 (self-rated general health)
r = -0.23 (number of chronic conditions)

Cohen's rule of thumb for correlations that correspond to effect size rules of 0.20 SD, 0.50 SD and 0.80 SD are as follows:

0.100 is small correlation
0.243 is medium correlation
0.371 is large correlation

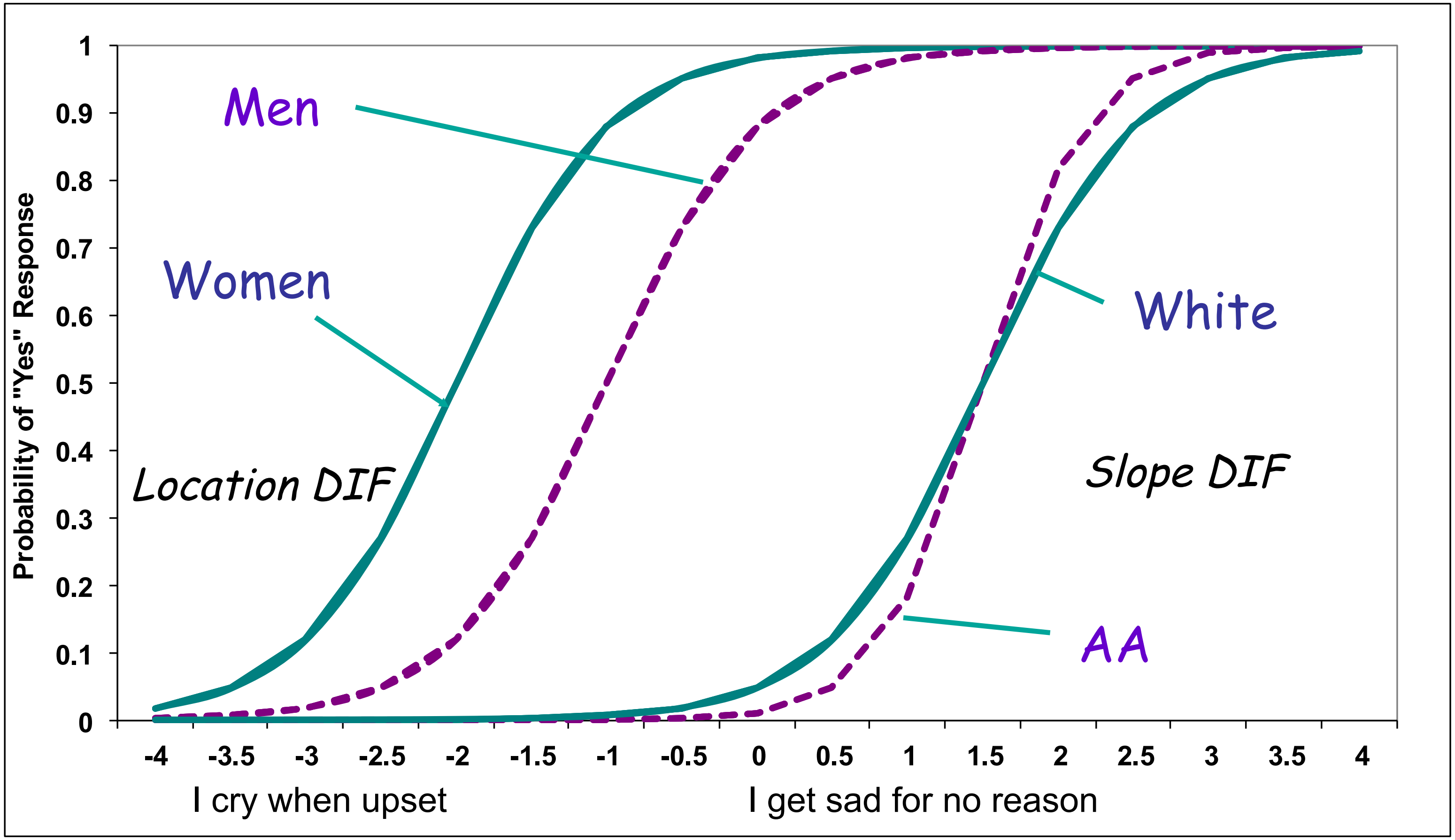
(r's of 0.10, 0.30 and 0.50 are often cited as small, medium and large, respectively).

Differential Item Functioning (DIF)

- Probability of choosing each response category should be the same for those who have the same estimated scale score, regardless of other characteristics

- Evaluation of DIF by subgroups

DIF (2-parameter model)



Higher Score = More Depressive Symptoms

Computer Adaptive Testing (CAT)

ARMY

Graduate Record Examinations®

NASD®

National Council
of State Boards of Nursing, Inc.

PROMIS®
Dynamic Tools to Measure Health Outcomes from the Patient Perspective

Reliability Target for Use of Measures with Individuals

- Reliability ranges from 0-1
  - 0.90 or above is goal

- SE = SD (1- reliability)$^{1/2}$

- Reliability = 1 – (SE/10)^{2}
  - Reliability = 0.90 when SE = 3.2
  - 95% CI = true score +/- 1.96 x SE

In the past 7 days …

I was grouchy [1st question]

- Never	[39]
- Rarely	[48]
- Sometimes	[56]
- Often	[64]
- Always	[72]

Estimated Anger = 56.1

SE = 5.7 (rel. = 0.68)

In the past 7 days …

I felt like I was ready to explode

[2nd question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 51.9

SE = 4.8 (rel. = 0.77)

In the past 7 days …

I felt angry [3rd question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 50.5

SE = 3.9 (rel. = 0.85)

In the past 7 days …

I felt angrier than I thought I should [4th question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 48.8

SE = 3.6 (rel. = 0.87)

In the past 7 days …

I felt annoyed [5th question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 50.1

SE = 3.2 (rel. = 0.90)

In the past 7 days …

I made myself angry about something just by thinking about it. [6th question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 50.2

SE = 2.8 (rel = 0.92) (95% CI: 44.7-55.7)

Recommended Reading

- Cappelleri, J. C., Lundy, J.J., & Hays, R. D. (2014). Overview of classical test theory and item response theory for quantitative assessment of items in developing patient-reported outcome measures. <u>Clinical Therapeutics</u>, <u>36</u> (5), 648-662

Thank You!

drhays@ucla.edu (310-794-2294). http://gim.med.ucla.edu/FacultyPages/Hays/