Assessing Psychometric Properties of Multi-Item Scales

Ron D. Hays, Ph.D.
UCLA GIM & HSR

February 17, 2017 (10:00-11:00 am)

https://global.gotomeeting.com/join/883684349

Physical Functioning

- Able to do a range of activities from basic (e.g., self-care) to advanced (e.g., running)

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

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+

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?

Listed from most to least difficult

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

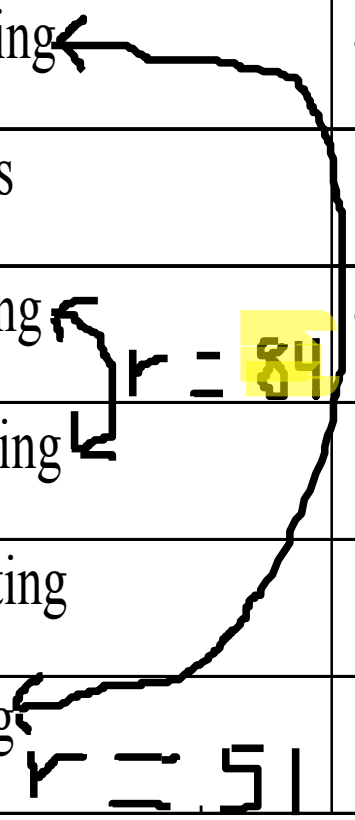
Higher Score is Better

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

Item (difficulty or unable to do)	Unable to do	Have difficulty	No difficulty
Walking **(1/3)**	4	27	69
Chairs **(1/5)**	3	19	78
Bathing **(1/7)**	4	11	85
Dressing **(1/8)**	3	9	88
Toileting **(1/11)**	3	6	91
Eating **(1/16)**	3	3	94

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

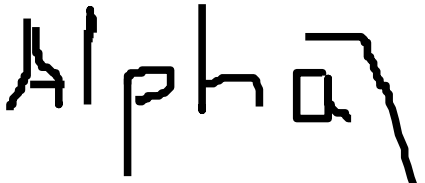
Item	Unable to do	Have difficulty	No difficulty
Walking	4	27	69
Chairs	3	19	78
Bathing	4	11	85
Dressing	3	9	88
Toileting	3	6	91
Eating	3	3	94



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

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



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	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (k-1)MS_{EMS}}$
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

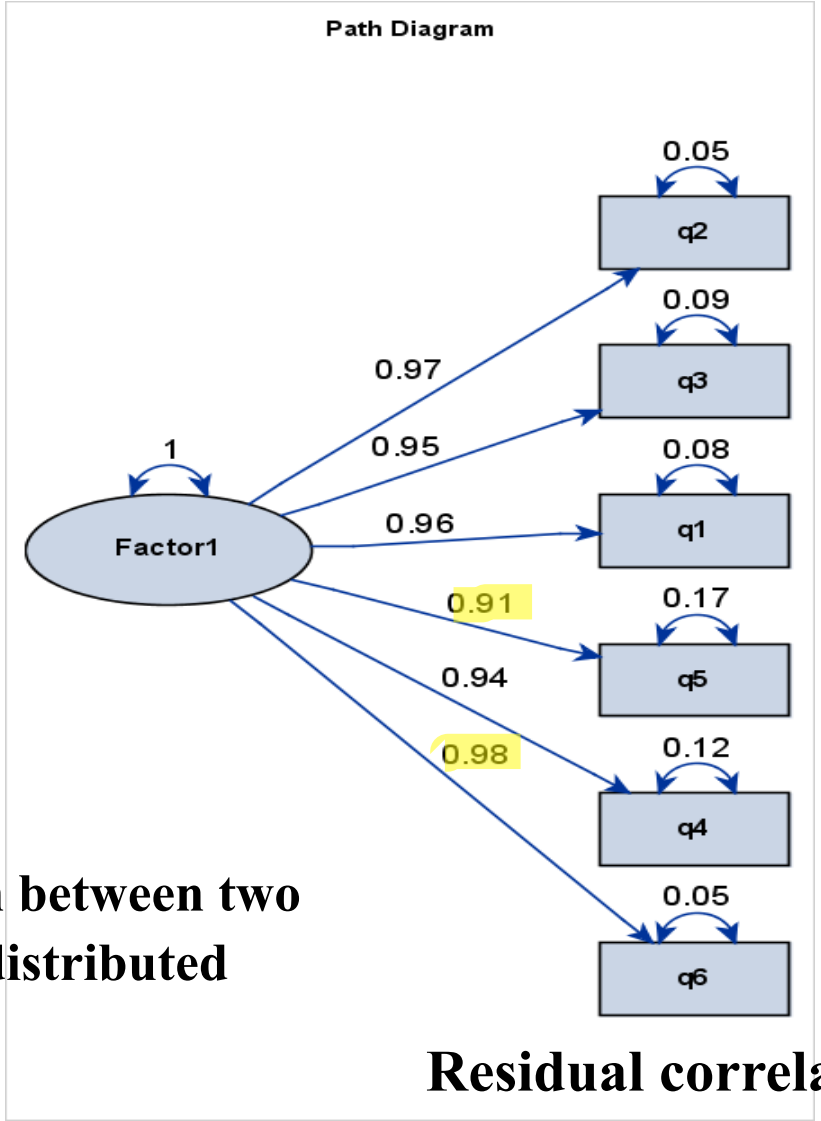
  $(MS_{bms} - MS_{ems})/MS_{bms}$

- Ordinal alpha = 0.98

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

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

Confirmatory Factor Analysis (Polychoric* Correlations)



Dressing

Eating

Bathing

Walking

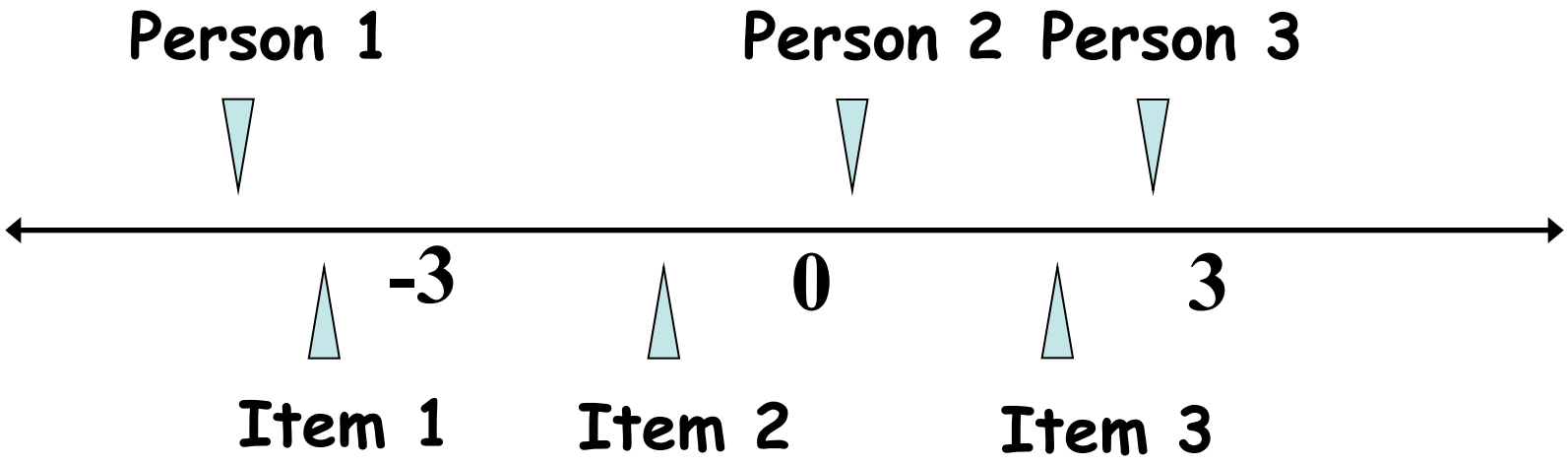
Chairs

Toileting

***Estimated correlation between two underlying normally distributed continuous variables.**

Residual correlations <= 0.04

People and Items on Same *z-score* metric



Item difficulty (p = 0.84 and 0.16)

Proportion of people endorsing the item (p) can be expressed in z distribution form:

$z = \ln (1-p)/p)/1.7 = (\ln (1-p) - \ln (p))/1.7$

$= (\ln (.16) - \ln (.84))/1.7$

$= (-1.83 + .17)/1.7$

$= -1.66/1.7$

$= \underline{-1.00}$

$z = \ln (0.84)/0.16)/1.7$

$= \underline{1.00}$

(-2 -> 2 is typical range)

Item Response Theory (IRT)

IRT graded response model estimates relationship between a person's response Y_i to the question (i) and his or her level on the latent construct (θ):

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

$\mathbf{e}^{\mathbf{a}(\theta\text{-}b)}/(1 + \mathbf{e}^{\mathbf{a}(\theta\text{-}b)})$

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

$\mathbf{a_i}$ = item discrimination.

Item Parameters (Graded Response Model)

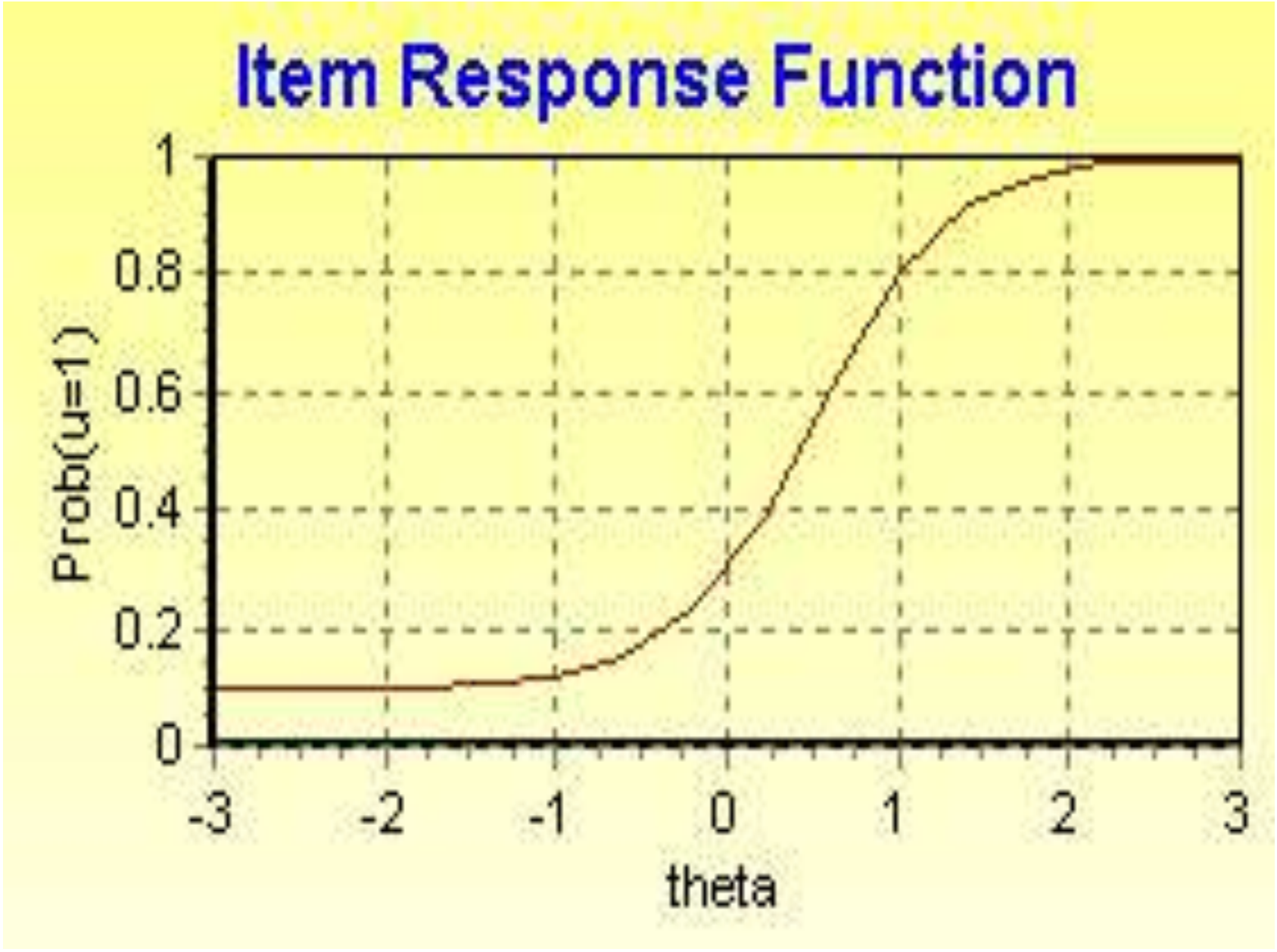
Physical Functioning	1st Threshold *Unable to do*	2nd Threshold *Have difficulty*	Slope (Discrimination)
Walking	-1.86	-0.55	4.63
Chairs	-1.91	-0.81	5.65
Bathing	-1.72	-1.02	6.34
Dressing	-1.78	-1.10	8.23
Toileting	-1.87	-1.27	7.23
Eating	-1.98	-1.53	4.87

Loadings and Item Parameters

	Loading	**Discrimination***	***Unable to do* vs *Have Difficulty***	***Have difficulty* vs *No Difficulty***
Walking	0.930 (6)	4.632 (6)	-1.861	-0.551
Chairs	0.950 (4)	5.652 (4)	-1.914	-0.806
Bathing	0.961 (3)	6.341 (3)	-1.719	-1.025
Dressing	0.977 (1)	8.228 (1)	-1.785	-1.101
Toileting	0.970 (2)	7.232 (2)	-1.872	-1.268
Eating	0.943 (5)	4.870 (5)	-1.983	-1.527

**Very low* (.01-.34), *low* (.35-.64), *moderate* (.65-1.34), *high* (1.35-1.69), and *very high* (> 1.70)

Baker, F. B. (2001). The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation


Item Response Function
Prob(u=1)
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
theta

R. M. Kaplan and D. P. Saccuzzo, Psychological Testing: Principles, Applications, and Issues (2nd Edition). Brooks/Cole Publishing Company1989 (page 152).

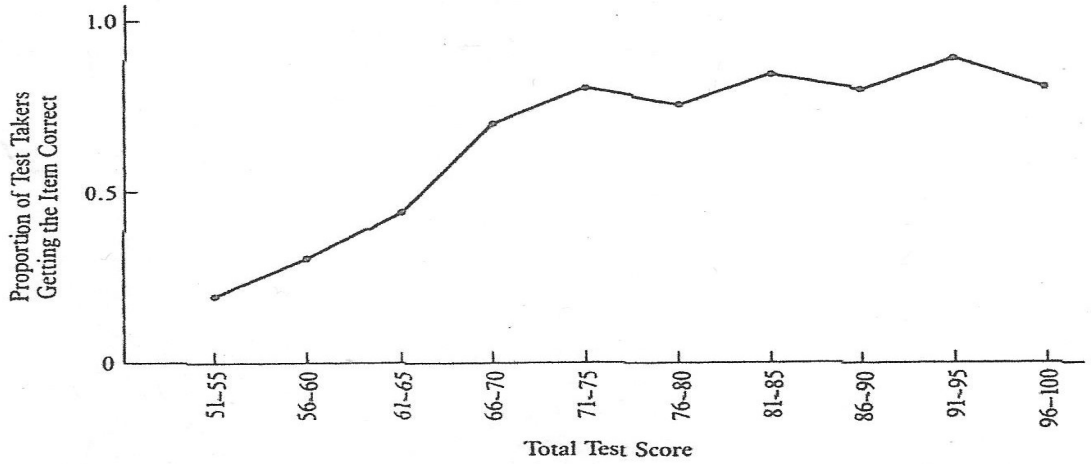
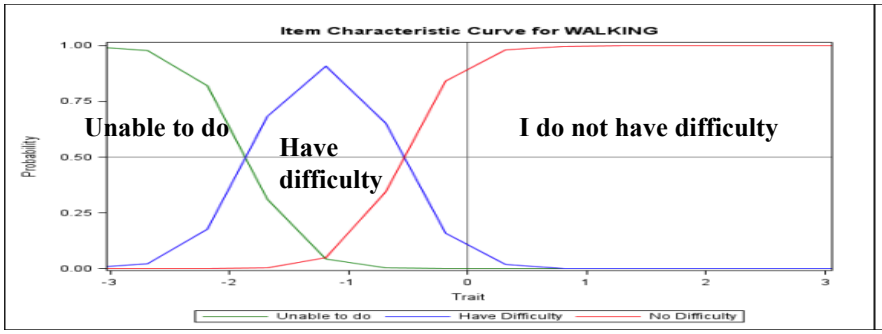
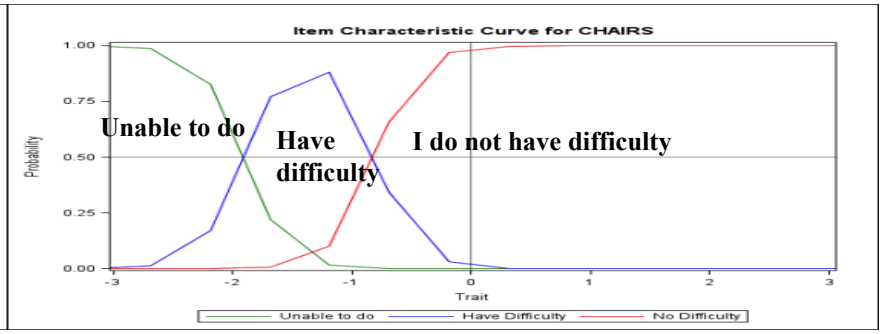
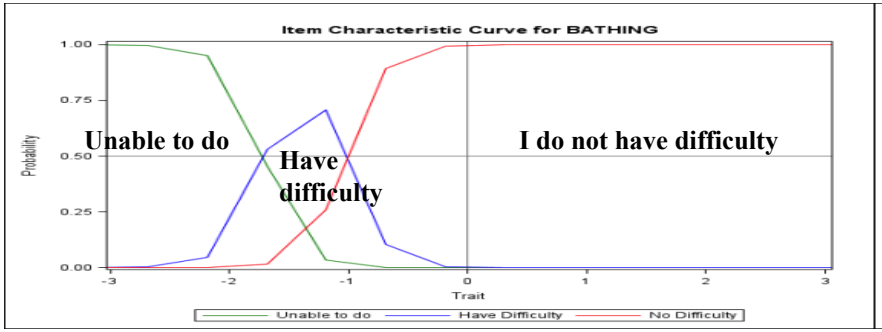


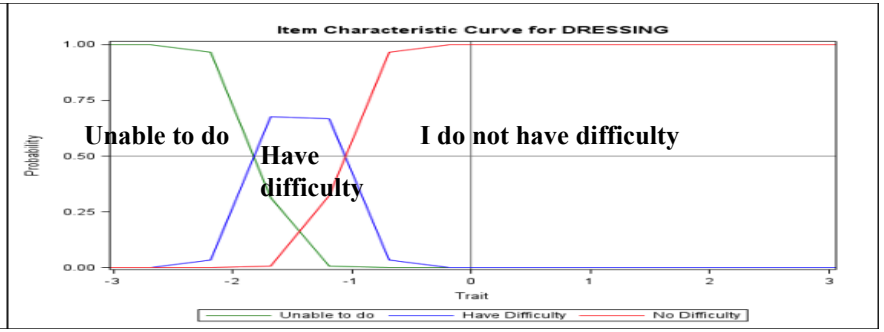
FIGURE 6-3 Item characteristic curve for a test item that discriminates well at low levels of performance but not at higher levels.

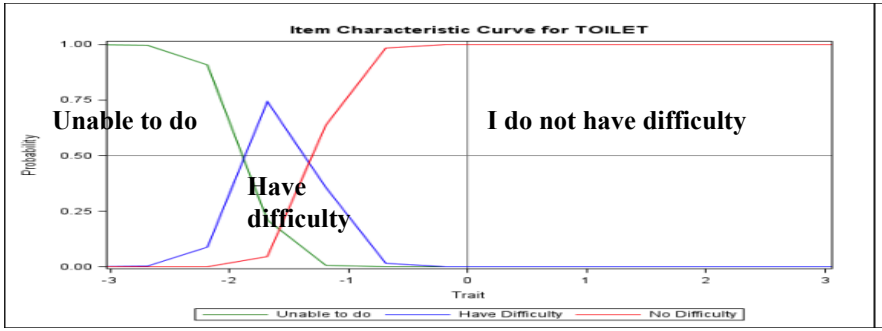
Item Characteristic Curves











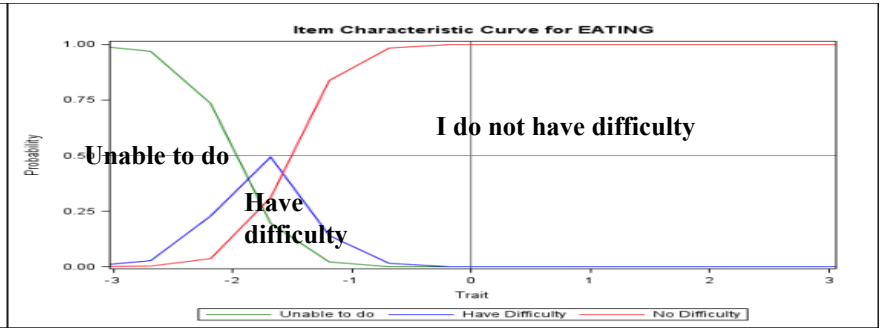
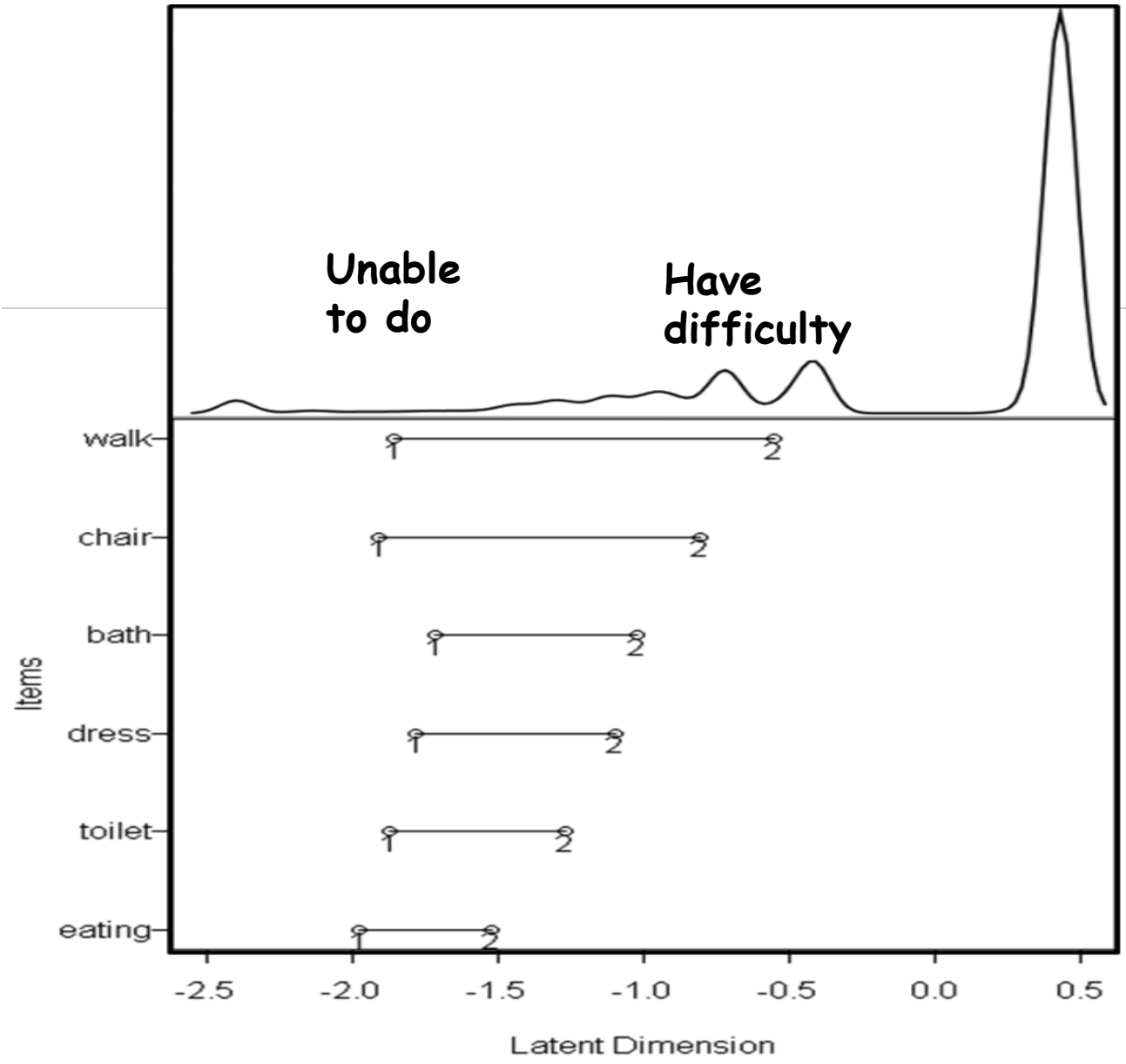


Figure 2. Person-Item Map

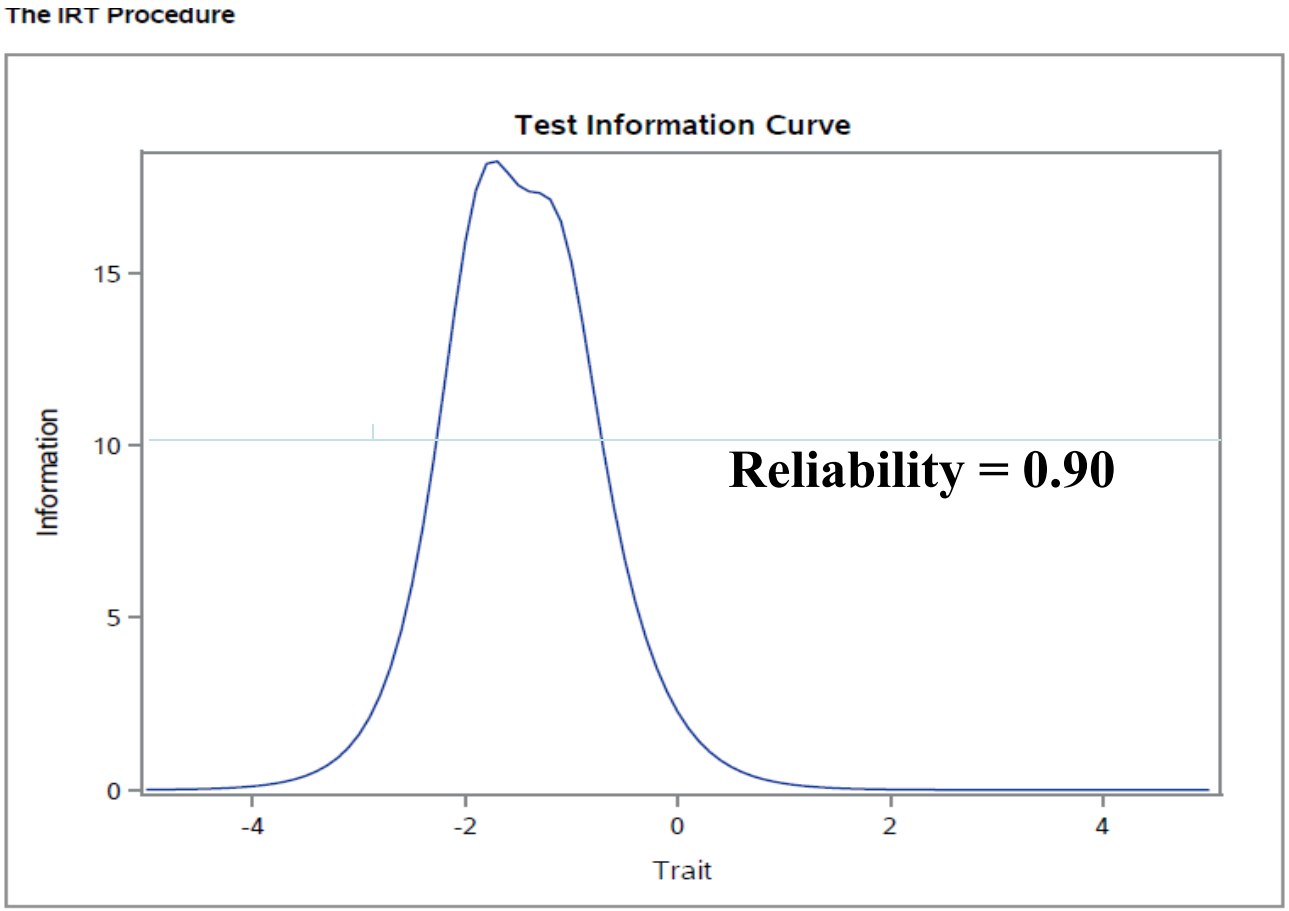


Simple-summated Scoring of Physical Functioning Scale

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

- Possible 6-item scale range: 0-12
  - Mean = 11 (2% floor, 65% ceiling)

Reliability = (Info – 1) / Info



Correlations with Other Variables

Physical Functioning	General Health	General Mental Health	Number of conditions
Simple-summated scoring	0.29	0.23	-0.16
Item response theory scoring	0.39	0.30	-0.23

Cohen's effect size rules of thumb (d = 0.2, 0.5, and 0.8): small = 0.100; medium = 0.243, and large = 0.371

$\underline{r} = \underline{d} / [(\underline{d}^2 + 4)^{.5}] = \underline{0.8} / [(0.8^2 + 4)^{.5}] = 0.8 / [(0.64 + 4)^{.5}] = 0.8 / [(4.64)^{.5}]$
$= 0.8 / 2.154 = \underline{0.371}$

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

	PhyF	RoleF
Walk	0.60*	0.20
Chairs	0.60*	0.20
Bathing	0.60*	0.20
School	0.20	0.60*
Job	0.20	0.60*
Family	0.20	0.60*

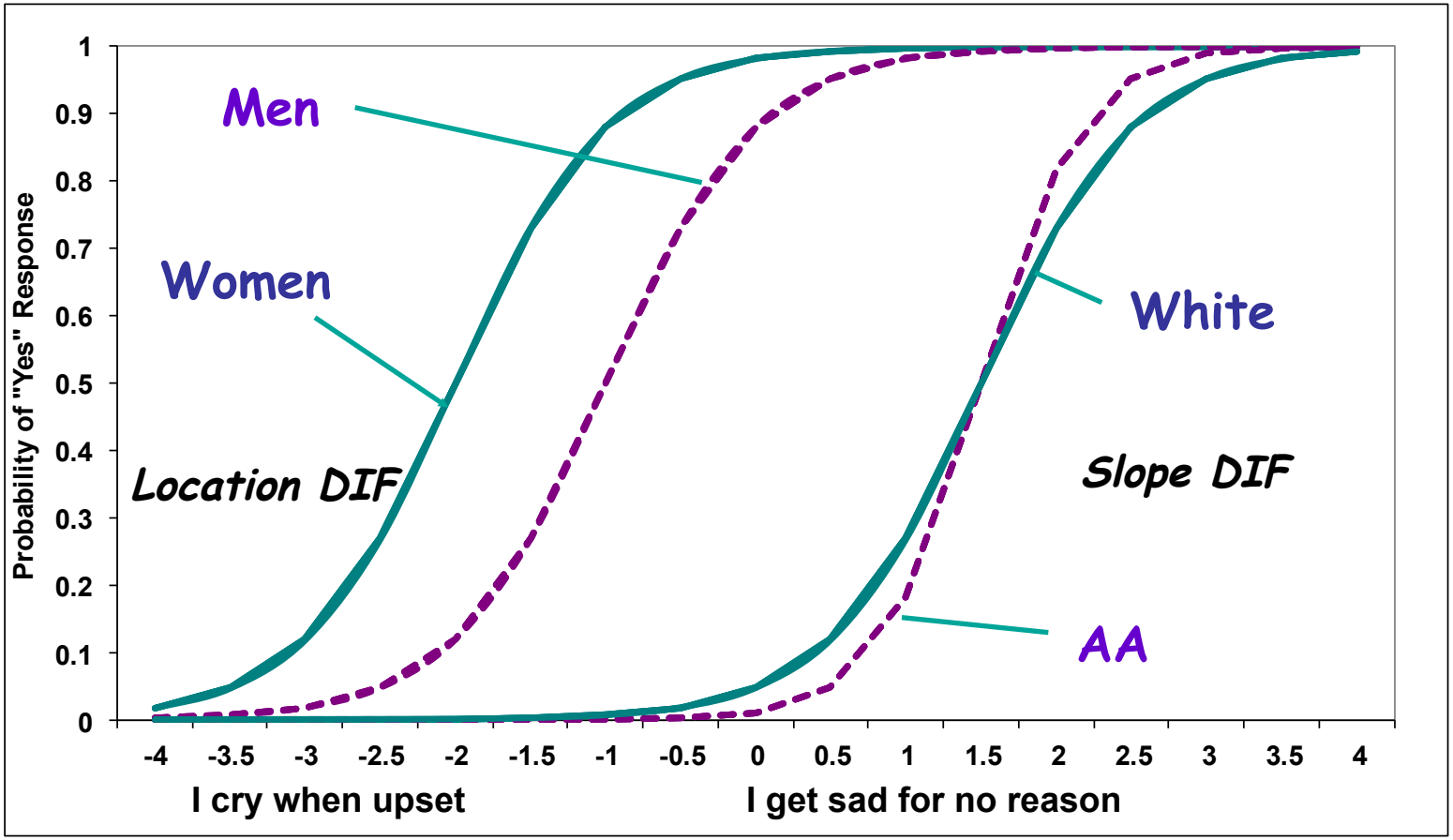
*Item-scale correlation, corrected for overlap.

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

	PhyF	RoleF
Walk	0.60*	0.60
Chairs	0.60*	0.60
Toilet	0.60*	0.60
School	0.60	0.60*
Job	0.60	0.60*
Family	0.60	0.60*

*Item-scale correlation, corrected for overlap.

DIF (2-parameter model)



Higher Score = More Depressive Symptoms

Person Fit

- Large negative Z_L values indicate misfit.

- One person in PROMIS project had $Z_L = -3.13$
- This person reported that they could do 13 physical functioning activities (including running 5 miles) without any difficulty, but
  - This person reported a little difficulty being out of bed for most of the day.

Questions?

DIVERSITY PROGRAM CONSORTIUM

Supported by the National Institutes of Health


drhays@ucla.edu

Questions (1)

- Same item stem but different response scales across sites.
- Item stem different across sites
  - Indicate to what extent you are confident that you can complete the following tasks.
  - Rate the confidence you have that your students can do the following tasks.

Questions (2)

- The Freshman Survey (TFS) and College Senior Survey (CSS): www.heri.ucla.edu
  - Handling of missing items
  - Calibration for within person change
  - Calibration for change across cohorts
    - Change from freshman to senior year
      - Cohort 1 (2015 and 2019)
      - Cohort 2 (2016 and 2020)
      - Cohort 3 (2017 and 2021)

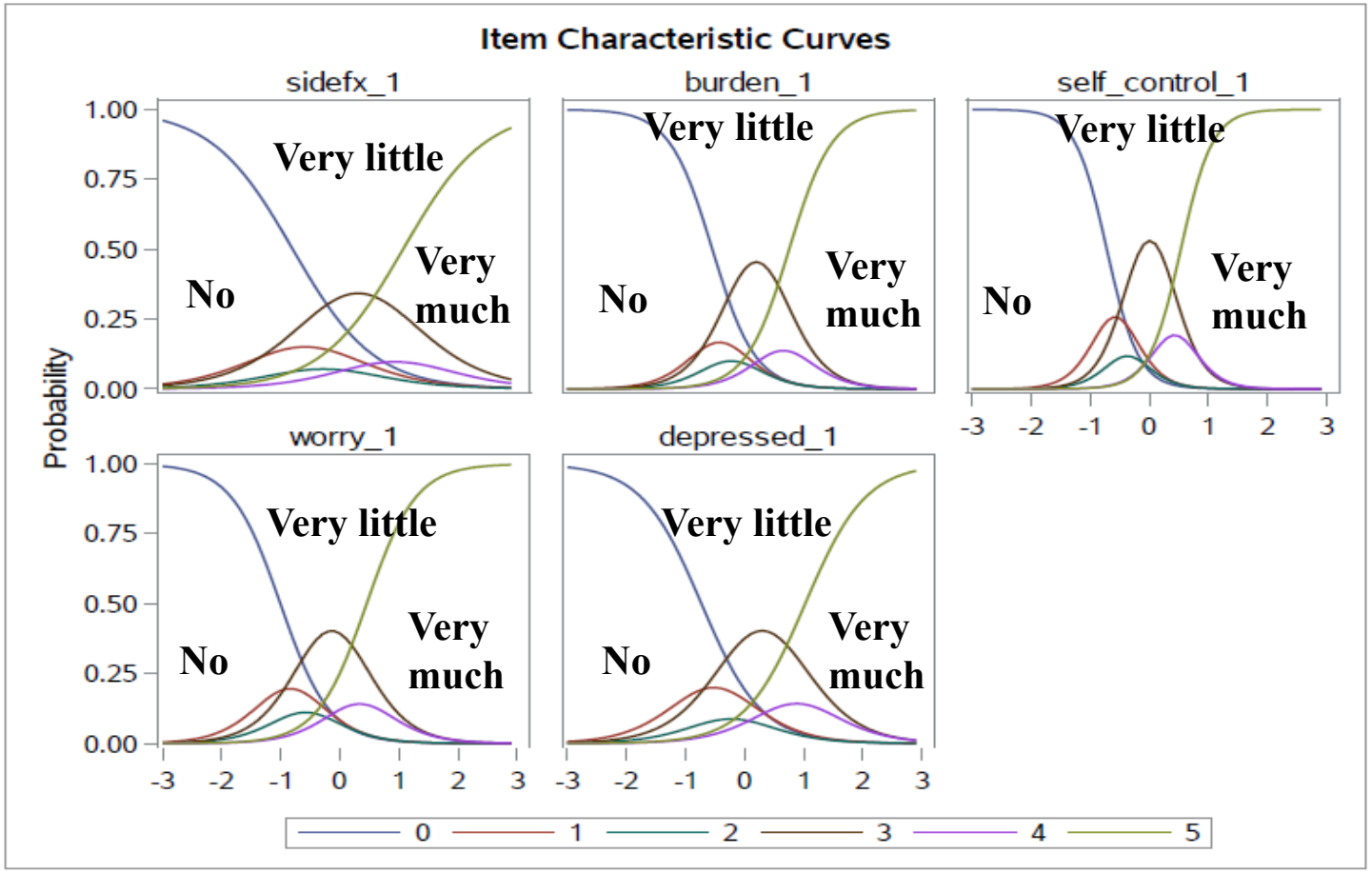
MINNESOTA LIVING WITH HEART FAILURE® QUESTIONNAIRE

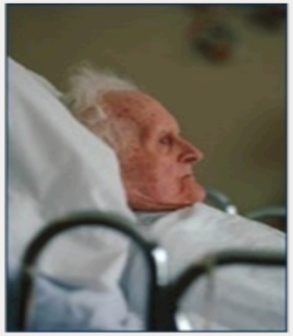
The following questions ask how much your heart failure (heart condition) affected your life during the past month (4 weeks). After each question, circle the 0, 1, 2, 3, 4 or 5 to show how much your life was affected. If a question does not apply to you, circle the 0 after that question.

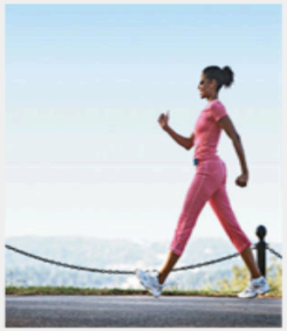
Did your heart failure prevent you from living as you wanted during the past month (4 weeks) by -	**No**			**Very Little**		**Very Much**
1. causing swelling in your ankles or legs?	0	1	2	3	4	5
2. making you sit or lie down to rest during the day?	0	1	2	3	4	5
3. making your walking about or climbing stairs difficult?	0	1	2	3	4	5
4. making your working around the house or yard difficult?	0	1	2	3	4	5
5. making your going places away from home difficult?	0	1	2	3	4	5
6. making your sleeping well at night difficult?	0	1	2	3	4	5
7. making your relating to or doing things with your friends or family difficult?	0	1	2	3	4	5
8. making your working to earn a living difficult?	0	1	2	3	4	5
9. making your recreational pastimes, sports or hobbies difficult?	0	1	2	3	4	5
10. making your sexual activities difficult?	0	1	2	3	4	5

Item Characteristic Curve for Emotional Health Scale

The IRT Procedure






-3 0 3

Physical Functioning Item Bank

Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Item *n*

Are you able to get in and out of bed?
Are you able to stand without losing your balance for 1 minute?
Are you able to walk from one room to another?
Are you able to walk a block on flat ground?
Are you able to run or jog for two miles?
Are you able to run five miles?

IRT Distortions

- "Parameter values are identical in separate subgroups or across different measurement conditions."

- It is the often misunderstood feature of parameter invariance that is frequently cited in introductory or advanced texts" (Rupp & Zumbo, 2006).

Interval-Level?

- "Modern day psychometric analyses such as Rasch analysis convert ordinal data to an interval scale so that response scores meet the criteria for measurement"

- Correlation (product-moment and ICC) between simple-summated scoring and IRT estimated score for physical functioning = 0.91

Ben Wright or Been Wrong?

- "Application of the Rasch model to the data set estimates a measure that can be considered valid."
- The "Rasch model is the only valid approach to measurement"
  - Bergan, 2013, *Rasch versus Birnbaum: New arguments in an old debate* (p. 3)

Computer Adaptive Testing (CAT)

2004

www.nihpromis.org

Reliability Target for Use of Measures with Individuals

- z-score (mean = 0, SD = 1)
- Reliability ranges from 0-1
  - 0.90 or above is goal
  - SE = SD $(1 - \text{reliability})^{1/2}$
  - Reliability = $1 - SE^2$
  - Reliability = 0.90 when SE = 0.32
- 95% CI = true score +/- 1.96 x SE

  (CI = -0.63 → 0.63 z-score when reliability = 0.90)

T-score Metric

- Mean = 50
- SD = 10
- Referenced to US General Pop.
- T = 50 + (z * 10)

www.nihpromis.org

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)

PROMIS Physical Functioning vs. "Legacy" Measures