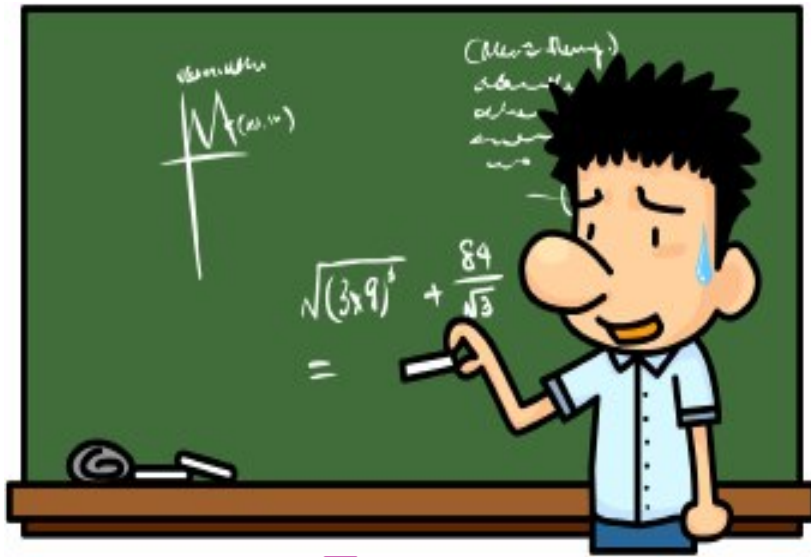# Evaluating the Psychometric Properties of Multi-Item Scales

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

October 13, 2017 (3:30-5 pm)
NCSP Instrument Design Workshop
http://gim.med.ucla.edu/FacultyPages/Hays/present/

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+

© Ron Leishman * www.ClipartOf.com/439705

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

r = .84

r = .51

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

Alpha Reliability Formulas

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

BMS = Between Ratee Mean Square   **N = n of ratees**

WMS = Within Mean Square   **k = n of items or raters**
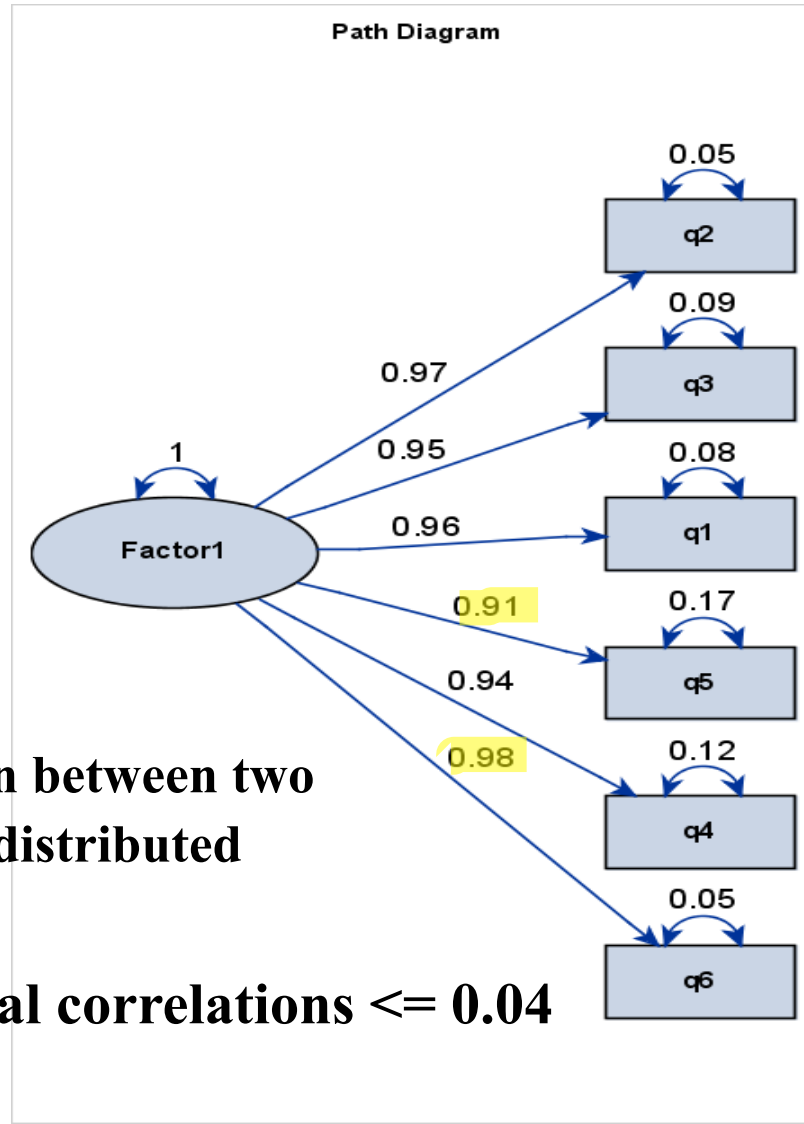
JMS = Item or Rater Mean Square

EMS = Ratee x Item (Rater) Mean Square

8

Internal Consistency Reliability (Coefficient Alpha)

- Coefficient alpha = <u>0.92</u>

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

- 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)

Path Diagram

Dressing

Eating

Bathing

Walking

Chairs

Toileting

*Estimated correlation between two underlying normally distributed continuous variables.
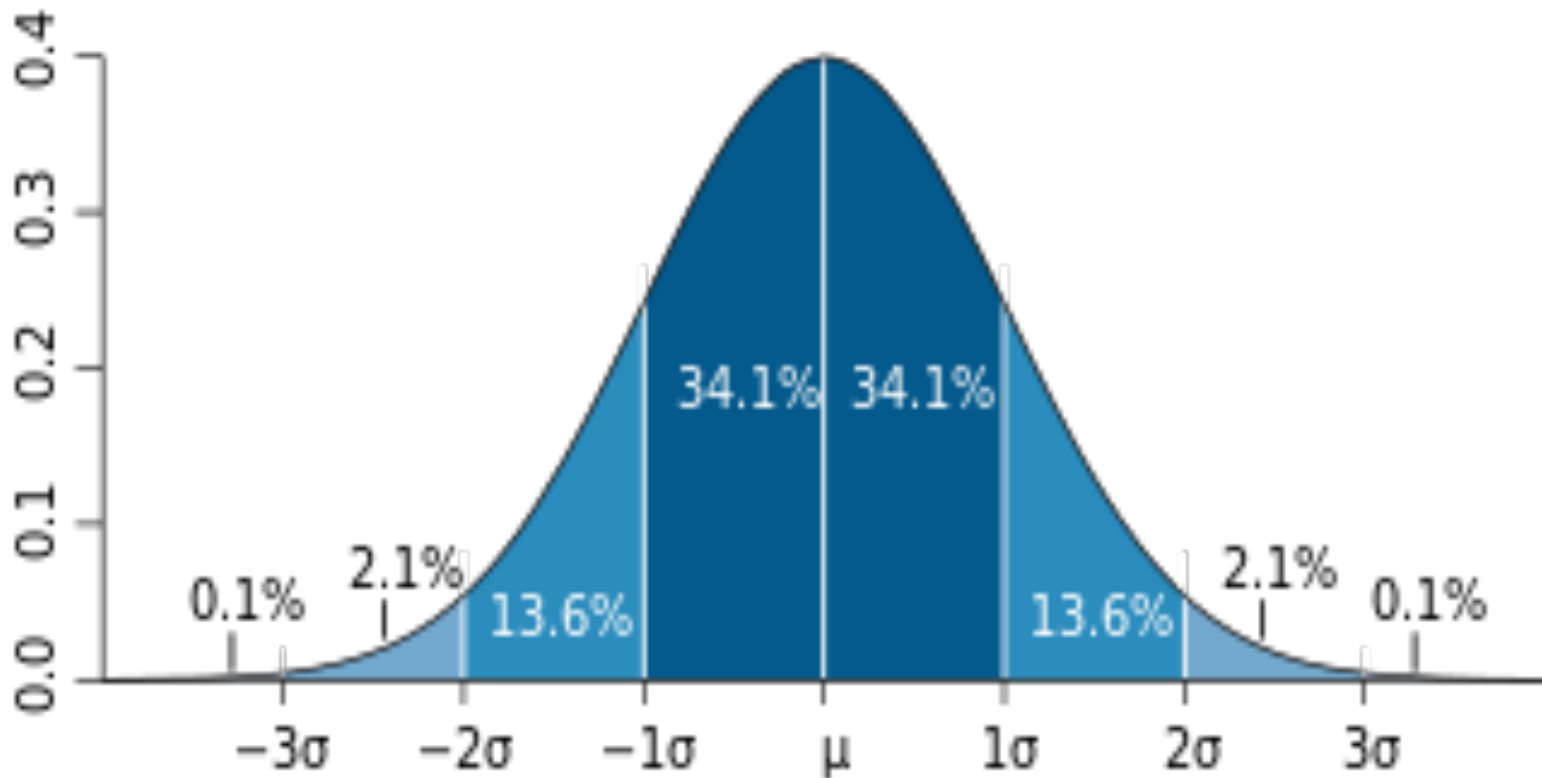
Residual correlations <= 0.04

CFA Fit Indices

- **Normed fit index:** $\dfrac{\chi^2_{null} - \chi^2_{model}}{\chi^2_{null}}$

- **Non-normed fit index:** $\dfrac{\dfrac{\chi^2_{null}}{df_{null}} - \dfrac{\chi^2_{model}}{df_{model}}}{\left[\dfrac{\chi^2_{null}}{df_{null}} - 1\right]}$

- **Comparative fit index:** $1 - \left[\dfrac{\chi^2_{model} - df_{model}}{\chi^2_{null} - df_{null}}\right]$

RMSEA = SQRT (λ^2 – df)/SQRT (df (N – 1))

CFI >=0.95 and RMSEA <=0.06

z-scores (raw – mean)/SD

Normal curve (68.2%, 95.4%, and 99.7%)

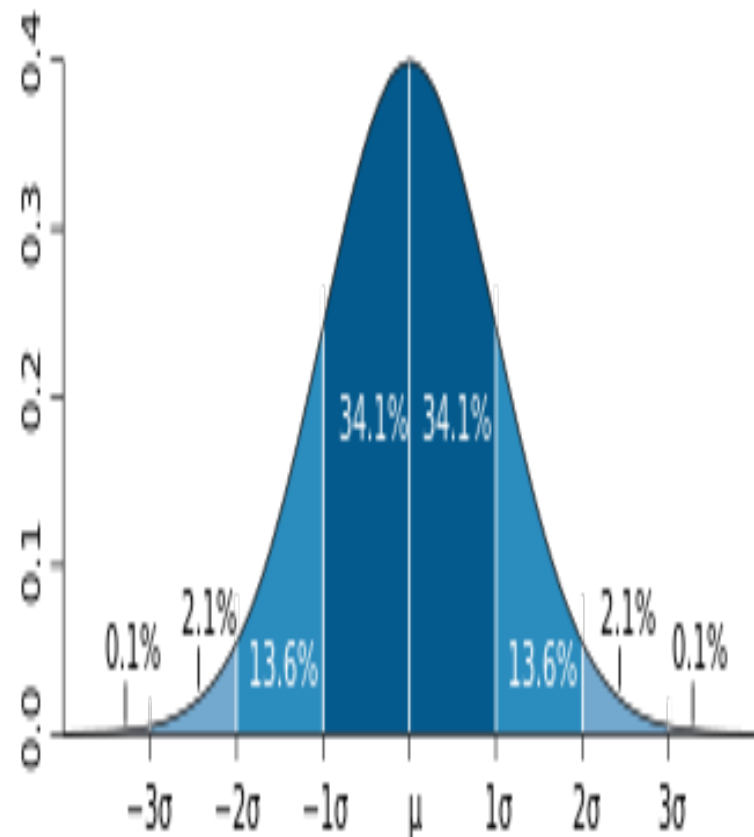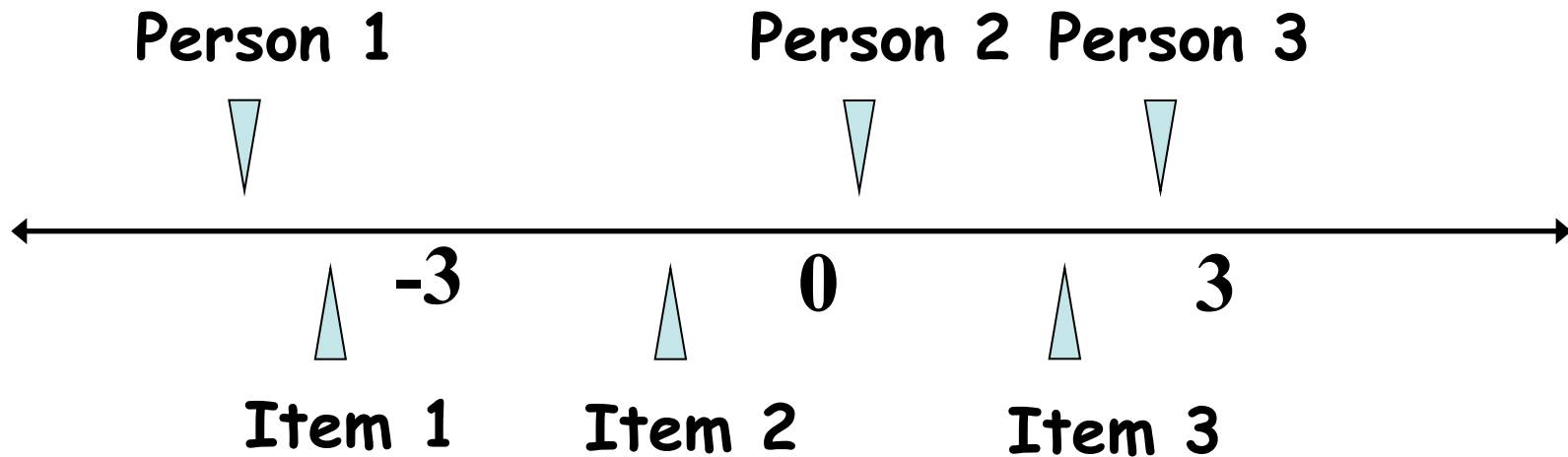# Item difficulty (p = 0.84 and 0.16)

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

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

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

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

$\qquad = -1.66/1.7$

$\qquad = \underline{-1.00}$

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

$\ = \underline{1.00}$

People and Items on Same *z-score* metric

Person 1   Person 2   Person 3

<——>
        -3   0   3

Item 1   Item 2   Item 3

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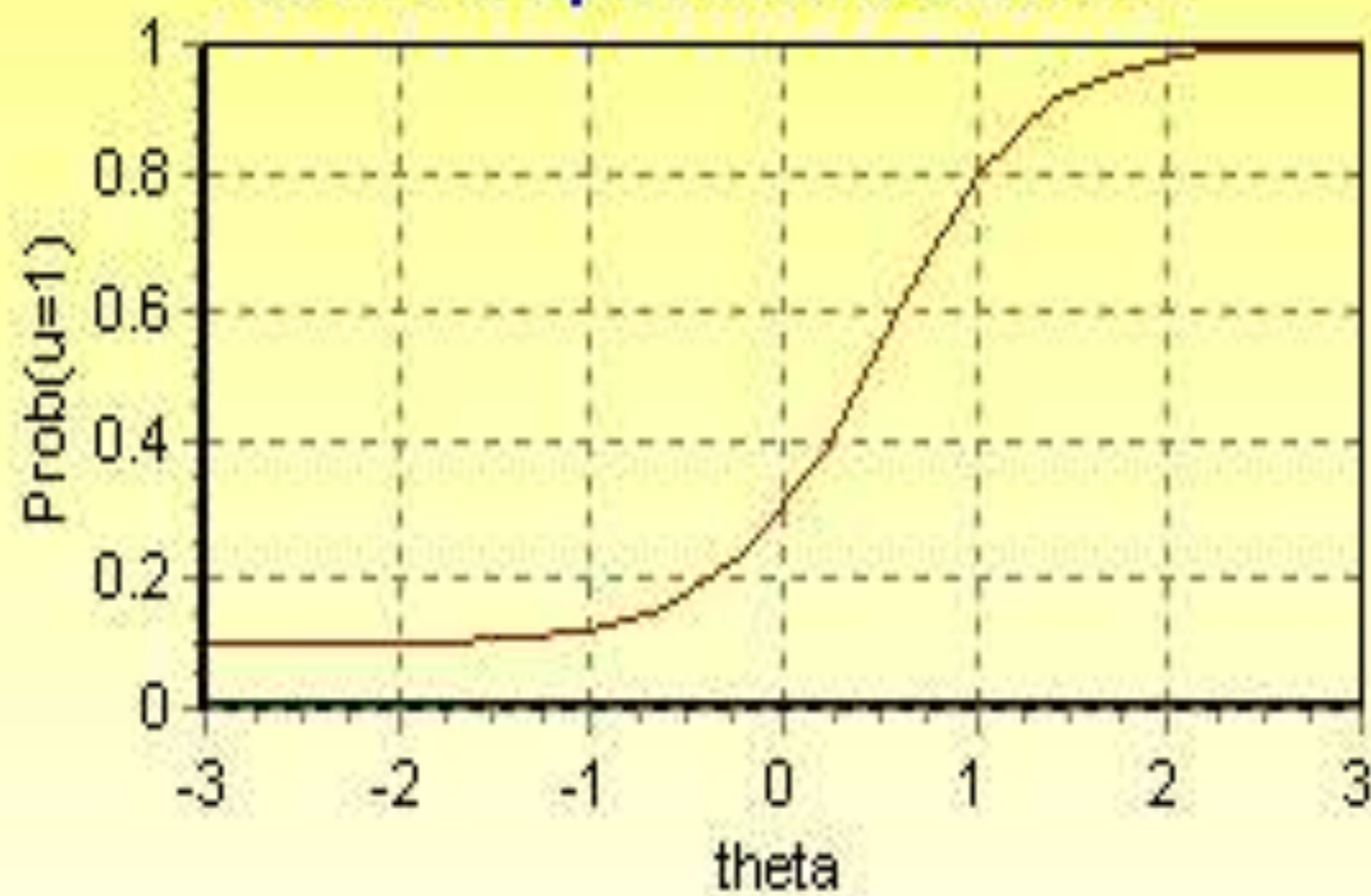
$$\Pr(Y_i \geq k) = \frac{1}{1 + \exp(-a_i\theta + b_{ik})}$$

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

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

a_i = item discrimination.

Item Response Function

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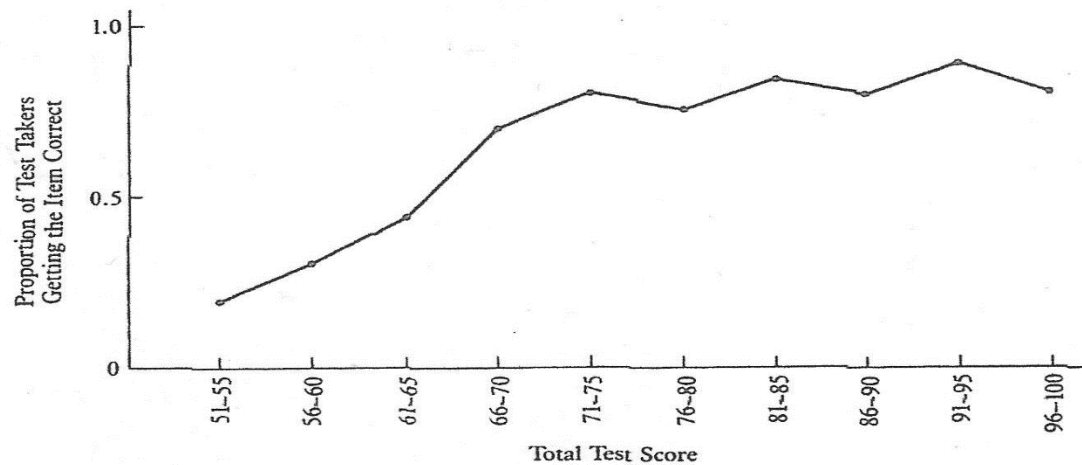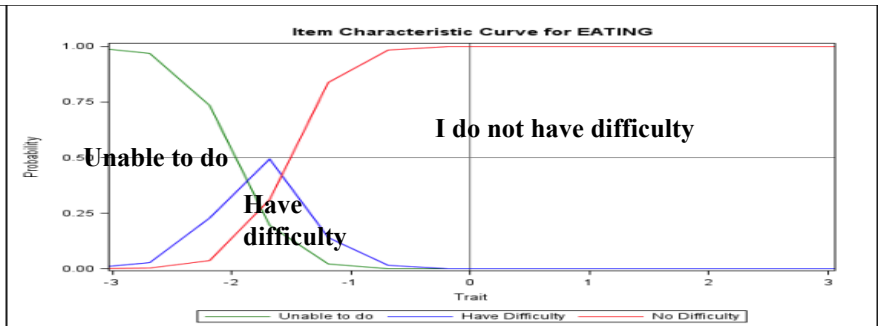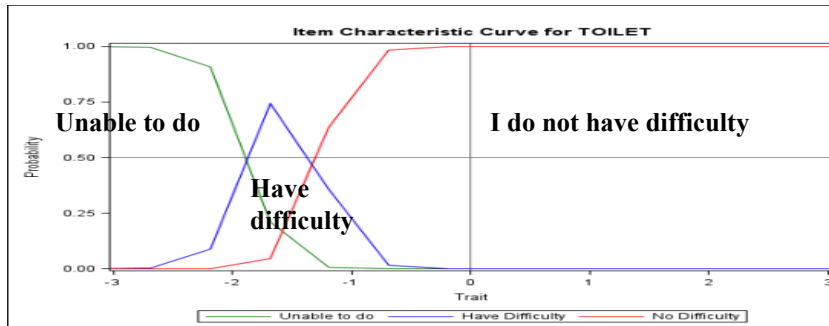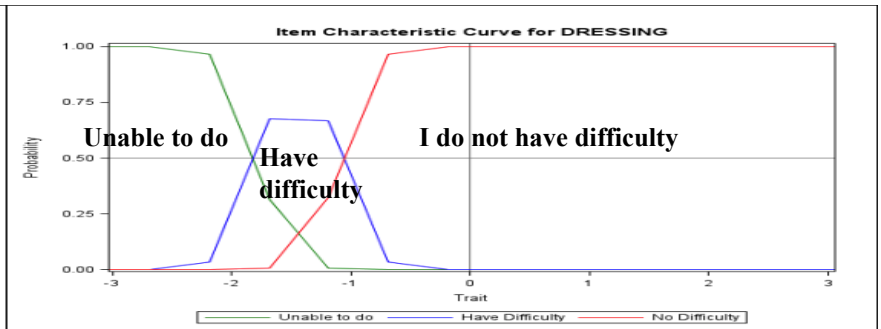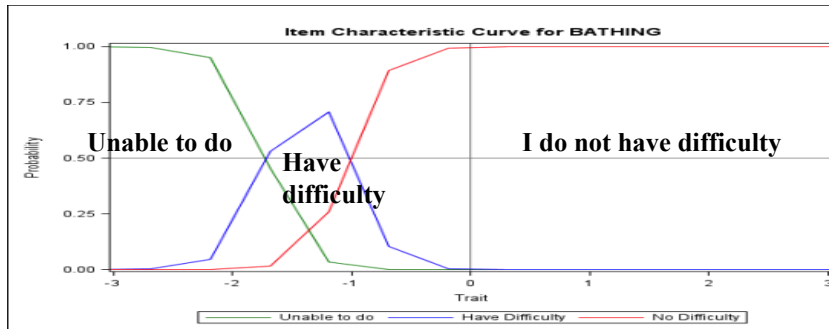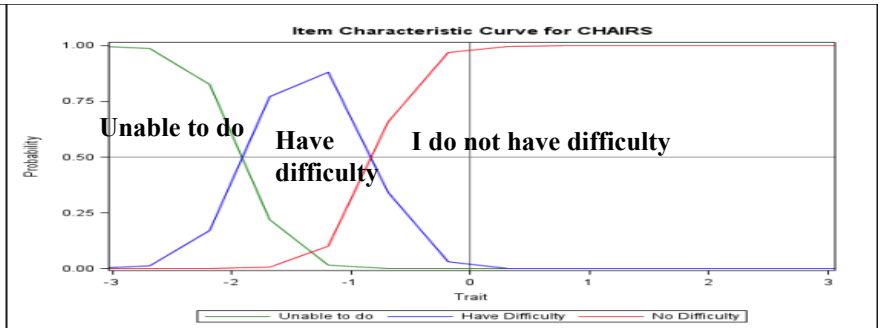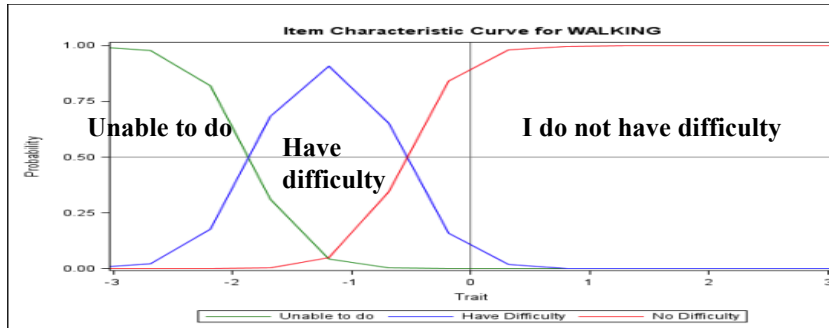

FIGURE 6-3   Item characteristic curve for a test item that discriminates well at low levels of performance but not at higher levels.

Item Characteristic Curves

# 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
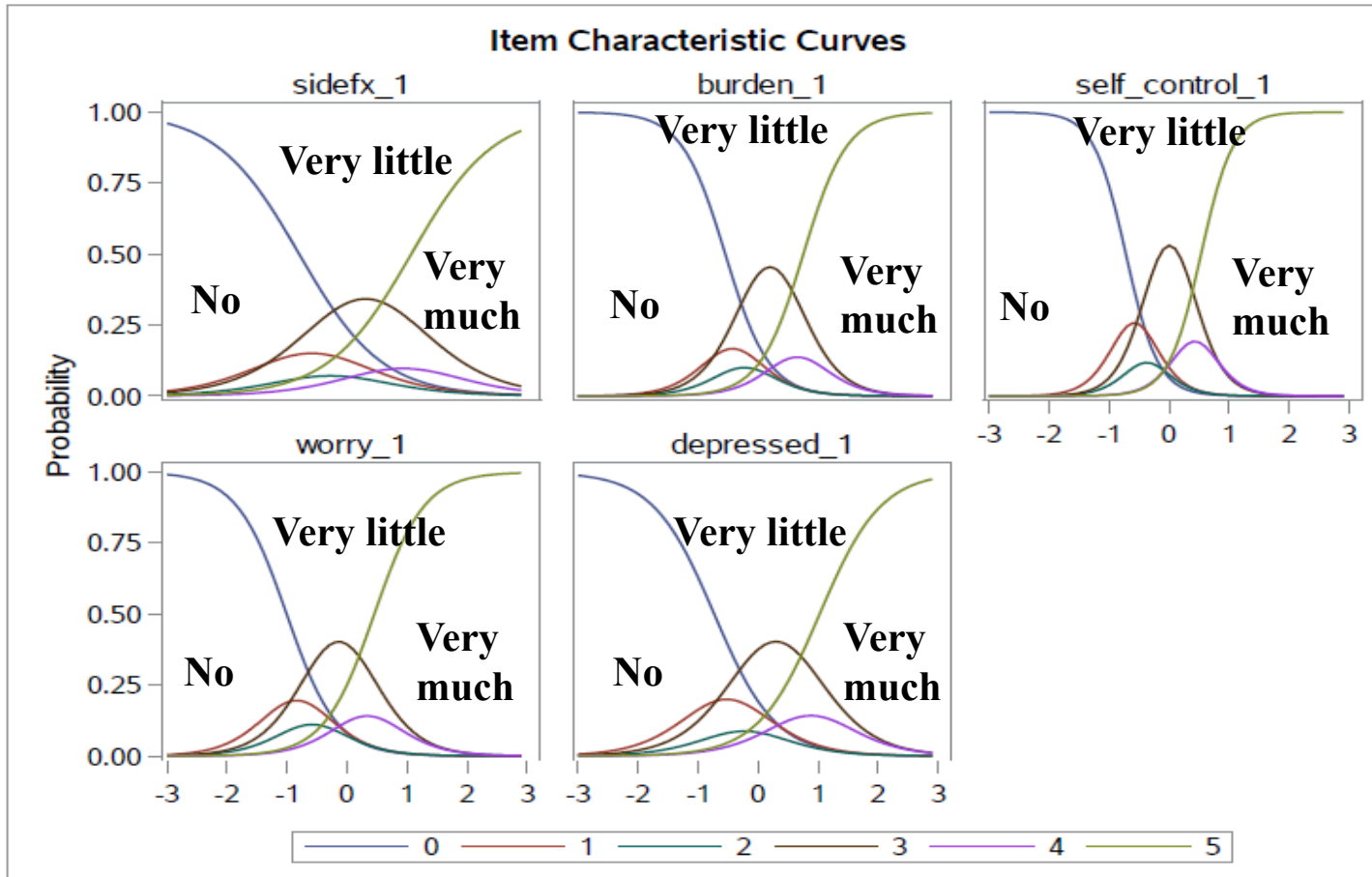
Loadings and Item Parameters

	Loading	Discrimination*	*Unable to do/ Have difficulty or no difficulty*	*Unable to do or have difficulty/ 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

Location of Walking Thresholds

Z-score (theta)		
0.5		
0.4		
0.3		
0.2		
0.1		
0		
-0.1		
-0.2		
-0.3	2nd threshold for walking	
-0.4		
-0.5		
-0.6	Unable to do or have difficulty	No difficulty
-0.7		
-0.8		
-0.9		
-1.0		
-1.1		
-1.2		
-1.3		
-1.4		
-1.5		
-1.6	1st threshold for walking	
-1.7		
-1.8		
-1.9	Unable to do	Have difficulty or No difficulty
-2.0		
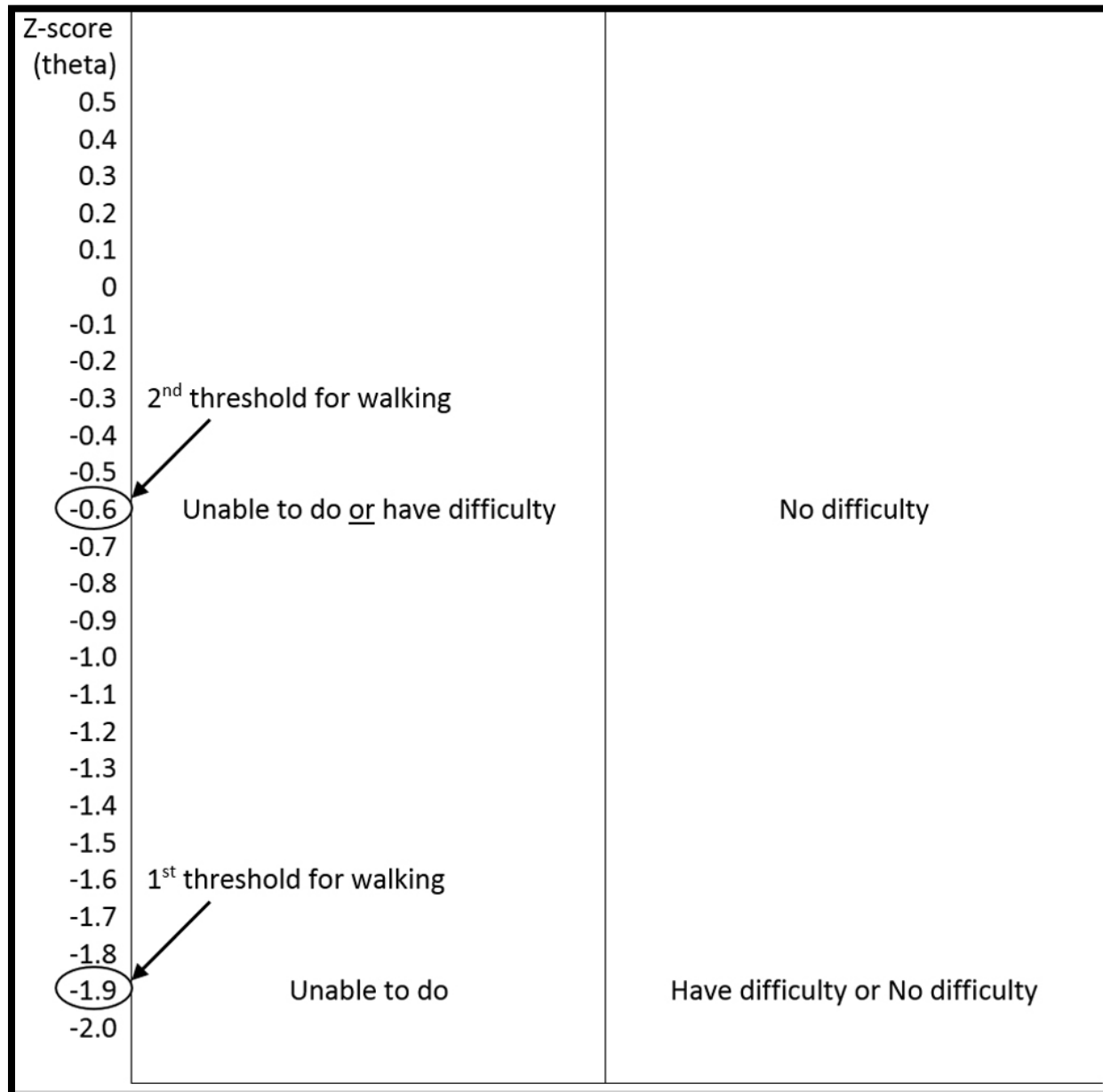
Location of Item Thresholds

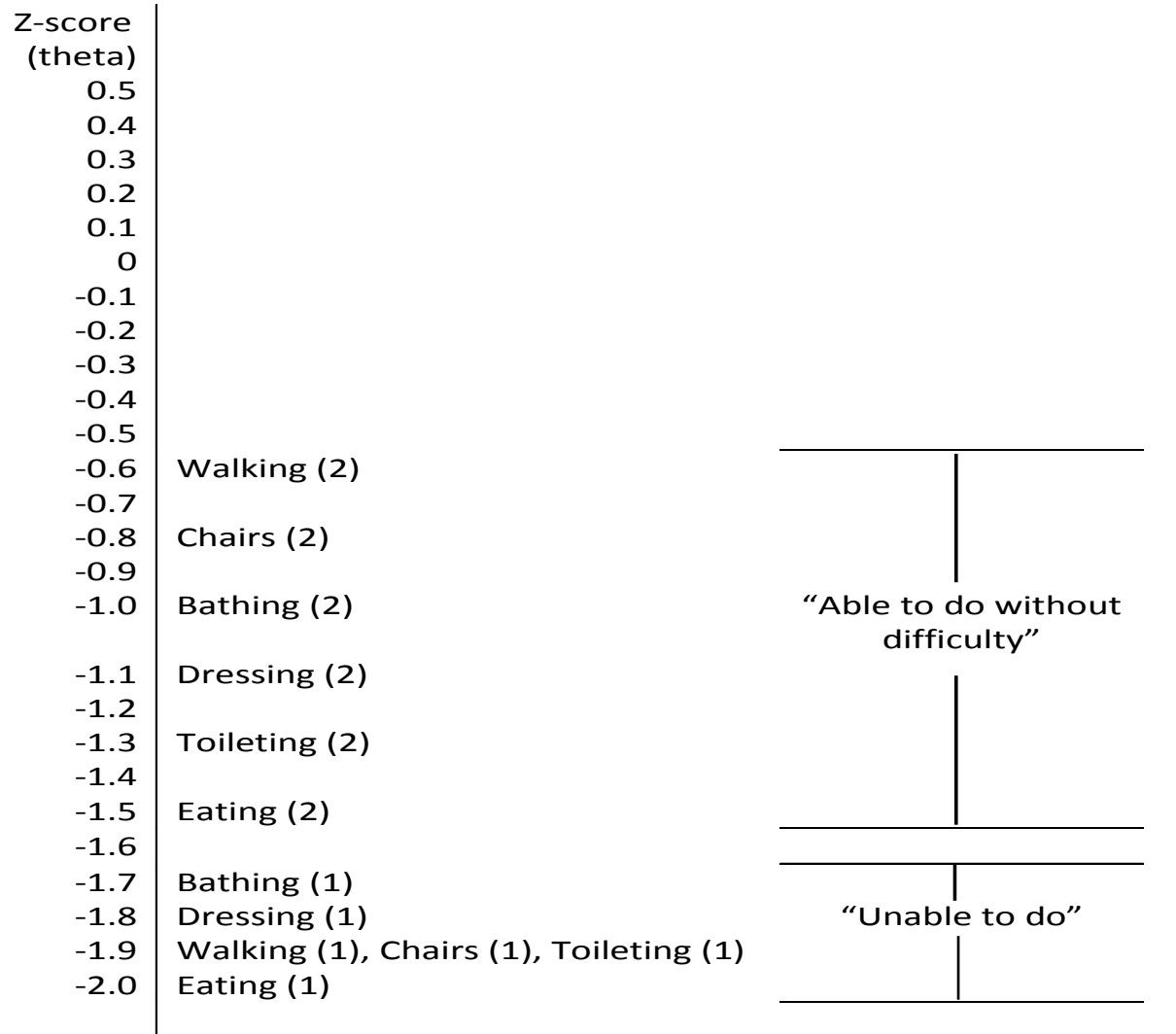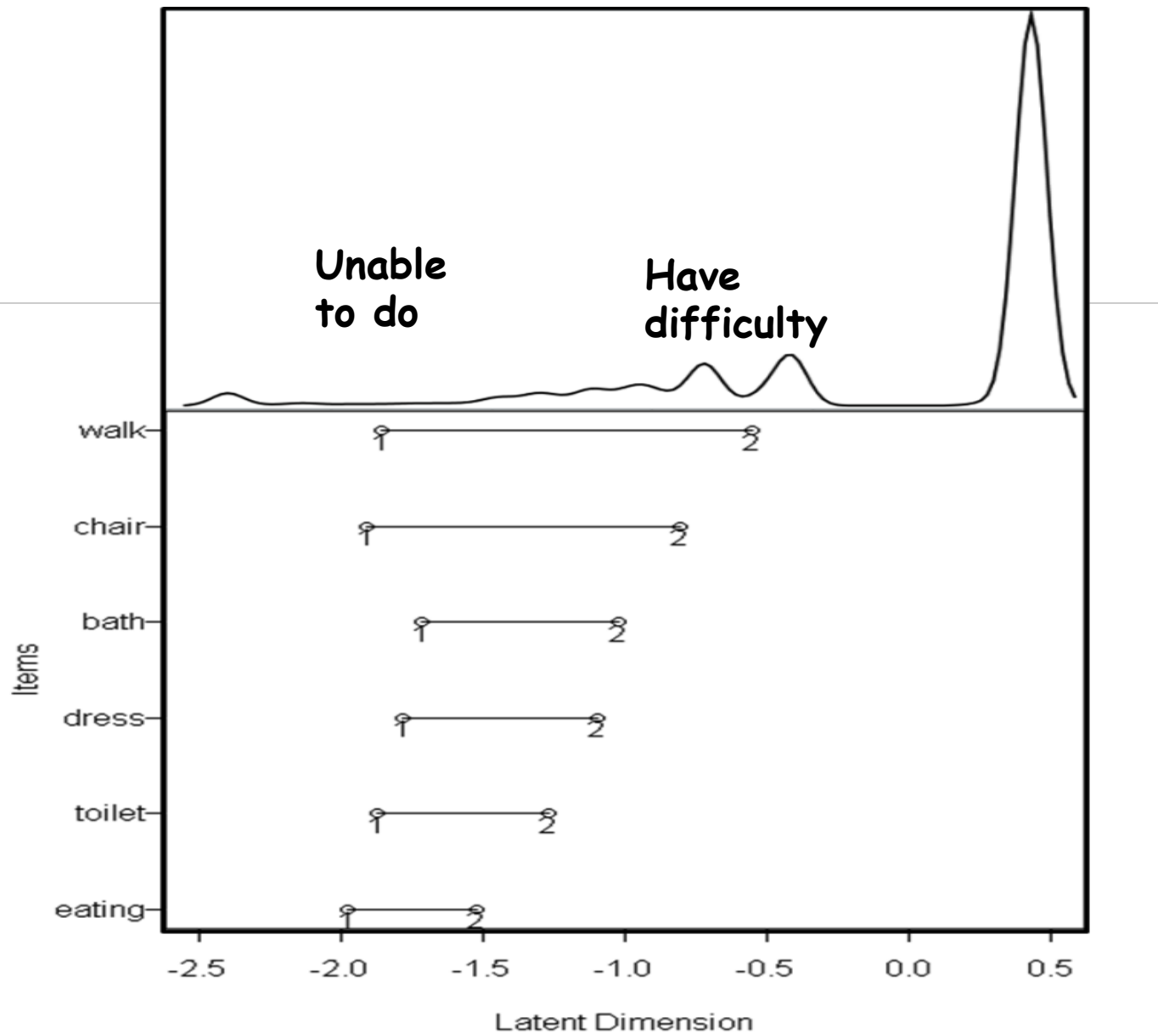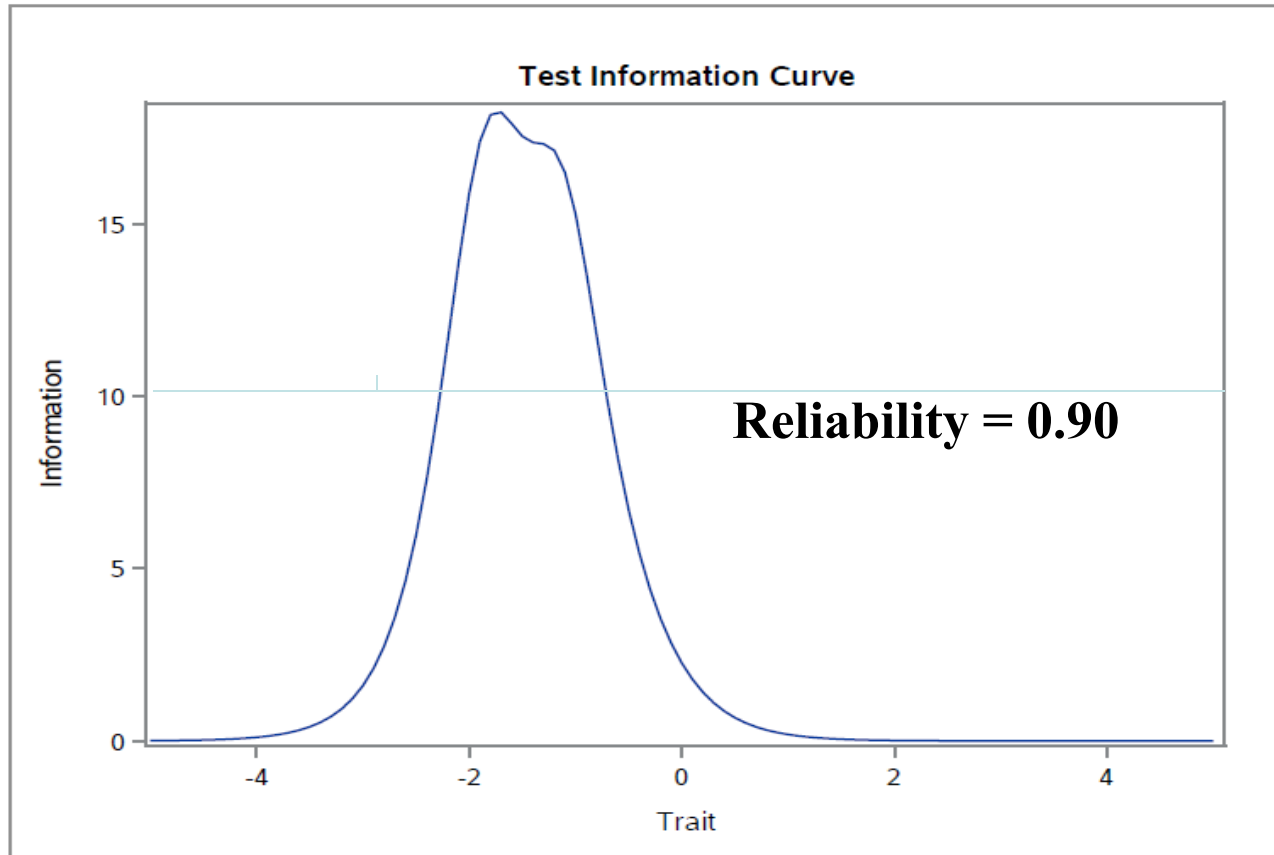| Z-score (theta)		
0.5		
0.4		
0.3		
0.2		
0.1		
0		
-0.1		
-0.2		
-0.3		
-0.4		
-0.5		
-0.6	Walking (2)	
-0.7		
-0.8	Chairs (2)	
-0.9		
-1.0	Bathing (2)	"Able to do without difficulty"
-1.1	Dressing (2)	
-1.2		
-1.3	Toileting (2)	
-1.4		
-1.5	Eating (2)	
-1.6		
-1.7	Bathing (1)	
-1.8	Dressing (1)	"Unable to do"
-1.9	Walking (1), Chairs (1), Toileting (1)	
-2.0	Eating (1)	

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

The IRT Procedure

Test Information Curve

Reliability = 0.90

(y-axis: Information, values 0, 5, 10, 15; x-axis: Trait, values -4, -2, 0, 2, 4)

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

DIF (2-parameter model)
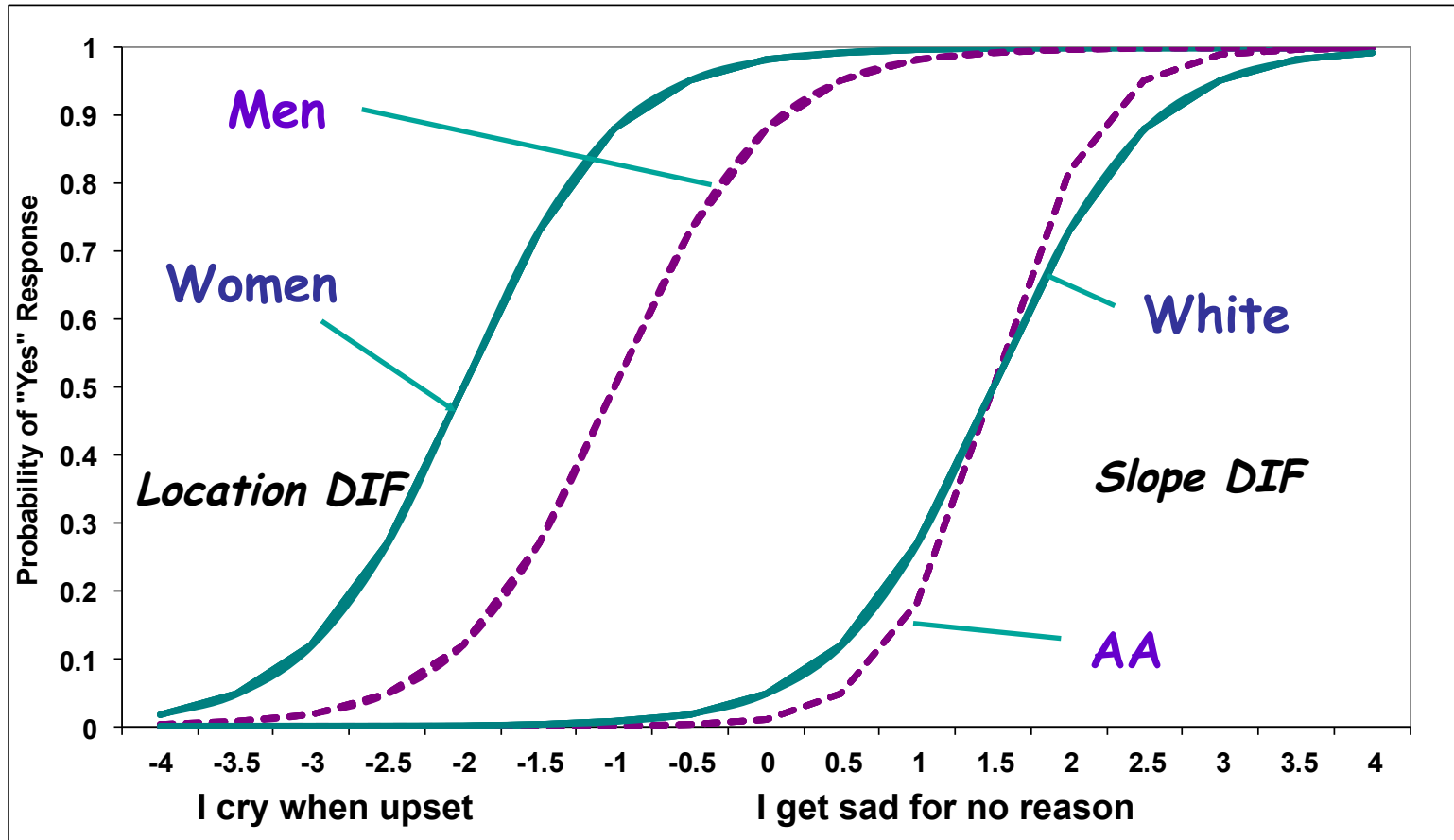

Higher Score = Better Physical Functioning

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) <u>without any difficulty</u>, but

  - This person reported <u>a little difficulty</u> being out of bed for most of the day.

Physical Functioning and Emotional Well-Being at Baseline for 54 Patients at UCLA-Center for East West Medicine
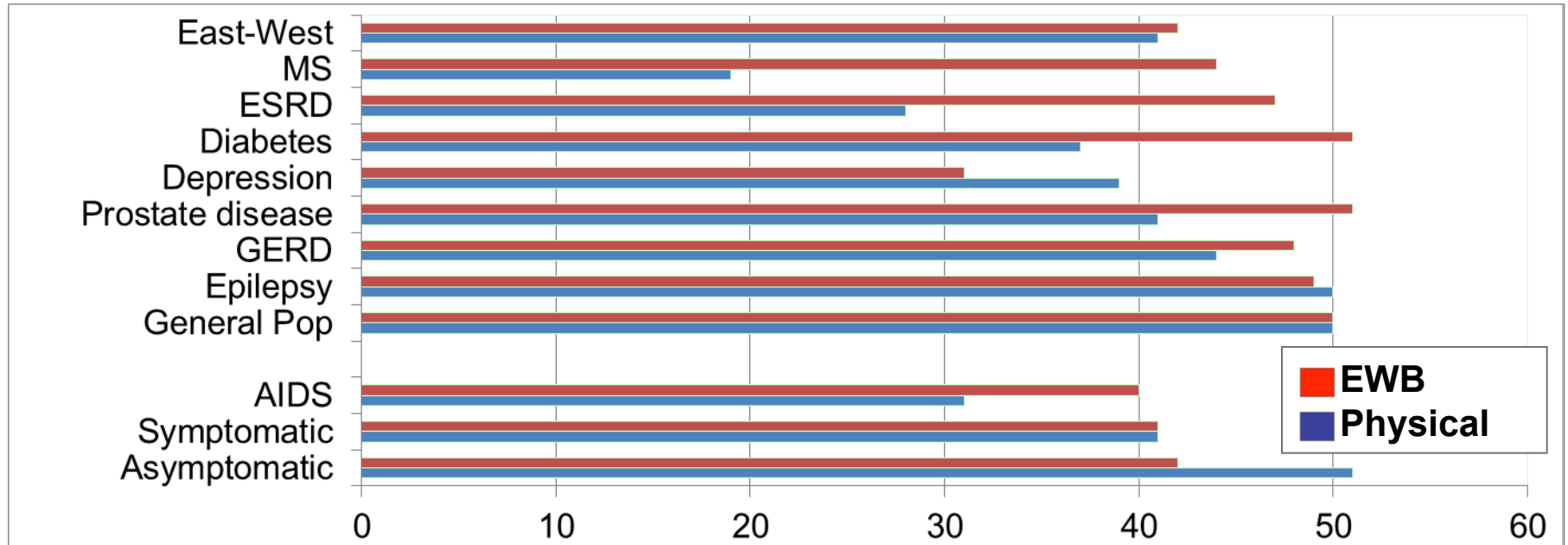
MS = multiple sclerois; ESRD = end-stage renal disease; GERD = gastroesophageal reflux disease.

Significant Improvement in all but 1 of SF-36 Scales (Change is in T-score metric)

	Change	t-test	prob.
PF-10	1.7	2.38	.0208
RP-4	4.1	3.81	.0004
BP-2	3.6	2.59	.0125
GH-5	2.4	2.86	.0061
EN-4	5.1	4.33	.0001
SF-2	4.7	3.51	.0009
RE-3	1.5	0.96	.3400 ←
EWB-5	4.3	3.20	.0023
PCS	2.8	3.23	.0021
MCS	3.9	2.82	.0067

Effect Size

$(\text{Follow-up} - \text{Baseline}) / SD_{baseline}$

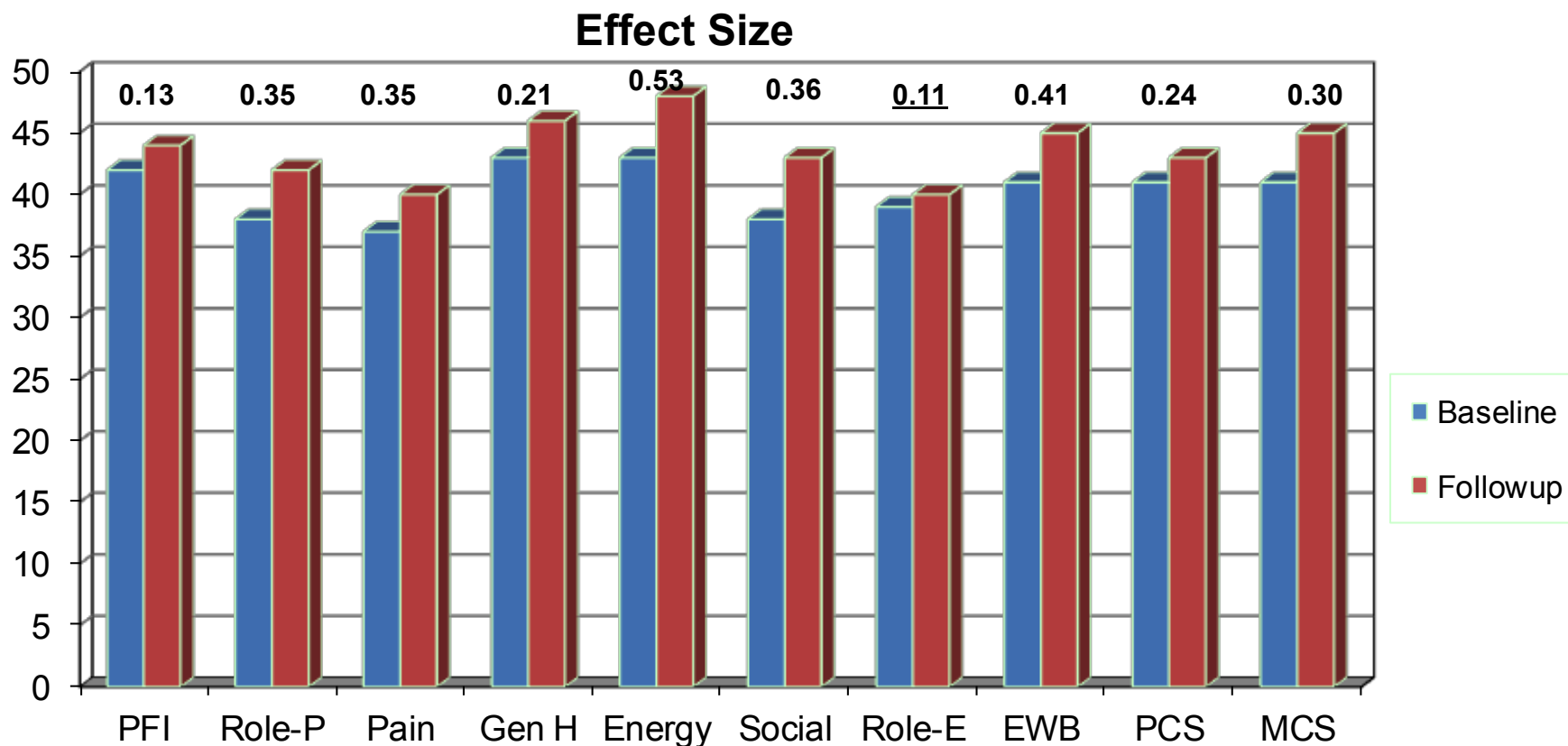*Cohen's Rule of Thumb:*

✓ ES = 0.20   Small

✓ ES = 0.50   Medium

✓ ES = 0.80   Large

32

Effect Sizes for Changes in SF-36 Scores

Effect Size

	PFI	Role-P	Pain	Gen H	Energy	Social	Role-E	EWB	PCS	MCS
Effect Size	0.13	0.35	0.35	0.21	0.53	0.36	0.11	0.41	0.24	0.30

Legend: ■ Baseline ■ Followup

PFI = Physical Functioning; Role-P = Role-Physical; Pain = Bodily Pain; Gen H=General Health; Energy = Energy/Fatigue; Social = Social Functioning; Role-E = Role-Emotional; EWB = Emotional Well-being; PCS = Physical Component Summary; MCS =Mental Component Summary.

Defining a Responder: Reliable Change Index (RCI)

$$\frac{X_2 - X_1}{(\sqrt{2})\,(SEM)}$$

$$SEM = SD_{bl} \times \sqrt{1 - r_{xx}}$$

Note: SD_{bl} = standard deviation at baseline
r_{xx} = reliability

Coefficient of Repeatibility
Amount of Change in Observed Score Needed To be Statistically Significant

$$\overline{(\sqrt{2})\,(\text{SD}_{bl})\sqrt{(1-r_{xx})}\,(1.96)}$$

$$= 2.77 * \text{SEM}$$

Note: SD_{bl} = standard deviation at baseline and r_{xx} = reliability

Amount of Change in Observed Score Needed for Significant Individual Change

Scale	Coefficient of Repeatability	Effect size	Cronbach's alpha
PF-10	8.4	0.67	0.94
RP-4	8.4	0.72	0.93
BP-2	10.4	1.01	0.87
GH-5	13.0	1.13	0.83
EN-4	12.8	1.33	0.77
SF-2	13.8	1.07	0.85
RE-3	9.7	0.71	0.94
EWB-5	13.4	1.26	0.79
PCS	7.1	0.62	0.94*
MCS	9.7	0.73	0.93*

* Mosier's formula (not coefficient alpha).

Amount of Change Needed for Significant Individual Change
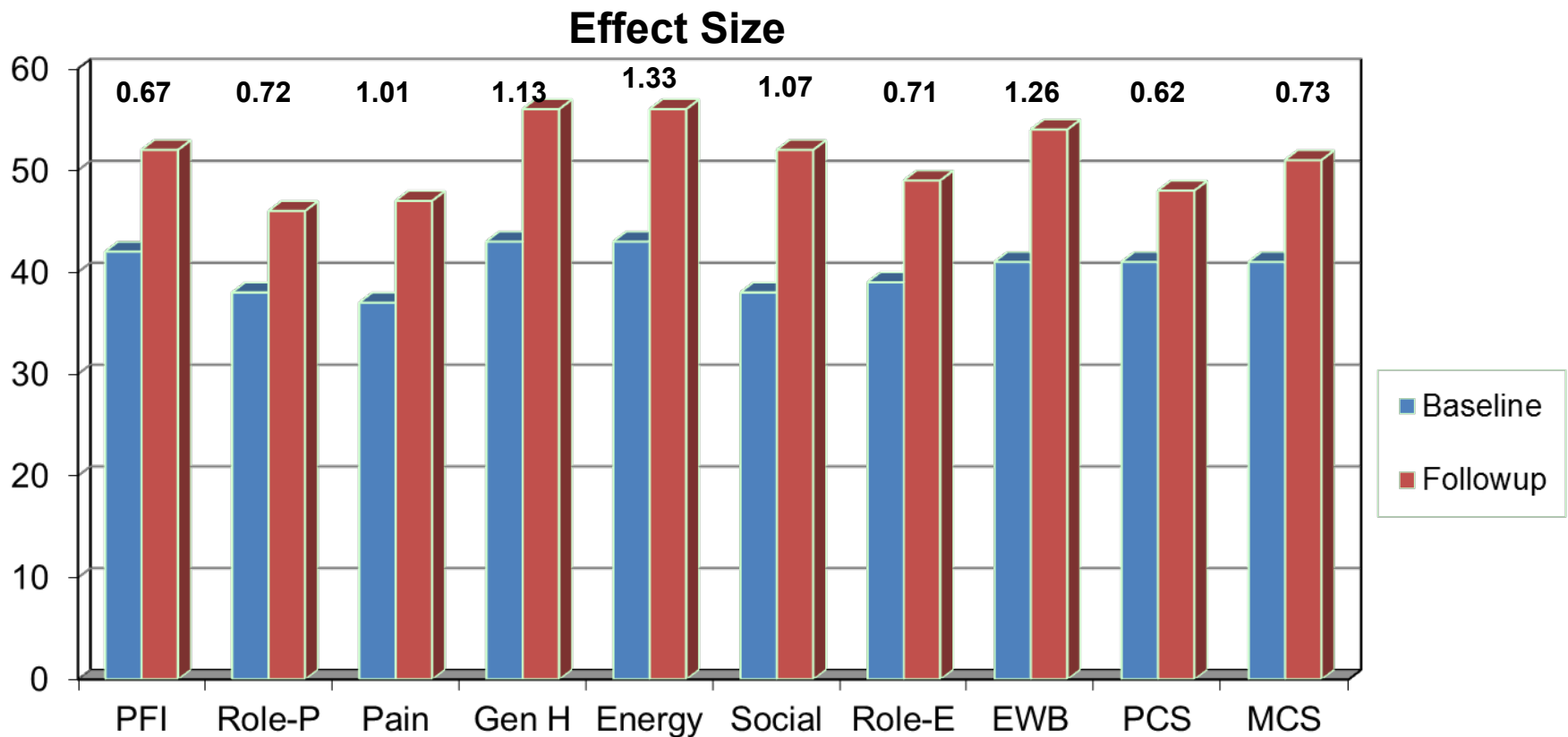
PFI = Physical Functioning; Role-P = Role-Physical; Pain = Bodily Pain; Gen H=General Health; Energy = Energy/Fatigue; Social = Social Functioning; Role-E = Role-Emotional; EWB = Emotional Well-being; PCS = Physical Component Summary; MCS =Mental Component Summary.

7-31% of People in Sample Improve Significantly

	% Improving	% Declining	Difference
PF-10	13%	2%	+ 11%
RP-4	31%	2%	+ 29%
BP-2	22%	7%	+ 15%
GH-5	7%	0%	+ 7%
EN-4	9%	2%	+ 7%
SF-2	17%	4%	+ 13%
RE-3	15%	15%	0%
EWB-5	19%	4%	+ 15%
PCS	24%	7%	+ 17%
MCS	22%	11%	+ 11%

drhays@ucla.edu

Appendix:
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

◀···· **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
  - $SE = SD (1- reliability)^{1/2}$
  - $Reliability = 1 - SE^2$
  - Reliability = 0.90 when <u>SE = 0.32</u>
- 95% CI = true score +/- 1.96 x SE

  (CI = <u>-0.63</u> → <u>0.63</u> z-score when reliability = 0.90)

T-score Metric

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

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)