Item Response Theory in Patient-Reported Outcomes Research

Ron D. Hays, Ph.D., UCLA

Society of General Internal Medicine

California-Hawaii Regional Meeting

February 2, 2013 (10:45-12:00 pm)

Covel Commons (Grand Horizon, Salon C)

Computer Adaptive Testing (CAT)

www.nihpromis.org

- Patient Reported Outcomes Measurement Information System (PROMIS®)
- Funded by the National Institutes of Health
- One domain captured is “anger”
  - Mood (irritability, frustration)
  - Negative social cognitions (interpersonal sensitivity, envy, disagreeableness)
  - Needing to control anger

Reliability (0-1)

- 0.70 or above for group comparisons
- 0.90 or above for individual assessment

z-scores (mean = 0 and SD = 1):

- Reliability = $1 - SE^2$
- So reliability = **0.90** when SE = 0.32

T-scores (mean = 50 and SD = 10):

$$T = 50 + (z * 10)$$

- Reliability = $1 - (SE/10)^2$
- So reliability = **0.90** when SE = 3.2

In the past 7 days …

I was grouchy [1st question]

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

Theta = 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

Theta = 51.9 SE = 4.8 (rel. = 0.77)

In the past 7 days …

I felt angry [3rd question]

- Never
- Rarely
- Sometimes
- Often
- Always

Theta = 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

Theta = 48.8 SE = 3.6 (rel. = 0.87)

In the past 7 days …

I felt annoyed [5th question]

- Never
- Rarely
- Sometimes
- Often
- Always

Theta = 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

Theta = 50.2 SE = 2.8 (rel = 0.92)

Theta, SEM, and 95% CI

- 56 and 6 (reliability = .68) W = 22
- 52 and 5 (reliability = .77) W = 19
- 50 and 4 (reliability = .85) W = 15
- 49 and 4 (reliability = .87) W = 14
- 50 and 3 (reliability = .90) W = 12
- 50 and <3 (reliability = .92) W = 11

Response Burden

- Paper and pencil rules of thumb
  - 3-5 items per minute
- PROMIS computer administration to general population
  - 8-12 items per minute
- Scleroderma patients at UCLA
  - 6 items per minute

Item Response Theory (IRT)

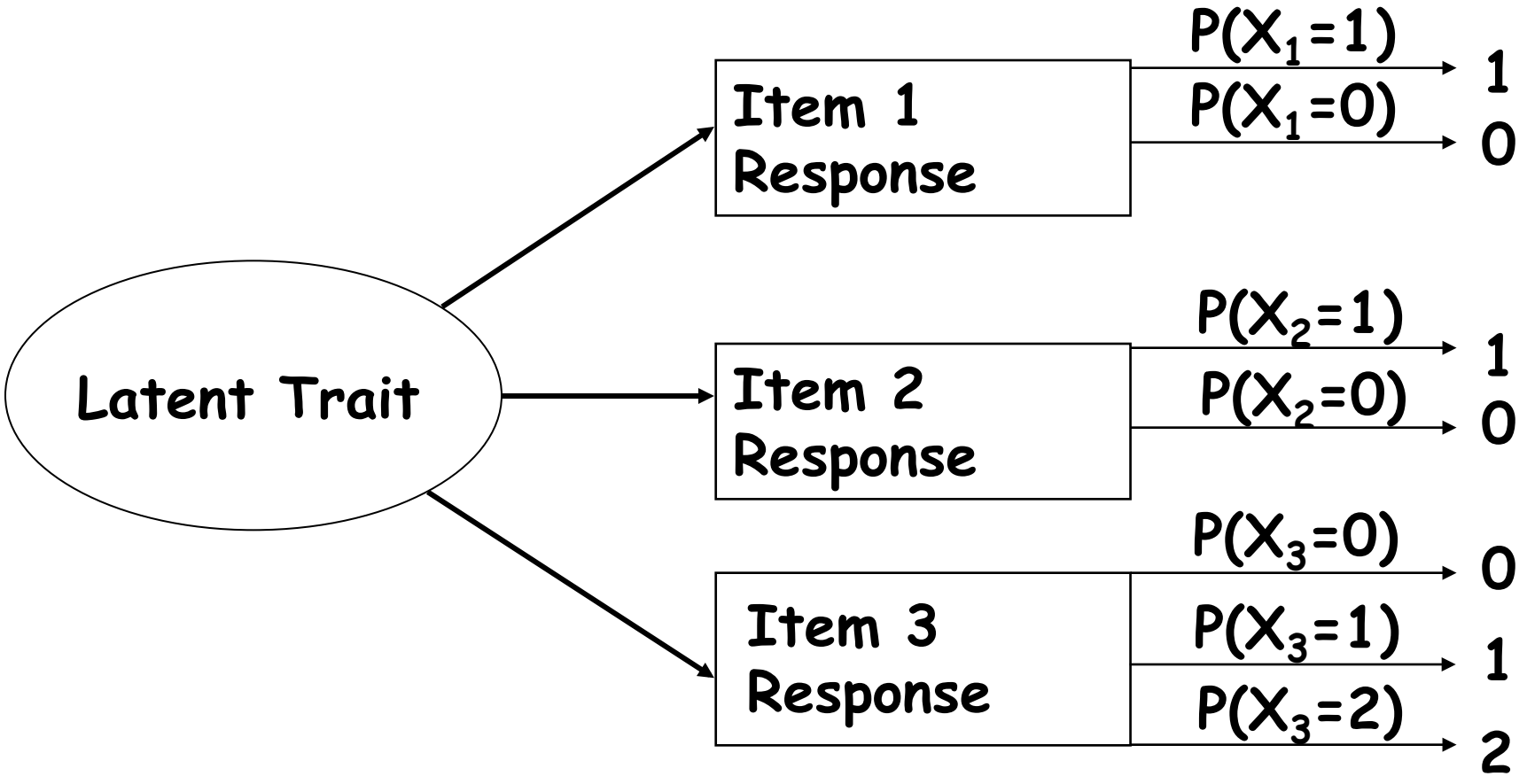
IRT models the relationship between a person's response Y_i to the question (i) and his or her level of the latent construct θ:

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

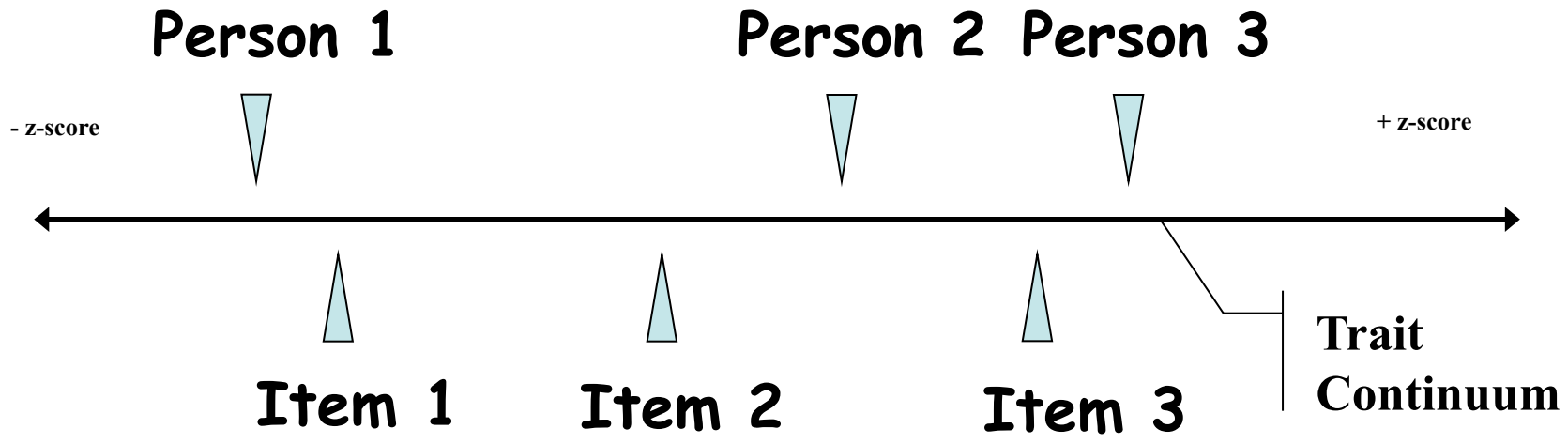
b_{ik} is how difficult it is to answer in category k or higher on item i

a_i is the item discrimination or slope parameter.

Latent Trait and Item Responses



Item Responses and Trait Levels



Person Scale Scores (θ)

- Level on attribute measured
- Average items together and compute z-score
- Mean = 0, SD = 1

$$z_X = \frac{(X - \bar{X})}{SD_X}$$

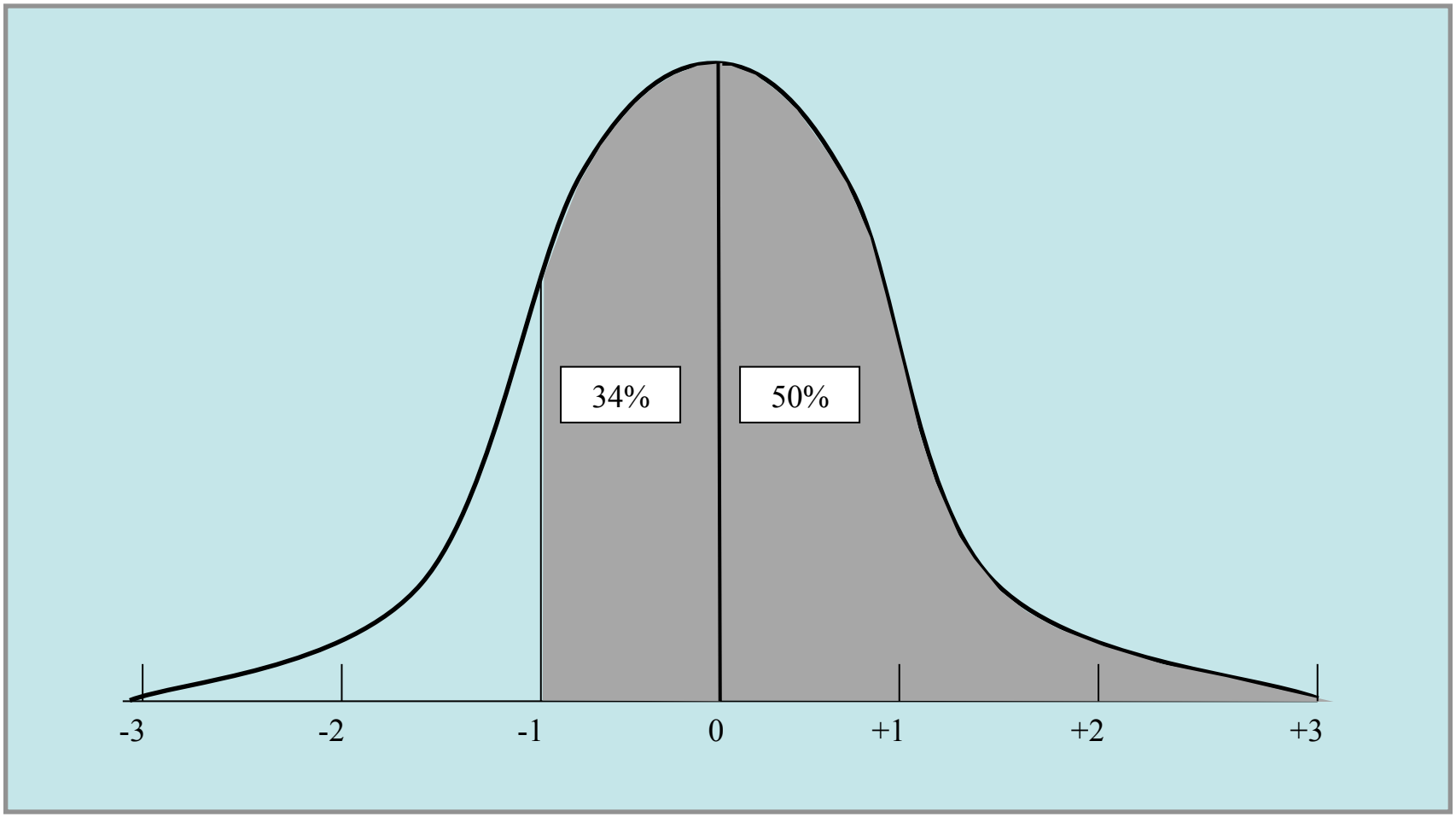
Item difficulty (p = 0.84)

Proportion of people endorsing the item (p) can be expressed as z:

$$Z = \ln (1-p)/p)/1.7 = (\ln (1-p) - \ln (p))/1.7$$
$$= (\ln (.16) - \ln (.84))/1.7$$
$$= (-1.83 + .17)/1.7$$
$$= -1.66/1.7$$
$$= \underline{-1.00}$$

(-2 -> 2 is typical range)

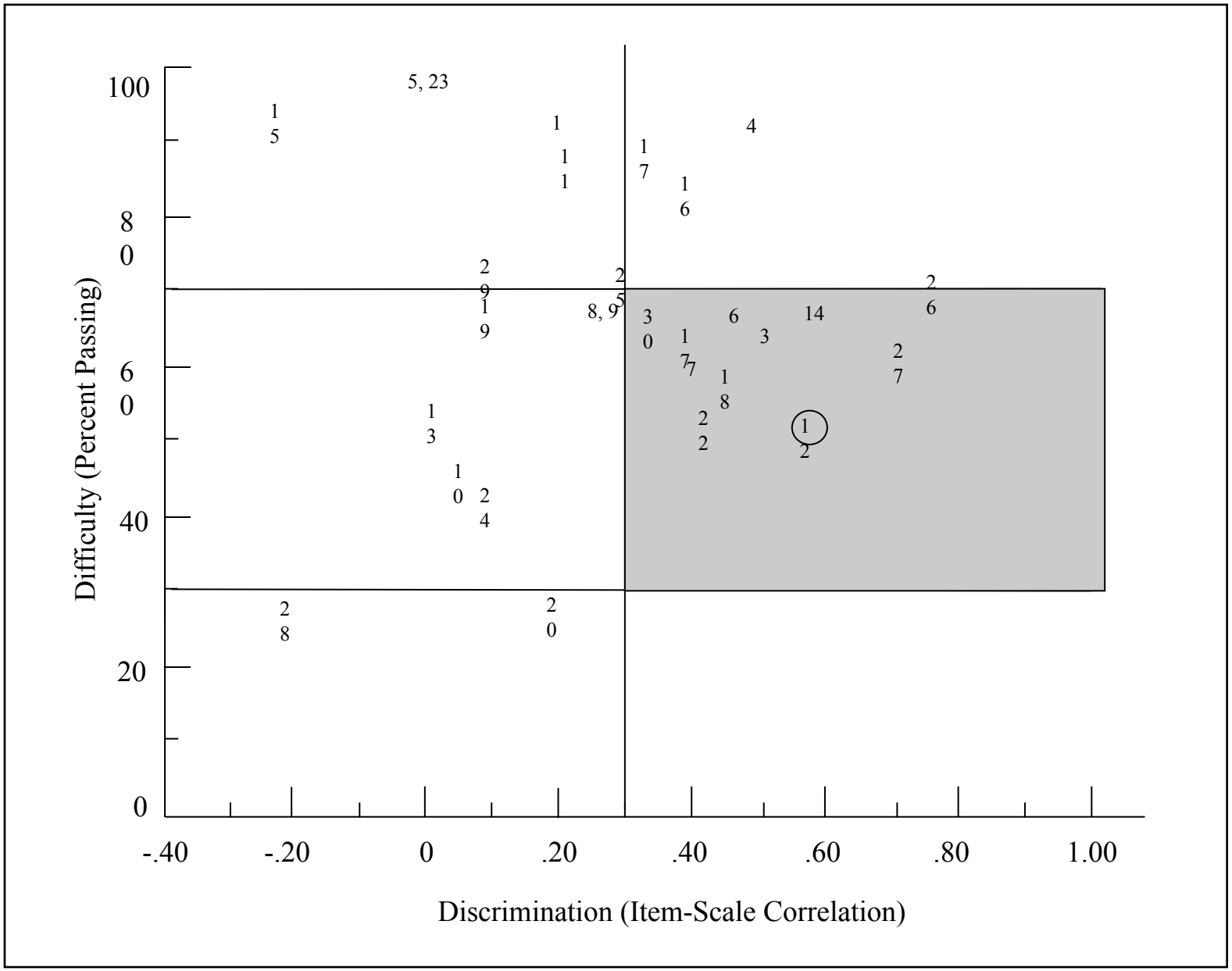
P-value transformation for an Item (*p* =.84)



Item Discrimination or Slope

Item-scale correlation can be expressed as z:

- $z = \frac{1}{2} [\ln (1 + r) - \ln (1-r)]$
- if r = 0.30, z = 0.31
- if r = 0.80, z = 1.10
- if r = 0.95, z = 1.83
- if r = 0.99, z = 2.65


Difficulty (Percent Passing)
100
80
60
40
20
0
-.40
-.20
0
.20
.40
.60
.80
1.00
Discrimination (Item-Scale Correlation)
5, 23
15
1
11
4
17
16
29
19
25
8, 9
26
30
6
3
14
17
7
18
27
22
12
13
10
24
28
20

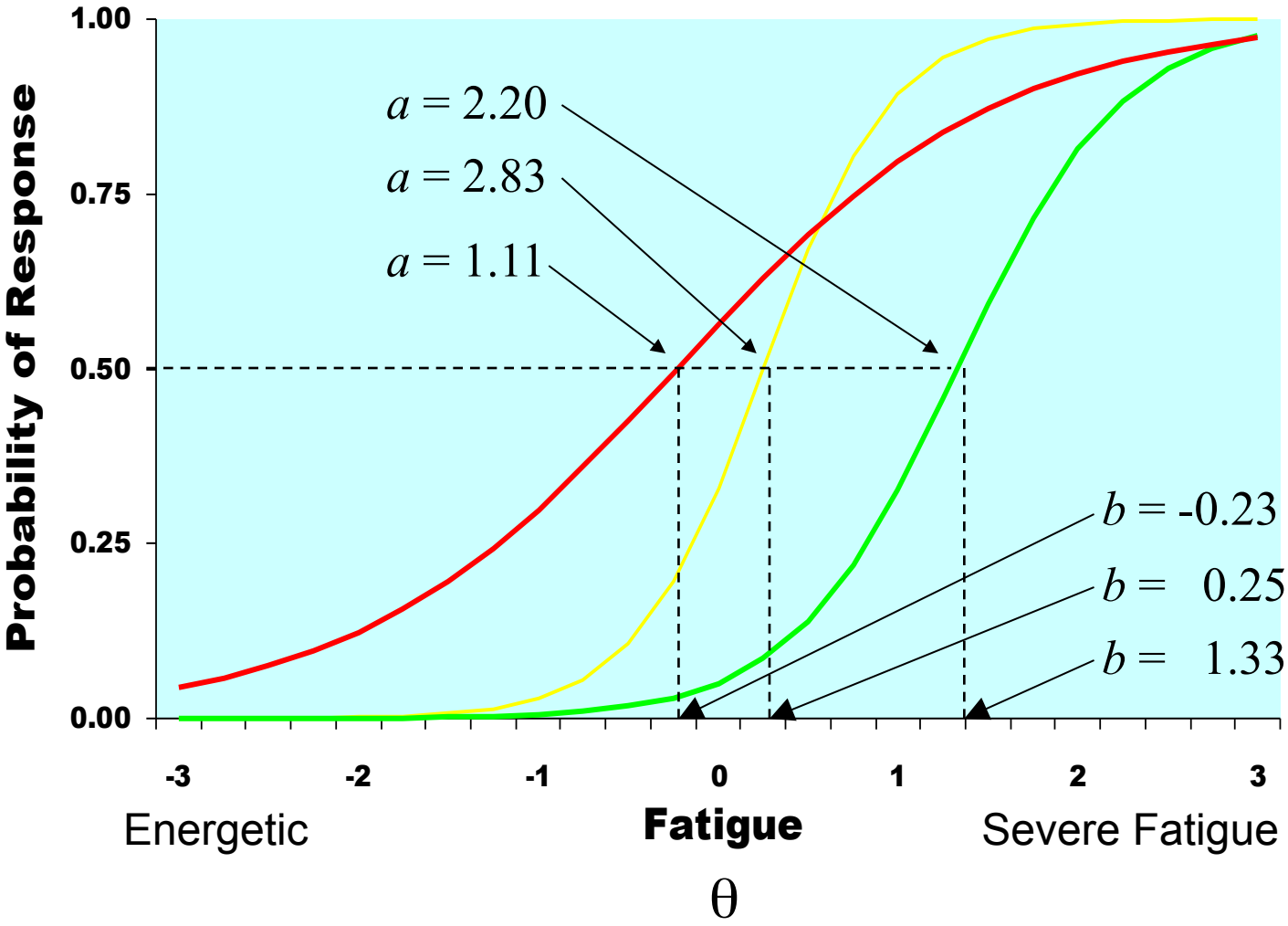
IRT Versus CTT

- Dichotomous and polytomous items
- Item parameters (difficulty and discrimination) estimated using logistic models instead of proportions and item-scale correlations
- IRT models
  - Rasch model/Graded response model
    - Difficulty parameter
    - Discrimination or slope parameter

2-Parameter Logistic IRT Model

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

$$a_i(\theta - b_i)$$



IRT Versus CTT

- Reliability (information) is conditional on where one is estimated to be on the underlying attribute
- Rather than estimated overall (coefficient alpha)

Information Conditional on Trait Level

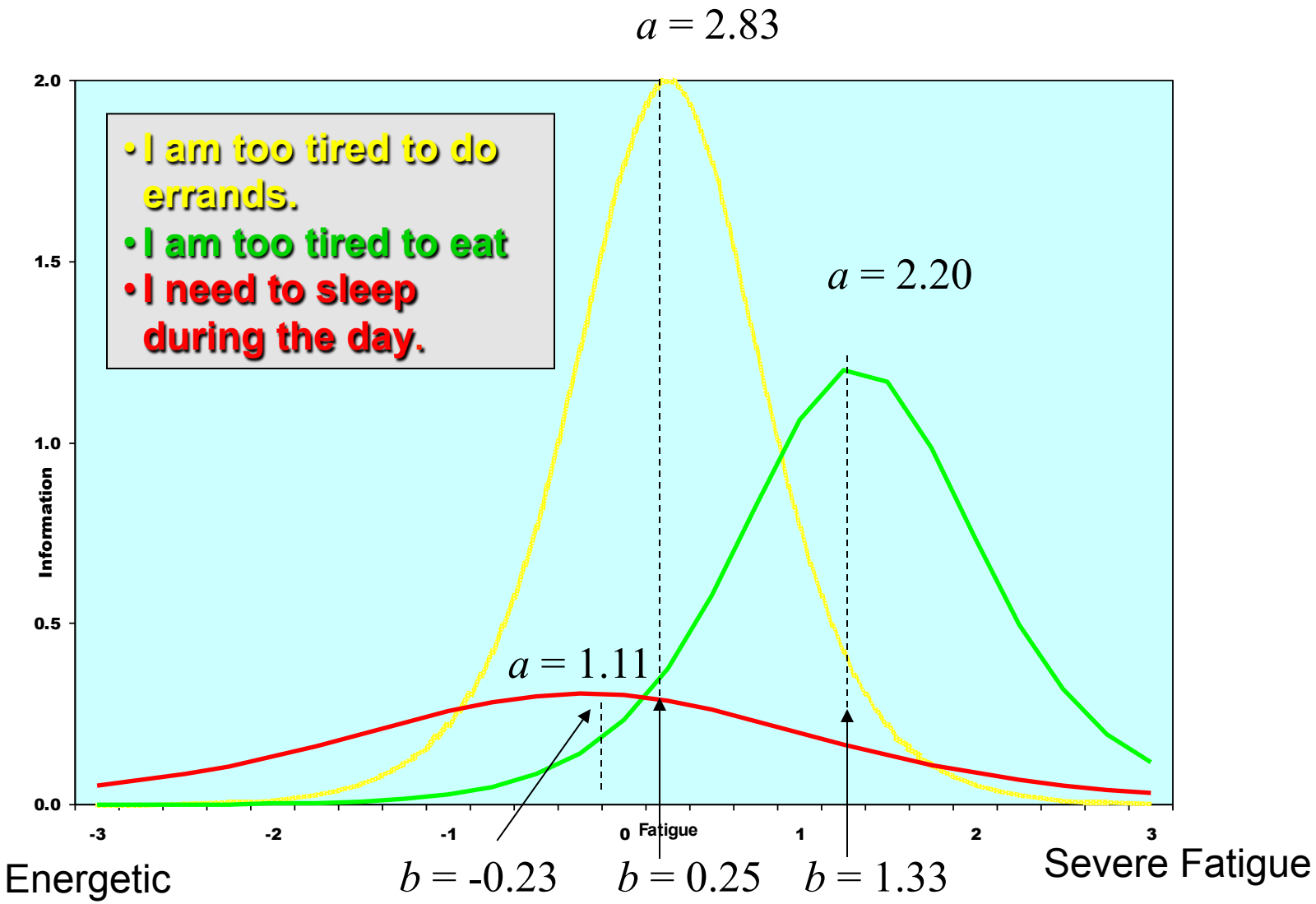
- Item information proportional to inverse of standard error of measurement:

$$SEM(\Theta) = \frac{1}{\sqrt{I(\Theta)}}$$

- Scale information is the sum over item information:

$$I(\Theta) = \sum_{i=1}^{n} I_i(\Theta)$$

Item Parameters indicate where items are most useful (informative) for distinguishing among respondents)



IRT Versus CTT

- Item parameters estimated by:
  - Marginal maximum likelihood estimation (MML)
- Level on attribute (person score or θ) estimated by:
  - ML (maximum likelihood)
  - MAP (Maximum a posterior scoring)
  - EAP (Expected a posterior scoring)

Scoring All Response Patterns Using Sum Score and Different IRT Models

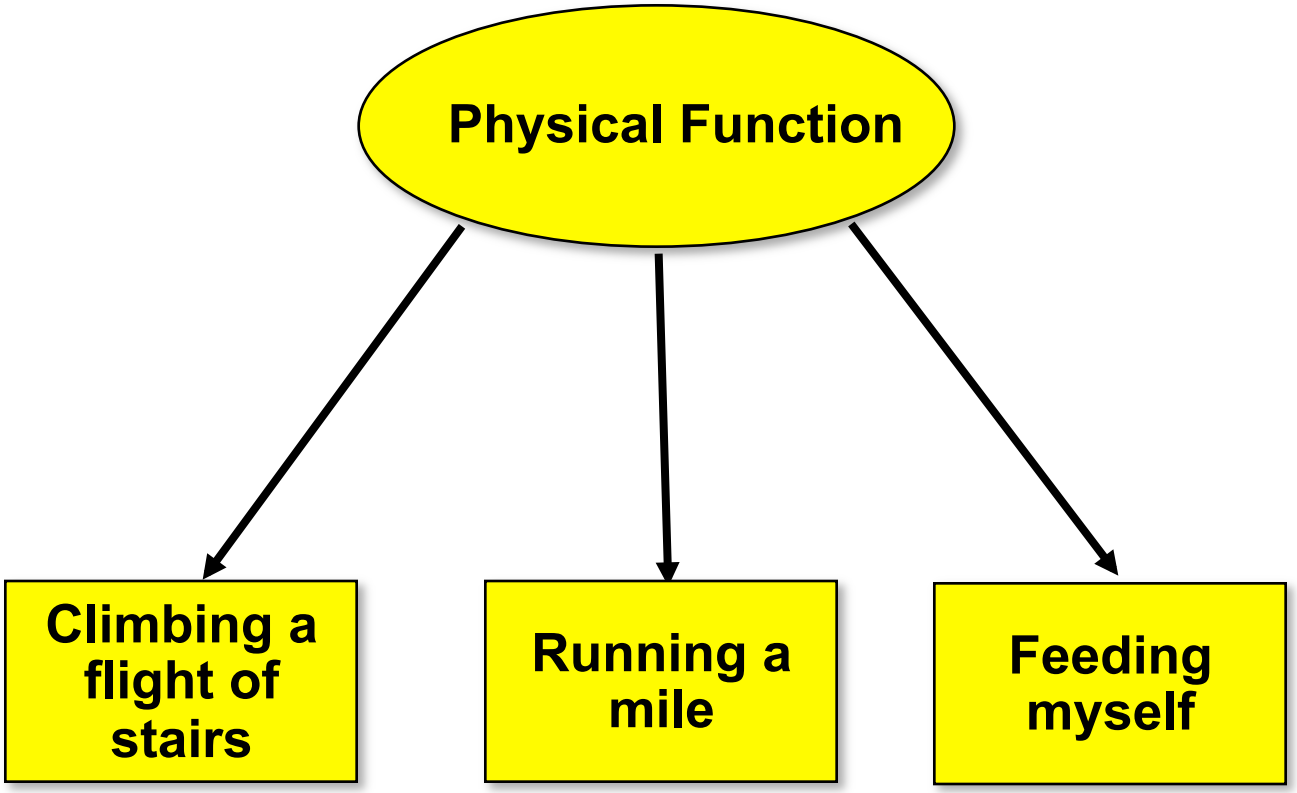
More Fatigued

#	Item Response Pattern 0 = false, 1 = true	Summed Score	1 PL IRT / Rasch Model M-L Estimate	2 PL IRT Model M-L Estimate
1	0 0 0 0	0	-0.84	-0.82
2	1 0 0 0	1	-0.22	-0.27
3	0 1 0 0	1	-0.22	-0.21
4	0 0 1 0	1	-0.22	-0.19
5	0 0 0 1	1	-0.22	-0.01
6	1 1 0 0	2	0.22	0.14
7	1 0 1 0	2	0.22	0.15
8	0 1 1 0	2	0.22	0.19
9	1 0 0 1	2	0.22	0.31
10	0 1 0 1	2	0.22	0.36
11	0 0 1 1	2	0.22	0.37
12	1 1 1 0	3	0.71	0.52
13	1 1 0 1	3	0.71	0.72
14	1 0 1 1	3	0.71	0.74
15	0 1 1 1	3	0.71	0.80
16	1 1 1 1	4	1.36	1.35

IRT Assumptions

- Dimensionality
  - Unidimensionality for typical models
- Local Independence
- Monotonicity
- Person fit

Hypothesized One-Factor Model



Sufficient Unidimensionality

- One-Factor Categorical Confirmatory Factor Analytic Model (e.g., using Mplus)
  - Polychoric correlations
  - Weighted least squares with adjustments for mean and variance
- Bifactor Model
  - General factor and group-specific factors

Local Independence

- After controlling for dominant factor(s), item pairs should not be associated.

- Evaluated by looking at size of residual correlations from one-factor model
  - Look for residual correlations > 0.20

- Avoid asking the same item multiple times.
  - "I' m generally sad about my life."
  - "My life is generally sad."

Graded Response Model Parameters for Global Physical Health

Item	a	b1	b2	b3	b4
Global01	7.37 (na)	-1.98 (na)	-0.97 (na)	0.03 (na)	1.13 (na)
Global03	7.65 (2.31)	-1.89 (-2.11)	-0.86 (-0.89)	0.15 (0.29)	1.20 (1.54)
Global06	1.86 (2.99)	-3.57 (-2.80)	-2.24 (-1.78)	-1.35 (-1.04)	-0.58 (-0.40)
Global07	1.13 (1.74)	-5.39 (-3.87)	-2.45 (-1.81)	-0.98 (-0.67)	1.18 (1.00)
Global08	1.35 (1.90)	-4.16 (-3.24)	-2.39 (-1.88)	-0.54 (-0.36)	1.31 (1.17)

Note: Parameter estimates for 5-item scale are shown first, followed by estimates for 4-item scale (in parentheses). na = not applicable

a = discrimination parameter; b1 = 1st threshold; b2 = 2nd threshold; b3 = 3rd threshold; b4 = 4th threshold

Global01: In general, would you say your health is …?

Global03: In general, how would you rate your physical health?

Global06: To what extent are you able to carry out your everyday physical activities?

Global07: How would you rate your pain on average?

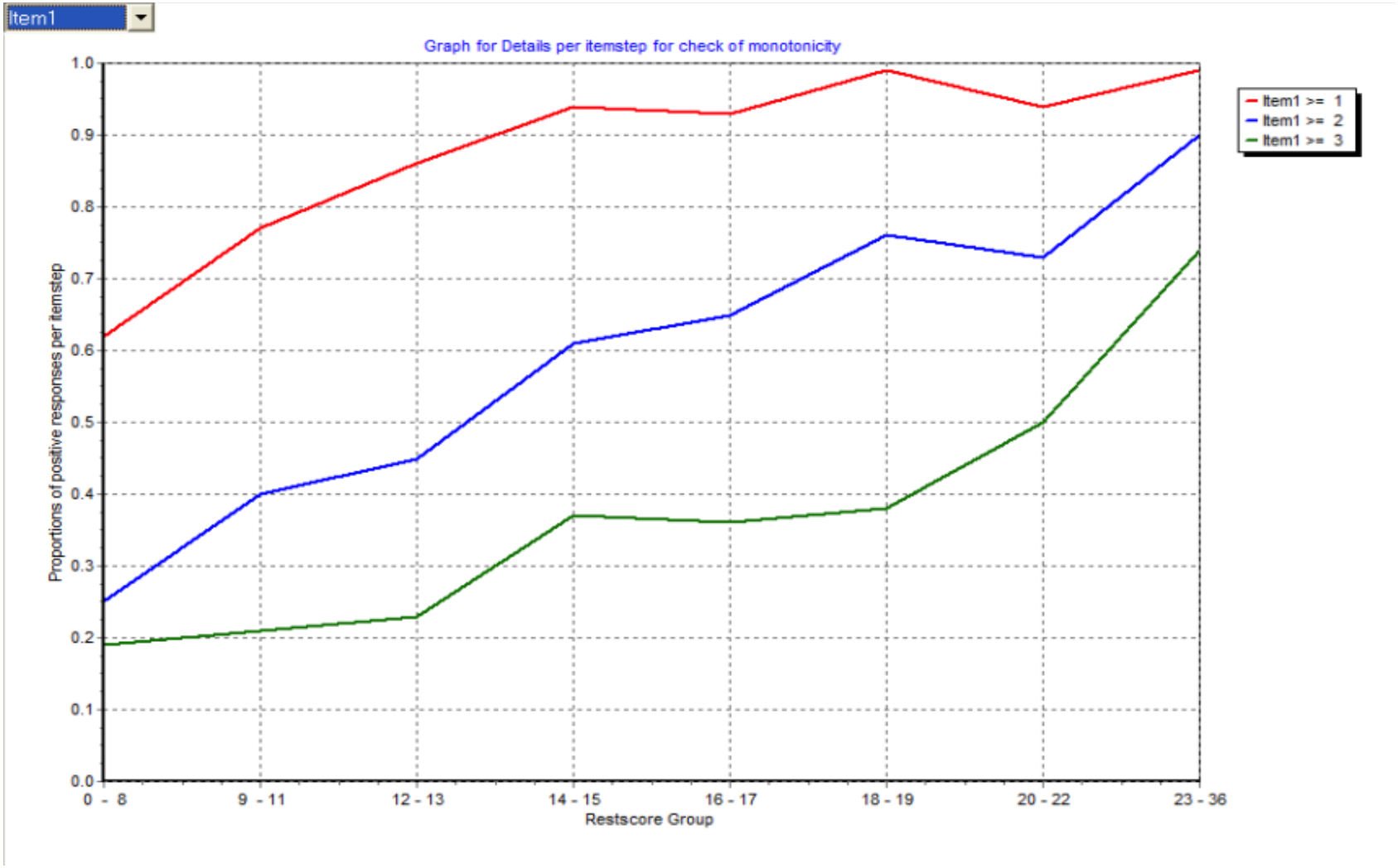
Global08: How would you rate your fatigue on average?

Monotonicity

- Probability of selecting a response category indicative of better health should increase as underlying health increases.

- Item response function graphs with
  - y-axis: proportion positive for item step
  - x-axis: raw scale score minus item score

Check of Monotonicity



IRT Model Fit

- Compare observed and expected response frequencies by item and response category
- Items that do not fit and less discriminating items identified and reviewed by content experts

Person Fit

- Large negative Z_L values indicate misfit.
  - Z_L has expected value of zero, with variance of one if responses are consistent with IRT model. Large negative Z_L values (>= -2.0) indicate misfit.
  -
- Person who responded to 14 PROMIS physical functioning items had a Z_L = -3.13
  - For 13 items the person could do the activity (including running 5 miles) *without any difficulty*.
  - But this person reported *a little difficulty* being out of bed for most of the day.

Person Fit

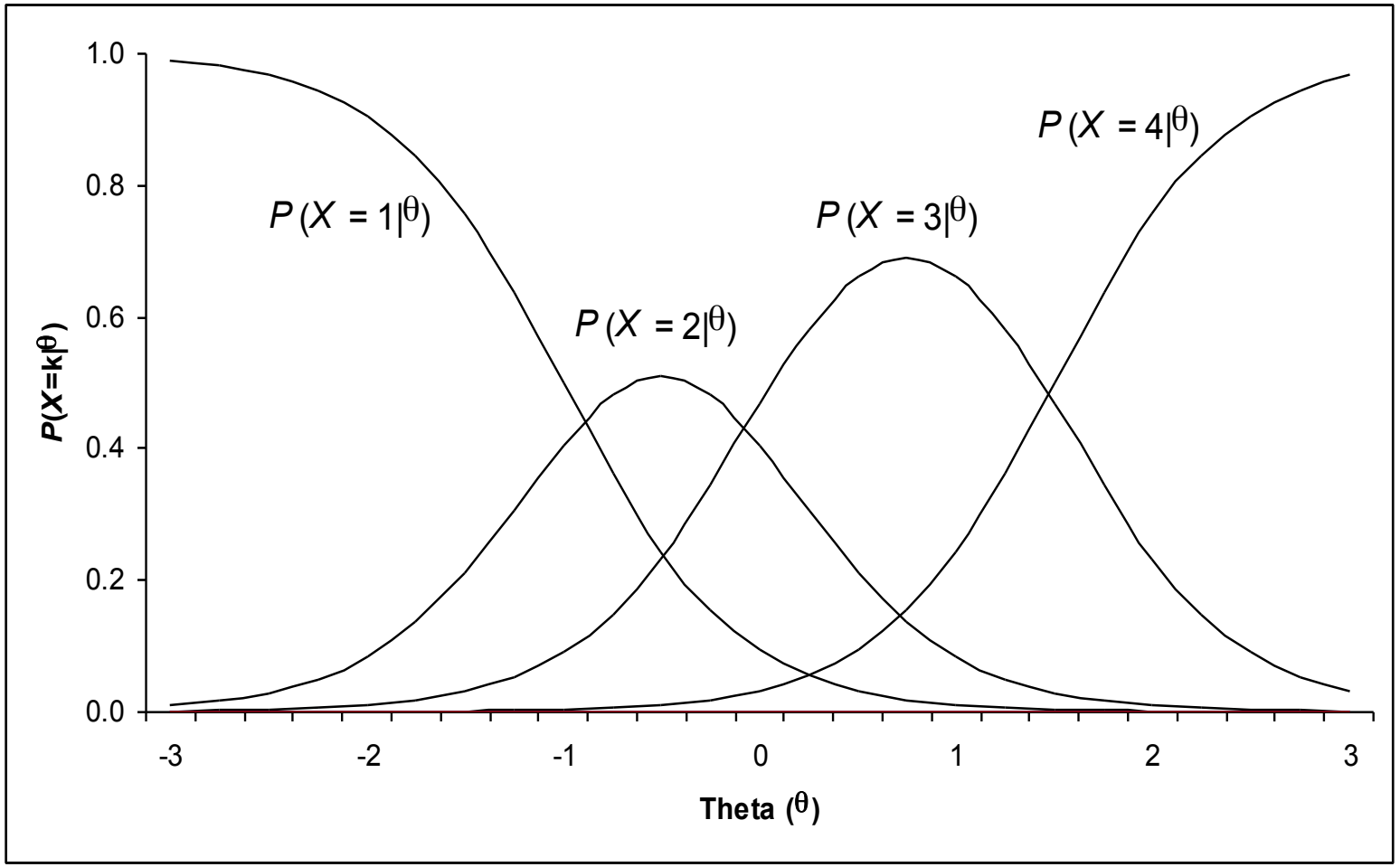
Item misfit significantly associated with

- *Less than high school education*
- *More chronic conditions*
- *Longer response time*

Nice Features of IRT

- Category response curves (CRCs)
- Information depending on theta
- Assessing differential item functioning

Samejima's Graded Response Model (Category Response Curves)



Posttraumatic Growth Inventory

Indicate for each of the statements below the degree to which this change occurred in your life as a result of your crisis.

(*Appreciating each day*)

(0) I did not experience this change as result of my crisis

(1) I experienced this change to a very small degree as a result of my crisis

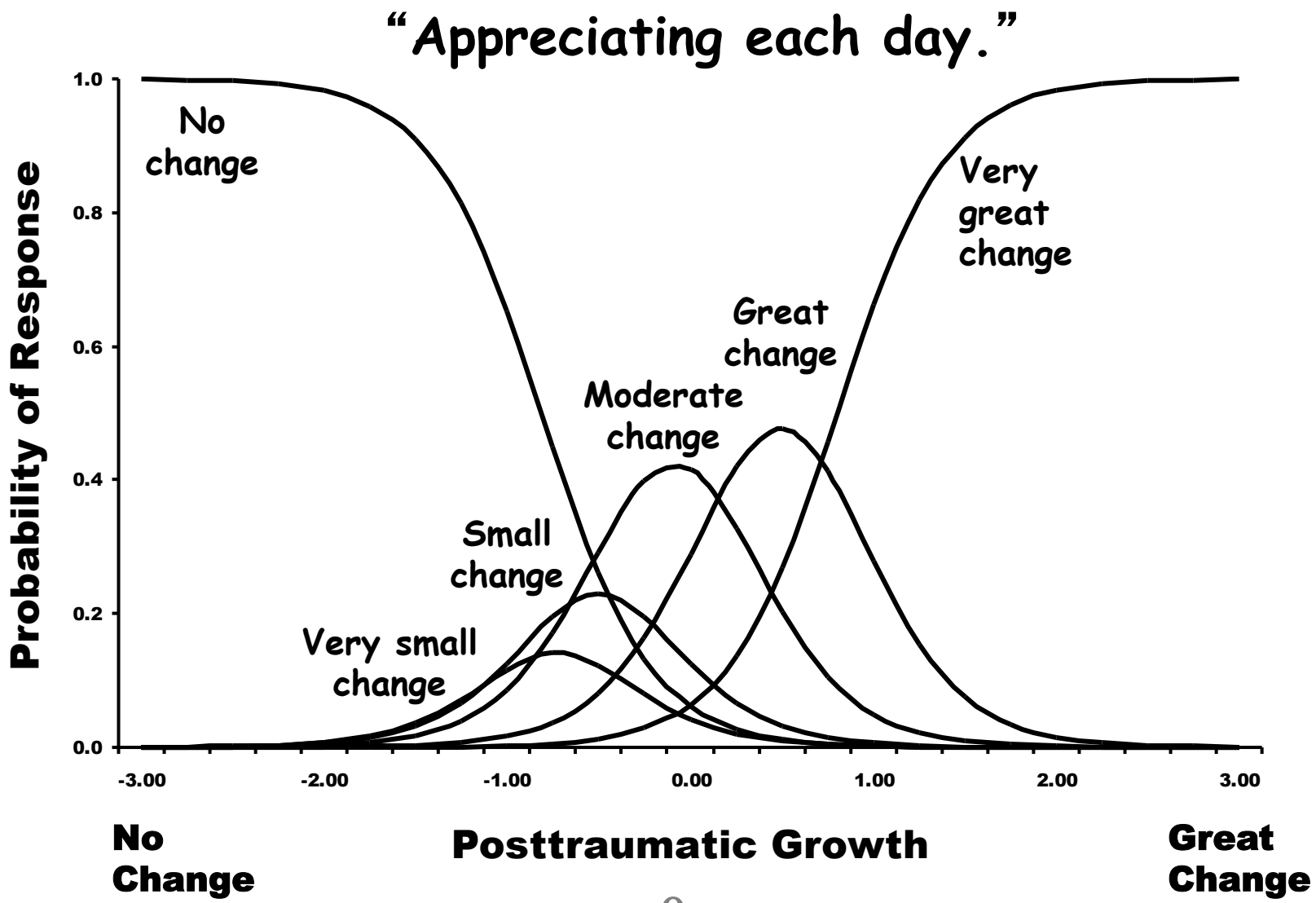
(2) I experienced this change to a small degree as a result of my crisis

(3) I experienced this change to a moderate degree as a result of my crisis

(4) I experienced this change to a great degree as a result of my crisis

(5) I experienced this change to a very great degree as a result of my crisis

Category Response Curves



Drop Response Options?

Indicate for each of the statements below the degree to which this change occurred in your life as a result of your crisis. (*Appreciating each day*)

(0) I did not experience this change as result of my crisis

(1) I experienced this change to a moderate degree as a result of my crisis

(2) I experienced this change to a great degree as a result of my crisis

(3) I experienced this change to a very great degree as a result of my crisis

Reword response options?

- Might be challenging to determine what alternative wording to use so that the replacements are more likely to be endorsed.

Keep as is?

- CAHPS global rating items
  - 0 = worst possible
  - 10 = best possible
- 11 response categories capture about 3 levels of information.
  - 10/9/8-0 or 10-9/8/7-0
- Scale is administered as is and then collapsed in analysis

Response Burden vs. Standard Error (SE)

- 3-5 items per minute rule of thumb for paper survey
  - 8 items per minute for dichotomous items
- Lowering SE means adding or replacing existing items with more informative ones at the target range of the continuum.

Reliability and Information

- Only as much response burden as needed for target level of reliability

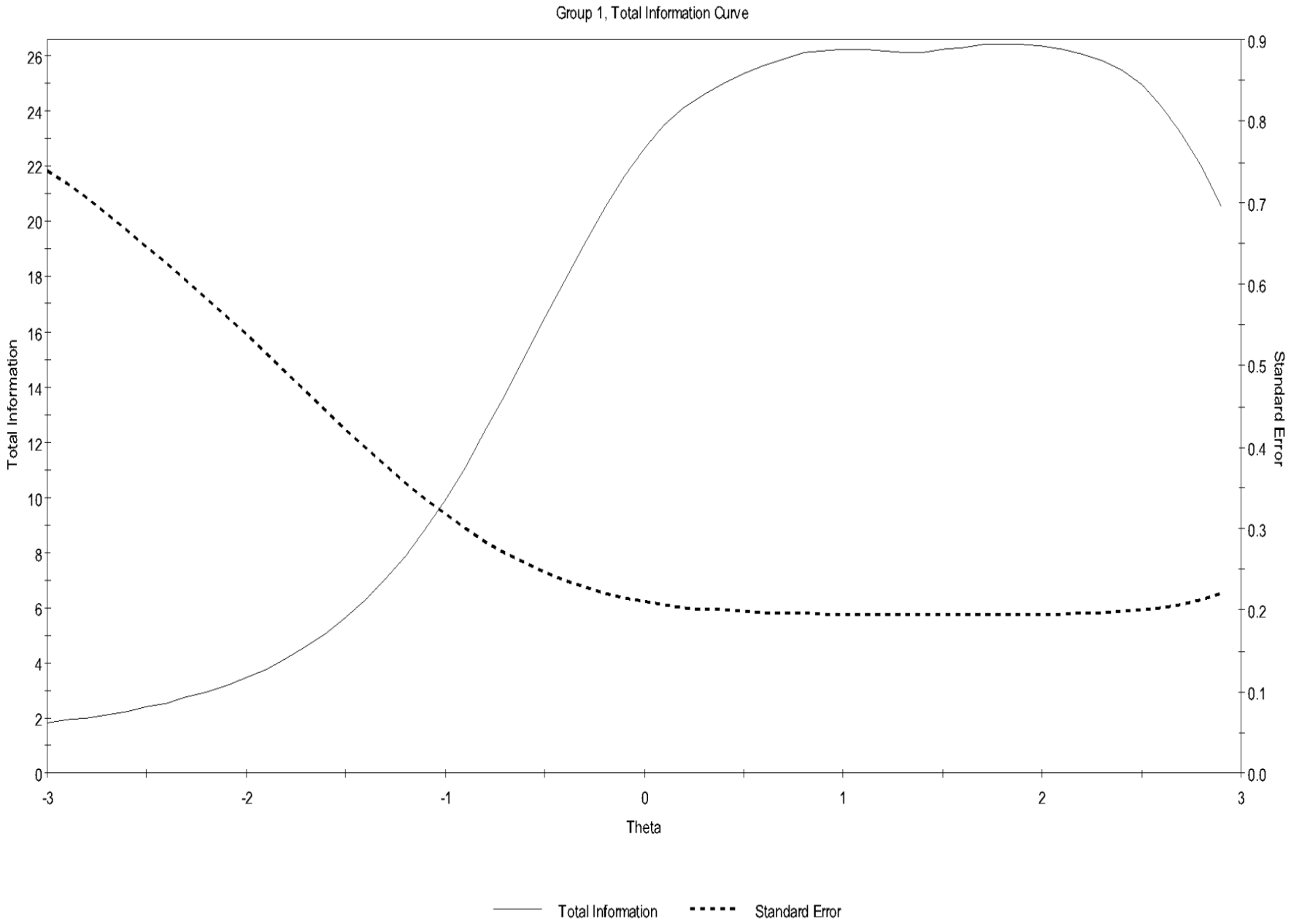
- CATs for patient-reported outcomes yield 0.90 reliability with about 5-6 items

- For z-scores (mean = 0 and SD = 1):
  - Reliability = $1 - SE^2$ = **0.90** (when SE = 0.32)
  - Information = $1/SE^2$ = **10** (when SE = 0.32)
  - Reliability = 1 – 1/information

Gastroesophageal Reflux

1. How often did you have regurgitation—that is, food or liquid coming back up into your throat or mouth without vomiting?
2. What was the most food or liquid you had come back into your mouth at one time?
3. During the time you were awake, how often did you gag on liquid or food coming back up into your throat?
4. How often were you awakened from sleep by choking on liquid or food coming back up into your throat?
5. How much did regurgitation bother you?
6. After eating a meal, how often did food or liquid come back into your throat or mouth without vomiting?
7. How often did you re-swallow food that came back into your throat?
8. How often did you feel like you were going to burp, but food or liquid came up instead?
9. How often did you feel like there was too much saliva in your mouth?
10. How frequently did you feel burning in the red area show in the picture—that is, behind the breastbone?
11. At its worst, how bad was the pain behind your breastbone?
12. How often did you know that you would have pain behind your breastbone before it happened?
13. How much did pain behind the breastbone interfere with your day-to-day activities?
14. How much did pain behind the breastbone bother you?
15. When you had pain behind the breastbone, how long did it usually last?
16. How often did you feel burning in your throat?
17. How often did you burp?
18. How much did burping bother you?
19. How often did you have hiccups?
20. How much did hiccups bother you?
21. How often did you feel like there was a lump in your throat?
22. How much did having a lump in your throat bother you?

Test Information for Reflux

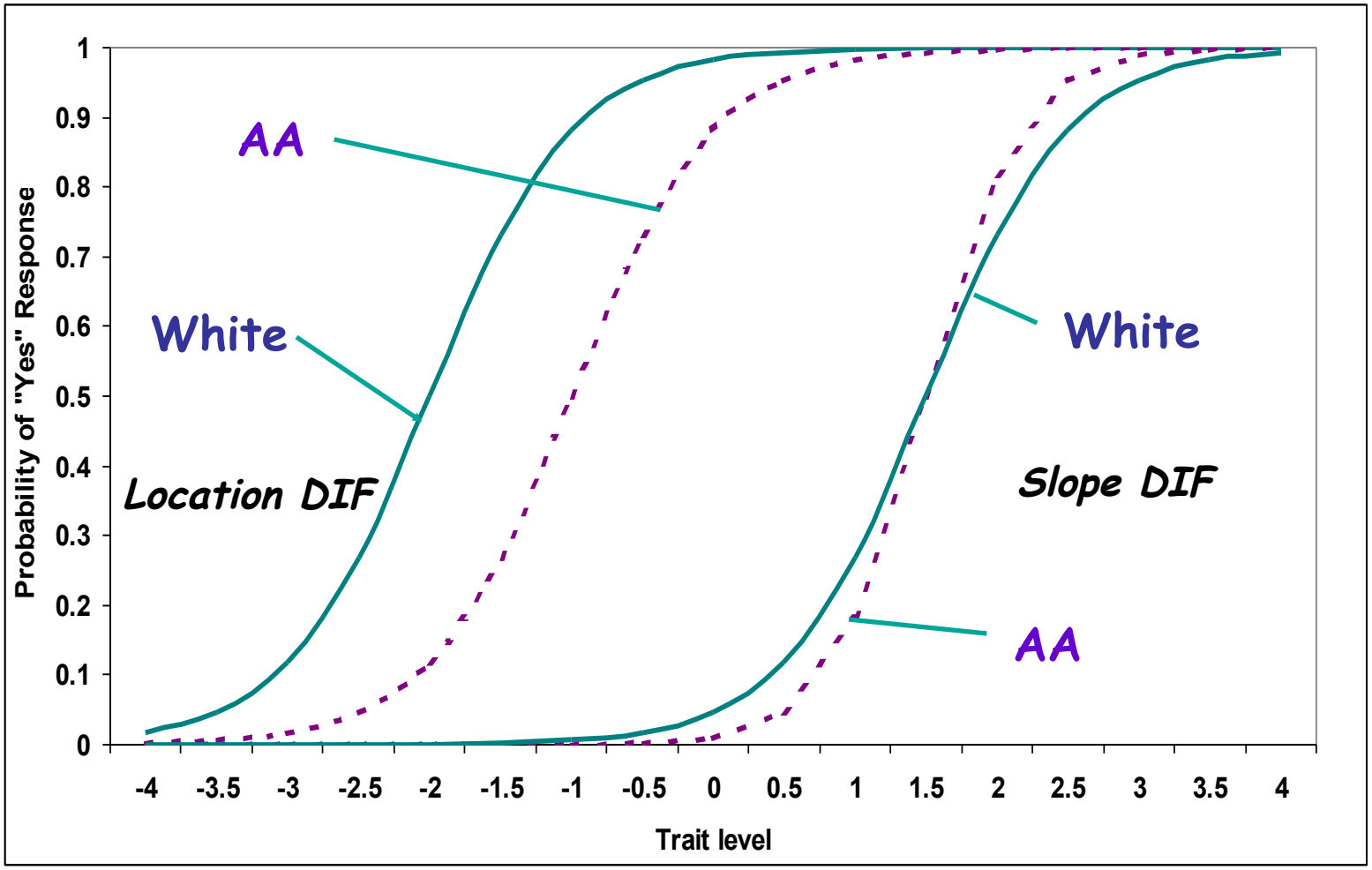


Differential Item Functioning (DIF)

- 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
  - Different subgroups
  - Mode differences

DIF (2-parameter model)



Location = uniform; Slope = non-uniform

Suggested Readings

- Cella, D., et al. (2010). Initial item banks and first wave testing of the Patient-Reported Outcomes Measurement Information System (PROMIS) network: 2005-2008. Journal of Clinical Epidemiology, 63 (11), 1179-1194.
- Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. New Jersey: Erlbaum.
- Gorman, S. Computerized adaptive testing with a military population. http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/go77221.pdf
- Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.
- Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st Century. Medical Care, 38, II-28-42.

Questions?

Contact Information:
drhays@ucla.edu (310-794-2294)

Powerpoint file available for downloading at:
http://gim.med.ucla.edu/FacultyPages/Hays/

Acknowledgment of Support

University of California, Los Angeles, Resource Center for Minority Aging Research (RCMAR)/ Center for Health Improvement in Minority Elderly (CHIME), NIH/NIA Grant P30-AG021684.