Multi-item Scale Evaluation

Ron D. Hays, Ph.D.
UCLA Division of General Internal Medicine/Health Services Research
drhays@ucla.edu
http://twitter.com/RonDHays
http://gim.med.ucla.edu/FacultyPages/Hays/

Questionnaire Design and Testing Workshop
11-18-11: 3-5 pm, Broxton 2nd Floor Conference Room

Responses of 5 People to 2 Items

ID	Poor (1)	Fair (2)	Good (3)	Very Good (4)	Excellent (5)
01					2
02				1	1
03		1		1	
04			1		1
05		2			

Cronbach's Alpha

01	55
02	45
03	42
04	35
05	22

Source	df	SS	MS
Respondents (BMS)	4	11.6	2.9
Items (JMS)	1	0.1	0.1
Resp. x Items (EMS)	4	4.4	1.1
Total	9	16.1	

$$\text{Alpha} = \frac{2.9 - 1.1}{2.9} = \frac{1.8}{2.9} = \boxed{0.62}$$

Computations

- Respondents SS

  $(10^2+9^2+6^2+8^2+4^2)/2 - 37^2/10 = \underline{11.6}$

- Item SS

  $(18^2+19^2)/5 - 37^2/10 = \underline{0.1}$

- Total SS

  $(5^2+ 5^2+4^2+5^2+4^2+2^2+3^2+5^2+2^2+2^2) - 37^2/10 = \underline{16.1}$

- Res. x Item SS= Tot. SS – (Res. SS+Item SS)

Reliability Minimum Standards

- 0.70 or above (for group comparisons)
- 0.90 or higher (for individual assessment)
  - SEM = SD (1- reliability)$^{1/2}$

Alpha for Different Numbers of Items and Average Correlation

Number of Items (k)	Average Inter-item Correlation ($\bar{r}$)					
	.0	.2	.4	.6	.8	1.0
2	0.00	0.33	0.57	0.75	0.89	1.00
4	0.00	0.50	0.73	0.86	0.94	1.00
6	0.00	0.60	0.80	0.90	0.96	1.00
8	0.00	0.67	0.84	0.92	0.97	1.00

$$\text{Alpha}_{st} = \frac{k * \bar{r}}{1 + (k - 1) * \bar{r}}$$

Intraclass Correlation and Reliability

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

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

Spearman-Brown Prophecy Formula

$$\text{alpha}_{y} = \left(\frac{N \bullet \text{alpha}_{x}}{1 + (N - 1) * \text{alpha}_{x}} \right)$$

N = how much longer scale y is than scale x

Clark, E. L. (1935). Spearman-Brown formula applied to ratings of personality traits. Journal of Educational Psychology, 26, 552-555.

Example Spearman-Brown Calculation

MHI-18

$$\frac{18/32\ (0.98)}{(1+(18/32\ -1)*0.98}$$

$$= 0.55125/0.57125 = 0.96$$

Spearman-Brown Estimates of Sample Needed for 0.70 Health-Plan Reliability

- Plan-level reliability estimates were significantly lower for African Americans than whites
  - Getting care quickly (118 vs. 82)
  - Getting needed care (110 vs. 76)
  - Provider communication (177 vs. 124)
  - Office staff courtesy (128 vs. 121)
  - Plan customer service (98 vs. 68)

M. Fongwa et al. (2006). Comparison of data quality for reports and ratings of ambulatory care by African American and White Medicare managed care enrollees. Journal of Aging and Health, 18, 707-721.

Item-scale correlation matrix

	Depress	Anxiety	Anger
Item #1	0.80*	0.20	0.20
Item #2	0.80*	0.20	0.20
Item #3	0.80*	0.20	0.20
Item #4	0.20	0.80*	0.20
Item #5	0.20	0.80*	0.20
Item #6	0.20	0.80*	0.20
Item #7	0.20	0.20	0.80*
Item #8	0.20	0.20	0.80*
Item #9	0.20	0.20	0.80*

*Item-scale correlation, corrected for overlap.

Item-scale correlation matrix

	Depress	Anxiety	Anger
Item #1	0.50*	0.50	0.50
Item #2	0.50*	0.50	0.50
Item #3	0.50*	0.50	0.50
Item #4	0.50	0.50*	0.50
Item #5	0.50	0.50*	0.50
Item #6	0.50	0.50*	0.50
Item #7	0.50	0.50	0.50*
Item #8	0.50	0.50	0.50*
Item #9	0.50	0.50	0.50*

*Item-scale correlation, corrected for overlap.

Patient Satisfaction Ratings in Medical Outcomes Study

	Tech.	Interp.	Comm.
Item #1	0.66*	0.63	0.67
Item #2	0.55*	0.54	0.50
Item #3	0.48*	0.41	0.44
Item #4	0.58	0.68*	0.63
Item #5	0.59	0.58*	0.61
Item #6	0.62	0.65*	0.67
Item #7	0.58	0.59	0.61*
Item #8	0.47	0.50	0.50*
Item #9	0.58	0.66	0.63*

***Item-scale correlation, corrected for overlap.**

Confirmatory Factor Analysis

- Factor loadings and correlations between factors
- Observed covariances compared to covariances generated by hypothesized model
    - Statistical and practical tests of fit

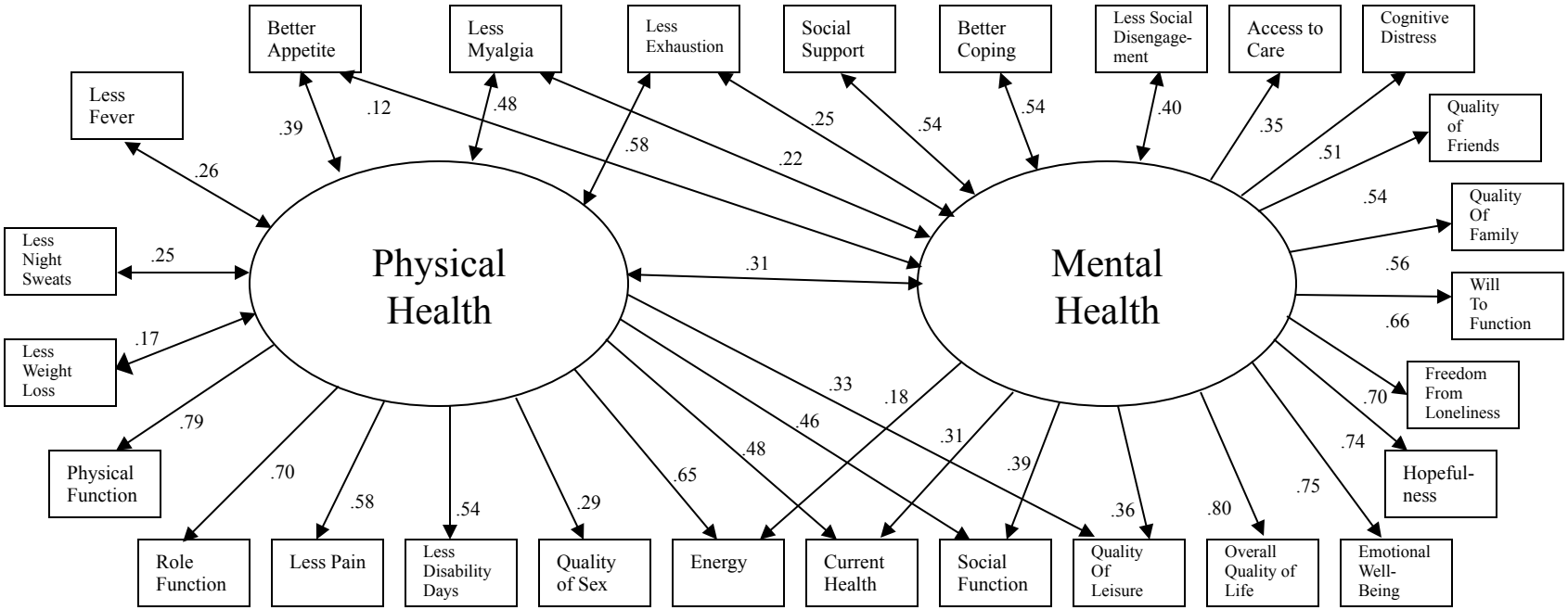
Fit Indices

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

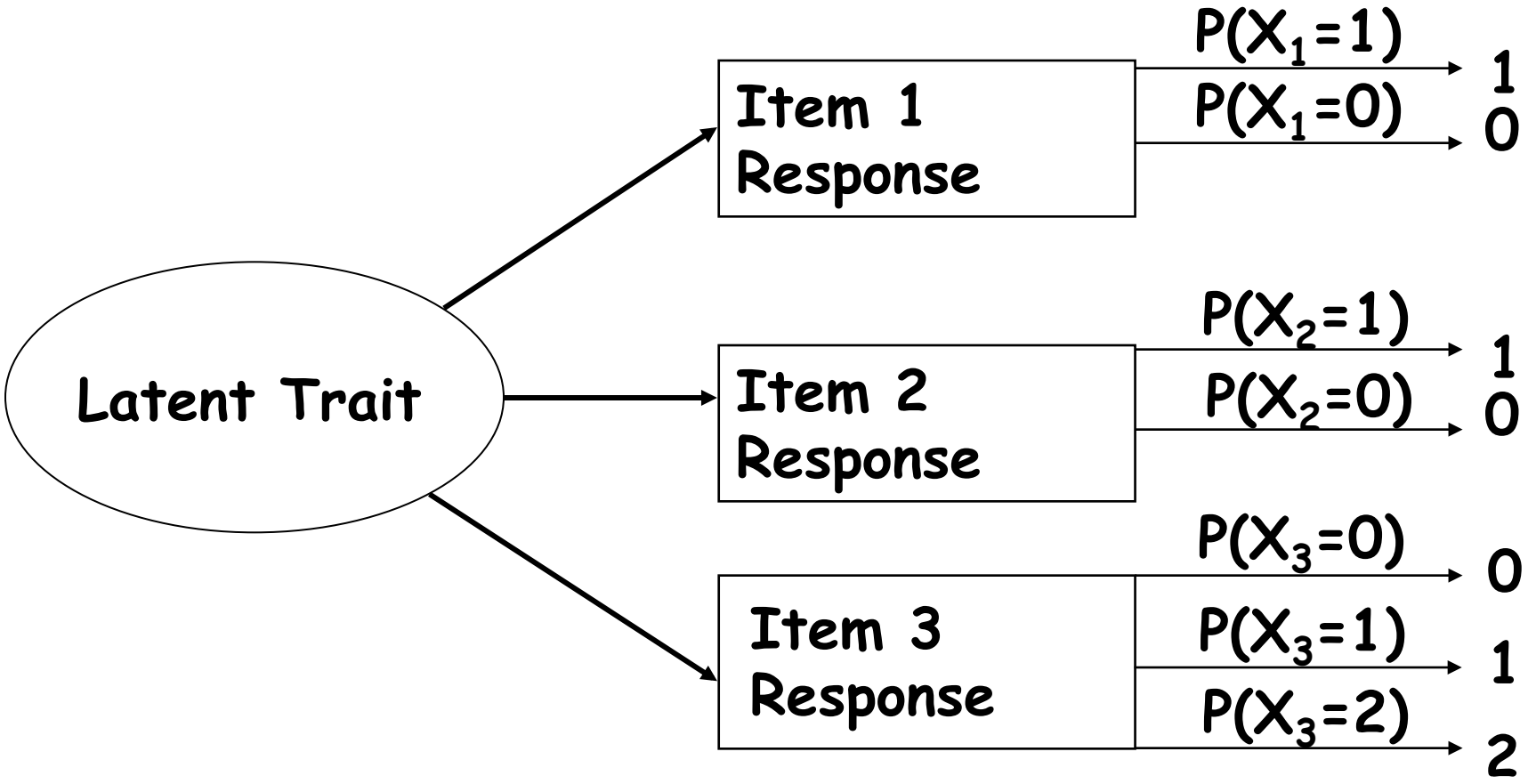
Hays, Cunningham, Ettl, Beck & Shapiro (1995, Assessment)

- 205 symptomatic HIV+ individuals receiving care at two west coast public hospitals
- 64 HRQOL items plus
  - 9 access, 5 social support, 10 coping, 4 social engagement and 9 HIV symptom items

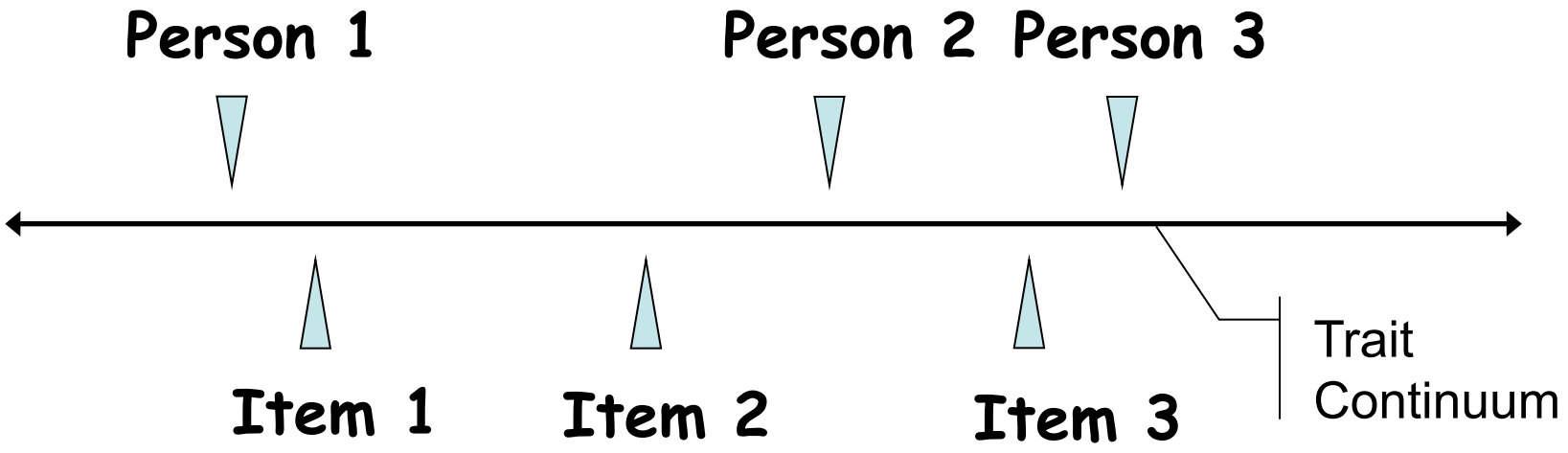
Confirmatory Factor Analysis Model of Physical and Mental Health



Latent Trait and Item Responses



Item Responses and Trait Levels



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 for the item (i) to have a score of k or more and the discrimination parameter a_i estimates the discriminatory power of the item.

If for one group versus another at the same level θ we observe systematically different probabilities of scoring k or above then we will say that item i displays DIF

Important IRT Features

- Category response curves
- Information/reliability
- Differential item functioning
- Person fit
- Computer-adaptive testing

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

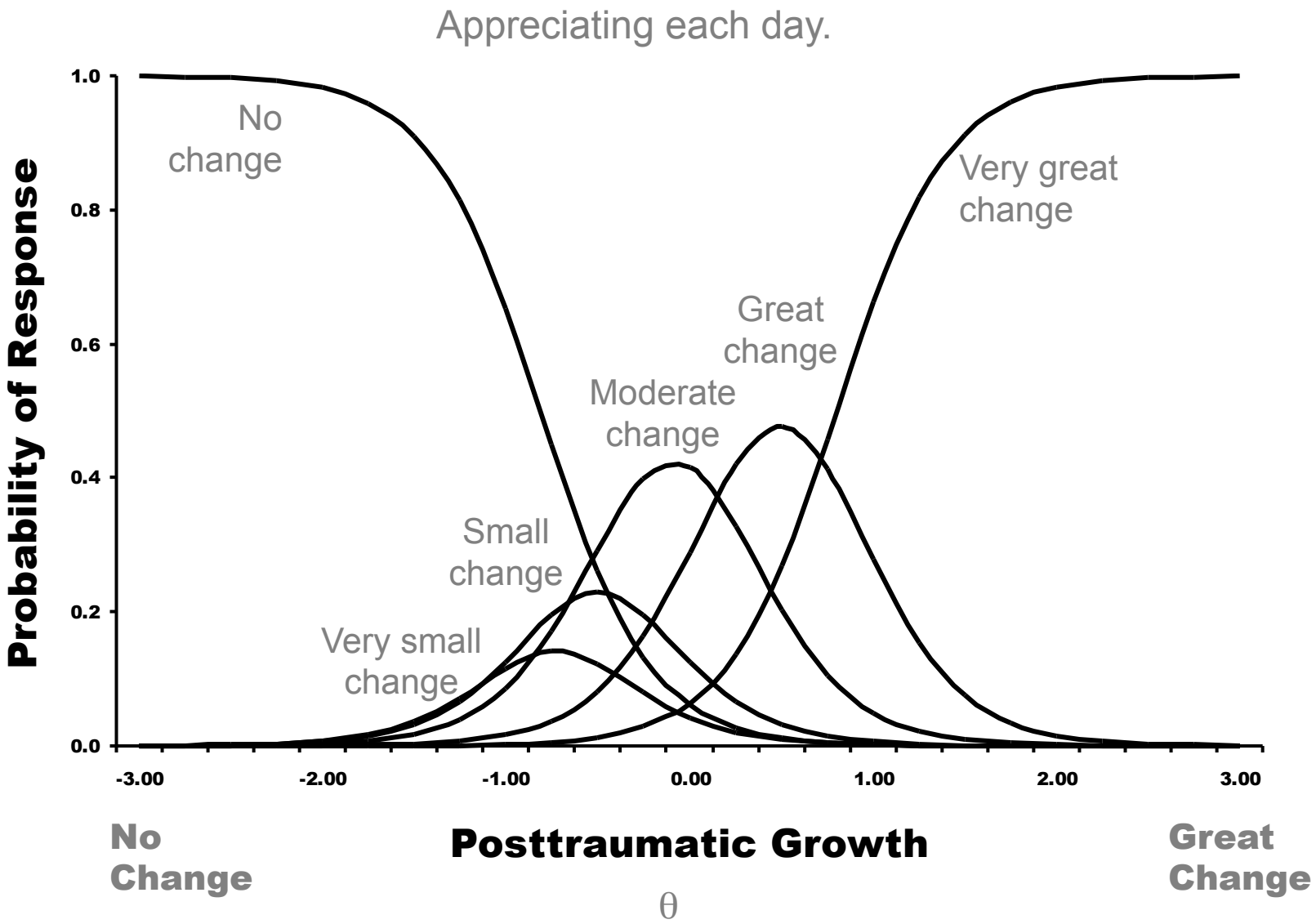
(2) I experienced this change to a small degree as a result of my crisis

(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



Category Response Curves (CRCs)

- Figure shows that 2 of 6 response options are never most likely to be chosen
  - No, **very small**, **small**, moderate, great, very great change

- One might suggest 1 or both of the response categories could be dropped or reworded to improve the response scale

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?

- 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

Information/Reliability

- 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

- Lowering the SE requires adding or replacing existing items with more informative items at the target range of the continuum.
  - But this is …

Easier said than done

- Limit on the number of ways to ask about a targeted range of the construct
- One needs to avoid asking the same item multiple times.
  - “I’m generally said about my life.”
  - “My life is generally sad.”
- Local independence assumption
  - Significant residual correlations

Item parameters (graded response model) for global physical health items in Patient-Reported Outcomes Measurement Information System

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

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?

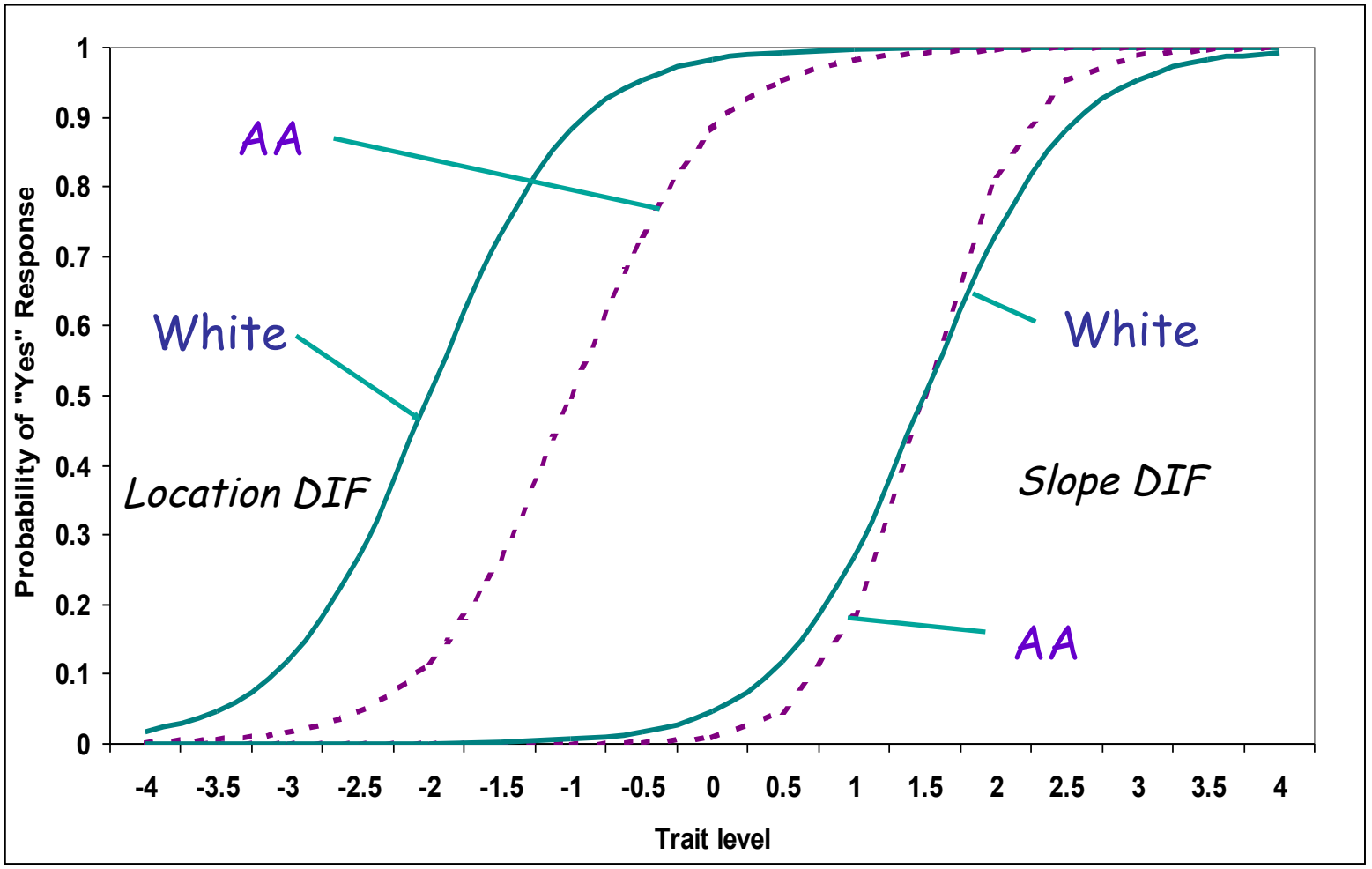
a = discrimination parameter; b1 = 1st threshold; b2 = 2nd threshold; b3 = 3rd threshold; b4 = 4th threshold

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 their other characteristics

- Evaluation of DIF
  - Different subgroups
  - Mode differences

Differential Item Functioning (2-Parameter Model)



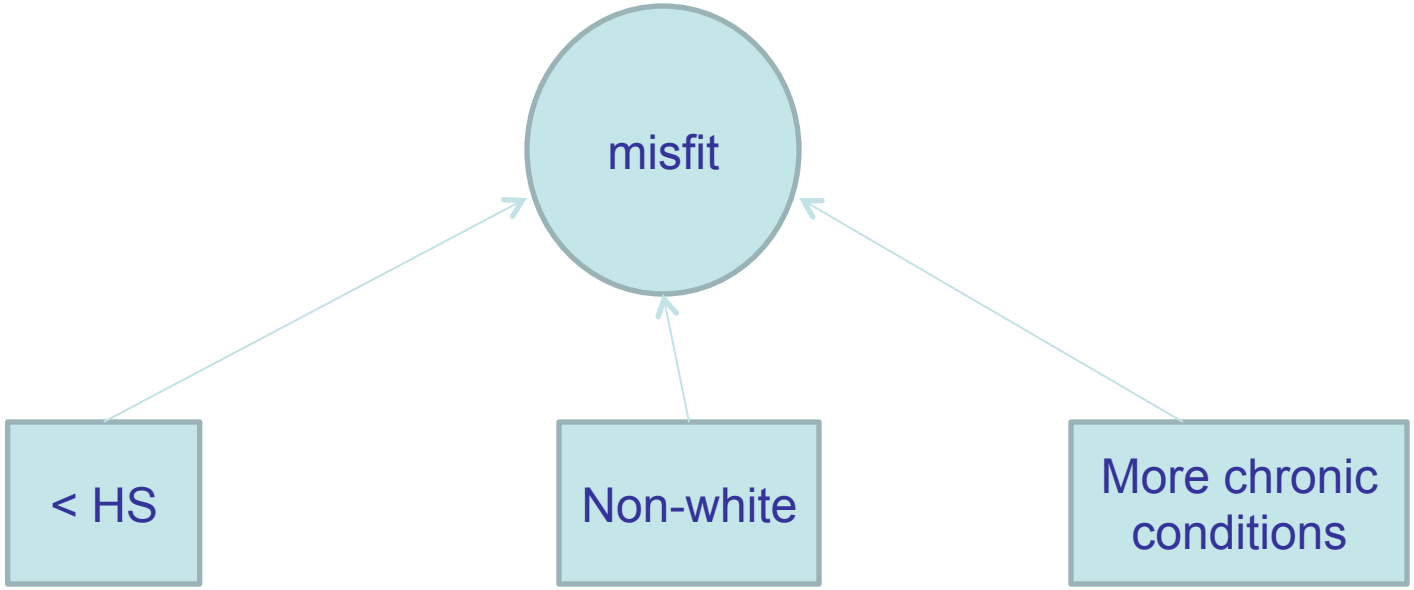
Location = uniform; Slope = non-uniform

Person Fit

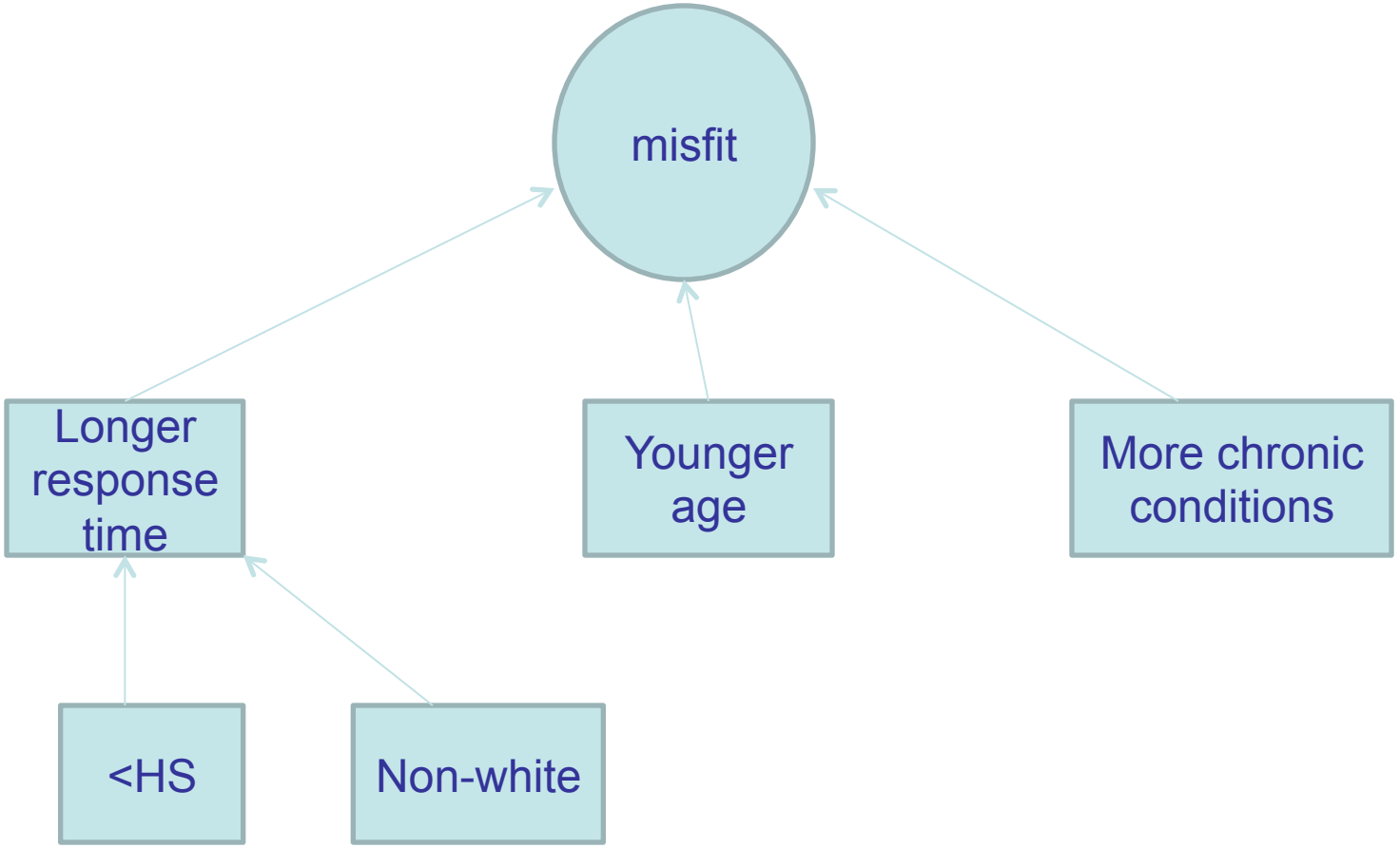
- Large negative Z_L values indicate misfit.

- Person responded to 14 items in physical functioning bank (Z_L = -3.13)
  - For 13 items the person could do the activity (including running 5 miles) without any difficulty.
  - However, this person reported *a little difficulty* being out of bed for most of the day.

Unique Associations with Person Misfit



Unique Associations with Person Misfit



Computer Adaptive Testing http://www.nihpromis.org/

- Patient-reported outcomes measurement information system (PROMIS) project
  - Item banks measuring patient-reported outcomes
  - Computer-adaptive testing (CAT) system

.

PROMIS Banks

- Emotional Distress
  - Depression (28)
  - Anxiety (29)
  - Anger (29)
- Physical Function (124)
- Pain
  - Behavior (39)
  - Impact (41)
- Fatigue (95)
- Satisfaction with Participation in Discretionary Social Activities (12)
- Satisfaction with Participation in Social Roles (14)
- Sleep Disturbance (27)
- Wake Disturbance (16)

Time to complete item

- 3-5 items per minute rule of thumb
  - 8 items per minute for dichotomous items
- Polimetrix panel sample
  - 12-13 items per minute (automatic advance)
  - 8-9 items per minute (next button)
- 6 items per minute among UCLA Scleroderma patients

Anger CAT (In the past 7 days)

I was grouchy [1st question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 56.1 SE = 5.7

In the past 7 days …

I felt like I was read to explode [2nd question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 51.9 SE = 4.8

In the past 7 days …

I felt angry [3rd question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 50.5 SE = 3.9

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

In the past 7 days …

I felt annoyed [5th question]

- Never
- Rarely
- Sometimes
- Often
- Always

- Theta = 50.1 SE = 3.2

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

Theta and SE estimates

- 56 and 6
- 52 and 5
- 50 and 4
- 49 and 4
- 50 and 3
- 50 and <3

CAT

- Context effects (Lee & Grant, 2009)
  - 1,191 English and 824 Spanish respondents to 2007 California Health Interview Survey
  - Spanish respondents self-rated health was worse when asked before compared to after questions about chronic conditions.

Language DIF Example

- Ordinal logistic regression to evaluate differential item functioning
  - Purified IRT trait score as matching criterion
  - McFadden's pseudo R^2 >= 0.02
- Thetas estimated in Spanish data using
  - English calibrations
  - Linearly transformed Spanish calibrations (Stocking-Lord method of equating)

Lordif

http://CRAN.R-project.org/package=lordif

Model 1 : logit $P(u_i >= k) = \alpha_k + \mathbf{\beta_1} * \mathbf{ability}$

Model 2 : logit $P(u_i >= k) = \alpha_k + \beta_1 *$ ability $+ \mathbf{\beta_2} * \mathbf{group}$

Model 3 : logit $P(u_i >= k) = \alpha_k + \beta_1 *$ ability $+ \beta_2 *$ group $+ \mathbf{\beta_3} * \mathbf{ability} * \mathbf{group}$

DIFF assessment (log likelihood values compared):

- Overall: Model 3 versus Model 1
- Non-uniform: Model 3 versus Model 2
- Uniform: Model 2 versus Model 1

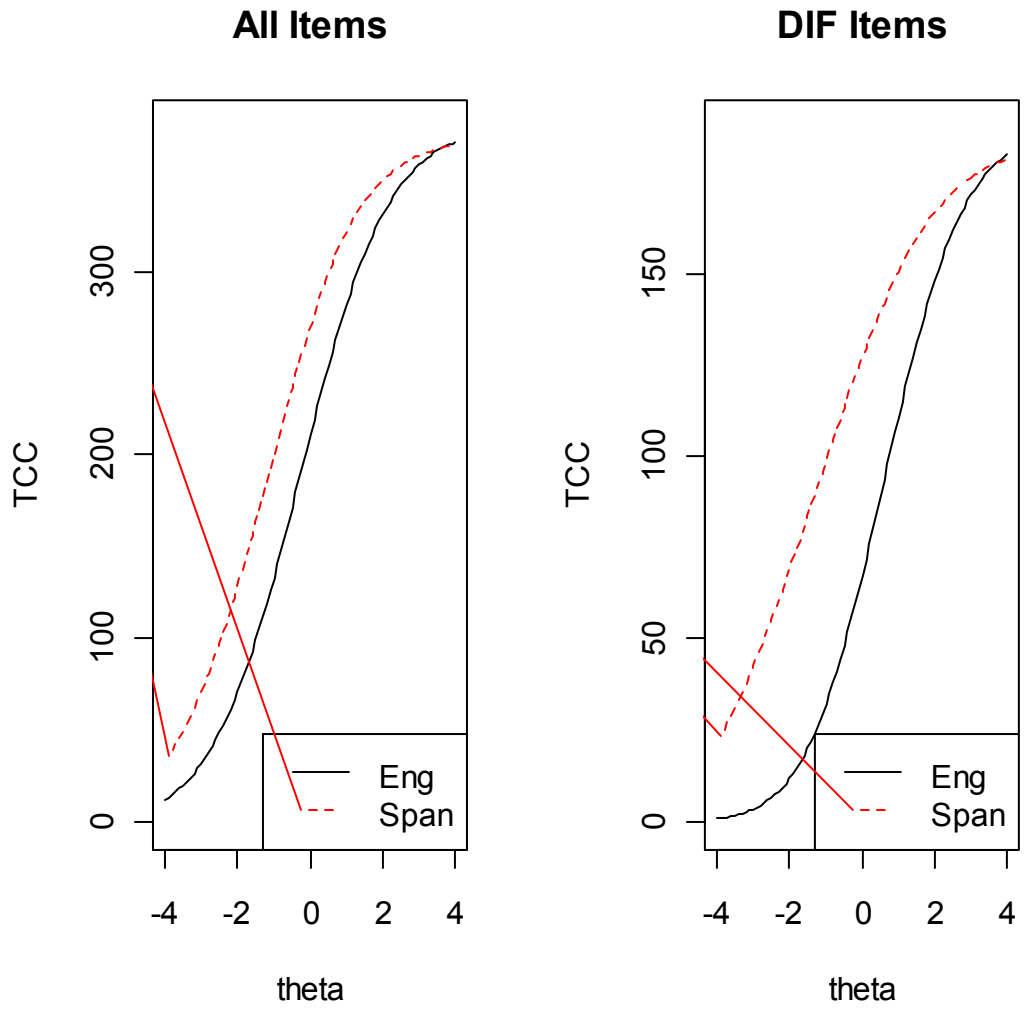
Sample Demographics

	English (n = 1504)	Spanish (n = 640)
% Female	52%	58%
% Hispanic	11%	100%
Education		
< High school	2%	14%
High school	18%	22%
Some college	39%	31%
College degree	41%	33%
Age	51 (SD = 18)	38 (SD = 11)

Results

- One-factor categorical model fit the data well (CFI=0.971, TLI=0.970, and RMSEA=0.052).
  - Large residual correlation of 0.67 between "Are you able to run ten miles" and "Are you able to run five miles?"
- 50 of the 114 items had language DIF
  - 16 uniform
  - 34 non-uniform

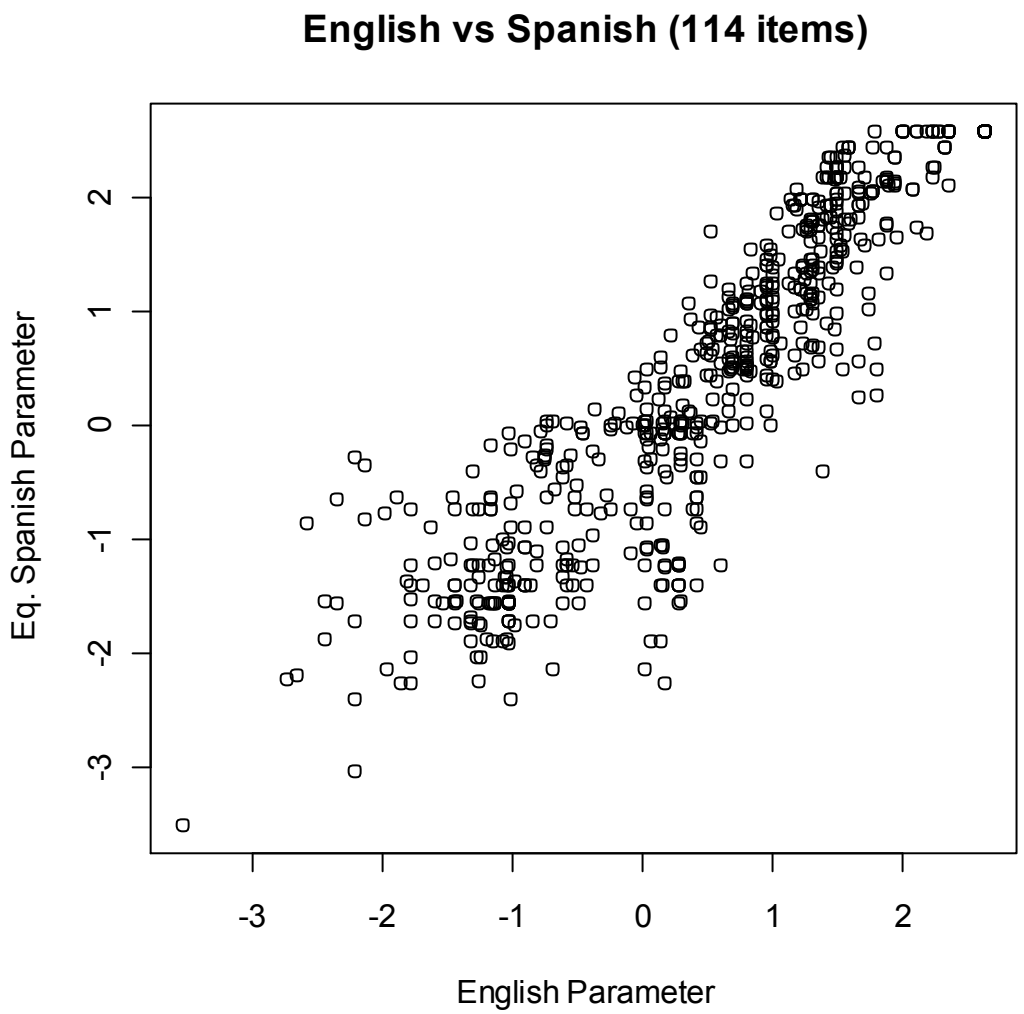
Impact of DIF on Test Characteristic Curves (TCCs)



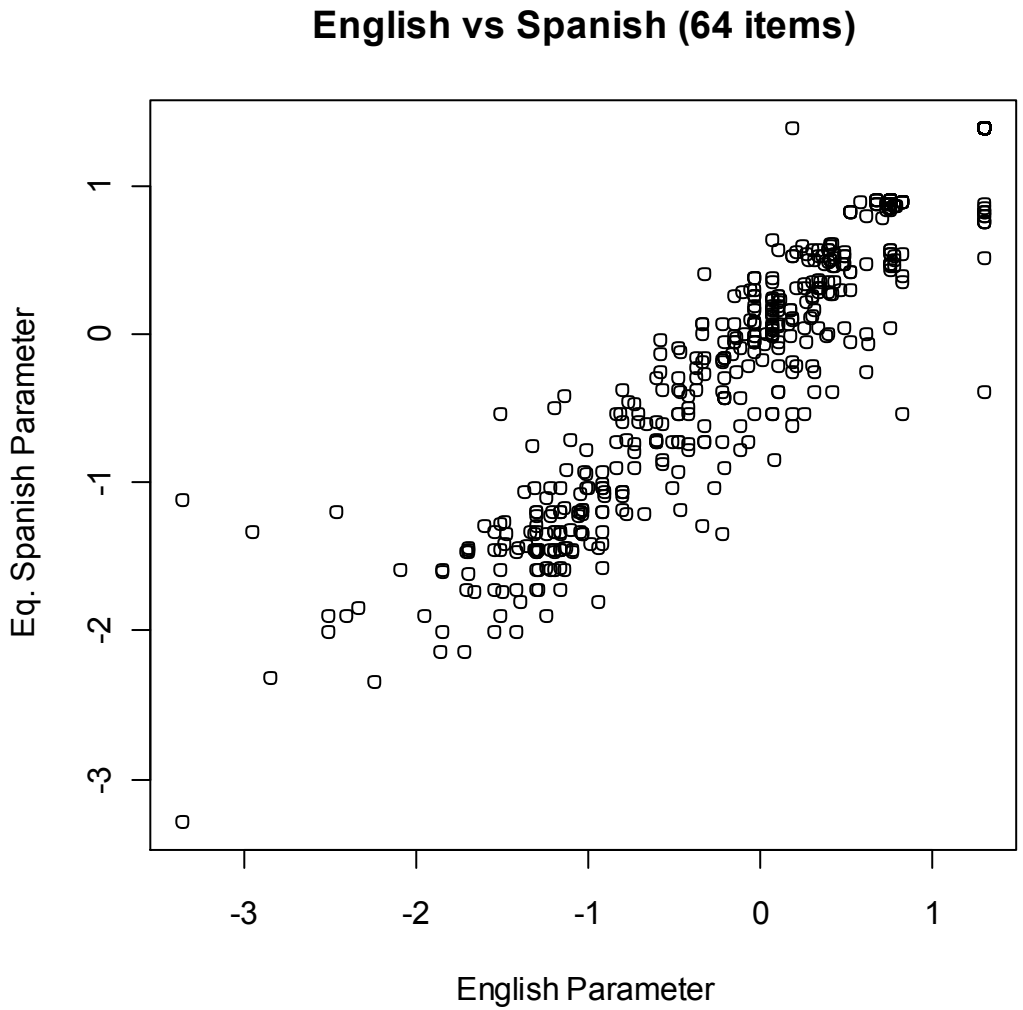
Stocking-Lord Method

- Spanish calibrations transformed so that their TCC most closely matches English TCC.
- $a^* = a/A$ and $b^* = A * b + B$
- Optimal values of A (slope) and B (intercept) transformation constants found through multivariate search to minimize weighted sum of squared distances between TCCs of English and Spanish transformed parameters
  - Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

CAT-based Theta Estimates Using English (x-axis) and Spanish (y-axis) Parameters for 114 Items in Spanish Sample (n = 640, ICC = 0.89)



CAT-based Theta Estimates Using English (x-axis) and Spanish (y-axis) Parameters for 64 non-DIF Items in Spanish Sample (n = 640, ICC = 0.96)



Thank you.