Evaluating Health-Related Quality of Life Measures

June 12, 2014 (1:00 – 2:00 PDT)

Kaiser Methods Webinar Series

Ron D.Hays, Ph.D.
drhays@ucla.edu

Burden of Kidney Disease Scale

How true or false is each of the following statements for you?

1. My kidney disease interferes too much with my life.
2. Too much of my time is spent dealing with kidney disease.
3. I feel frustrated dealing with my kidney disease.
4. I feel like a burden on my family.

- Definitely True = 100
- Mostly True = 75
- Don't Know = 50
- Mostly false = 25
- Definitely false = 0

FDA PRO Process
(released February 2006)

Aspects of Good Health-Related Quality of Life Measures

Aside from being practical..

1. Same people get same scores
2. Different people get different scores and differ in the way you expect
3. Measure is interpretable
4. Measure works the same way for different groups (age, gender, race/ethnicity)

Aspects of Good Health-Related Quality of Life Measures

Aside from being practical..

1. Same people get same scores— Reliable
2. Different people get different scores and differ in the way you expect
3. Measure is interpretable
4. Measure works the same way for different groups (age, gender, race/ethnicity)

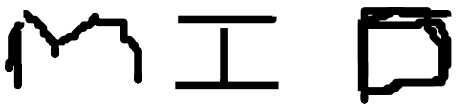
Aspects of Good Health-Related Quality of Life Measures

Aside from being practical..

1. Same people get same scores
2. Different people get different scores and differ in the way you expect — Valid
3. Measure is interpretable
4. Measure works the same way for different groups (age, gender, race/ethnicity)

Aspects of Good Health-Related Quality of Life Measures

Aside from being practical..

1. Same people get same scores
2. Different people get different scores and differ in the way you expect
3. Measure is interpretable 
4. Measure works the same way for different groups (age, gender, race/ethnicity)

Aspects of Good Health-Related Quality of Life Measures

Aside from being practical..

1. Same people get same scores
2. Different people get different scores and differ in the way you expect
3. Measure is interpretable
4. Measure works the same way for different groups (age, gender, race/ethnicity)

Validity

Does scale represent what it is supposed to be measuring?

- 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?

Relative Validity Example

Sensitivity of measure to important (clinical) difference

	Severity of Kidney Disease				
	None	Mild	Severe	F-ratio	Relative Validity
Burden of Disease #1	87	90	91	2	--
Burden of Disease #2	74	78	88	10	5
Burden of Disease #3	77	87	95	20	10

Evaluating Construct Validity

Scale	Age (years)			
(Better) Physical Functioning	(-)			

Evaluating Construct Validity

Scale	Age (years)			
(Better) Physical Functioning	Medium (-)			

Evaluating Construct Validity

Scale	Age (years)			
(Better) Physical Functioning	Medium (-)			

Effect size (ES) = D/SD

D = Score difference
SD = SD

Small (0.20), medium (0.50), large (0.80)

Evaluating Construct Validity

Scale	Age (years)			
(Better) Physical Functioning	Medium (-) r ~0.24			

Cohen effect size rules of thumb (d = 0.20, 0.50, and 0.80):

Small r = 0.100; medium r = 0.243; large r = 0.371

$$r = d / [(d^2 + 4)^{.5}]$$
$$= 0.80 / [(0.80^2 + 4)^{.5}] = 0.80 / [(0.64 + 4)^{.5}] = 0.80 / [(4.64)^{.5}] = 0.80 / 2.154$$
$$= 0.371$$

Evaluating Construct Validity

Scale	Age (years)	Obese yes = 1, no = 0	Kidney Disease yes = 1, no = 0	In Nursing home yes = 1, no = 0
(Better) Physical Functioning	Medium (-)	Small (-)	Large (-)	Large (-)

Cohen effect size rules of thumb (d = 0.20, 0.50, and 0.80):

Small r = 0.100; medium r = 0.243; large r = 0.371

$$r = d / [(d^2 + 4)^{.5}]$$
$$= 0.80 / [(0.80^2 + 4)^{.5}] = 0.80 / [(0.64 + 4)^{.5}] = 0.80 / [(4.64)^{.5}] = 0.80 / 2.154$$
$$= 0.371$$

Evaluating Construct Validity

Scale	Age (years)	Obese yes = 1, no = 0	Kidney Disease yes = 1, no = 0	In Nursing home yes = 1, no = 0
(Better) Physical Functioning	Medium (-)	Small (-)	Large (-)	Large (-)
(More) Depressive Symptoms	?	Small (+)	?	Small (+)

Cohen effect size rules of thumb (d = 0.20, 0.50, and 0.80):

Small r = 0.100; medium r = 0.243; large r = 0.371

$$\underline{r} = \underline{d} / [(\underline{d}^2 + 4)^{.5}]$$
$$= \underline{0.80} / [(0.80^2 + 4)^{.5}] = 0.80 / [(0.64 + 4)^{.5}] = 0.80 / [(4.64)^{.5}] = 0.80 / 2.154$$
$$= \underline{0.371}$$

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

Questions?

Responsiveness to Change

- HRQOL measures should be responsive to interventions that change HRQOL
- Need external indicators of change (Anchors)
  - Clinical measure
    - "improved" group = 100% reduction in seizure frequency
    - "unchanged" group = <50% change in seizure frequency
  - Retrospective self- or provider-report of change
    - *Much better, A little better, Same, A little worse, Much worse*
- Anchor correlated with change on target measure at 0.371 or higher

Responsiveness Index

Effect size (ES) = D/SD

D = raw score change in "changed" (improved) group

SD = baseline SD

- Small: 0.20->0.49
- Medium: 0.50->0.79
- Large: 0.80 or above

Responsiveness Indices

(1) Effect size (ES) = D/SD

(2) Standardized Response Mean (SRM) = D/SD†

(3) Guyatt responsiveness statistic (RS) = D/SD‡

D = raw score change in “changed” group;

SD = baseline SD;

SD† = SD of D;

SD‡ = SD of D among “unchanged”

Amount of Expected Change Varies

SF-36 physical function score mean = 87 (SD = 20)
Assume I have a score of 100 at baseline

- Hit by **Bike** causes me to be
  - *limited a lot* in vigorous activities
  - *limited a lot* in climbing several flights of stairs
  - *limited a little* in moderate activities

SF-36 physical functioning score drops to 75 (-1.25 SD)

- Hit by **Rock** causes me to be
  - *limited a little* in vigorous activities

SF-36 physical functioning score drops to 95 (- 0.25 SD)

Partition Change on Anchor

- A lot better
- A little better
- No change
- A little worse
- A lot worse

Use Multiple Anchors

- 693 RA clinical trial participants evaluated at baseline and 6-weeks post-treatment.
- Five anchors:
  1. Self-report (global) by patient
  2. Self-report (global) by physician
  3. Self-report of pain
  4. Joint swelling (clinical)
  5. Joint tenderness (clinical)

Kosinski, M. et al. (2000). Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. Arthritis and Rheumatism, 43, 1478-1487.

Patient and Physician Global Reports

How are you (is the patient) doing, considering all the ways that RA affects you (him/her)?

- Very good (asymptomatic and no limitation of normal activities)
- Good (mild symptoms and no limitation of normal activities)
- Fair (moderate symptoms and limitation of normal activities)
- Poor (severe symptoms and inability to carry out most normal activities)
- Very poor (very severe symptoms that are intolerable and inability to carry out normal activities

--> Improvement of 1 level **over time**

Global Pain, Joint Swelling and Tenderness

- 0 = no pain, 10 = severe pain
- Number of swollen and tender joints

-> 1-20% improvement over time

Effect Sizes for SF-36 PF
Change Linked to Minimal Change in Anchors

Scale	Self-R	Clin.-R	Pain	Swell	Tender	Mean
Physical Function	.35	.33	.34	.26	.32	.32

Effect Sizes for SF-36
Changes Linked to Minimal Change in Anchors

Scale	Self-R					
PF	.35					
Role-P	.56					
Pain	.83					
GH	.20					
EWB	.39					
Role-E	.41					
SF	.43					
EF	.50					
PCS	.49					
MCS	.42					

Effect Sizes (mean = 0.34) for SF-36 Changes Linked to Minimal Change in Anchors

Scale	Self-R	Clin.-R	Pain	Swell	Tender	Mean
PF	.35	.33	.34	.26	.32	.32
Role-P	.56	.52	.29	.35	.36	.42
Pain	.83	.70	.47	.69	.42	.62
GH	.20	.12	.09	.12	.04	.12
EWB	.39	.26	.25	.18	.05	.23
Role-E	.41	.28	.18	.38	.26	.30
SF	.43	.34	.28	.29	.38	.34
EF	.50	.47	.22	.22	.35	.35
PCS	.49	.48	.34	.29	.36	.39
MCS	.42	.27	.19	.27	.20	.27

Reliability

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

- ✓ Inter-rater (rater)
  - ✓ Need 2 or more raters of the thing being measured
- ✓ Internal consistency (items)
  - ✓ Need 2 or more items
- ✓ Test-retest (administrations)
  - ✓ Need 2 or more time points

Ratings of Performance of Six Kaiser Presentations by Two Raters

[1 = Poor; 2 = Fair; 3 = Good; 4 = Very good; 5 = Excellent]

1= Karen Kaiser (Good, Very Good)
2= Adam Ant (Very Good, Excellent)
3= Rick Dees (Good, Good)
4= Ron Hays (Fair, Poor)
5= John Adams (Excellent, Very Good)
6= Jane Error (Fair, Fair)

(*Target = 6 presenters; assessed by 2 raters*)

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

01 13
01 24
02 14
02 25
03 13
03 23
04 12
04 21
05 15
05 24
06 12
06 22

Two-Way Random Effects
(Reliability of Ratings of Presentations)

Source	df	SS	MS
Presenters (BMS)	5	15.67	3.13
Raters (JMS)	1	0.00	0.00
Pres. x Raters (EMS)	5	2.00	0.40
Total	11	17.67	

$$\text{2-way R} = \frac{6\,(3.13 - 0.40)}{6\,(3.13) + 0.00 - 0.40} = \boxed{0.89}$$

ICC = 0.80

Responses of Presenters to Two Questions about Their Health

1= Karen Kaiser (Good, Very Good)
2= Adam Ant (Very Good, Excellent)
3= Rick Dees (Good, Good)
4= Ron Hays (Fair, Poor)
5= John Adams (Excellent, Very Good)
6= Jane Error (Fair, Fair)

(*Target = 6 presenters; assessed by 2 items*)

Two-Way Mixed Effects (Cronbach's Alpha)

01	34
02	45
03	33
04	21
05	54
06	22

Source	df	SS	MS
Presenters (BMS)	5	15.67	3.13
Items (JMS)	1	0.00	0.00
Pres. x Items (EMS)	5	2.00	0.40
Total	11	17.67	

$$\text{Alpha} = \frac{3.13 - 0.40}{3.13} = \frac{2.93}{3.13} = \boxed{0.87}$$

ICC = 0.77

Reliability Minimum Standards

- 0.70 or above (for group comparisons)
- 0.90 or higher (for individual assessment)

  - SEM = SD (1- reliability)$^{1/2}$
  - 95% CI = true score +/- 1.96 x SEM

    - if z-score = 0, then CI: -.62 to +.62 when reliability = 0.90
      - Width of CI is 1.24 z-score units

Guidelines for Interpreting Kappa

Conclusion	Kappa	Conclusion	Kappa
Poor	< .40	Poor	< 0.0
Fair	.40 - .59	Slight	.00 - .20
Good	.60 - .74	Fair	.21 - .40
Excellent	> .74	Moderate	.41 - .60
		Substantial	.61 - .80
		Almost perfect	.81 - 1.00
Fleiss (1981)		Landis and Koch (1977)	

Questions?

Sufficient Unidimensionality

- One-Factor Categorical Confirmatory Factor Analytic Model (e.g., using Mplus)
  - Polychoric correlations; weighted least squares with adjustments for mean and variance
- Fit Indices
  - Comparative Fit Index, etc.

Local Independence

- After controlling for dominant factor(s), item pairs should not be associated.
  - Look for residual correlations > 0.20

- Local dependence often caused by asking the same question multiple times.
  - "I'm generally sad about my life."
  - "My life is generally sad."

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.

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

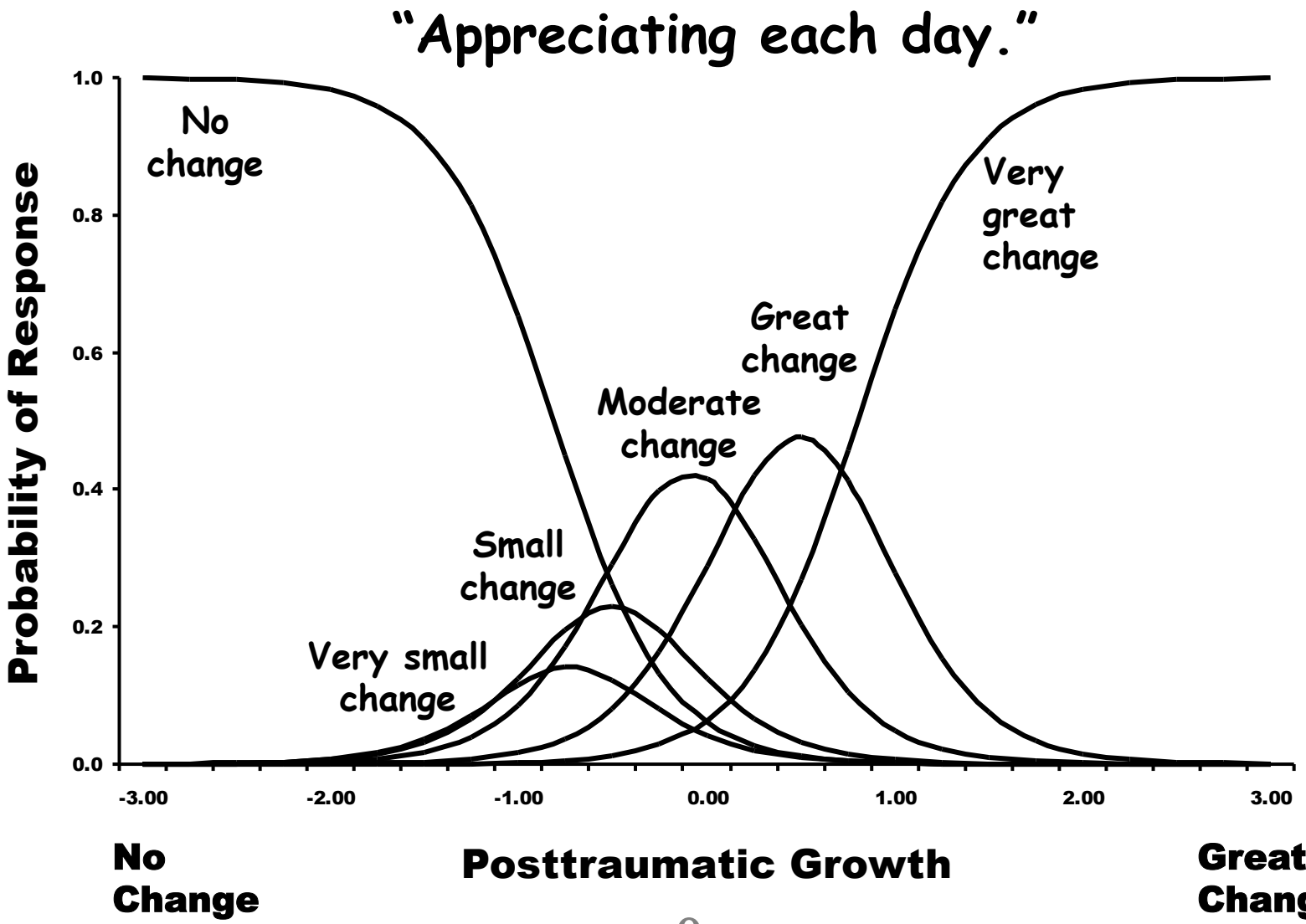
(2) I experienced this change to a small degree as a result of my crisis

(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

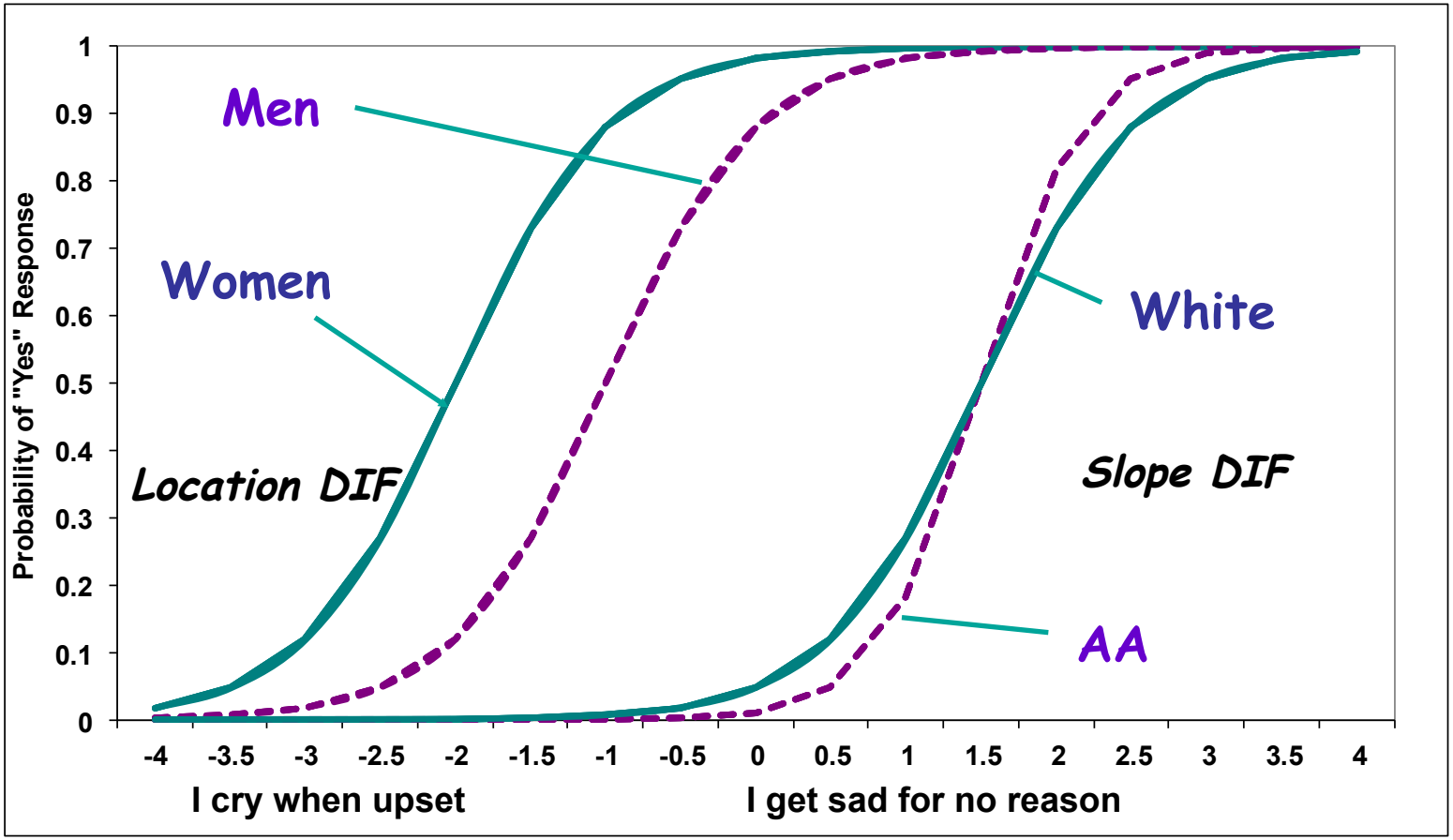


Differential Item Functioning (DIF)

- 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)



Higher Score = More Depressive Symptoms

Thank You!

drhays@ucla.edu (310-794-2294).

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