Evaluating Multi-Item Scales

Health Services Research Design (HS 225B)

January 26, 2015, 1:00-3:00pm

71-257 CHS

Listed below are a few statements about your relationships with others. How much is each statement TRUE or FALSE for you

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
1. I am always courteous even to people who are disagreeable.	1	2	3	4	5
2. There have been occasions when I took advantage of someone.	1	2	3	4	5
3. I sometimes try to get even rather than forgive and forget.	1	2	3	4	5
4. I sometimes feel resentful when I don't get my way.	1	2	3	4	5
5. No matter who I'm talking to, I'm always a good listener.	1	2	3	4	5

Scoring Multi-Item Scales

- Average or sum all items in the same scale.

- Transform average or sum to
  - 0 (worse) to 100 (best) possible range
  - z-score (mean = 0, SD = 1)
  - T-score (mean = 50, SD = 10)

Linear Transformations

$$X = \frac{\text{(original score - minimum) *100}}{\text{(maximum - minimum)}}$$

$$Z_X = \frac{(X - \overline{X})}{SD_X}$$

$$Y = \text{target mean} + (\text{target SD} * Zx)$$

Listed below are a few statements about your relationships with others. How much is each statement TRUE or FALSE for you

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
1. I am always courteous even to people who are disagreeable.	100	75	50	25	0
2. There have been occasions when I took advantage of someone.	0	25	50	75	100
3. I sometimes try to get even rather than forgive and forget.	0	25	50	75	100
4. I sometimes feel resentful when I don't get my way.	0	25	50	75	100
5. No matter who I'm talking to, I'm always a good listener.	100	75	50	25	0

Create T-score

z-score = (score – 36)/31

T-score = (10 * z-score) + 50

z-score = (100- 36)/31 = 2.06

T-score = 71

Aspects of Good 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 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)

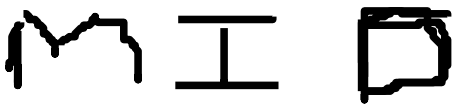
Aspects of Good Measures

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 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 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)

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 HPM 225B Lectures by Two Raters

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

1= Marjorie Kagawa-Singer (Good, Very Good)
2= Tom Belin (Very Good, Excellent)
3= Rick Dees (Good, Good)
4= Ron Hays (Fair, Poor)
5= Jack Needleman (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
WMS = Within Mean Square
JMS = Item or Rater Mean Square
EMS = Ratee x Item (Rater) Mean Square

N = n of ratees
k = n of items or raters

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= Marjorie Kagawa-Singer (Good, Very Good)
2= Tom Belin (Very Good, Excellent)
3= Rick Dees (Good, Good)
4= Ron Hays (Fair, Poor)
5= Jack Needleman (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	

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

$$\text{ICC} = 0.77$$

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

Cross-Tab of Ratings

Rater 2 \ Rater 1	P	F	G	VG	E	Total
P	0	1				1
F		1				1
G			1			1
VG			1	0	1	2
E				1	0	1
Total	0	2	2	1	1	6

Calculating KAPPA

$$\mathbf{P_C} = \frac{(0 \times 1) + (2 \times 1) + (2 \times 1) + (1 \times 2) + (1 \times 1)}{(6 \times 6)} = 0.19$$

$$\mathbf{P_{obs.}} = \frac{2}{6} = 0.33$$

$$\mathbf{Kappa} = \frac{0.33 - 0.19}{1 - 0.19} = \mathbf{0.17}$$

Weighted Kappa Linear (Quadratic)

	P	F	G	VG	E
P	1	.75 (.937)	.50 (.750)	.25 (.437)	0
F	.75 (.937)	1	.75 (.937)	.50 (.750)	.25 (.437)
G	.50 (.750)	.75 (.937)	1	.75 (.937)	.50 (.750)
VG	.25 (.437)	.50 (.750)	.75 (.937)	1	.75 (.937)
E	0	.25 (.437)	.5 (.750)	.75 (.937)	1

$W_l = 1 - (i/ (k - 1))$

$W_q = 1 - (i^2 / (k - 1)^2)$

i = number of categories ratings differ by
k = n of categories

All Kappas

$$\mathbf{P_C} = \frac{(0 \times 1) + (2 \times 1) + (2 \times 1) + (1 \times 2) + (1 \times 1)}{(6 \times 6)} = 0.19$$

$$\mathbf{P_{obs.}} = \frac{2}{6} = 0.33$$

$$\mathbf{Kappa} = \frac{0.33 - 0.19}{1 - 0.19} = \mathbf{0.17}$$

Linear weighted kappa = 0.52
Quadratic weighted kappa = 0.77

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)	

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.

Confirmatory Factor Analysis

	Depress	Anxiety	Anger
Item #1	0.80*	0.00	0.00
Item #2	0.80*	0.00	0.00
Item #3	0.80*	0.00	0.00
Item #4	0.00	0.80*	0.00
Item #5	0.00	0.80*	0.00
Item #6	0.00	0.80*	0.00
Item #7	0.00	0.00	0.80*
Item #8	0.00	0.00	0.80*
Item #9	0.00	0.00	0.80*

*Factor loading.

Confirmatory Factor Analysis Fit Indices

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

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

- Comparative fit index: $1 - \frac{\chi^2_{model} - df_{model}}{\chi^2_{null} - df_{null}}$

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

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

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

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

Questions?

Responsiveness to Change

- Measures should be responsive to interventions that change the underlying construct
- 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 Physical Function 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

Item Response Theory (IRT)

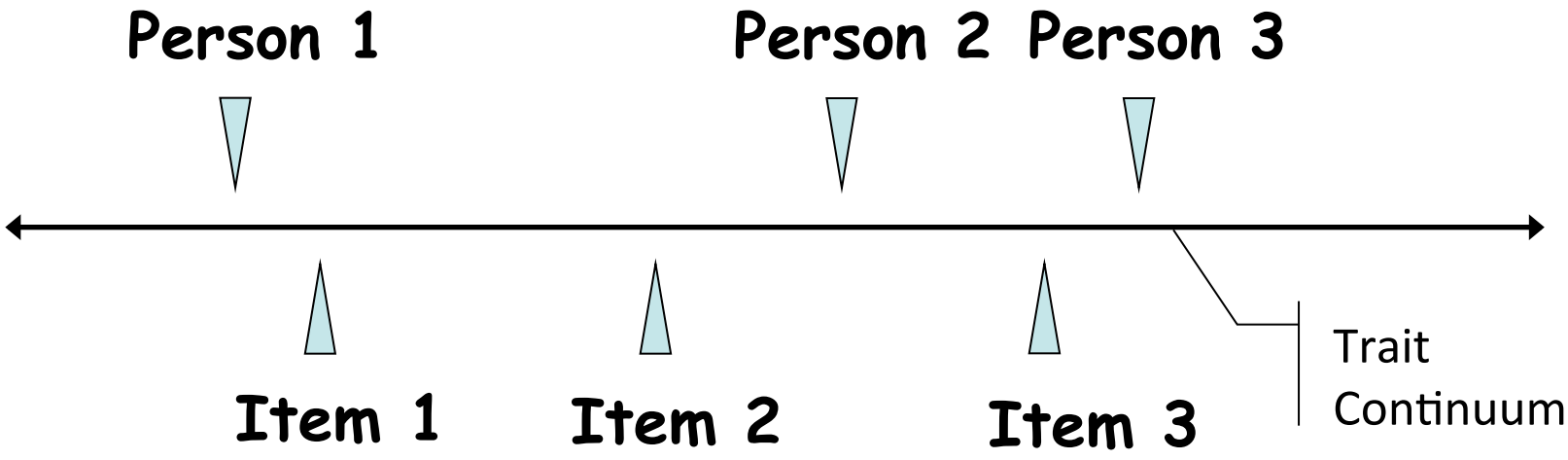
IRT models the relationship between a person's response Y_i to the question (i) and his or her level of the latent construct θ being measured by positing

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

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

a_i is an estimate of the discriminatory power of the item.

Item Responses and Trait Levels



www.nihpromis.org

Computer Adaptive Testing (CAT)

Reliability Target for Use of Measures with Individuals

- Reliability ranges from 0-1
  - 0.90 or above is goal
  - SEM = SD $(1- \text{reliability})^{1/2}$
  - 95% CI = true score +/- 1.96 x SEM
    - if true z-score = 0, then CI: -.62 to +.62
      - Width of CI is 1.24 z-score units

- Reliability = 0.90 when SE = 3.2
  - T-scores (mean = 50, SD = 10)
  - Reliability = $1 - (SE/10)^2$

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

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 …

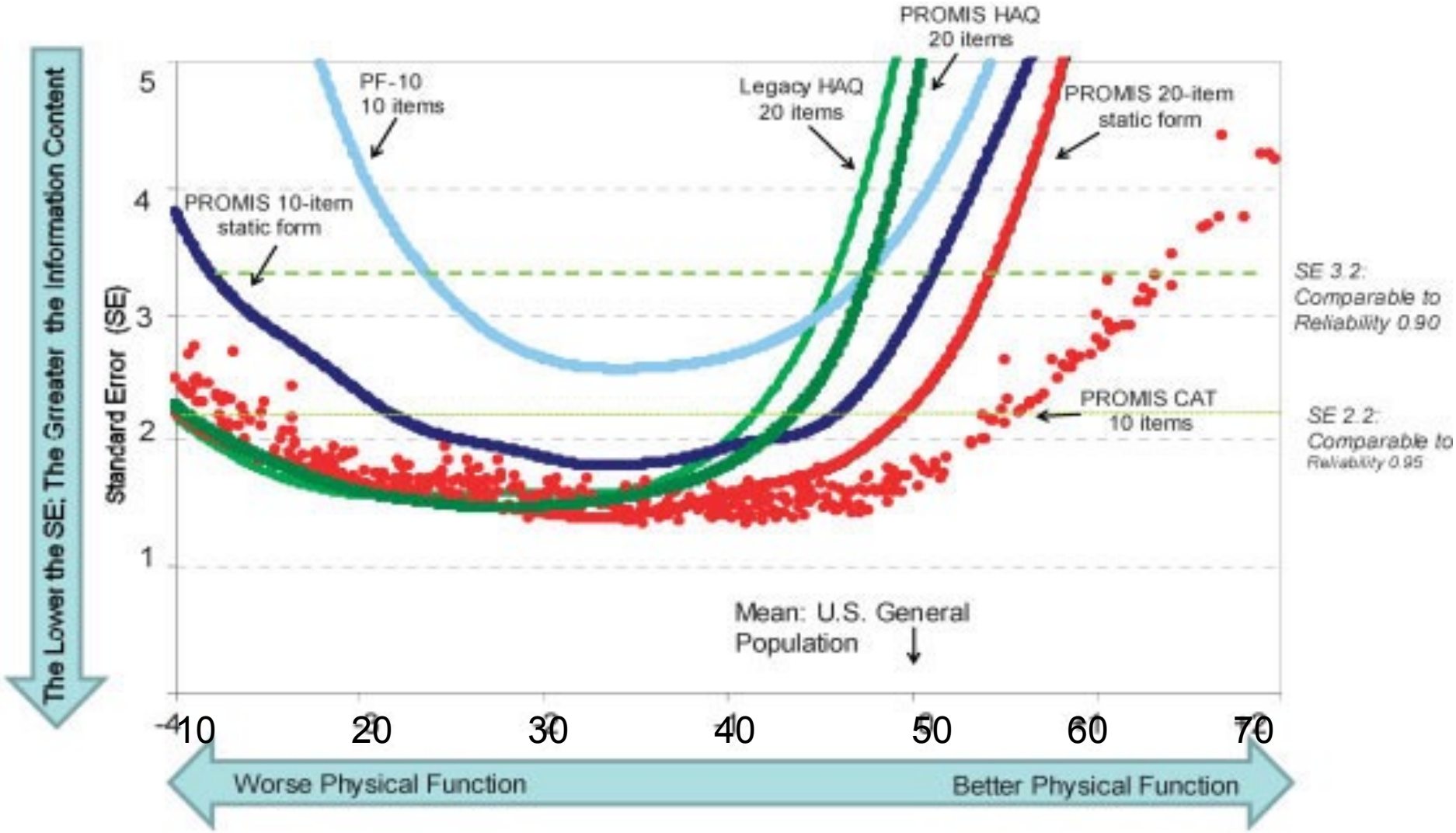
I made myself angry about something just by thinking about it. [6th question]

- Never
- Rarely
- Sometimes
- Often
- Always

Estimated Anger = 50.2

SE = 2.8 (rel = 0.92)

PROMIS Physical Functioning vs. “Legacy” Measures



Thank you.

Powerpoint file is freely available at:

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

Contact information:

drhays@ucla.edu 310-794-2294