Evaluating Self-Report Data Using Psychometric Methods

Ron D. Hays, PhD (hays@rand.org)

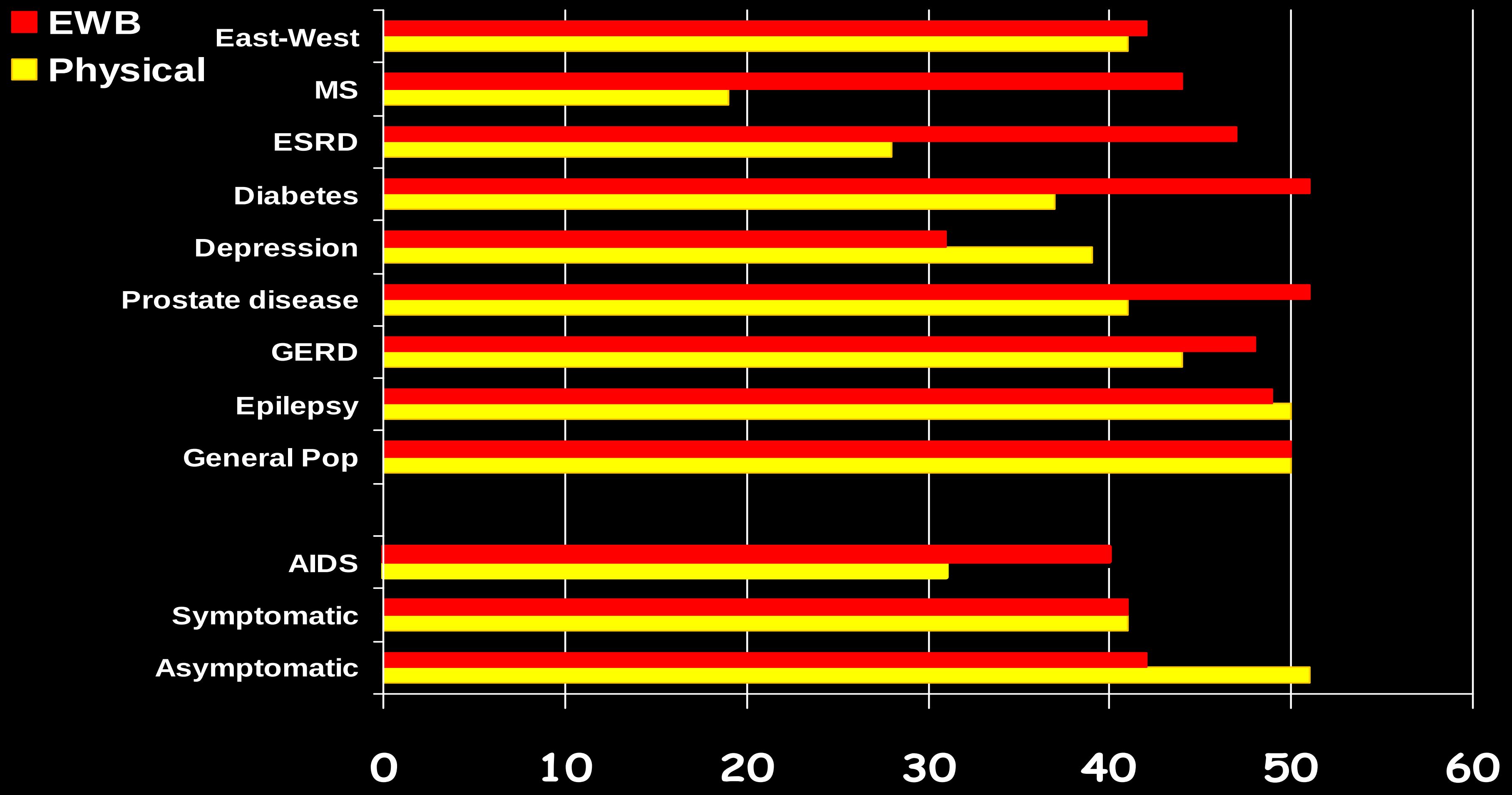
February 7, 2007 (3:30-6:30pm)

HS 249F

Individual Change

- Interest in
  - Knowing how many patients benefit from group intervention, or
  - Tracking progress on individual patients
- Sample
  - 54 patients
  - Average age = 56; 84% white; 58% female
- Method
  - Self-administered SF-36 version 2 at baseline and at end of therapy (about 6 weeks later).

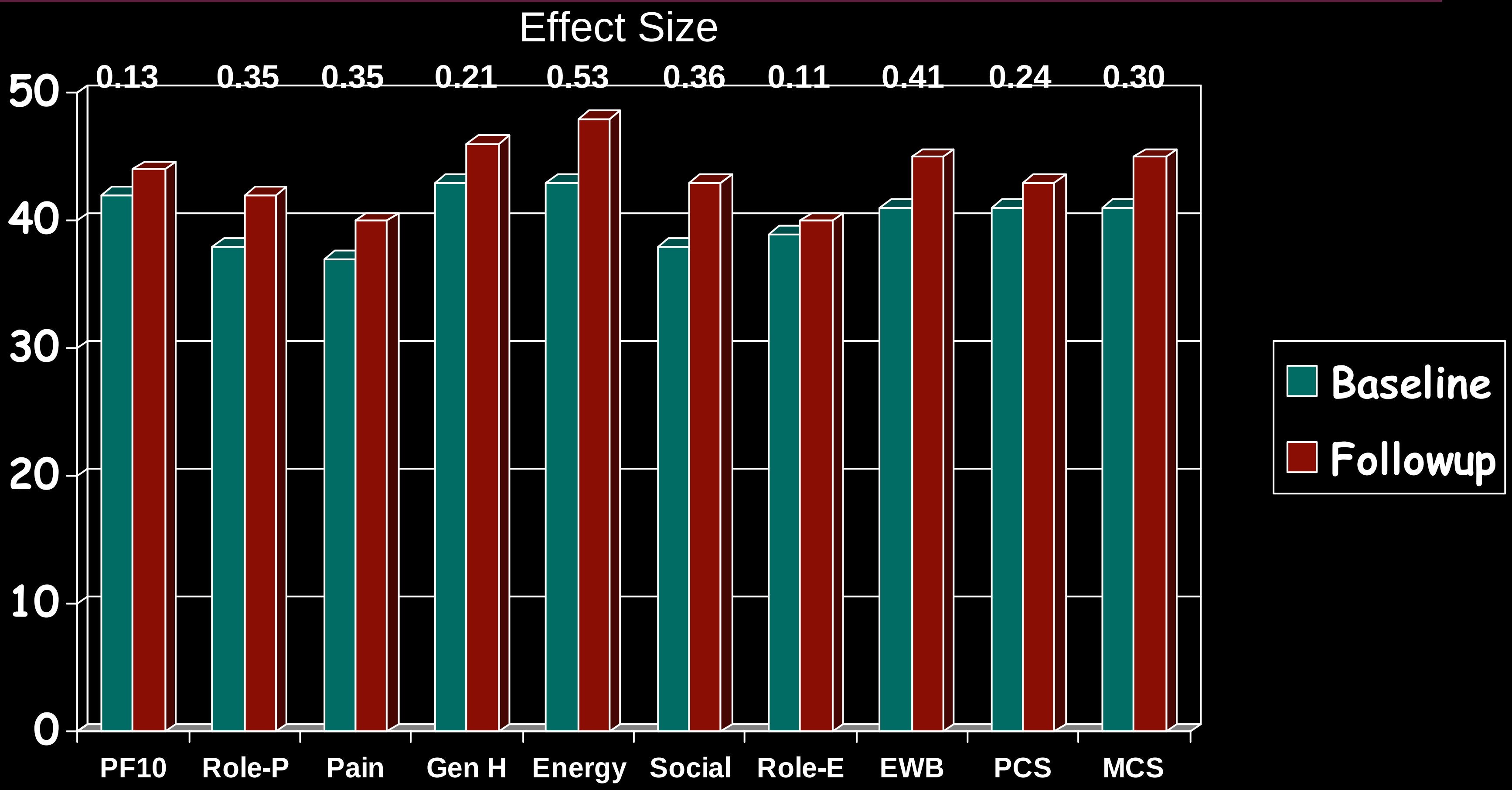
Physical Functioning and Emotional Well-Being at Baseline for 54 Patients at UCLA-Center for East West Medicine



RAND HEALTH

Hays et al. (2000), American Journal of Medicine

Change in SF-36 Scores Over Time



t-test for within group change

- $X_D/(SD_d/n^{½})$

X_D = is mean difference, SD_d = standard deviation of difference

RAND HEALTH

Significance of Group Change (T-scores)

	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

Reliable Change Index

$(X_2 - X_1)/ (SEM * SQRT [2])$

$SEM = SD_b * (1 - reliability)^{1/2}$

Amount of Change in Observed Score Needed for Significant Individual Change

	RCI	Effect size
PF-10	8.4	0.67
RP-4	8.4	0.72
BP-2	10.4	1.01
GH-5	13.0	1.13
EN-4	12.8	1.33
SF-2	13.8	1.07
RE-3	9.7	0.71
EWB-5	13.4	1.26
PCS	7.1	0.62
MCS	9.7	0.73

Significant Change for 54 Cases

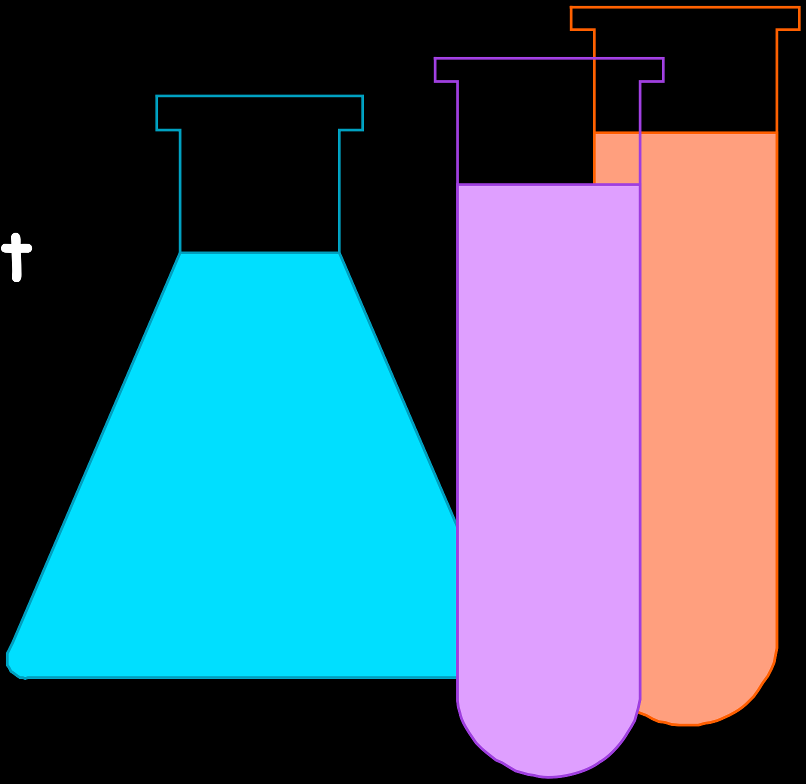
	% Improving	% Declining	Difference
PF-10	13%	2%	+ 11%
RP-4	31%	2%	+ 29%
BP-2	22%	7%	+ 15%
GH-5	7%	0%	+ 7%
EN-4	9%	2%	+ 7%
SF-2	17%	4%	+ 13%
RE-3	15%	15%	0%
EWB-5	19%	4%	+ 15%
PCS	24%	7%	+ 17%
MCS	22%	11%	+ 11%

Multiple Steps in Developing Good Survey

- Review literature
- Expert input (patients and clinicians)
- Define constructs you are interested in
- Draft items (item generation)
- Pretest
  - Cognitive interviews
  - Field and pilot testing
- Revise and test again
- Translate/harmonize across languages

What's a Good Measure?

- Same person gets same score (reliability)
- Different people get different scores (validity)
- People get scores you expect (validity)
- It is practical to use (feasibility)

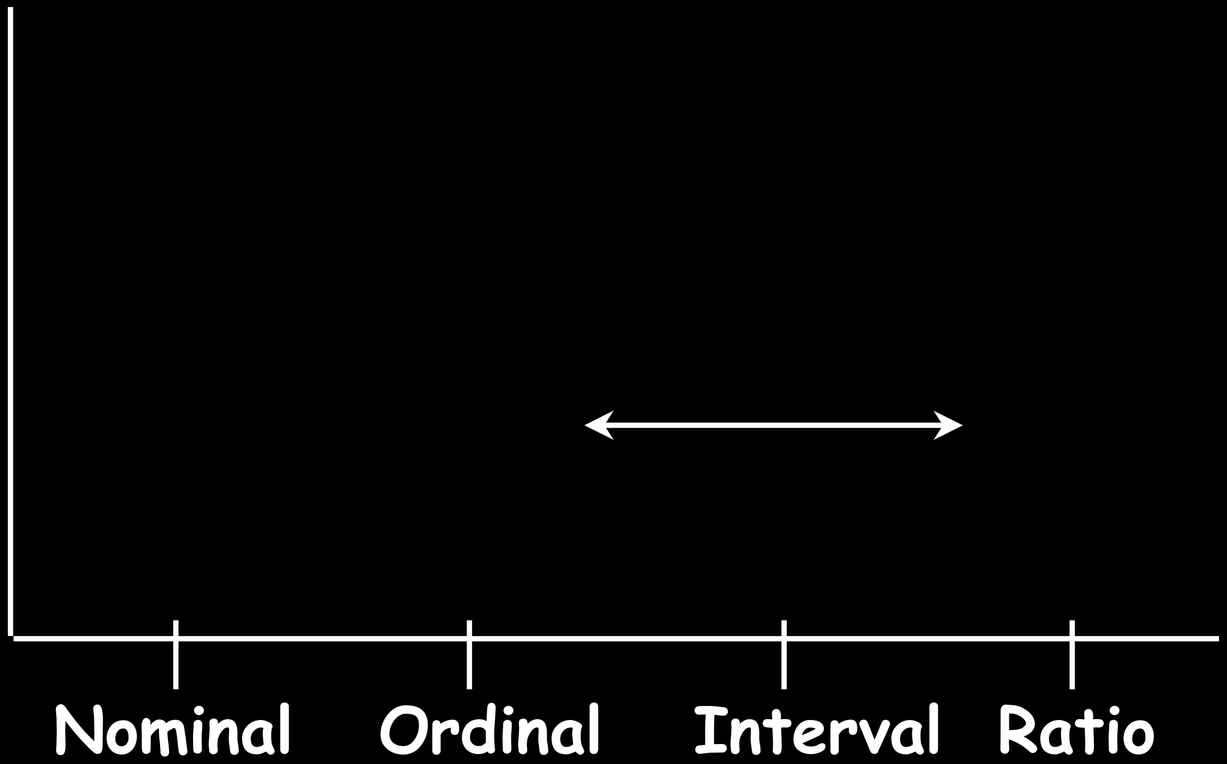


RAND HEALTH

Scales of Measurement and Their Properties

	Property of Numbers		
Type of Scale	Rank Order	Equal Interval	Absolute 0
Nominal			
Ordinal	+		
Interval	+	+	
Ratio	+	+	+

Measurement Range for Health Outcome Measures



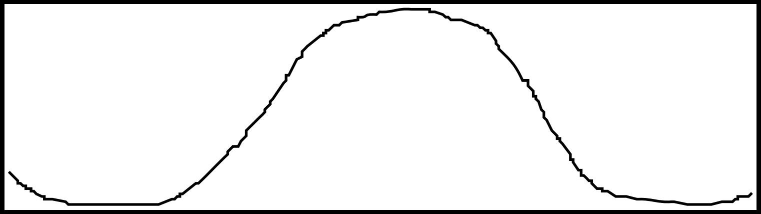
RAND HEALTH

Indicators of Acceptability

- Unit non-response
- Item non-response
- Administration time

Variability

- All scale levels are represented
- Distribution approximates bell-shaped "normal"



Measurement Error

observed = true score + systematic error + random error

(bias)

Four Types of Data Collection Errors

- Coverage Error

  Does each person in population have an equal chance of selection?

- Sampling Error

  Are only some members of the population sampled?

- Nonresponse Error

  Do people in the sample who respond differ from those who do not?

- Measurement Error

  Are inaccurate answers given to survey questions?

Flavors of Reliability

- Test-retest (administrations)
- Intra-rater (raters)
- Internal consistency (items)

Test-retest Reliability of MMPI 317-362
r = 0.75

		MMPI 317		
		True	False	
MMPI 362	True	169	15	184
	False	21	95	116
		190	110	

I am more sensitive than most other people.

Kappa Coefficient of Agreement (Corrects for Chance)

$$\text{kappa} = \frac{(\text{observed - chance})}{(1\ \text{- chance})}$$

Example of Computing KAPPA

		Rater A					
		1	2	3	4	5	Row Sum
	1	1	1				2
	2		2				2
Rater B	3			2			2
	4				2		2
	5					2	2
Column Sum		1	3	2	2	2	10

Example of Computing KAPPA (Continued)

$$P_c = \frac{(1 \times 2) + (3 \times 2) + (2 \times 2) + (2 \times 2) + (2 \times 2)}{(10 \times 10)} = \boxed{0.20}$$

$$P_{obs.} = \frac{9}{10} = \boxed{0.90}$$

$$\text{Kappa} = \frac{0.90 - 0.20}{1 - 0.20} = \boxed{0.87}$$

RAND HEALTH

Guidelines for Interpreting Kappa

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

Intraclass Correlation and Reliability

Model	Reliability	Intraclass Correlation
One-Way	$\dfrac{MS_{BMS} - MS_{WMS}}{MS_{BMS}}$	$\dfrac{MS_{BMS} - MS_{WMS}}{MS_{BMS} + (K-1)MS_{WMS}}$
Two-Way Fixed	$\dfrac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\dfrac{MS_{BMS} - MS_{EMS}}{MS_{EMS} + (K-1)MS_{EMS}}$
Two-Way Random	$\dfrac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\dfrac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)MS_{EMS} + K(MS_{JMS} - MS_{EMS})/N}$

Summary of Reliability of Plant Ratings

	Baseline		Follow-up	
	R_{TT}	R_{II}	R_{TT}	R_{II}
One-Way Anova	0.97	0.95	0.97	0.94
Two-Way Random Effects	0.97	0.95	0.97	0.94
Two-Way Fixed Effects	0.98	0.96	0.98	0.97

Source	Label	Baseline MS
Plants	BMS	628.667
Within	WMS	17.700
Raters	JMS	57.800
Raters X Plants	EMS	13.244

Raw Data for Ratings of Height (1/16 inch) of Houseplants (A1, A2, etc.) by Two Raters (R1, R2)

Plant		Baseline Height	Follow-up Height	Experimental Condition
A1				
	R1	120	121	1
	R2	118	120	
A2				
	R1	084	085	2
	R2	096	088	
B1				
	R1	107	108	2
	R2	105	104	
B2				
	R1	094	100	1
	R2	097	104	
C1				
	R1	085	088	2
	R2	091	096	

Ratings of Height of Houseplants (Cont.)

Plant		Baseline Height	Follow-up Height	Experimental Condition
C2				
	R1	079	086	1
	R2	078	092	
D1				
	R1	070	076	1
	R2	072	080	
D2				
	R1	054	056	2
	R2	056	060	
E1				
	R1	085	101	1
	R2	097	108	
E2				
	R1	090	084	2
	R2	092	096	

Reliability of Baseline Houseplant Ratings

Ratings of Height of Plants: 10 plants, 2 raters

Baseline Results

Source	DF	SS	MS	F
Plants	9	5658	628.667	35.52
Within	10	177	17.700	
Raters	1	57.8	57.800	
Raters x Plants	9	119.2	13.244	
Total	19	5835		

Sources of Variance in Baseline Houseplant Height

Source	dfs	MS	
Plants (N)	9	628.67	(BMS)
Within	10	17.70	(WMS)
Raters (K)	1	57.80	(JMS)
Raters x Plants	9	13.24	(EMS)
Total	19		

Cronbach's Alpha

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	

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

Alpha for Different Numbers of Items and Homogeneity

	Average Inter-item Correlation ($\bar{r}$)					
Number of Items (k)	.0	.2	.4	.6	.8	1.0
2	.000	.333	.572	.750	.889	1.000
4	.000	.500	.727	.857	.941	1.000
6	.000	.600	.800	.900	.960	1.000
8	.000	.666	.842	.924	.970	1.000

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

Spearman-Brown Prophecy Formula

$$\text{alpha}_y = \left(\frac{N \cdot \text{alpha}_x}{1 + (N - 1) * \text{alpha}_x} \right)$$

N = how much longer scale y is than scale x

Example Spearman-Brown Calculations

MHI-18

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

$= 0.55125/0.57125 = 0.96$

Number of Items and Reliability for Three Versions of the Mental Health Inventory (MHI)

Measure	Number of Items	Completion time (min.)	Reliability
MHI-32	32	5-8	.98
MHI-18	18	3-5	.96
MHI-5	5	1 or less	.90

Data from McHorney et al. 1992

Reliability Minimum Standards

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

Reliability of a Composite Score

$$\text{Mosier} = 1 - \frac{\Sigma(w_j^2)(S_j^2) - \Sigma(w_j^2)(S_j^2)(\alpha_j)}{\Sigma(w_j^2)(S_j^2) + 2\Sigma(w_j)(w_K)(S_j)(S_K)(r_{jK})}$$

w_j = weight given to component J

w_K = weight given to component K

S_j = standard deviation of J

α_j = reliability of J

r_{jK} = correlation between J and K

Hypothetical Multitrait/Multi-Item Correlation Matrix

	Trait #1	Trait #2	Trait #3
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.

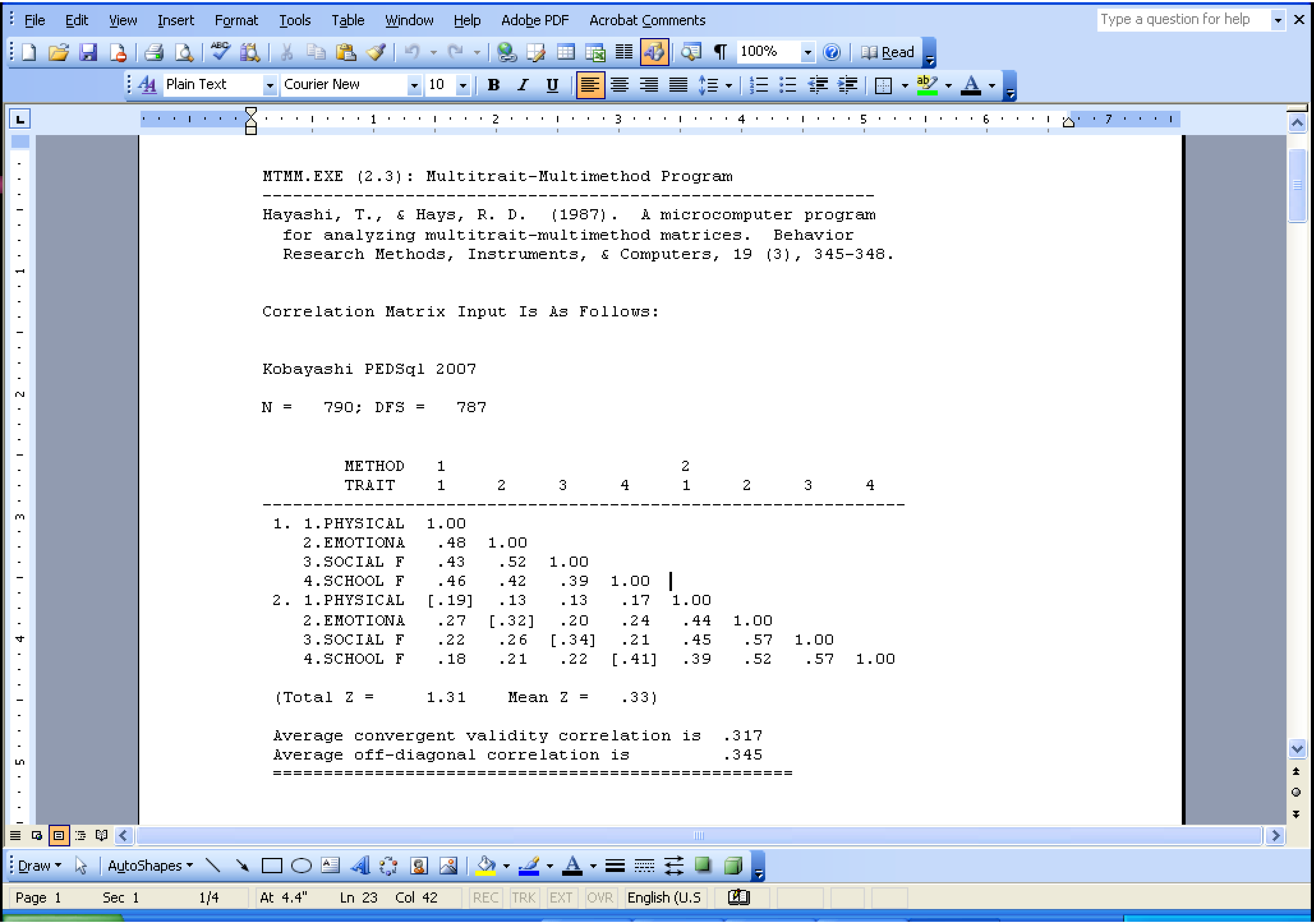
Multitrait/Multi-Item Correlation Matrix for Patient Satisfaction Ratings

	Technical	Interpersonal	Communication	Financial
Technical				
1	0.66*	0.63†	0.67†	0.28
2	0.55*	0.54†	0.50†	0.25
3	0.48*	0.41	0.44†	0.26
4	0.59*	0.53	0.56†	0.26
5	0.55*	0.60†	0.56†	0.16
6	0.59*	0.58†	0.57†	0.23
Interpersonal				
1	0.58	0.68*	0.63†	0.24
2	0.59†	0.58*	0.61†	0.18
3	0.62†	0.65*	0.67†	0.19
4	0.53†	0.57*	0.60†	0.32
5	0.54	0.62*	0.58†	0.18
6	0.48†	0.48*	0.46†	0.24

Note – Standard error of correlation is 0.03. Technical = satisfaction with technical quality. Interpersonal = satisfaction with the interpersonal aspects. Communication = satisfaction with communication. Financial = satisfaction with financial arrangements. *Item-scale correlations for hypothesized scales (corrected for item overlap). †Correlation within two standard errors of the correlation of the item with its hypothesized scale.

Construct Validity

- **Does measure relate to other measures in ways consistent with hypotheses?**
- **Responsiveness to change** including minimally important difference



```
MTMM.EXE (2.3): Multitrait-Multimethod Program
------------------------------------------------------------
Hayashi, T., & Hays, R. D.  (1987).  A microcomputer program
  for analyzing multitrait-multimethod matrices.  Behavior
  Research Methods, Instruments, & Computers, 19 (3), 345-348.


Correlation Matrix Input Is As Follows:


Kobayashi PEDSql 2007

N =   790; DFS =   787


        METHOD   1                       2
        TRAIT    1     2     3     4     1     2     3     4
---------------------------------------------------------------
 1. 1.PHYSICAL  1.00
    2.EMOTIONA   .48  1.00
    3.SOCIAL F   .43   .52  1.00
    4.SCHOOL F   .46   .42   .39  1.00
 2. 1.PHYSICAL  [.19]  .13   .13   .17  1.00
    2.EMOTIONA   .27  [.32]  .20   .24   .44  1.00
    3.SOCIAL F   .22   .26  [.34]  .21   .45   .57  1.00
    4.SCHOOL F   .18   .21   .22  [.41]  .39   .52   .57  1.00

 (Total Z =     1.31    Mean Z =   .33)

 Average convergent validity correlation is  .317
 Average off-diagonal correlation is         .345
 ===================================================
```

Construct Validity for Scales Measuring Physical Functioning

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

Responsiveness to Change and Minimally Important Difference (MID)

- HRQOL measures should be responsive to interventions that changes HRQOL

- Need external indicators of change (Anchors)
  - mean change in HRQOL scores among people who have changed ("minimal" change for MID).

Self-Report Indicator of Change

- Overall has there been any change in your asthma since the beginning of the study?

  Much improved; Moderately improved; Minimally improved

  No change

  Much worse; Moderately worse; Minimally worse

Clinical Indicator of Change

- “changed” group = seizure free (100% reduction in seizure frequency)

- “unchanged” group = <50% change in seizure frequency

Responsiveness Indices

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

(2) Standardized Response Mean (SRM) = $D/SD^{\dagger}$

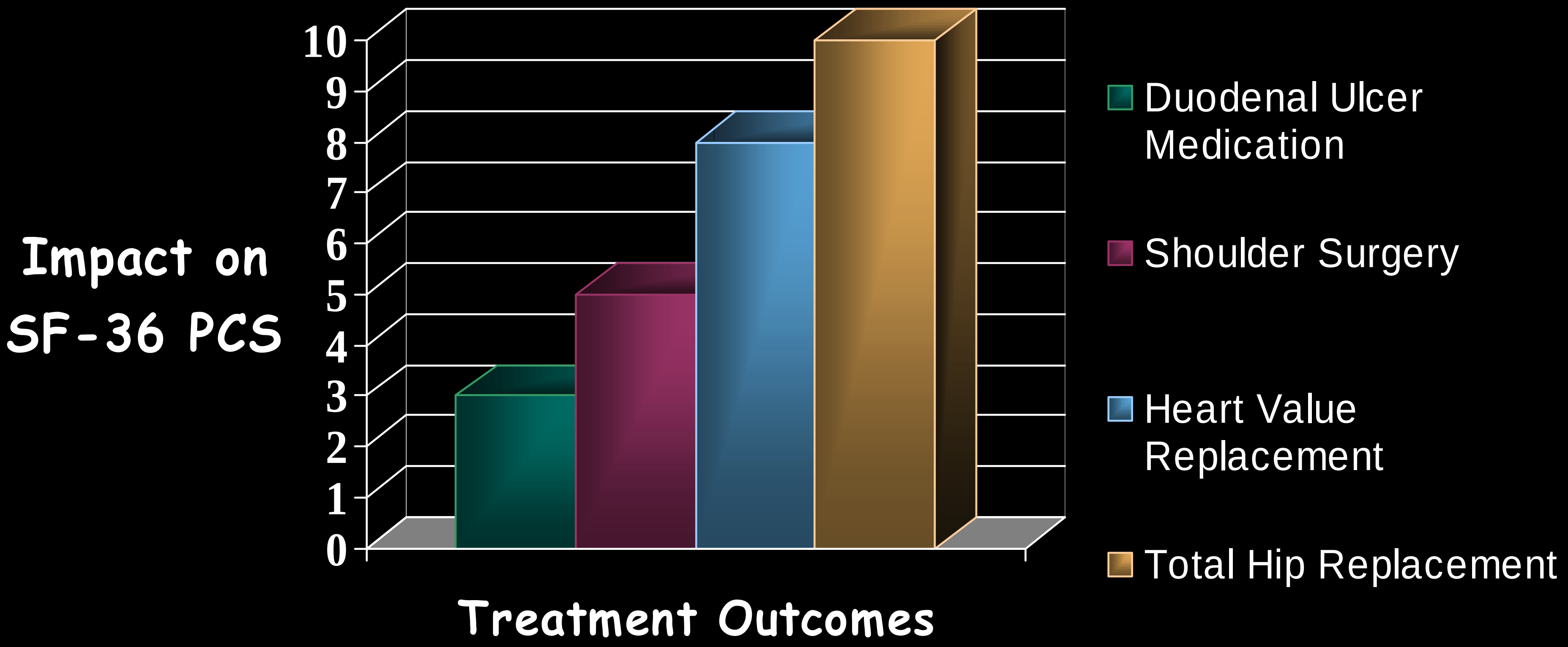
(3) Guyatt responsiveness statistic (RS) = $D/SD^{\ddagger}$

D = raw score change in “changed” group;
SD = baseline SD;
$SD^{\dagger}$ = SD of D;
$SD^{\ddagger}$ = SD of D among “unchanged”

Effect Size Benchmarks

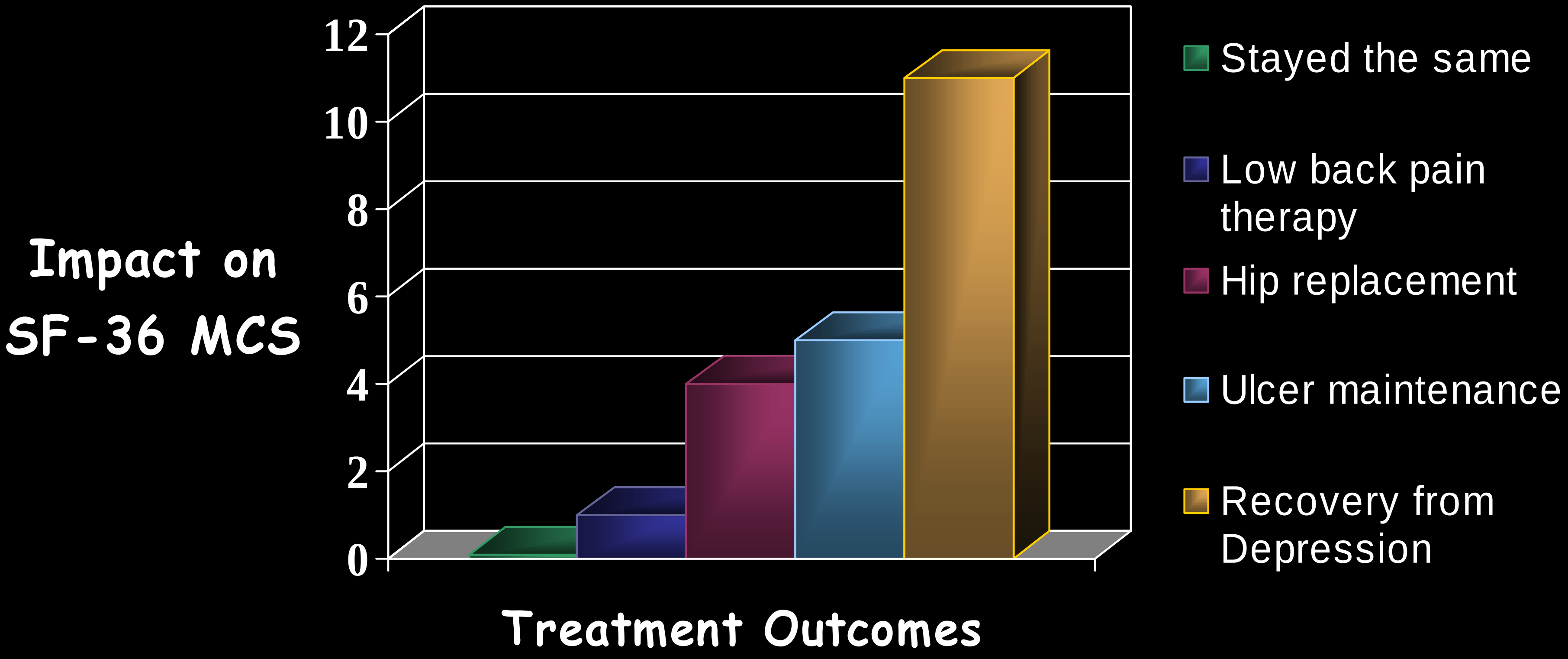
- Small: 0.20->0.49
- Moderate: 0.50->0.79
- Large: 0.80 or above

Treatment Impact on PCS



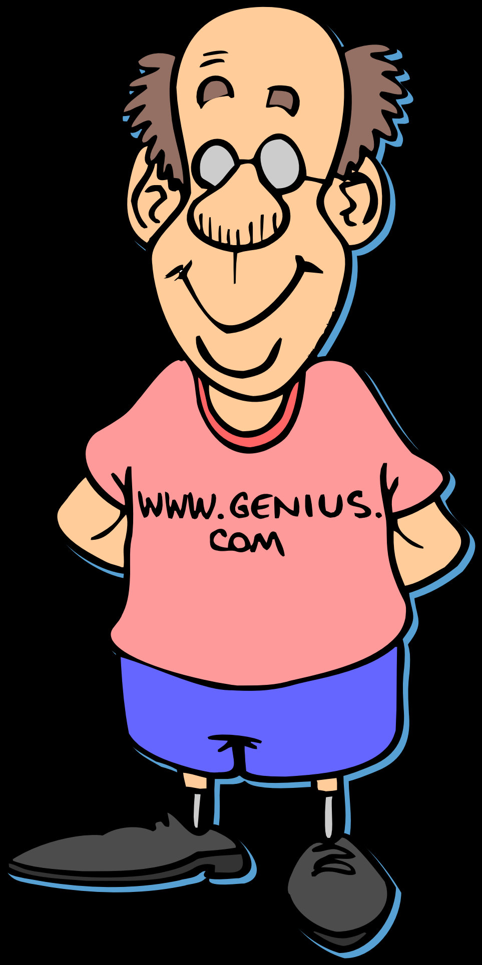
RAND HEALTH

Treatment Impact on MCS

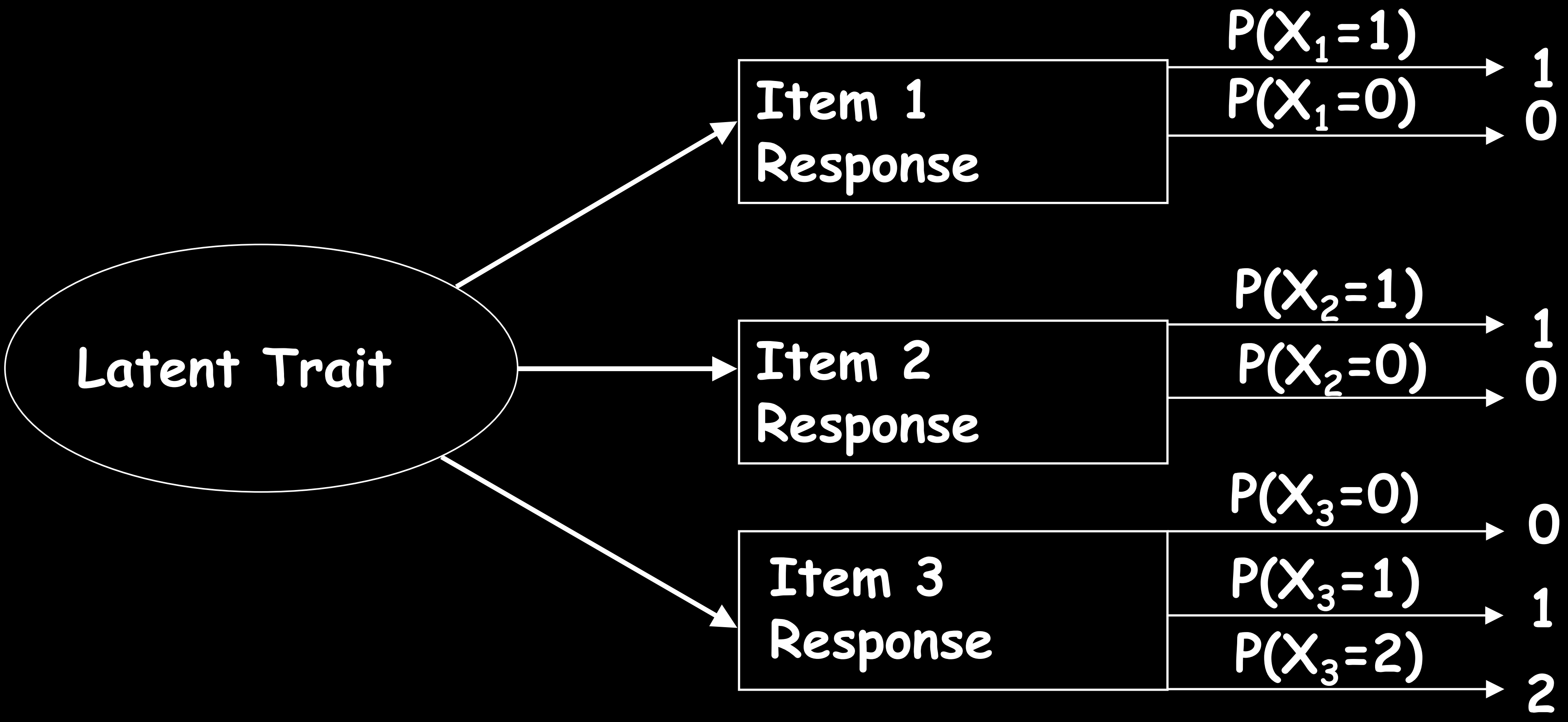


RAND HEALTH

IRT

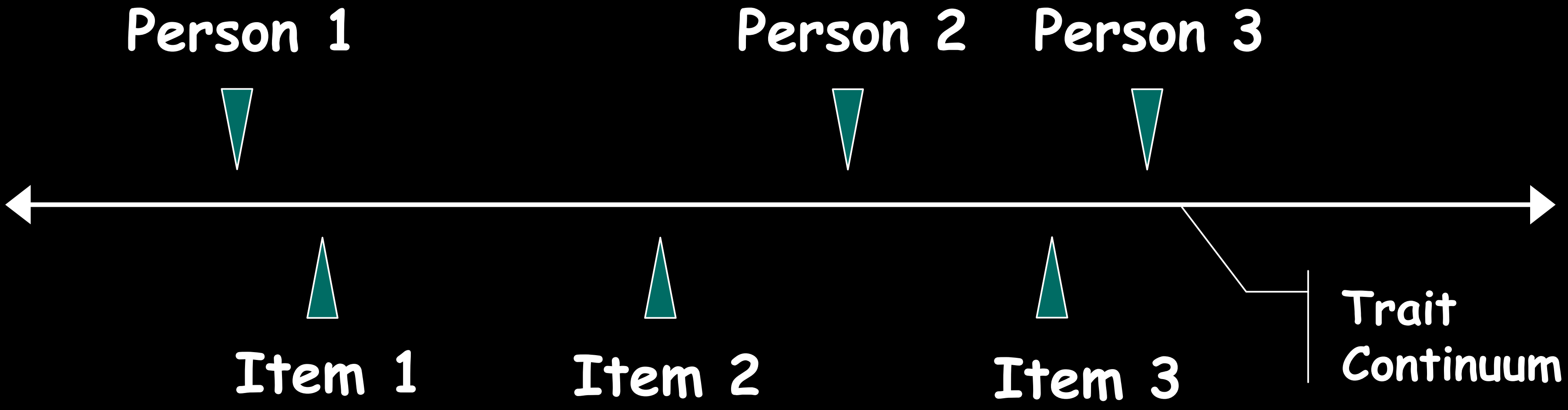


Latent Trait and Item Responses

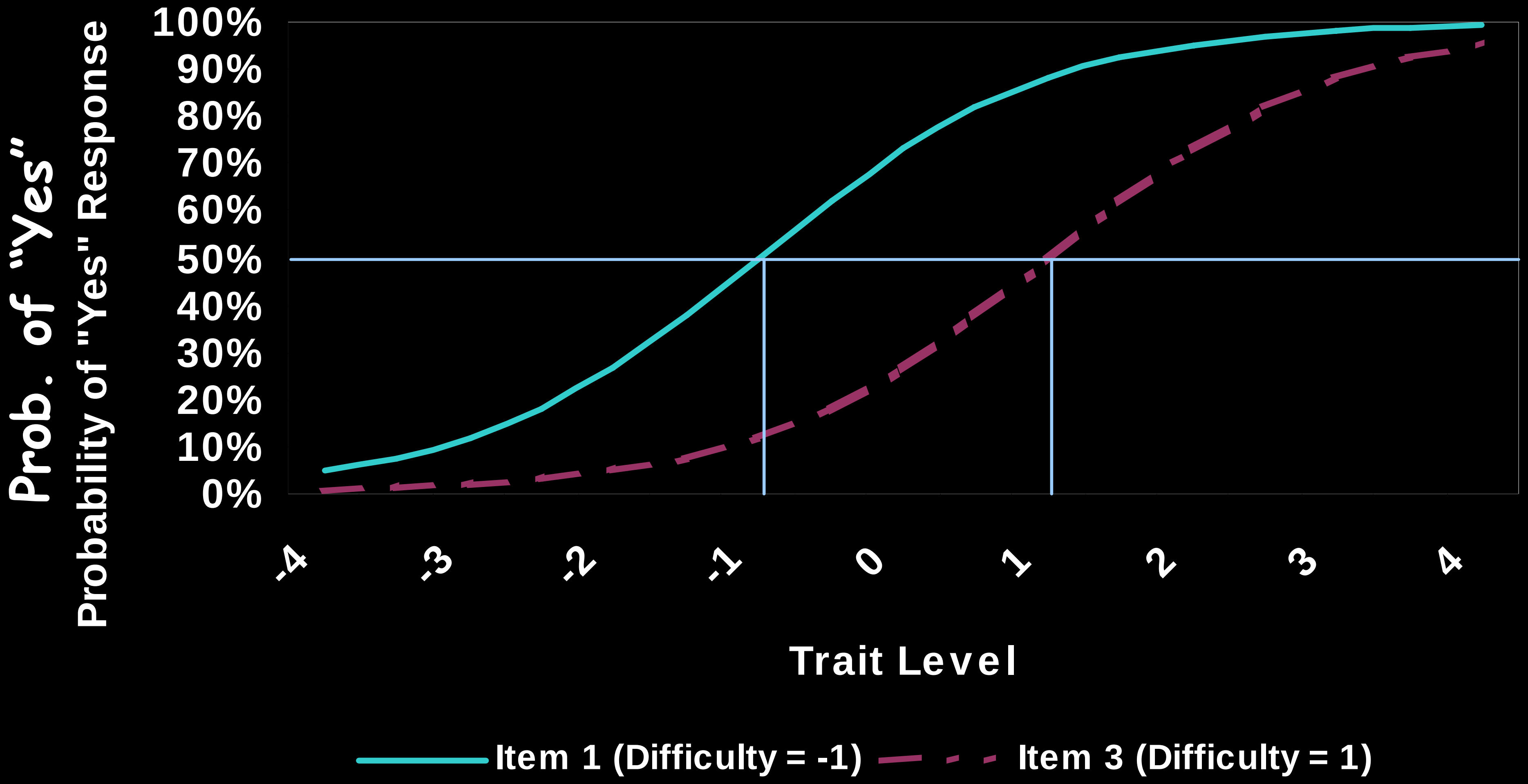


RAND HEALTH

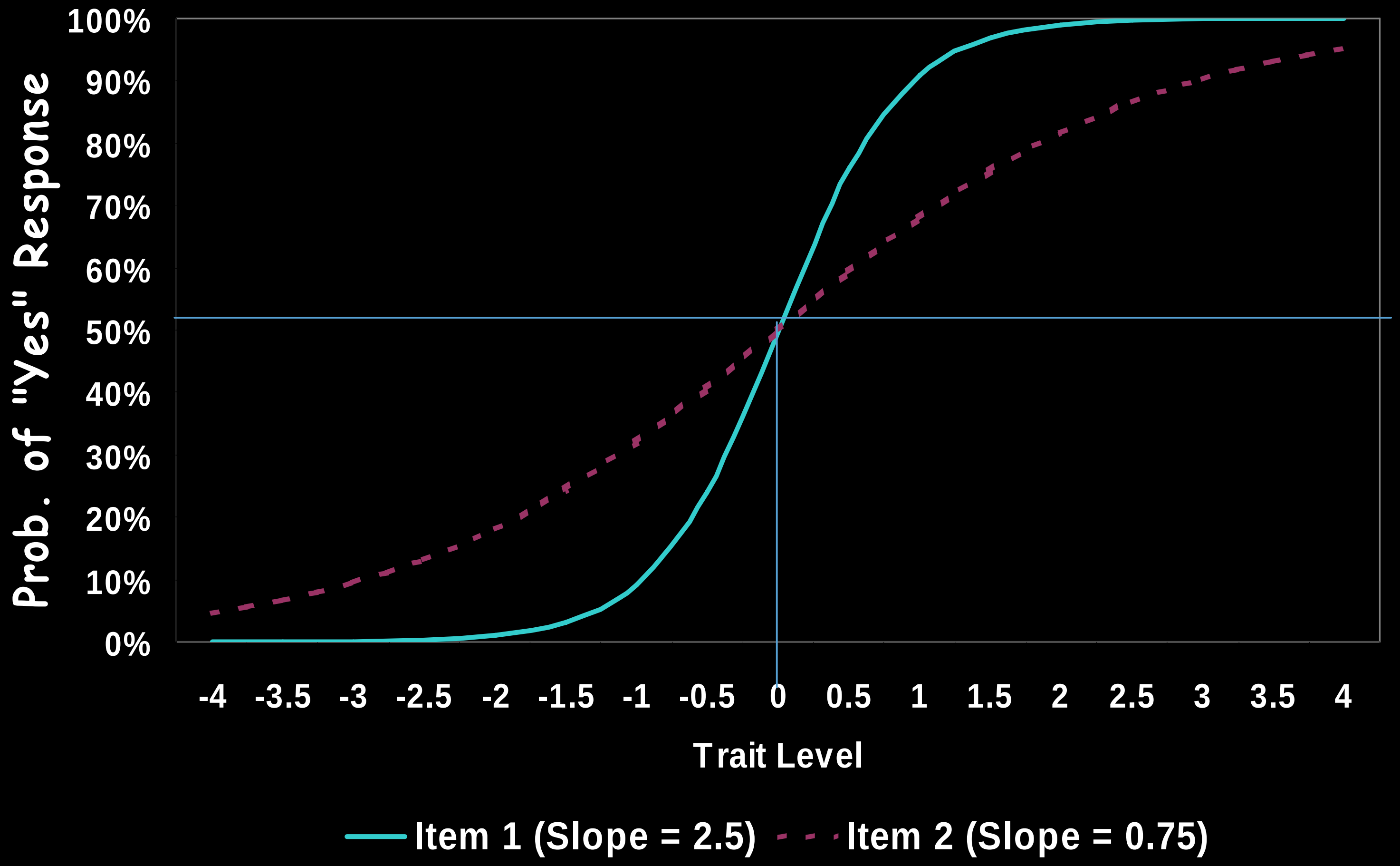
Item Responses and Trait Levels



Item Characteristic Curves (1-Parameter Model)

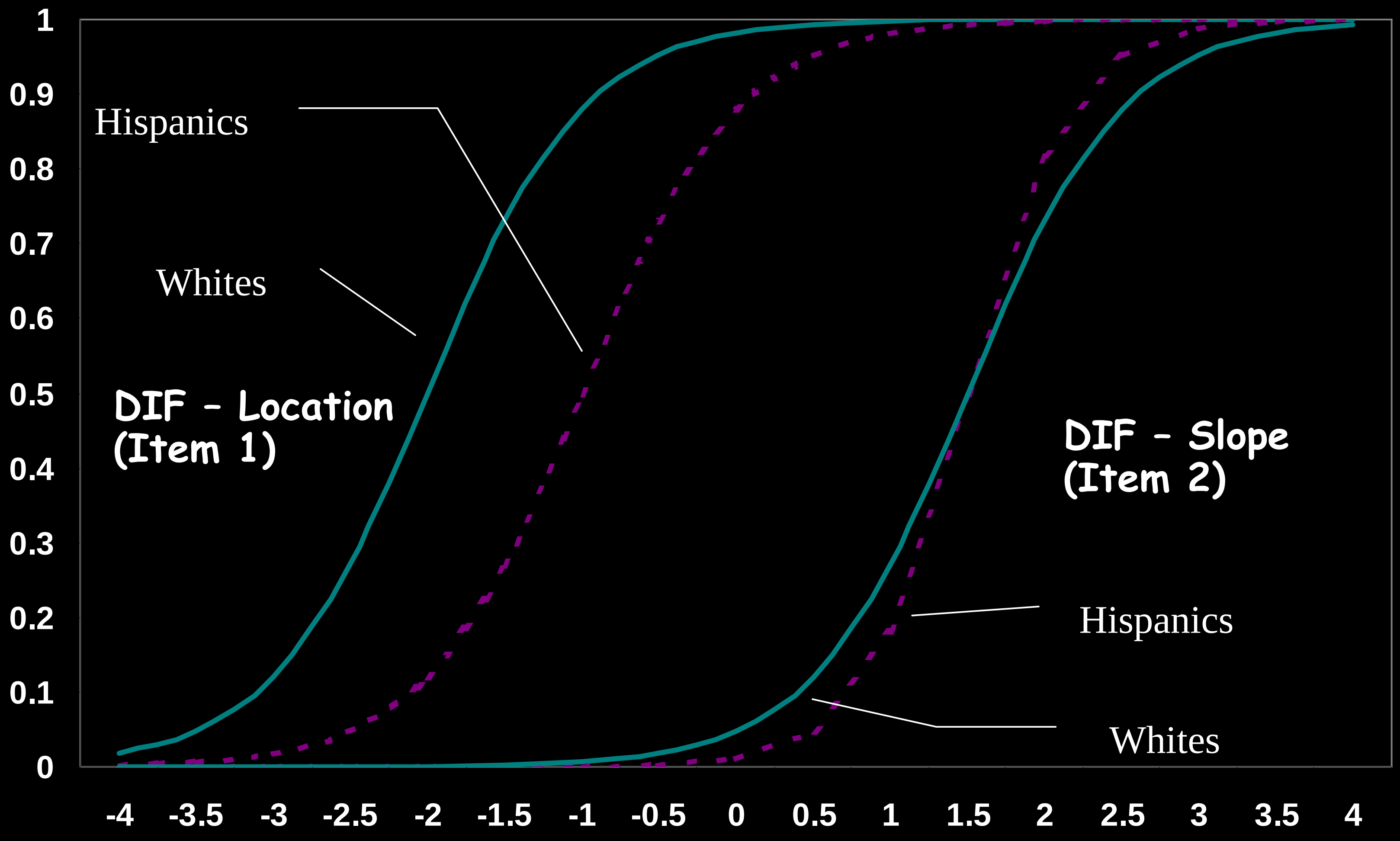


Item Characteristic Curves (2-Parameter Model)



RAND HEALTH

Dichotomous Items Showing DIF
(2-Parameter Model)



RAND HEALTH