Evaluating Self-Report Data Using Psychometric Methods

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

February 5, 2003 (3:00-6:00pm)

Four Types of Data Collection Errors

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

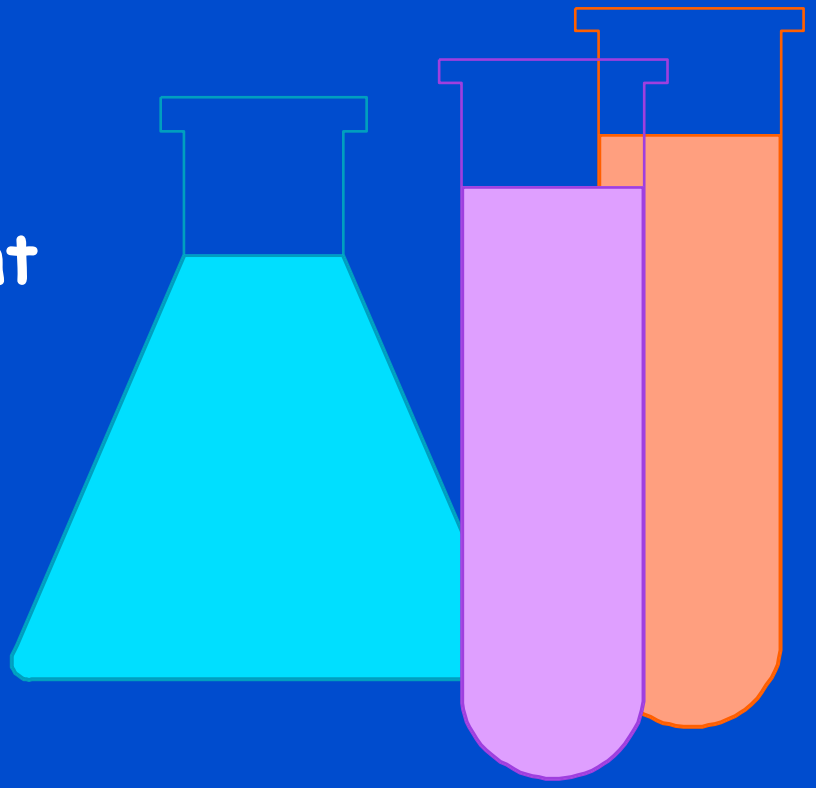
- Sampling Error
  Are only some members of the population sampled?

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

- Measurement Error
  Are inaccurate answers given to survey questions?

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)



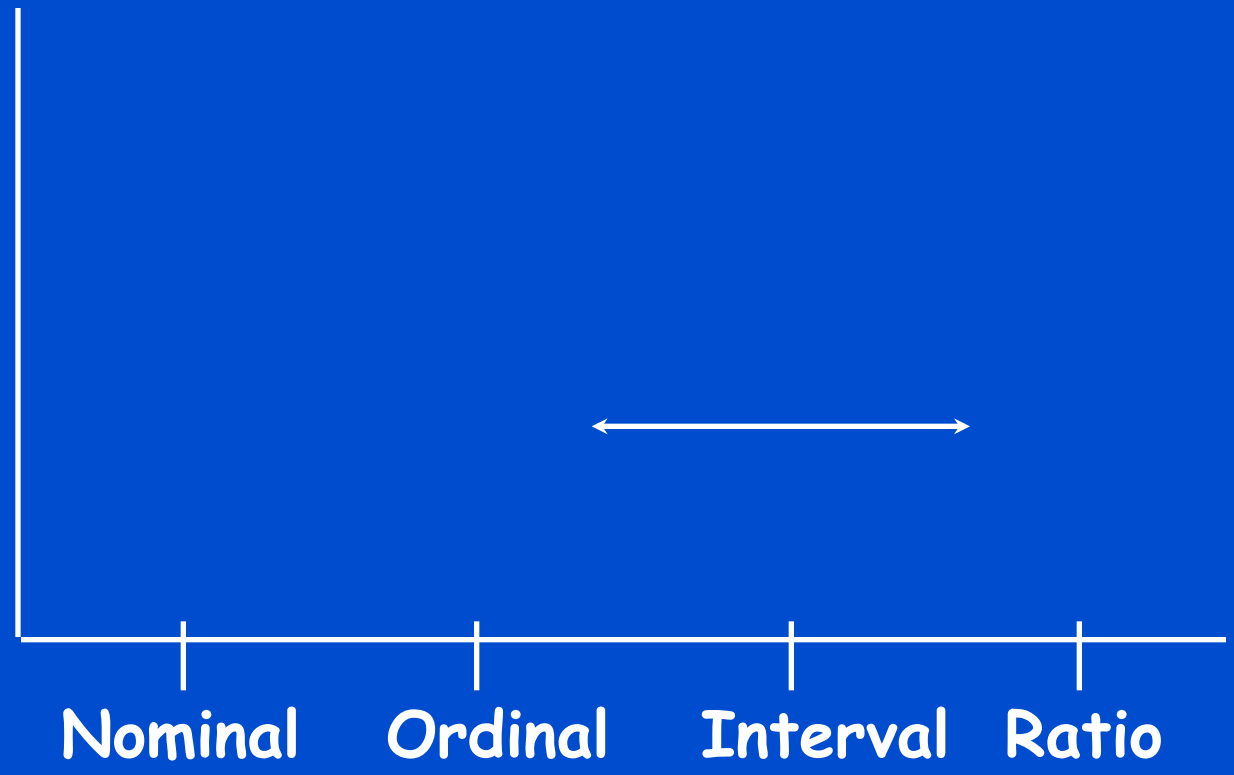
How Are Good Measures Developed?

- 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

Scales of Measurement and Their Properties

Type of Scale	Property of Numbers: Rank Order	Equal Interval	Absolute 0
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

Measurement Range for Health Outcome Measures



Indicators of Acceptability

- Response rate
- Administration time
- Missing data (item, scale)

Variability

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

Measurement Error

observed = true score + systematic error + random error

(bias)

Flavors of Reliability

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

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

MMPI 362 \ MMPI 317	True	False	
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)

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

Example of Computing KAPPA

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

Example of Computing KAPPA (Continued)

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

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

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

Guidelines for Interpreting Kappa

Conclusion	Kappa
Poor	< .40
Fair	.40 - .59
Good	.60 - .74
Excellent	> .74

Fleiss (1981)

Conclusion	Kappa
Poor	< 0.0
Slight	.00 - .20
Fair	.21 - .40
Moderate	.41 - .60
Substantial	.61 - .80
Almost perfect	.81 - 1.00

Landis and Koch (1977)

Ratings of Height of Houseplants

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		

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	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{EMS} + (K-1)MS_{EMS}}$
Two-Way Random	$\frac{N(MS_{BMS} - MS_{EMS})}{NMS_{BMS} + MS_{JMS} - MS_{EMS}}$	$\frac{MS_{BMS} - MS_{EMS}}{MS_{BMS} + (K-1)MS_{EMS} + K(MS_{JMS} - MS_{EMS})/N}$

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

$$ICC(1,1) = \frac{BMS - WMS}{BMS + (K - 1) * WMS}$$

$$ICC(2,1) = \frac{BMS - EMS}{BMS + (K - 1) * EMS + K(JMS - EMS)/n}$$

$$ICC(3,1) = \frac{BMS - EMS}{BMS + (K - 1) * EMS}$$

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 by Number of Items and Inter-item Correlations

$$alpha_{st} = \frac{K\,\bar{r}}{1 + (K - 1)\,\bar{r}}$$

K = number of items in scale

Alpha for Different Numbers of Items and Homogeneity

Number of Items (K)	Average Inter-item Correlation ($\bar{r}$)					
	.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

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

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

Reliability Minimum Standards

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

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.

http://www.tricare.osd.mil/tricaresurveys/rel_val.html

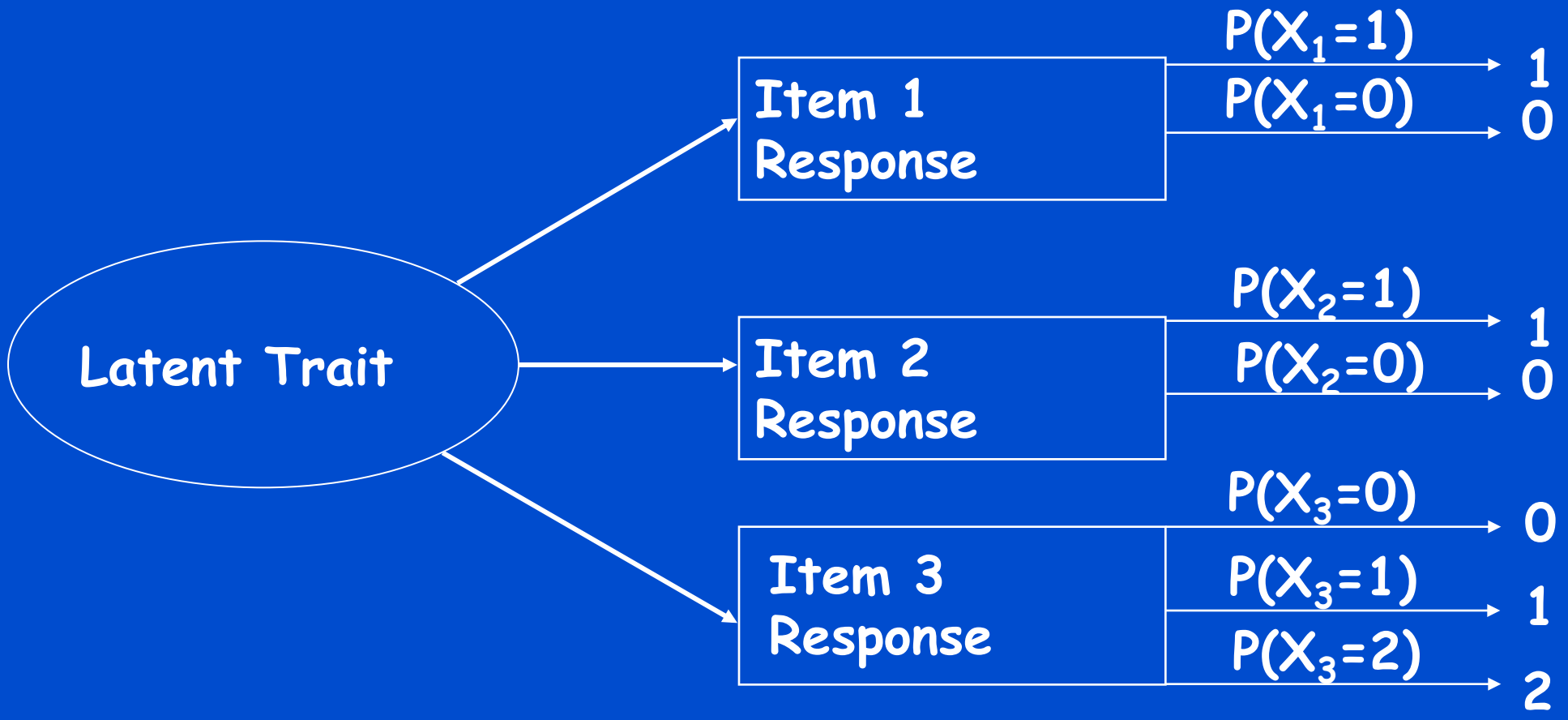
	Item Internal Consistency Validity Correlation		
Questions	Interpersonal Relationship	Quality of Medical Care	Access to Medical Care
Q3A	.82*	0.79	0.61
Q3B	.85*	0.82	0.6
Q3E	.89*	0.88	0.6
Q3F	.84*	0.84	0.59
Q3G	.84*	0.84	0.63
Q3C	0.89	.87*	0.6
Q3D	0.88	.85*	0.59
Q3H	0.81	.88*	0.6
Q3I	0.81	.90*	0.6
Q3J	0.89	.91*	0.63
Q07	0.51	0.5	.69*
Q09	0.59	0.56	.64*
Q10A	0.53	0.5	.79*
Q10B	0.58	0.57	.82*
Q10C	0.58	0.57	.80*
Q11	0.59	0.57	.81*
* Denotes questions include the Scale			

IRT

What are IRT Models?

Mathematical equations that relate observed survey responses to a persons location on an unobservable latent trait (i.e., intelligence, patient satisfaction).

Latent Trait and Item Responses



IRT Model Assumptions

- Unidimensionality
  - One construct measured by items in scale.
- Local Independence
  - Items uncorrelated when latent trait(s) have been controlled for.

Types of IRT Models

- Unidimensional and multidimensional
- Dichotomous and polytomous
- Parameterization
  - One parameter: difficulty (location)
  - Two Parameter: difficulty and slope (discrimination)

Item difficulty

Transform proportion of people endorsing the item (p) to correspond to (1-p)th percentile from z distribution

$$Z = \ln (1-p)/p)/1.7 = (\ln (1-p) - \ln (p))/1.7$$

$$= (\ln (.228) - \ln (.772))/1.7$$

$$= (-1.47840965 + .258770729)/1.7$$

$$= -1.21963892/1.7$$

$$= \underline{-0.72}$$

Item Discrimination

- Item-scale correlation, corrected for item overlap
  - $Z' = \frac{1}{2} [\ln (1 + r) - \ln (1-r)]$
  - if r = 0.30, z = 0.31
  - if r = 0.80, z = 1.10
  - if r = 0.95, z = 1.83

(0.5 -> 2 is typical range)

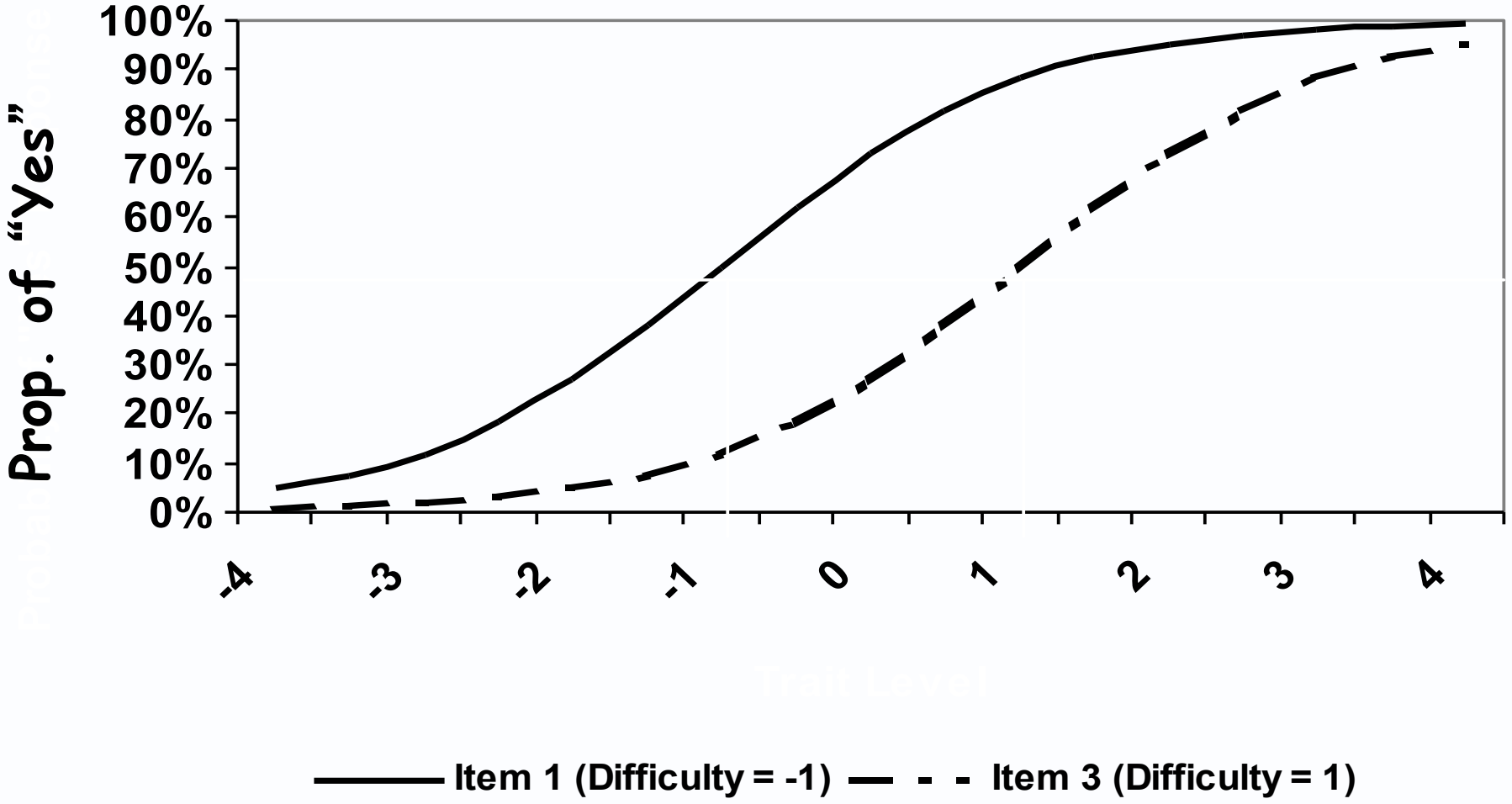
1-Parameter Logistic Model for (Dichotomous Outcomes)

$$P_i(\Theta) = \frac{e^{(\Theta - b_i)}}{1 + e^{(\Theta - b_i)}}$$

$P_i(\Theta)$ Probability that a randomly selected respondent with ability Θ (trait level) answers item i correctly.

b_i Item i difficulty.

Item Characteristic Curves (1-Parameter Model)



2-Parameter Logistic Model
(Dichotomous Outcomes)

$$P_i(\Theta) = \frac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i(\Theta - b_i)}}$$

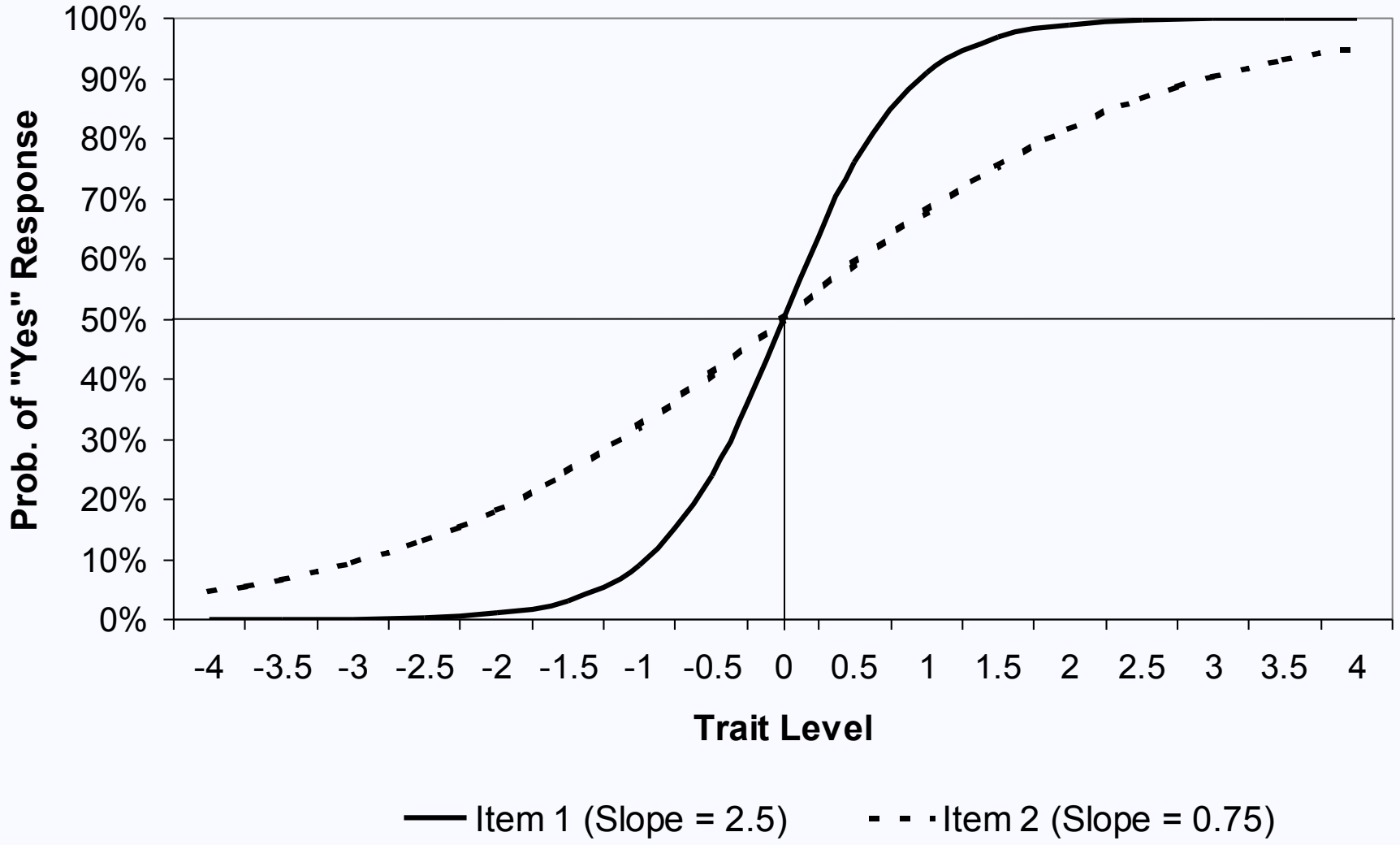
$P_i(\Theta)$ Probability that a randomly selected respondent with ability Θ (trait level) answers item i correctly.

b_i Item i difficulty.

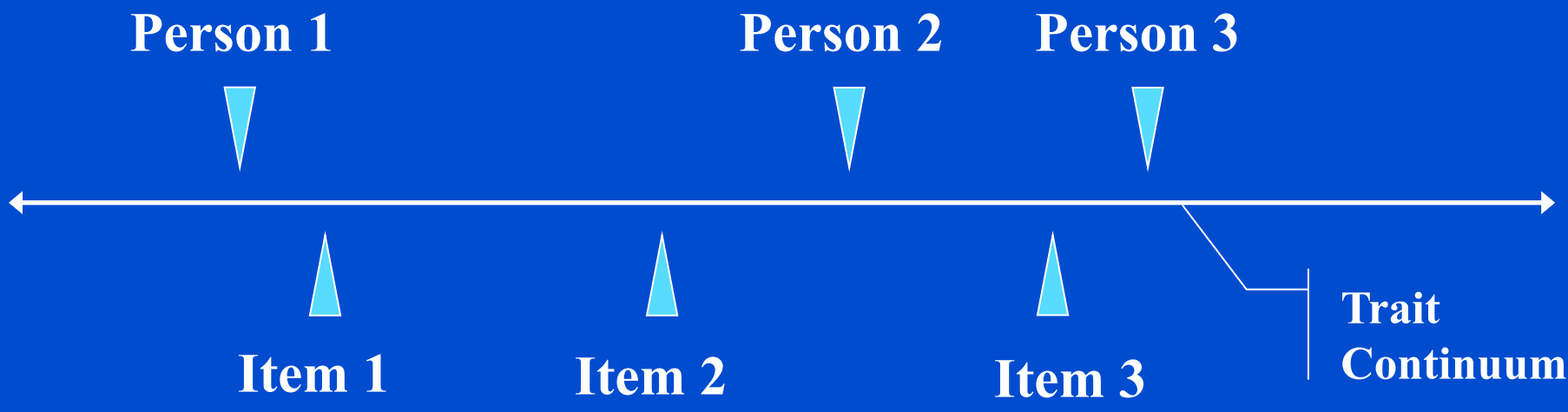
a_i Item i slope.

D Scaling constant.

Item Characteristic Curves (2-Parameter Model)



Item Responses and Trait Levels



Information Conditional on Trait Level

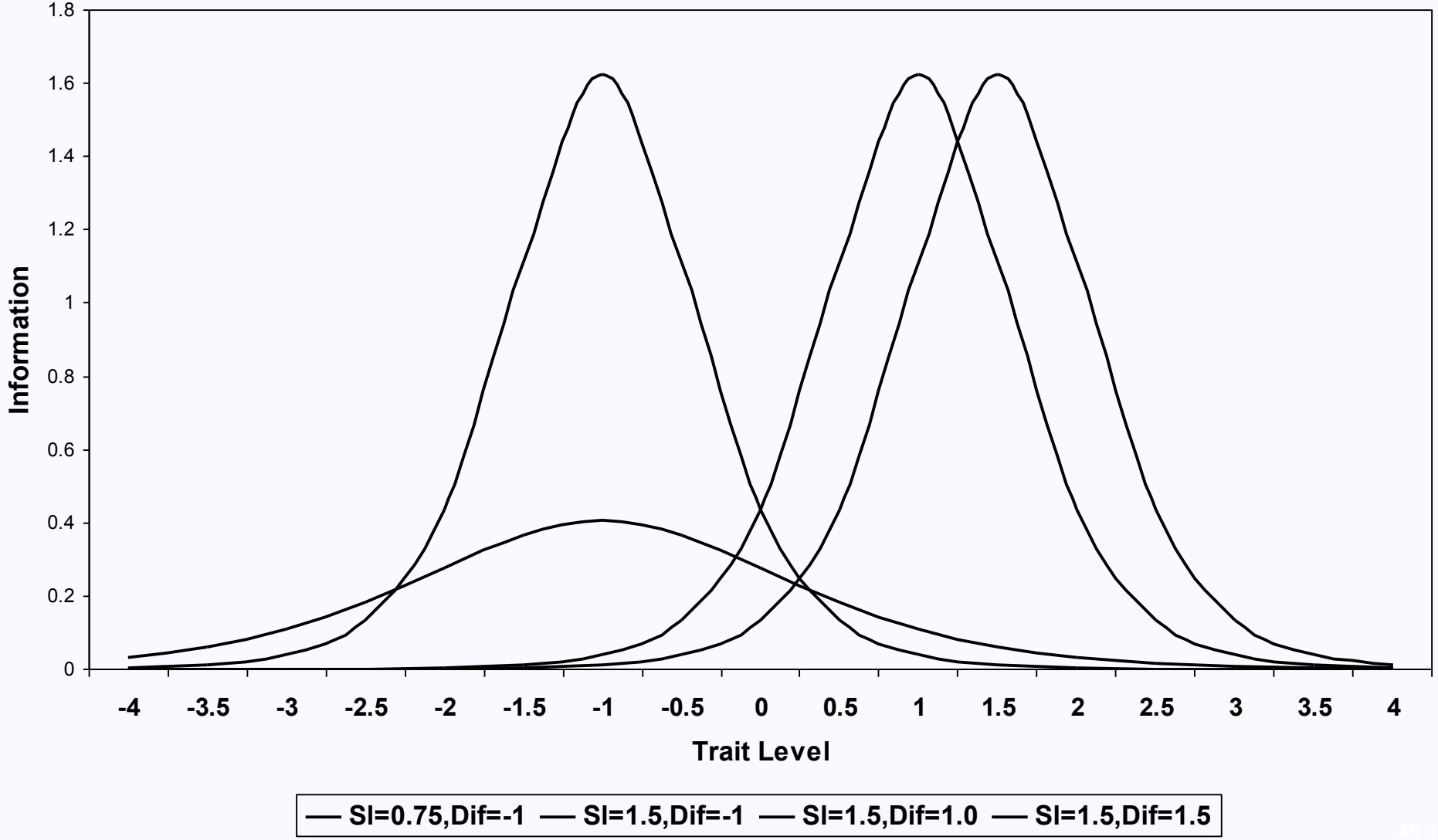
- Item information proportional to inverse of standard error:

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

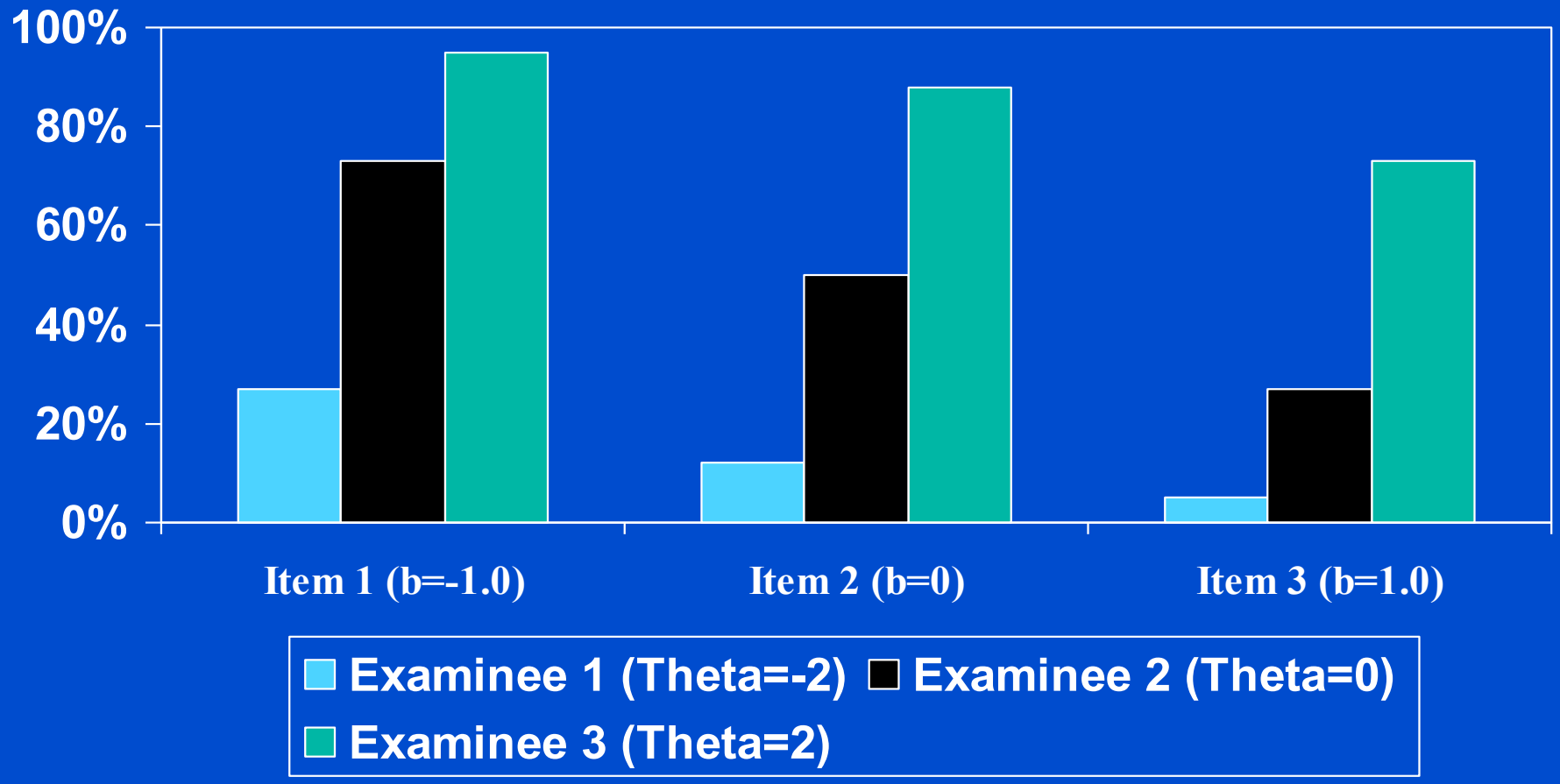
- Scale/Test information is the sum over item information:

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

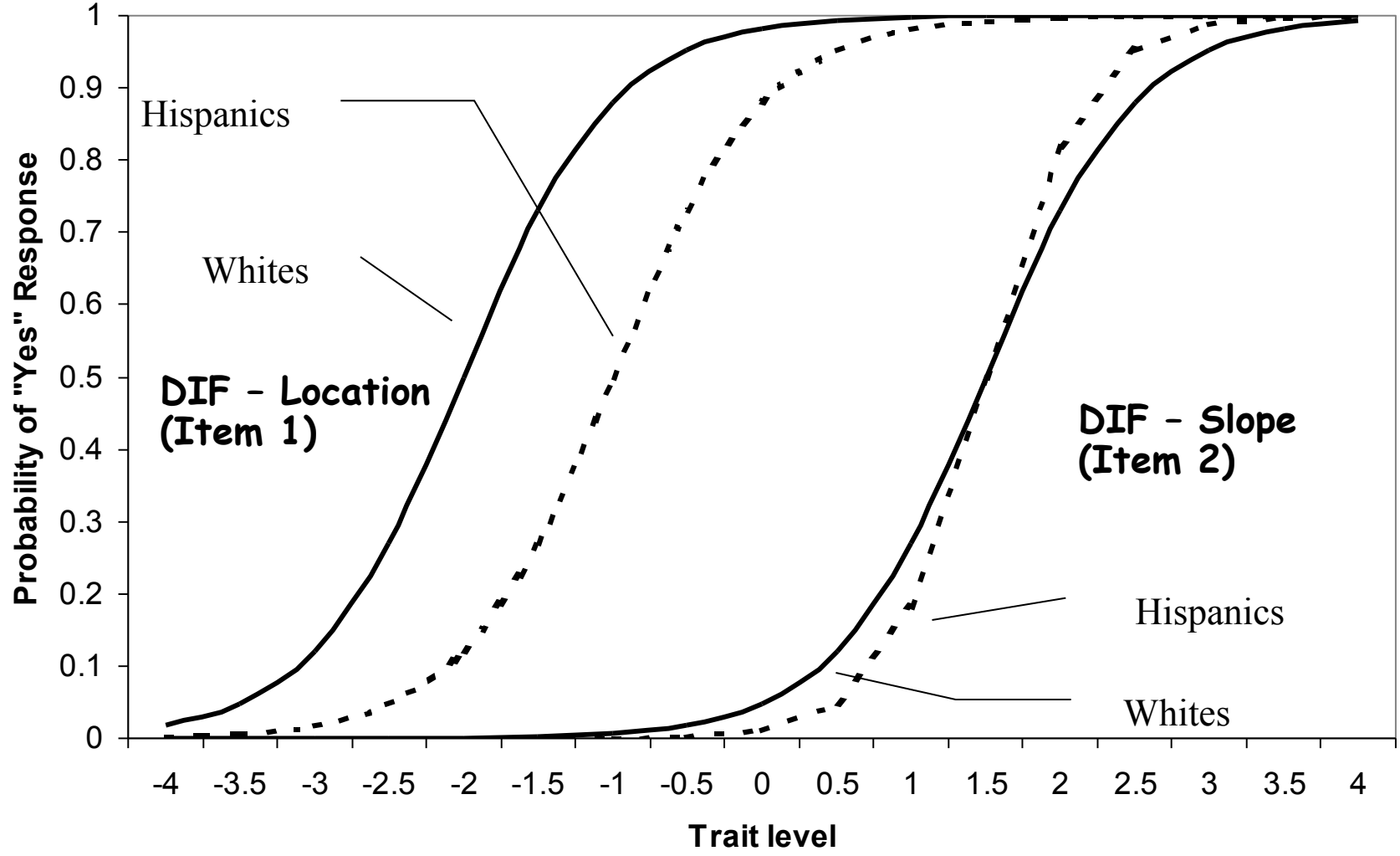
Item Information (2-parameter model)



Linking Item Content to Trait Estimates



Dichotomous Items Showing DIF
(2-Parameter Model)



Forms of Validity

- Content
- Criterion
- Construct Validity

Construct Validity

- Does measure relate to other measures in ways consistent with hypotheses?
- Responsiveness to change

Relative Validity Analyses

- Form of "known groups" validity
- Relative sensitivity of measure to important clinical difference
- One-way between group ANOVA

Relative Validity Example

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

Responsiveness to Change and Minimally Important Difference

- HRQOL measures should be responsive to interventions that changes HRQOL

- Evaluating responsiveness requires assessment of HRQOL
  - pre-post intervention of known efficacy
  - at two times in tandem with gold standard

Two Essential Elements

- External indicator of change (Anchors)
  - mean change in HRQOL scores among people who have a "minimal" change in HRQOL.
- Amount of HRQOL change

External Indicator of Change (A)

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

External Indicator of Change (B)

Rate your overall condition. This rating should encompass factors such as social activities, performance at work or school, seizures, alertness, and functional capacity; that is, your overall quality of life.

7 response categories; ranging from no impairment to extremely severe impairment

External Indicator of Change (C)

- "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) = $D/SD^{\dagger}$

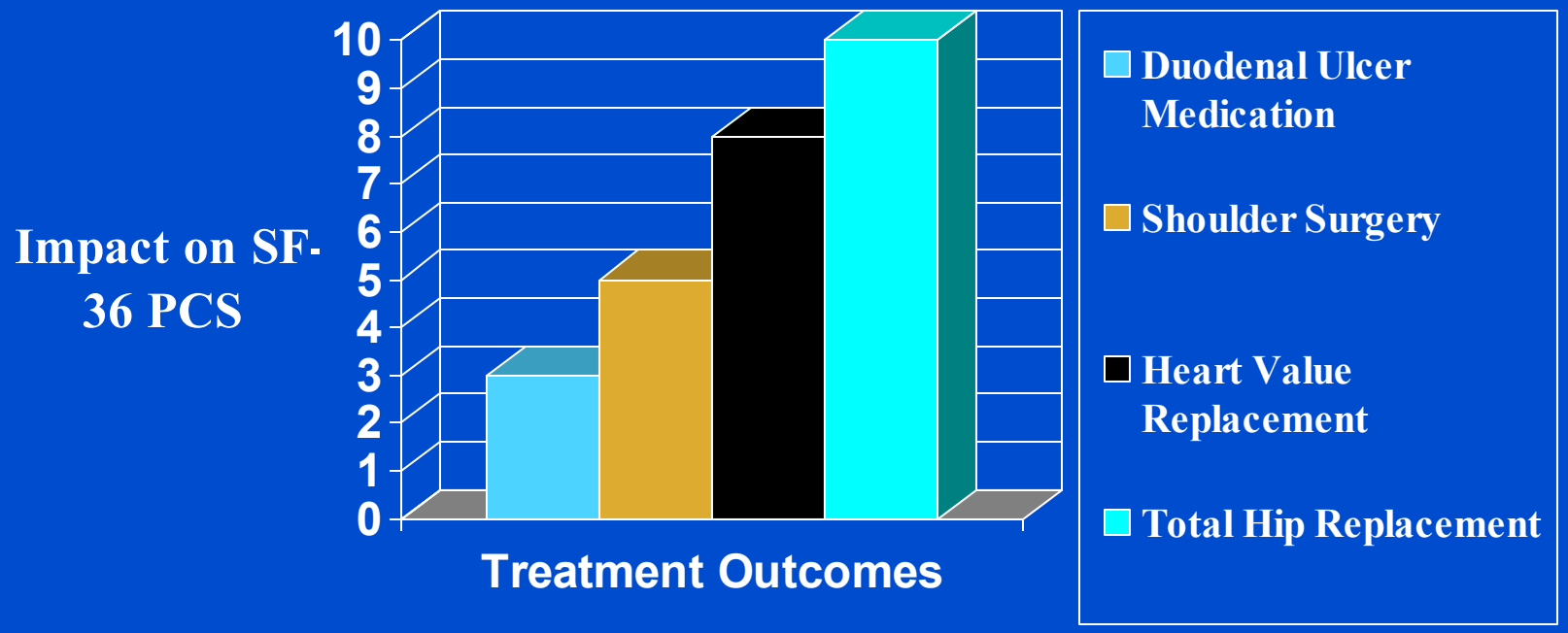
(3) Guyatt responsiveness statistic (RS) = $D/SD^{\ddagger}$

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



Treatment Impact on MCS