Item Response Theory and Computerized Adaptive Testing

Richard C. Gershon, PhD
Department of Medical Social Sciences
Feinberg School of Medicine
Northwestern University
gershon@northwestern.edu
May 20, 2011

PROMIS

Outline

- Item Response Theory
  - versus Classical Test Theory
- Uses of IRT
  - Item Banking
  - Short Forms
  - Computerized Adaptive Tests

Psychometrics

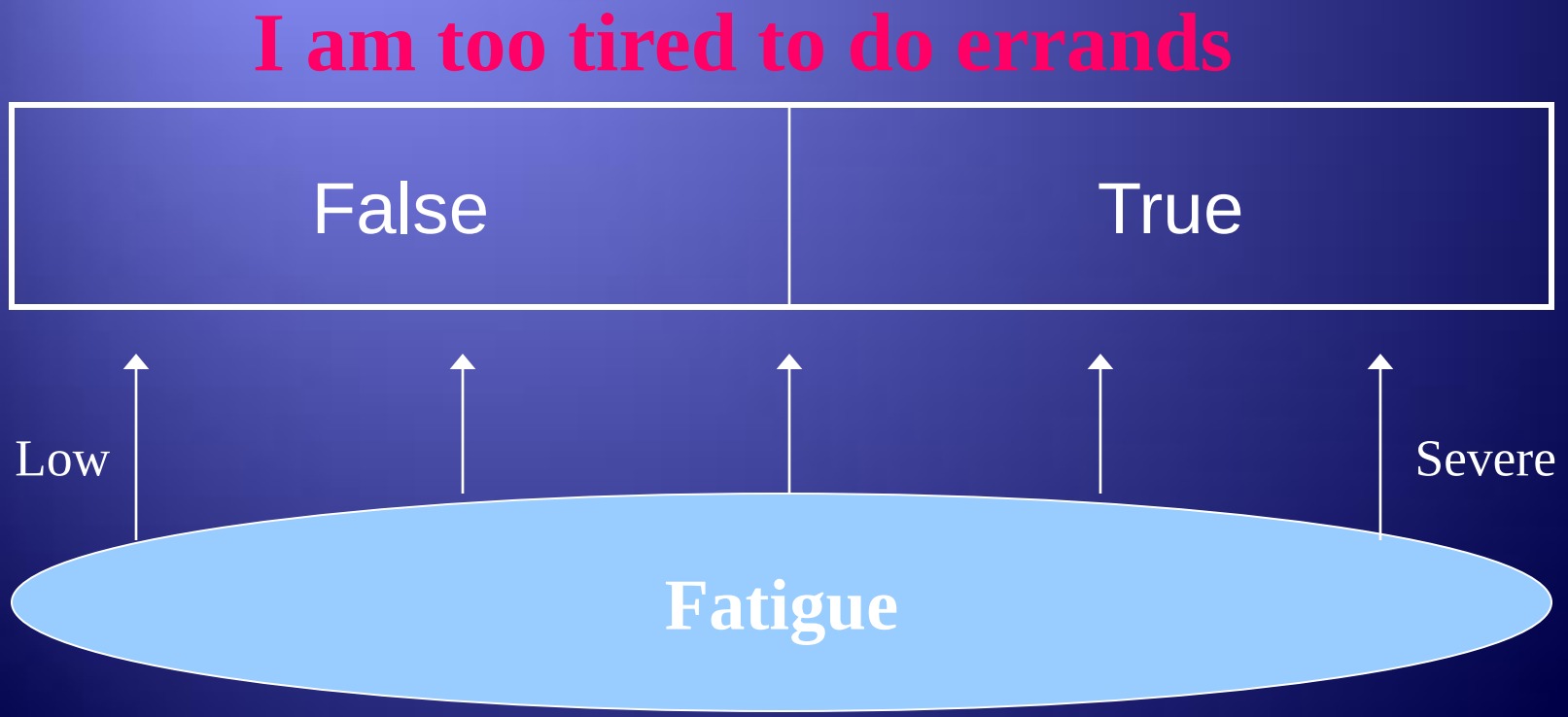
- **Psychometrics** is the field of study concerned with the theory and technique of educational and psychological measurement, which includes the measurement of knowledge (achievement), abilities, attitudes, and personality traits.
  - Source-Wikipedia

Requirements for Measurement

- Measurement requires the concept of an underlying trait that can be expressed in terms of **more** *or* **less**
- Test items are the operational definition of the underlying trait
- Test items
  - can be ordered from easy to hard
- Test takers
  - can be ordered from less able to more able

IRT Modeling is Latent Trait Modeling

- A latent trait is an *unobservable* latent dimension that is thought to give rise to a set of observed item responses.



Latent Traits (con't)

- These latent traits (constructs, variables, θ) are measured on a **continuum** of severity.

Advantages of Using IRT

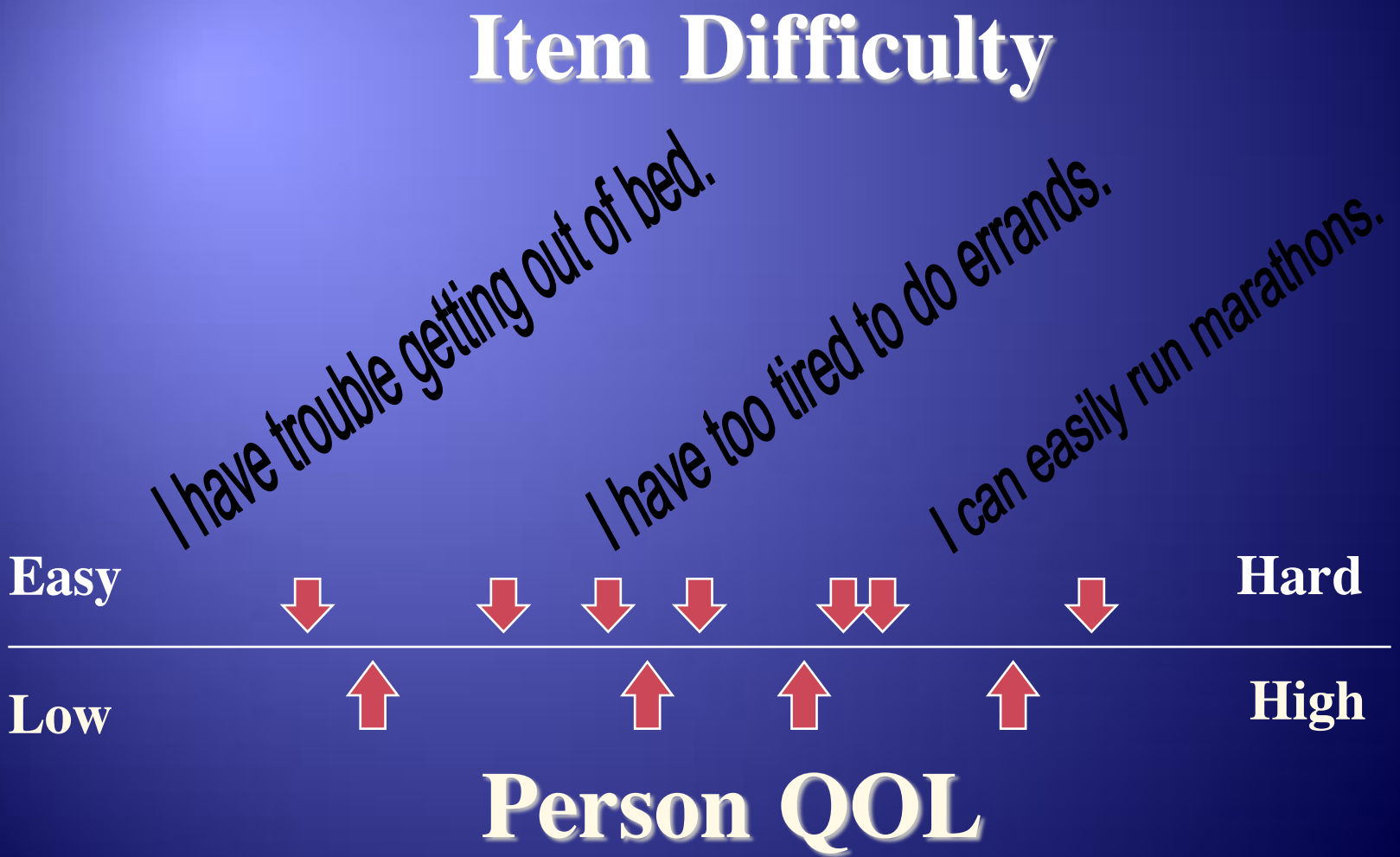
- Equal Interval Measure
- Test-takers and items are represented on the same scale
- Item calibrations are independent of the test-takers used for calibration
- Candidate ability estimates are independent of the particular set of items used for estimation
- Measurement precision is estimated for each person and each item

Test-takers and Items are Represented on the Same Scale

- Item Difficulty = Severity = Measure = Theta = Item Calibration = Location

- Person Ability = Measure = Theta = Person Calibration = Location

Test-takers and Items are Represented on the Same Scale



More Basic Terms

- Discrimination = the degree to which an item discriminates person ability
- Item Information = the area where an item discriminates
- Test Information = the area where the test discriminates

Last, but not Least

- Item "Parameters" =
  - IRT statistics about an item
  - Primary: Item Difficulty
  - Often: Item Discrimination
  - Sometimes: Guessing
  - Lots of other "ugly looking numbers"

Even More New Terms

- Differential Item Functioning (DIF)
  - Does an item have different item parameters for different subgroups?
    - Gender
    - Race
    - Age
    - Disease

The Three Main IRT Models

- Rasch model one parameter logistic model (**1PL**)
- Two parameter logistic model (**2PL**)
- Three parameter logistic model (**3PL**)

How to choose an appropriate IRT Model

OR

My religion is better than your religion!

WARNING!

- You are about to see about to see mathematical formulas!

One Parameter Logistic Model

$$P_{1,0} = \frac{e^{(ability - difficulty)}}{1 + e^{(ability - difficulty)}}$$

When the difficulty of a given item exactly matches the Examinee's ability level, then the person has 50% chance of answering that item correctly:

$$P_{1,0} = \frac{e^{(0)}}{1 + e^{(0)}} = \frac{1}{2} = .50$$

One Parameter Logistic Model

- Only option for small sample sizes
- Often the real model underlying a test labeled as three parameter
- Less costly
- "The simple solution is always the best"

Two Parameter Logistic Model

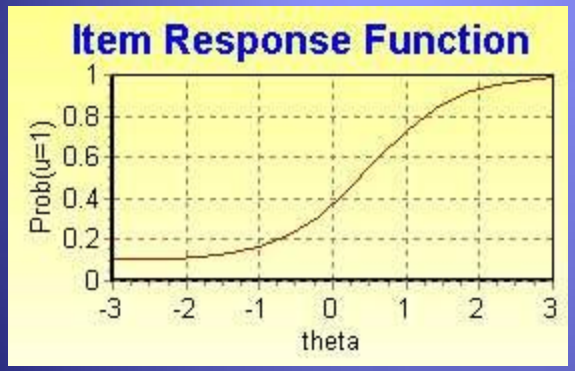
$$P_{1,0} = \frac{e^{a\,(ability - b)}}{1 + e^{a\,(ability - b)}}$$

Two parameters

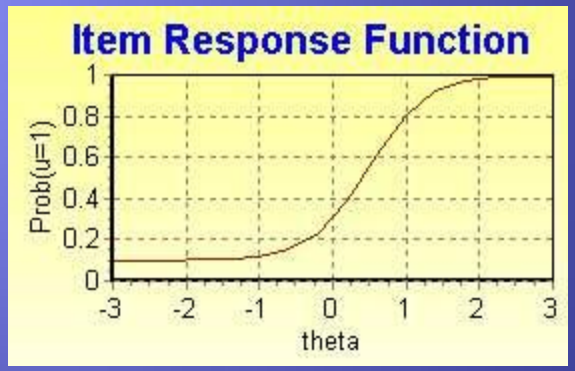
a=Discrimination

b=Item Difficulty

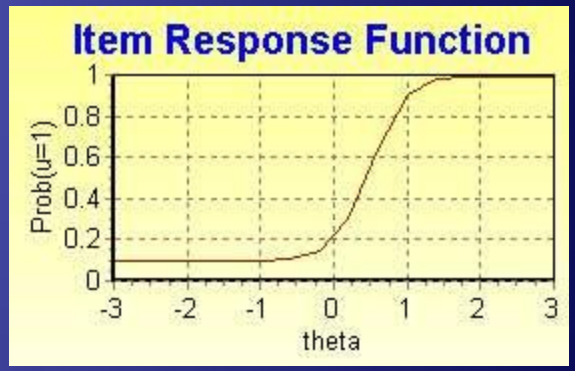
Two Parameter Examples



a=.5,b=.5,c=.1



a=1.5,b=.5,c=.1



a=2.5,b=.5,c=.1

Three Parameter Logistic Model

$$P_{1,0} = c + (1-c)\frac{e^{a\,(ability - b)}}{1 + e^{a\,(ability - b)}}$$

Three parameters

a= Discrimination

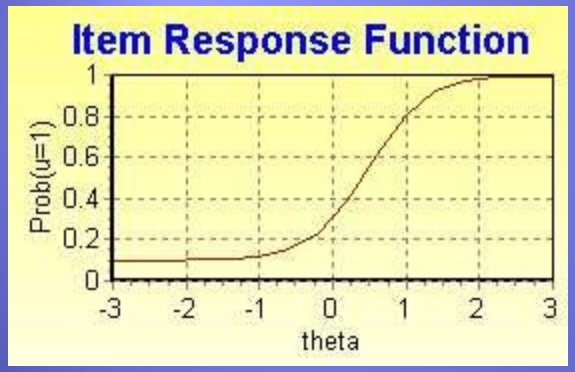
b= Item Difficulty

c= Guessing

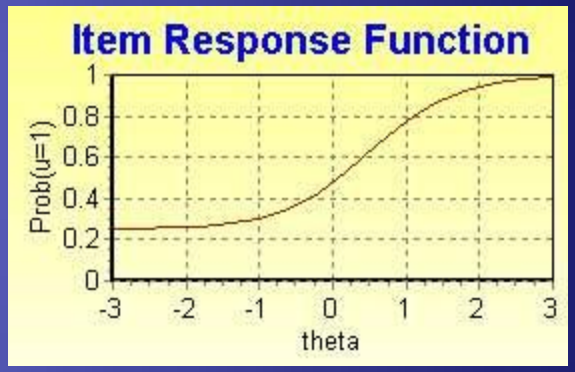
Three Parameter Logistic Model (3PL)

- Requires a large sample size
- Significant research demonstrating that theoretically 3PL is better, but practically has little advantage over 1PL
- "Most accepted theoretical model"

Three Parameter Examples



a=1.5,b=.5,c=.1



a=2.5,b=.5,c=.25

Polytomous Models

One Parameter

- Rating Scale Model
- Partial Credit Model

Two Parameter

- Graded Response Model
- Generalized Partial Credit Model

Multi-dimensional Models

- There are also IRT models which consider more than one unidimensional trait at a time

How does IRT differ from conventional test theory?

Classical Test Theory

- An individual takes an assessment
- Their total score on that assessment is used for comparison purposes
- High Score – The person is higher on the trait
- Low Score-The person is lower on the trait

Item Response Theory

- Each individual item can be used for comparison purposes
- Person endorses better rating on "hard items"- The person is higher on the trait
- Person endorses worse rating on "easy items" - The person is lower on the trait
- Items that measure the same construct can be aggregated into longer assessments

Reliability

CTT

- Reliability is based upon the total test.
- Regardless of patient "ability", reliability is the same.

IRT

- Reliability is calculated for each patient "ability" and varies across the continuum.
- Typically, there is better reliability in the middle of the distribution.

Validity

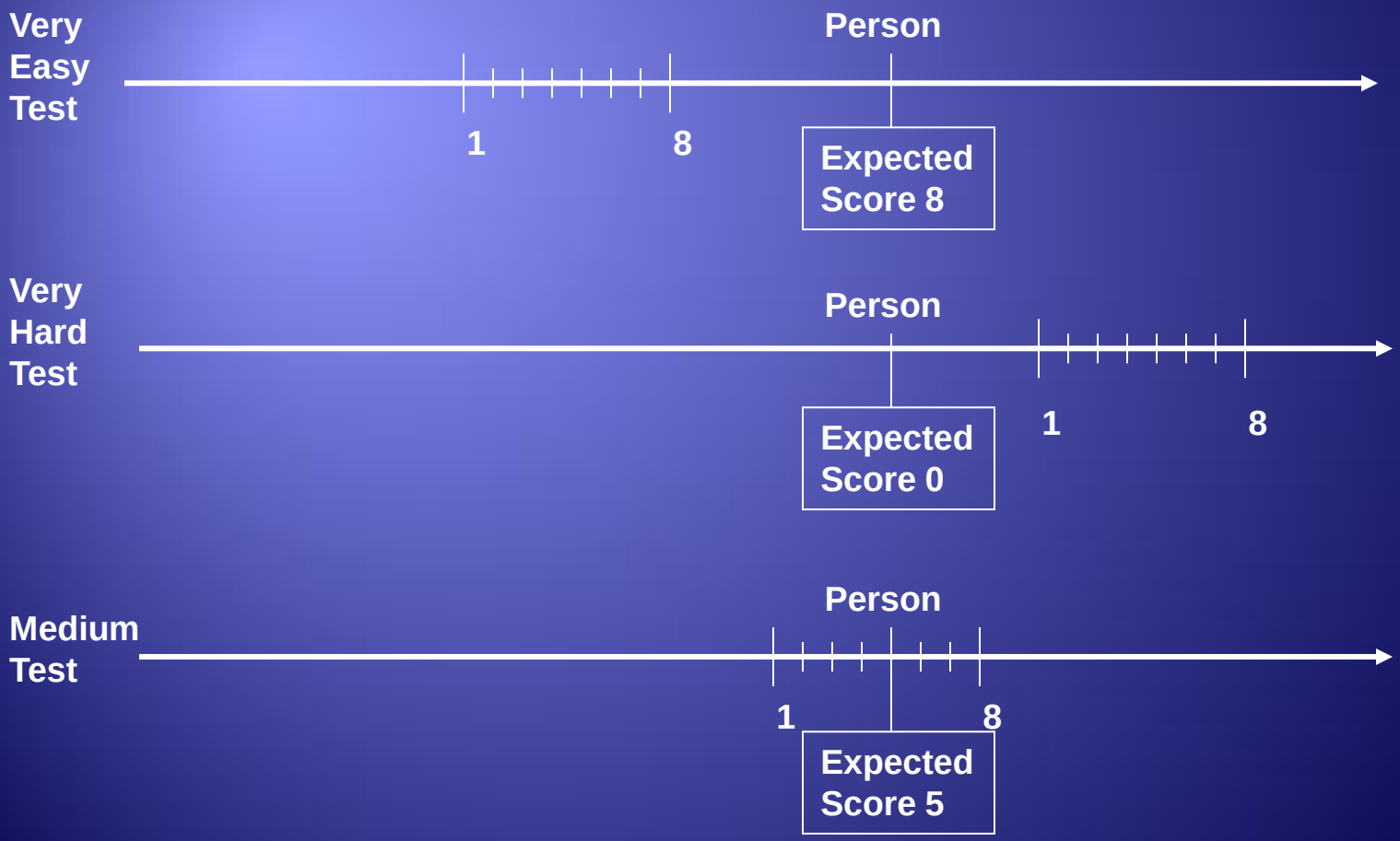
CTT

- Validity is based upon the total test.
- Typically, validity would need to be re-assessed if the instrument is modified in any way.

IRT

- Validity is assessed for the entire item bank.
- Subsets of items (full length tests, short forms and CAT) all inherit the validity assessed for the original item bank.

How Scores Depend on the Difficulty of Test Items



Reprinted with permission from: Wright, B.D. & Stone, M. (1979) *Best test design,* Chicago: MESA Press, p. 5.

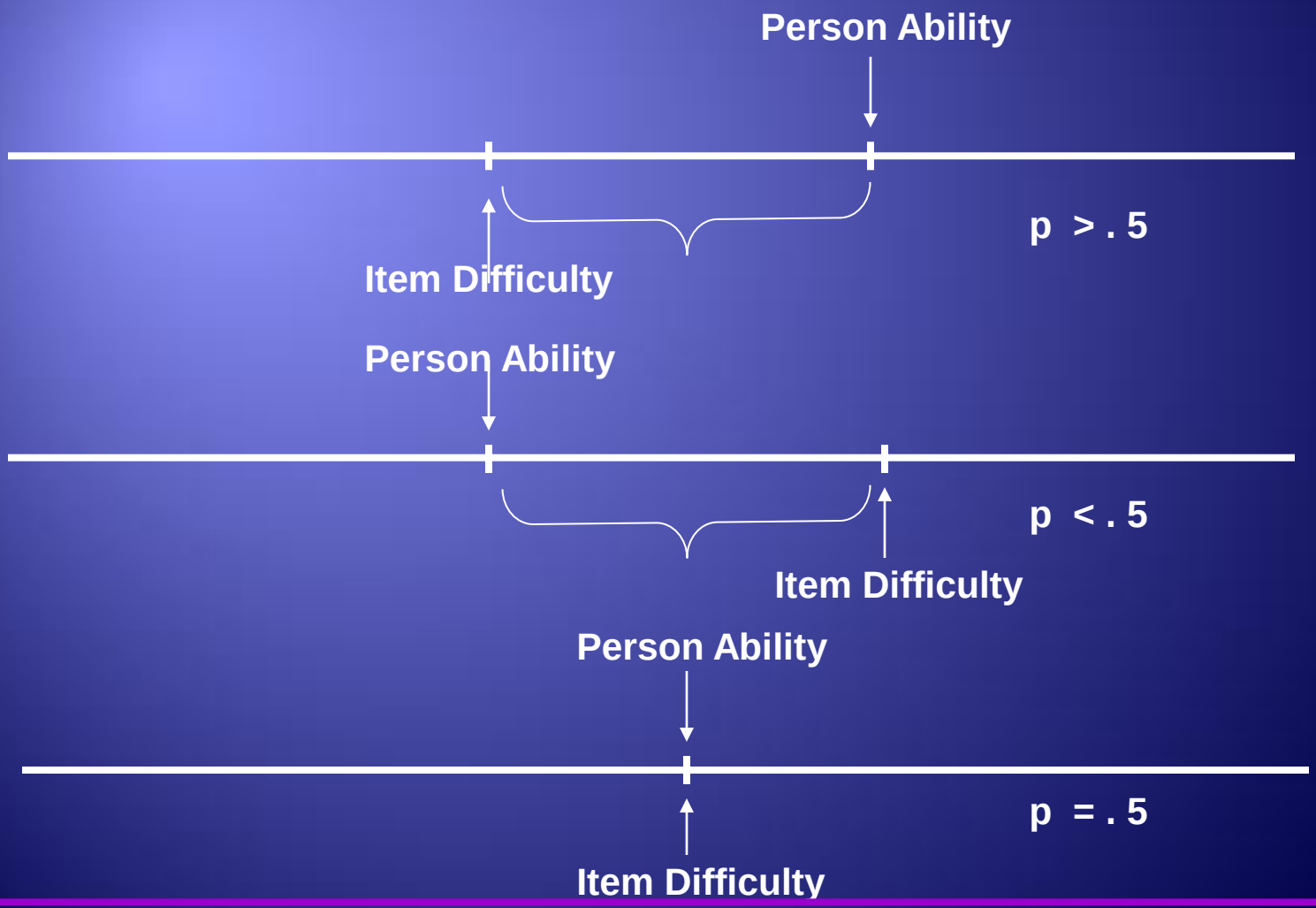
Raw Scores vs. IRT Measures
IRT has Equal Interval Measurement

4 Item Test

Raw:

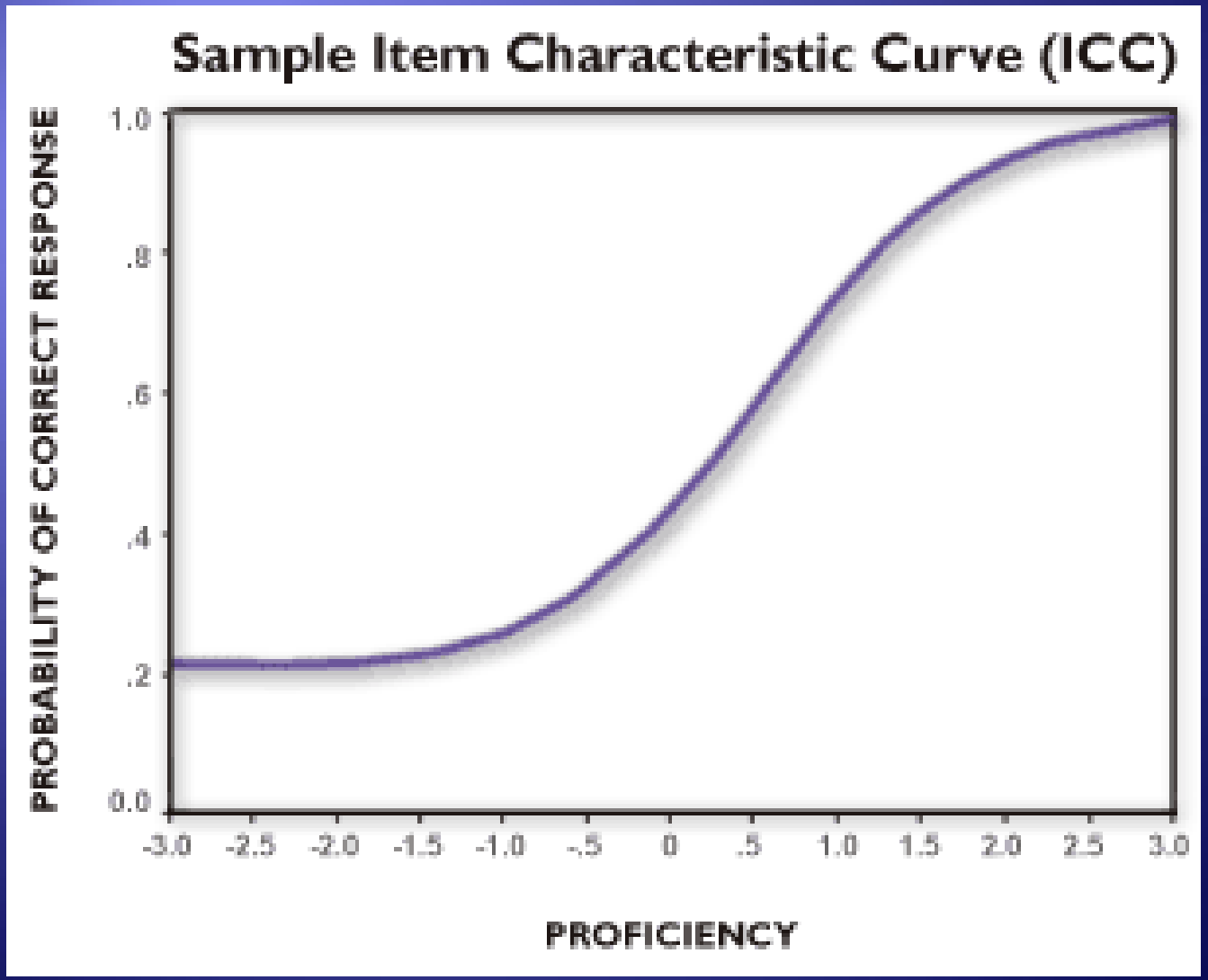
Logit Measures:

How Differences Between Person Ability and Item Difficulty Ought to Affect the Probability of a Correct Response



Reprinted with permission from: Wright, B.D. & Stone, M. (1979) *Best test design,* Chicago: MESA Press, p. 13.

The Item Characteristic Curve



I Have a Lack of Energy

Traditional Test Theory

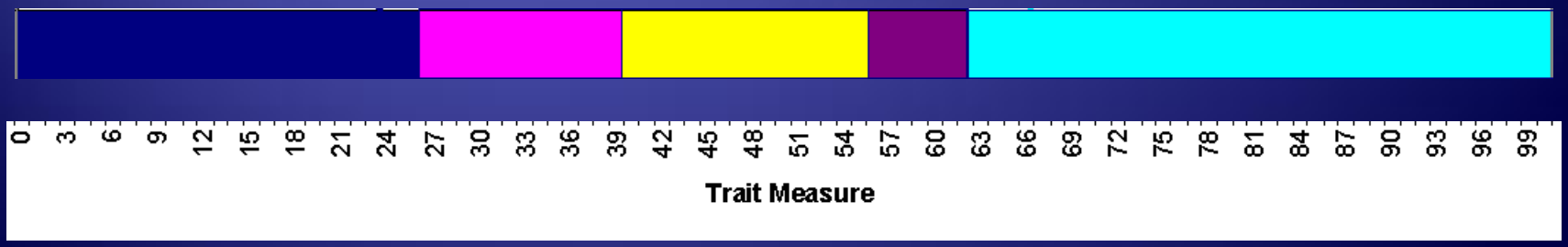
0 = Very Much 1 = Quite a Bit 2 = Somewhat 3 = A Little Bit 4 = Not at All

I Have a Lack of Energy

Traditional Test Theory

0 = Very Much 1 = Quite a Bit 2 = Somewhat 3 = A Little Bit 4 = Not at All

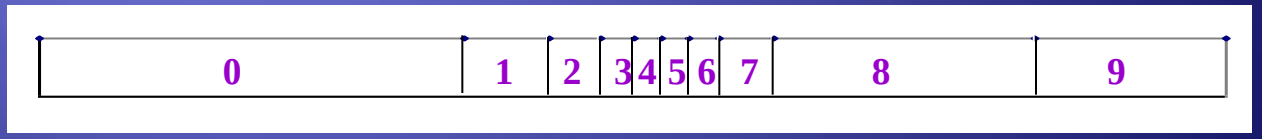
Item Response Theory

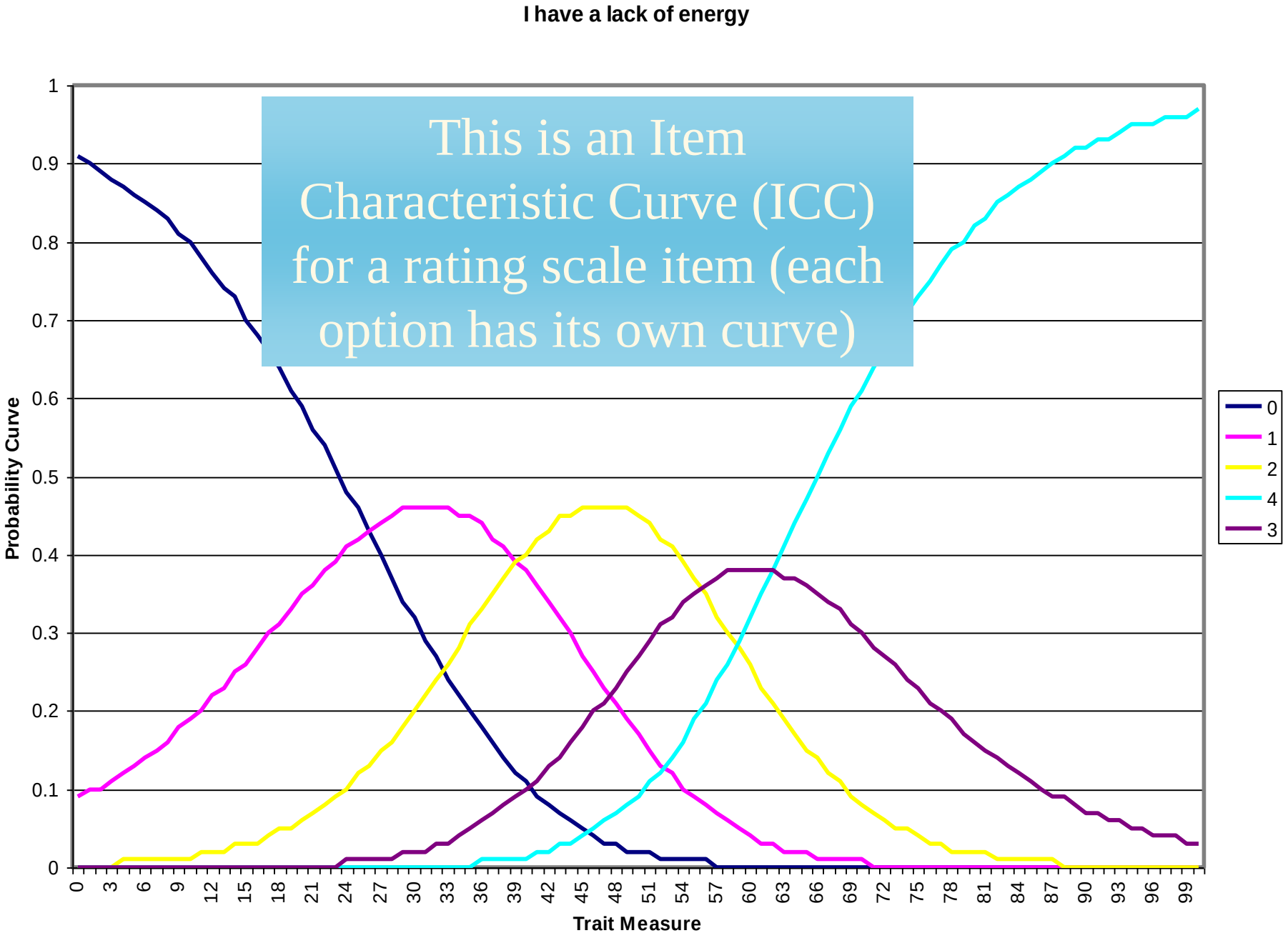


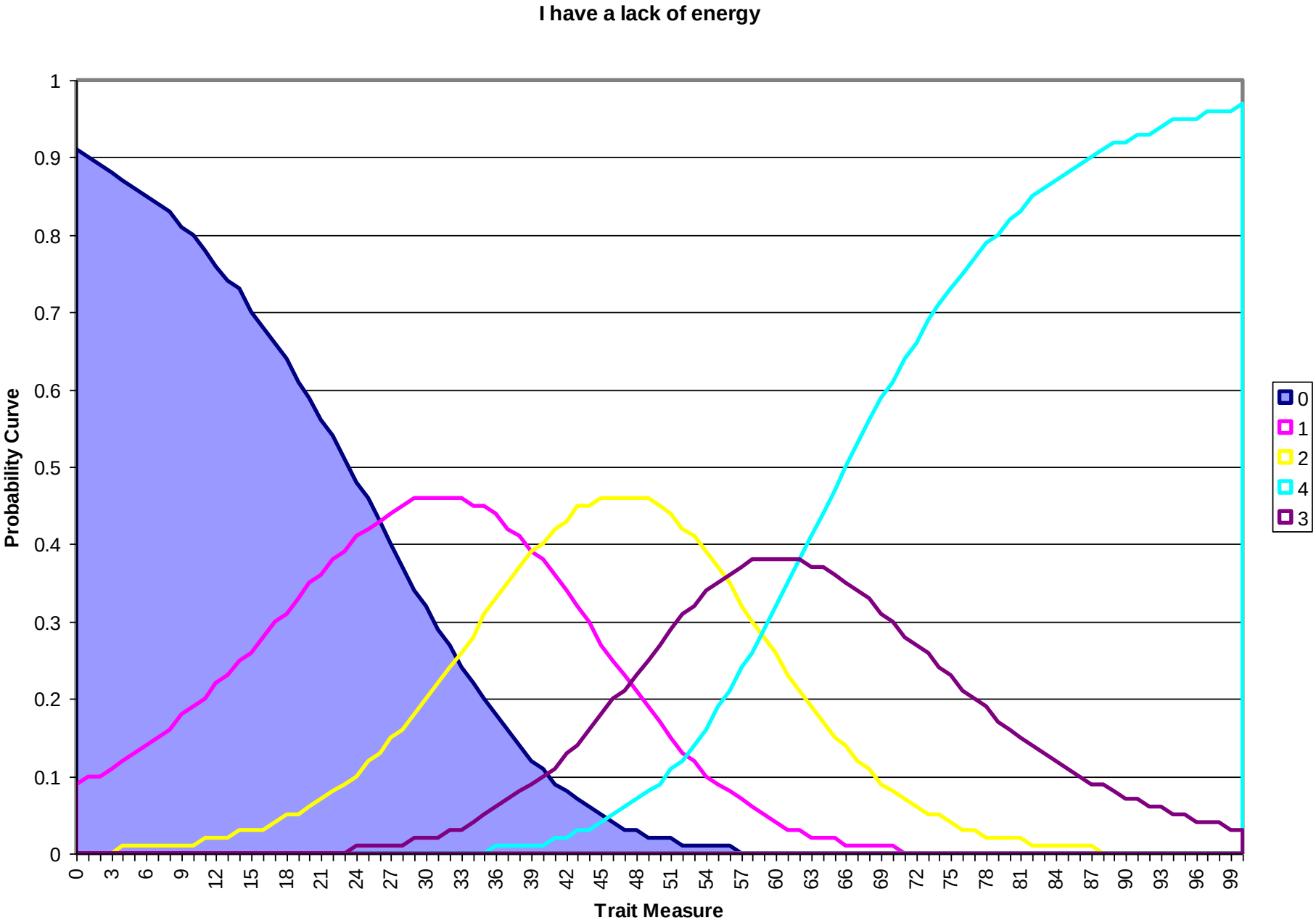
The IRT “Reality” of a 10 Point Rating-Scale Item

No Pain

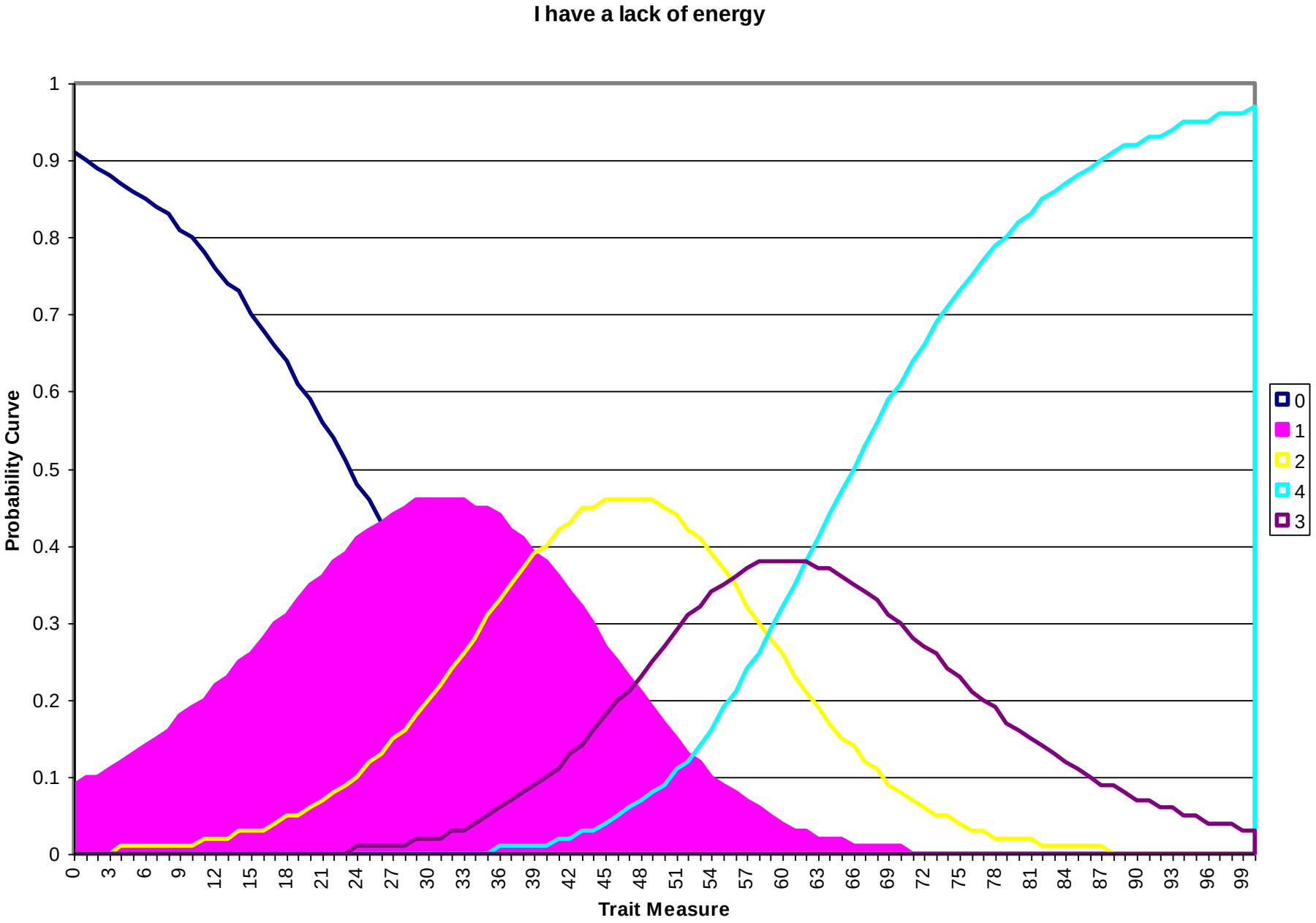
Worst Pain



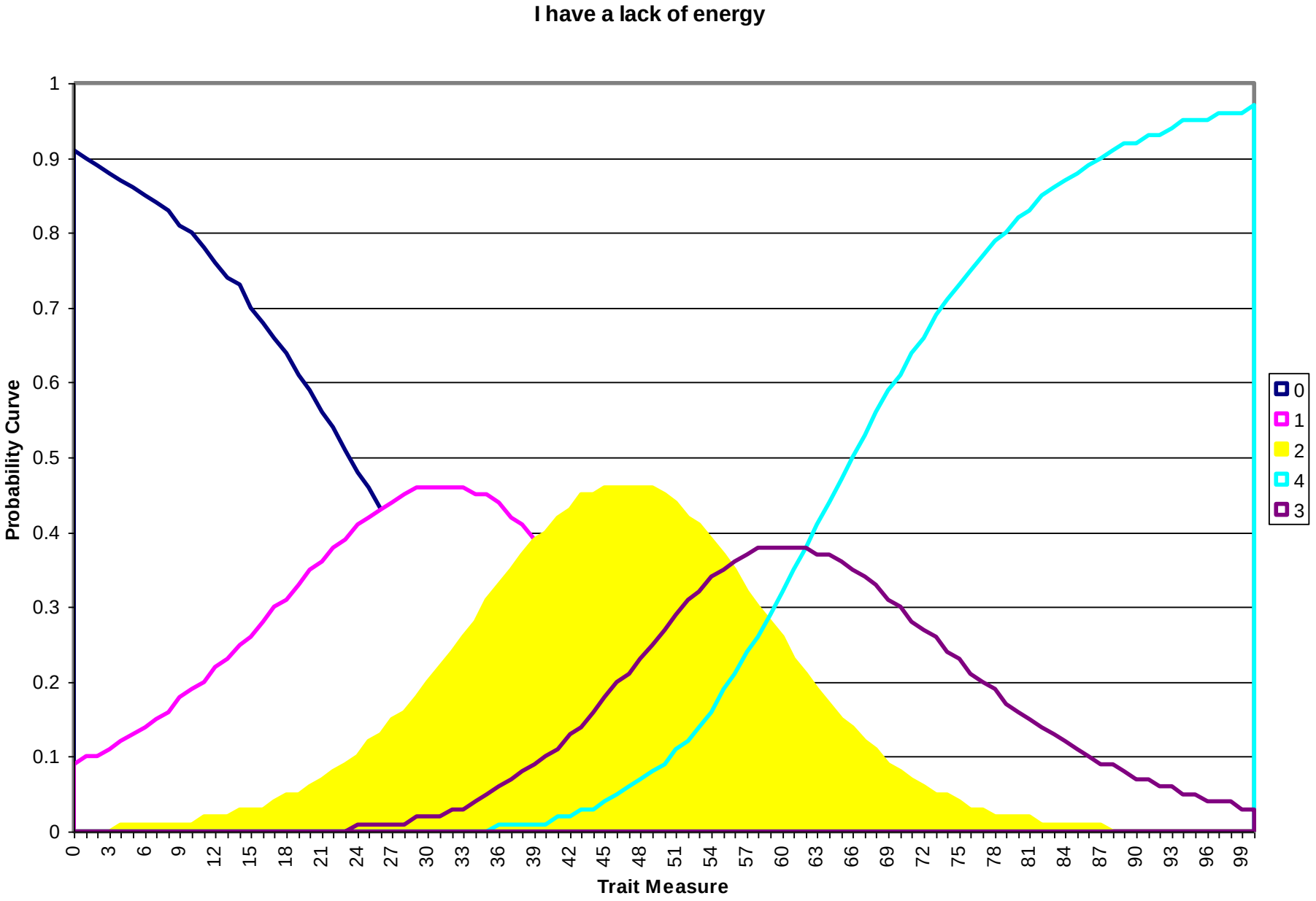

I have a lack of energy
This is an Item Characteristic Curve (ICC) for a rating scale item (each option has its own curve)
Probability Curve
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0 3 6 9 12 15 18 21 24 27 30 33 36 39 42 45 48 51 54 57 60 63 66 69 72 75 78 81 84 87 90 93 96 99
Trait Measure
0
1
2
4
3
0 = Very Much 1 = Quite a Bit 2 = Somewhat 3 = A Little Bit 4 = Not at All



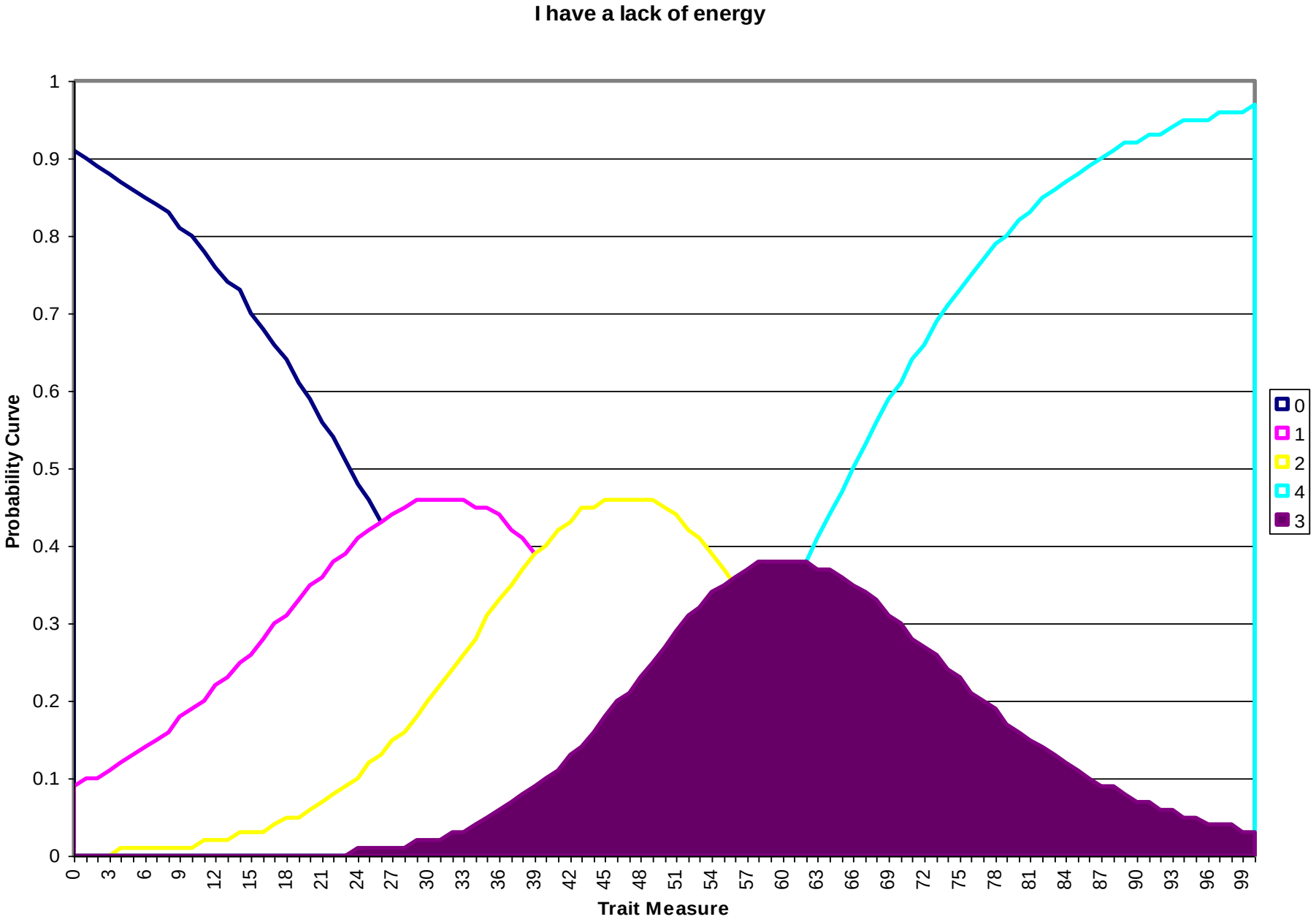
0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All



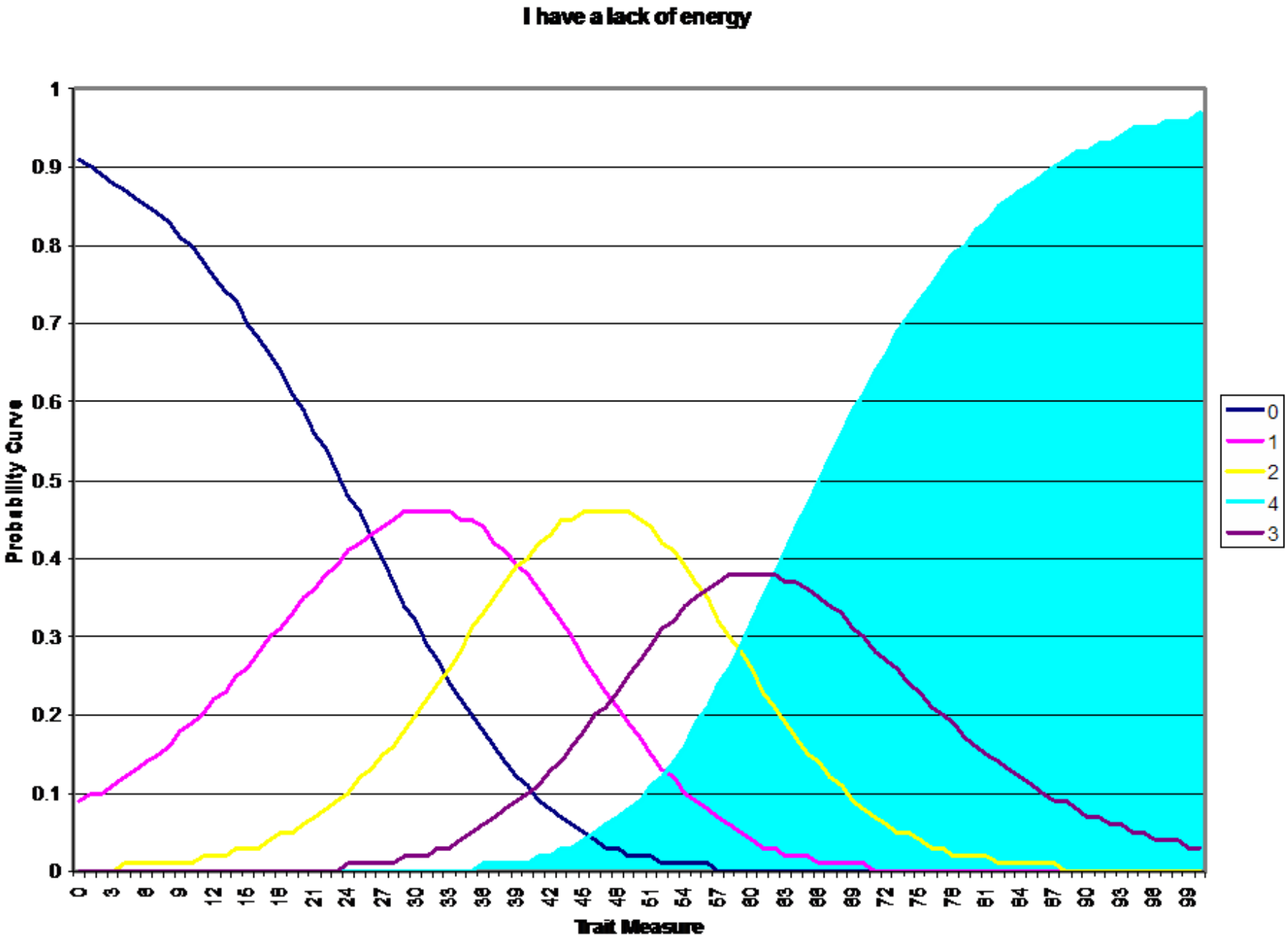
0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All



0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All



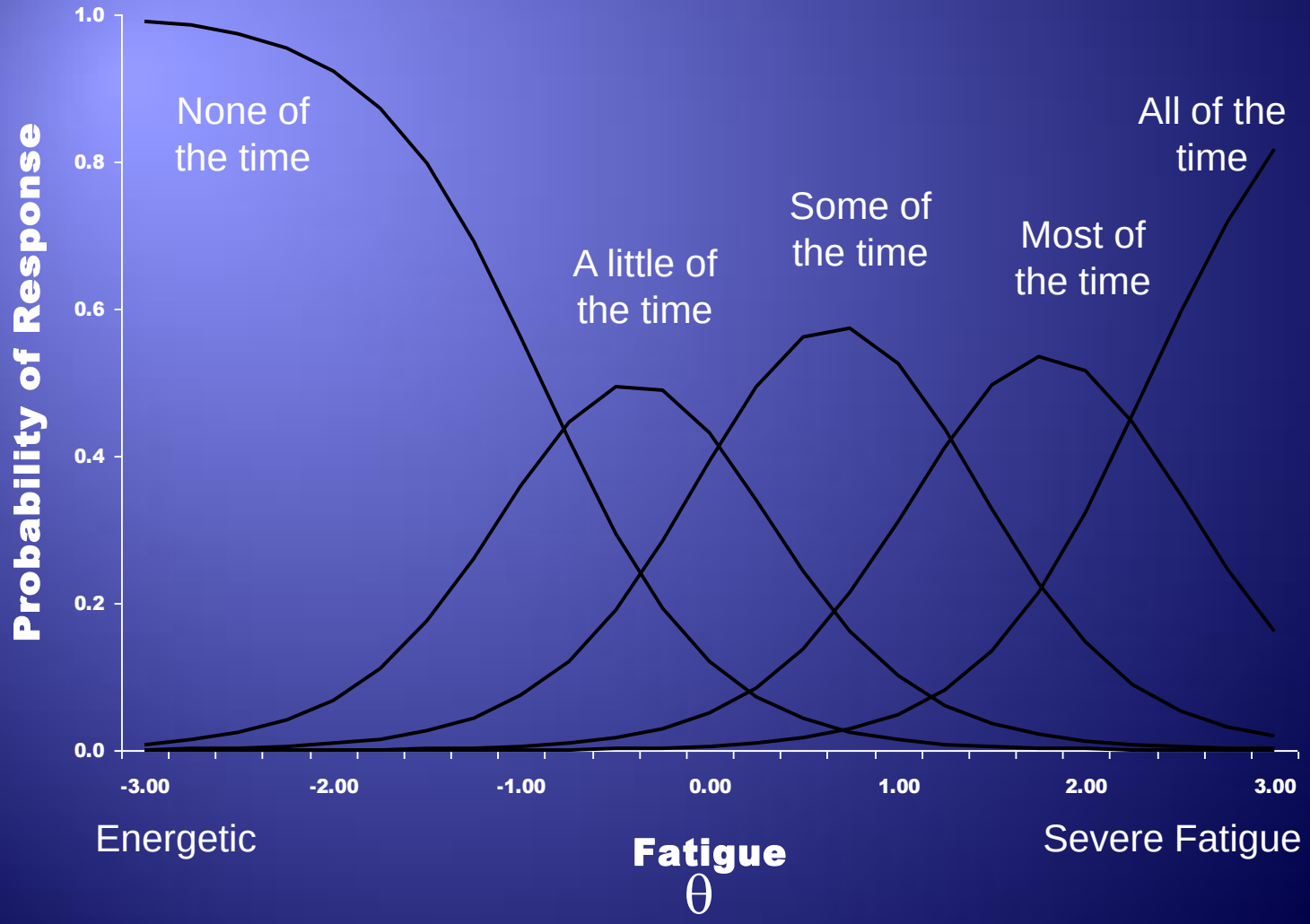
0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All



0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All

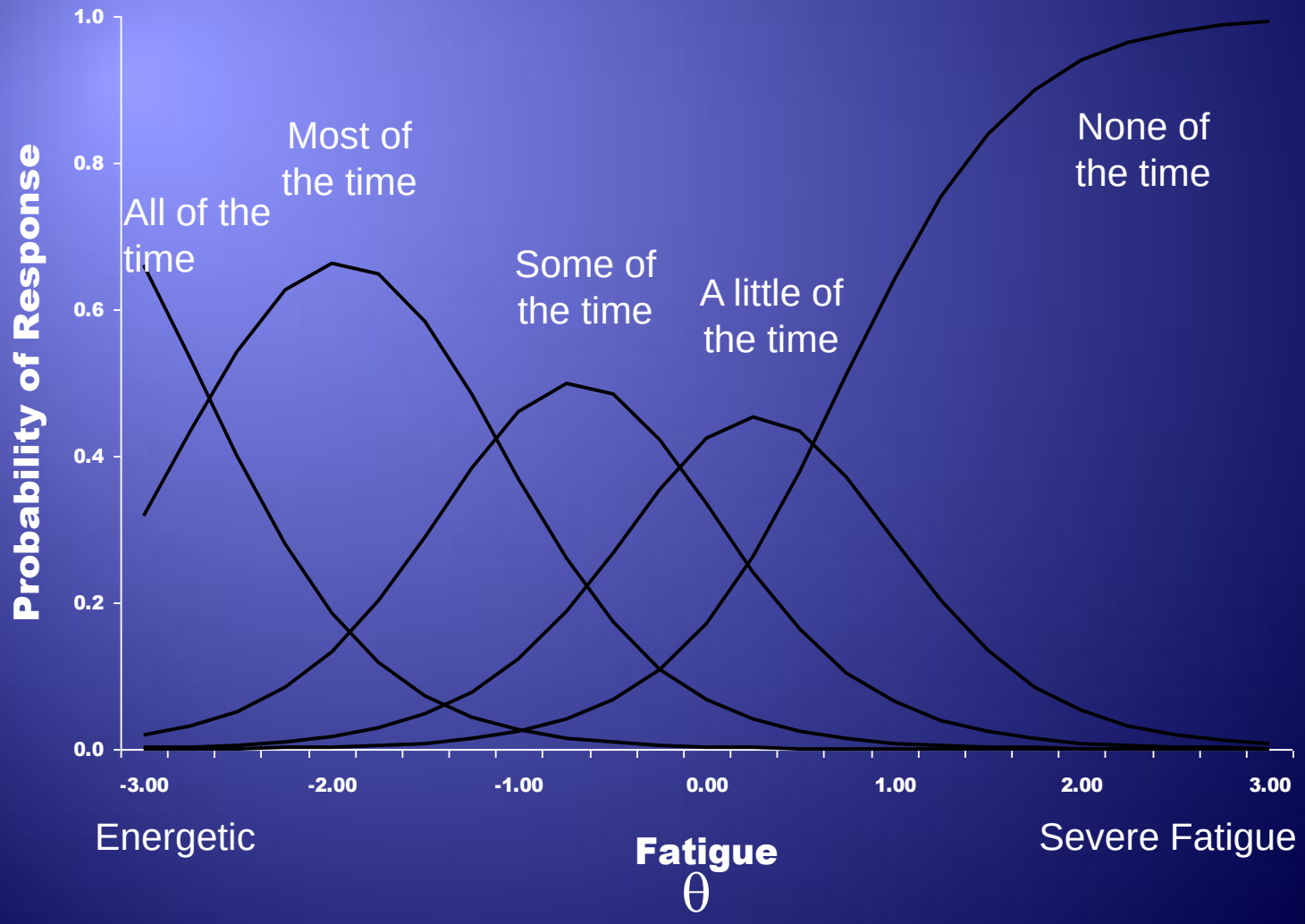
IRT Polytomous Responses

I have been too tired to feel happy.



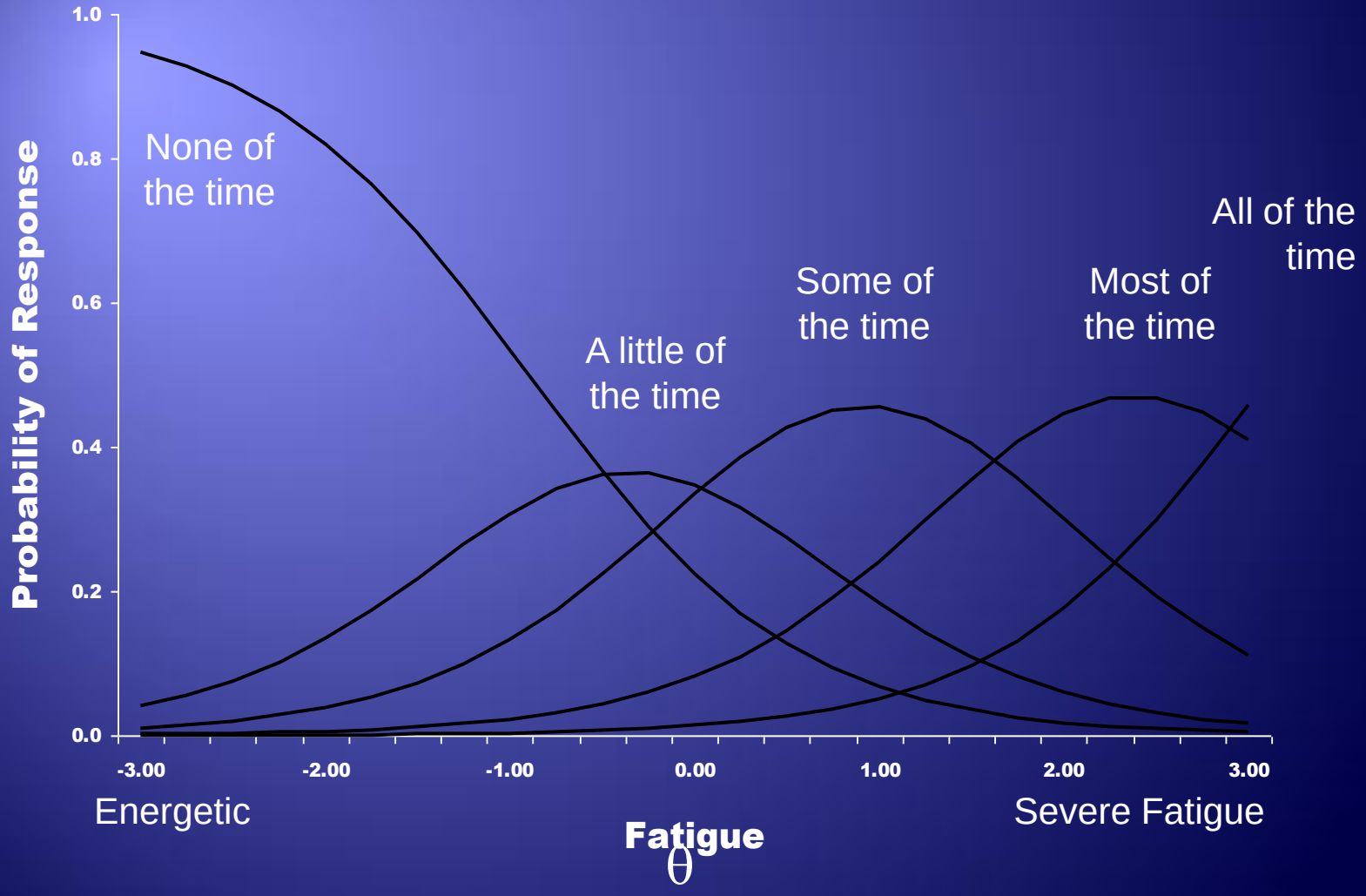
IRT Polytomous Responses

I have felt energetic.



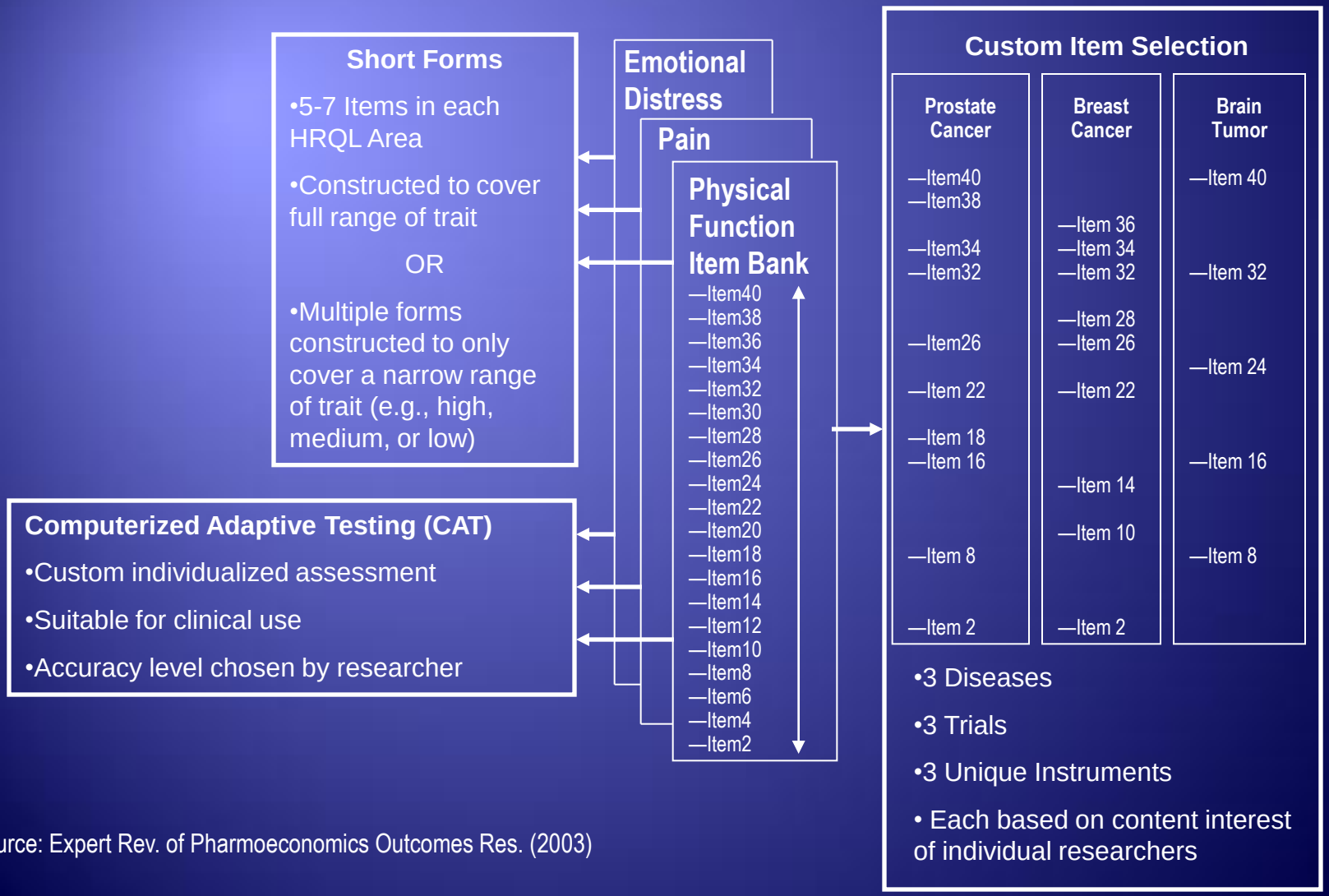
IRT Polytomous Responses

I have been too tired to read.



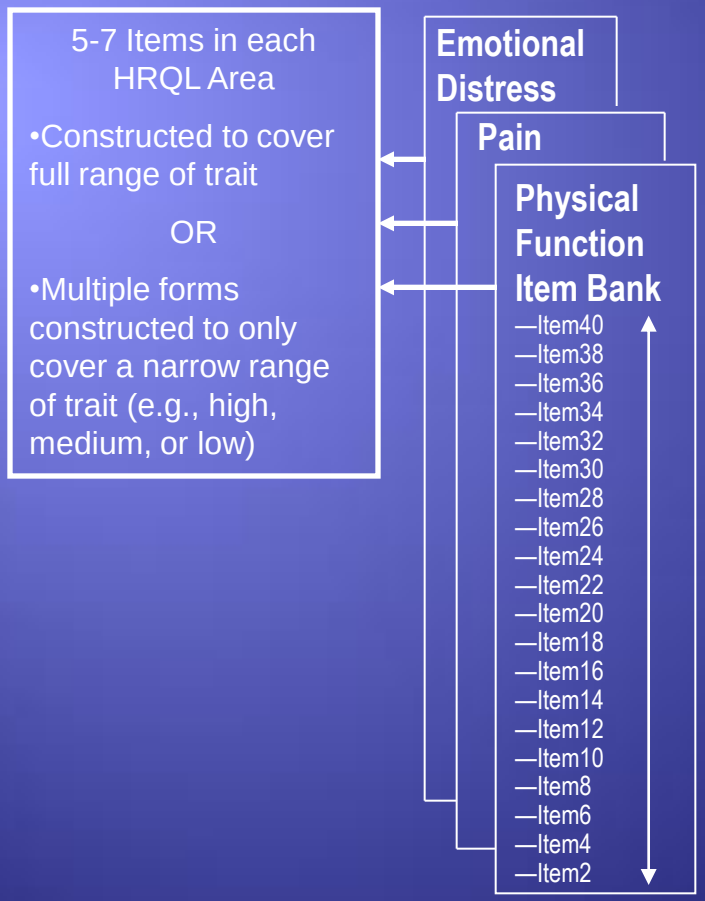
ITEM BANKING

Calibrated Item Banks can be used to Create Numerous Instrument Types



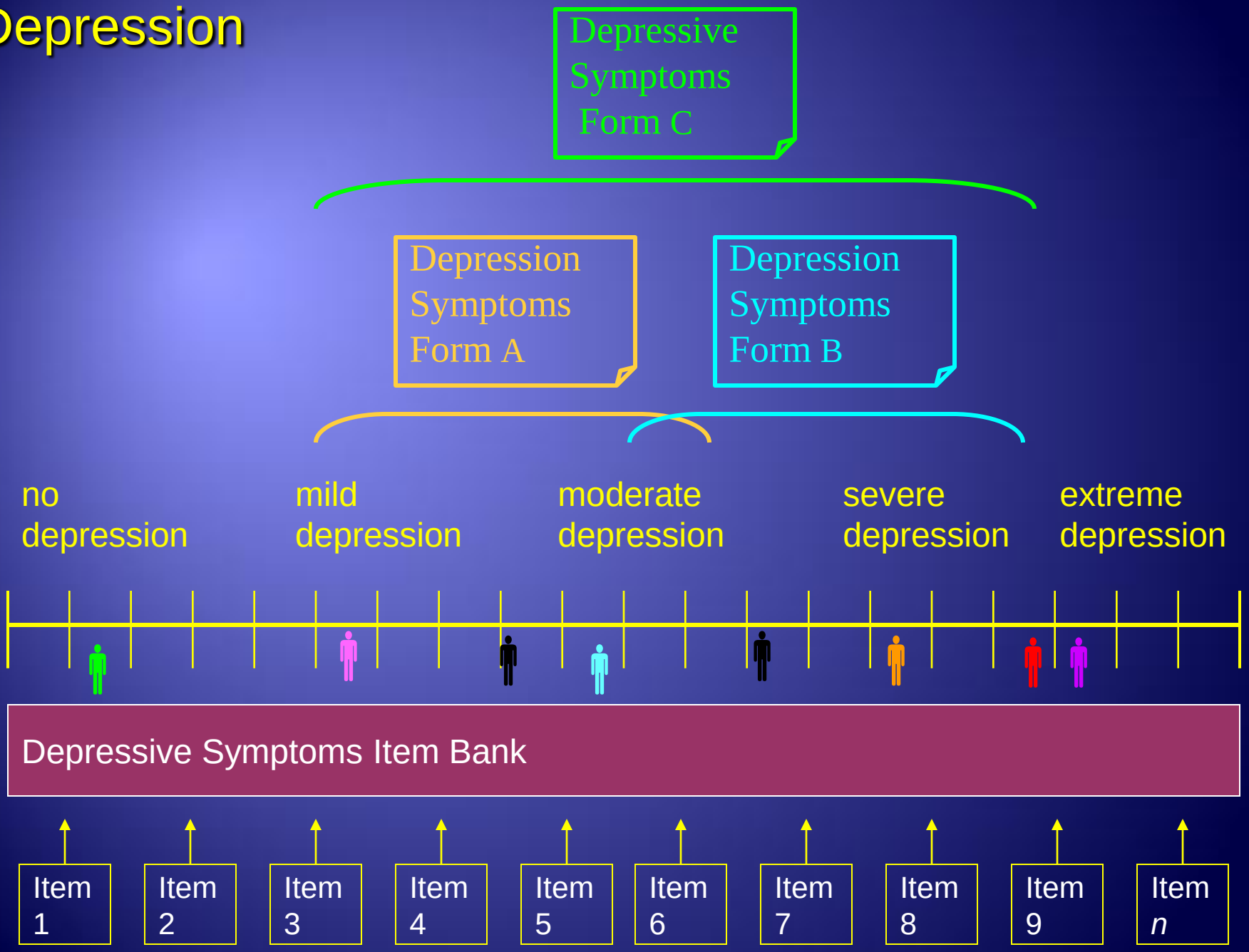
Source: Expert Rev. of Pharmoeconomics Outcomes Res. (2003)

Short Forms

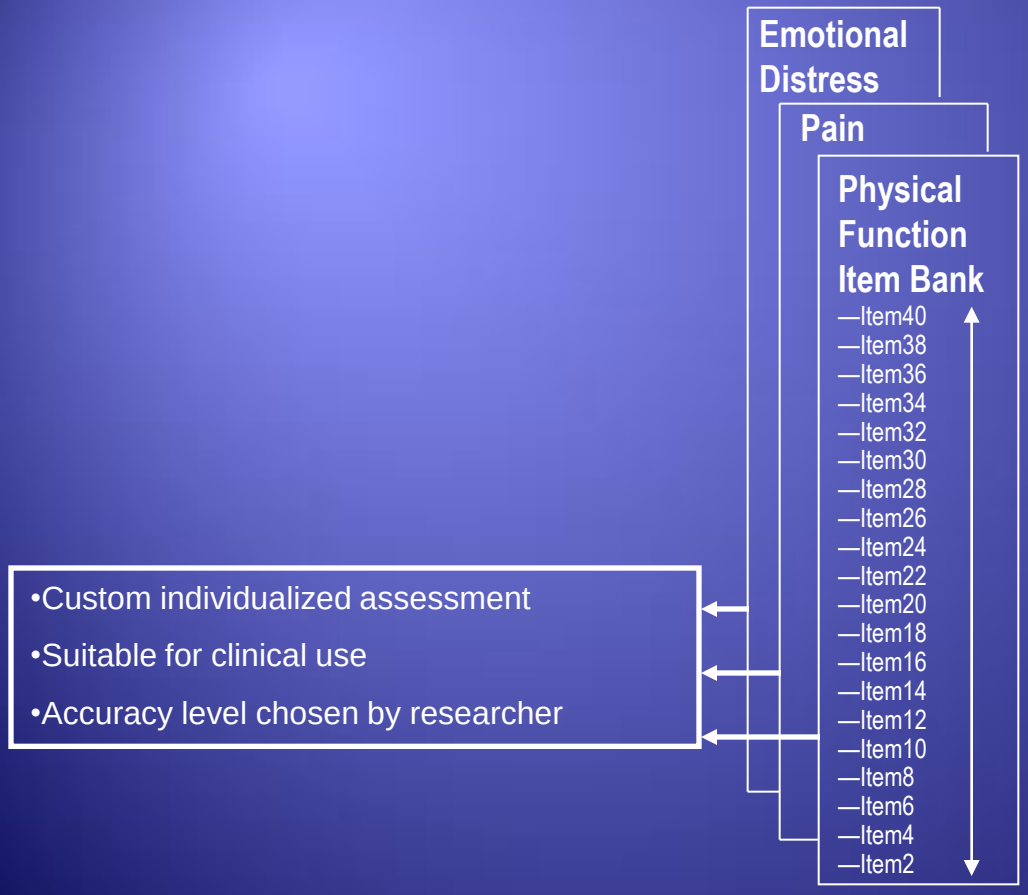


Source: Expert Rev. of Pharmoeconomics Outcomes Res. (2003)

Depression

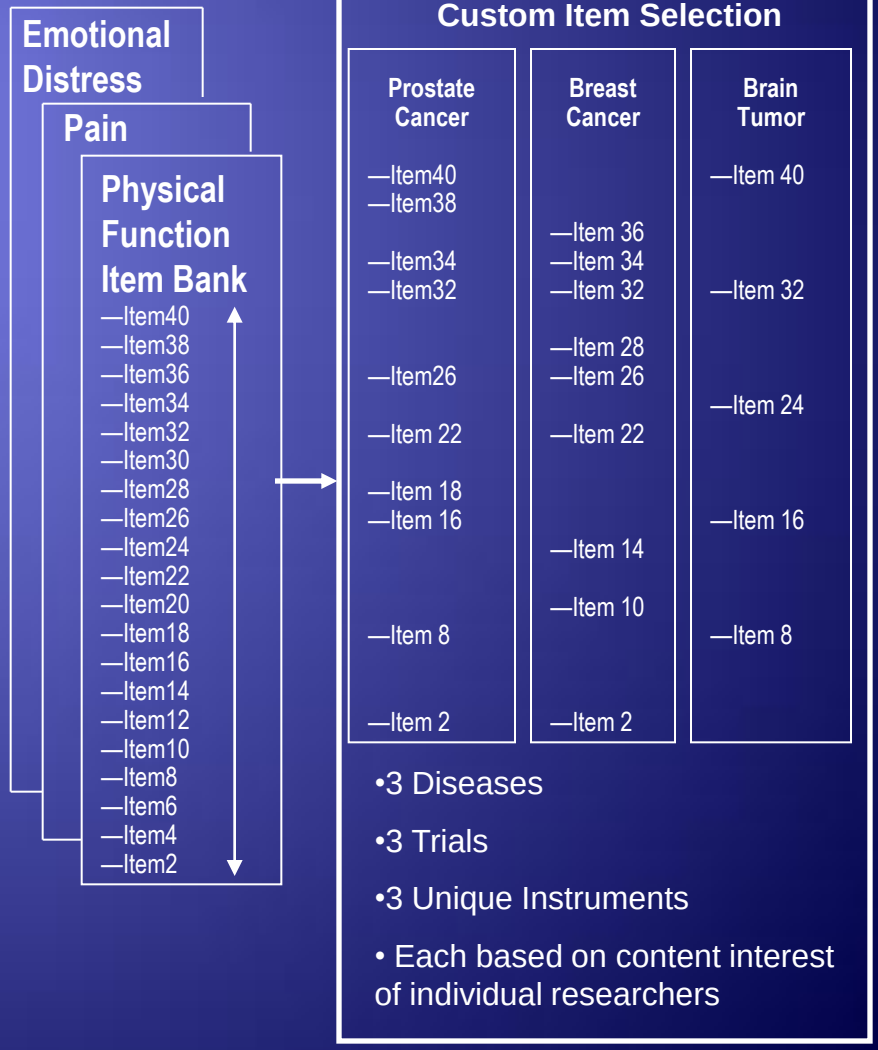

Depressive Symptoms Form C
Depression Symptoms Form A
Depression Symptoms Form B
no depression
mild depression
moderate depression
severe depression
extreme depression
Depressive Symptoms Item Bank
Item 1
Item 2
Item 3
Item 4
Item 5
Item 6
Item 7
Item 8
Item 9
Item *n*

Computerized Adaptive Testing (CAT)



Source: Expert Rev. of Pharmoeconomics Outcomes Res. (2003)

Custom Item Selection



Source: Expert Rev. of Pharmoeconomics Outcomes Res. (2003)

In Summary, Calibrated Item Banks can be used to:

- Create a standard static instrument
- Construct short forms
- Enable CAT
- Select items based on unique content interests and formulate custom short-form or full-length instruments

In every case, using a validated, pre-calibrated item bank allows any of these instruments to be pre-validated and produce standardized scores on the same scale

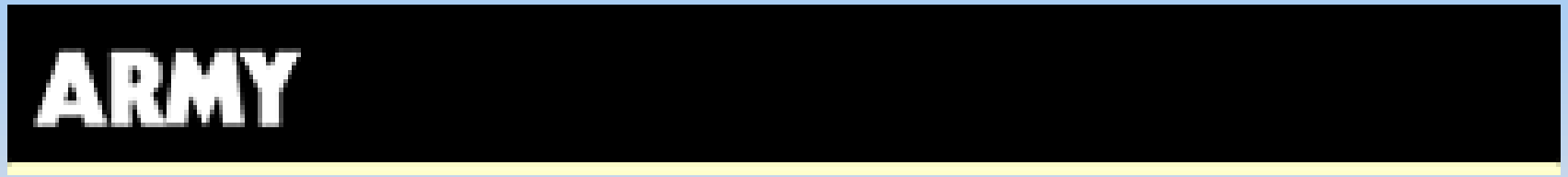
Computerized Adaptive Testing

What is Computerized Adaptive Testing?

- Shorter
- Targeting
- Computerized Algorithm

CAT in the Military

Armed Services Vocational Aptitude Battery (ASVAB)



CAT for Certification

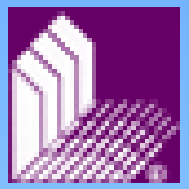
AANA

American Association of Nurse Anesthetists

CompTIA.

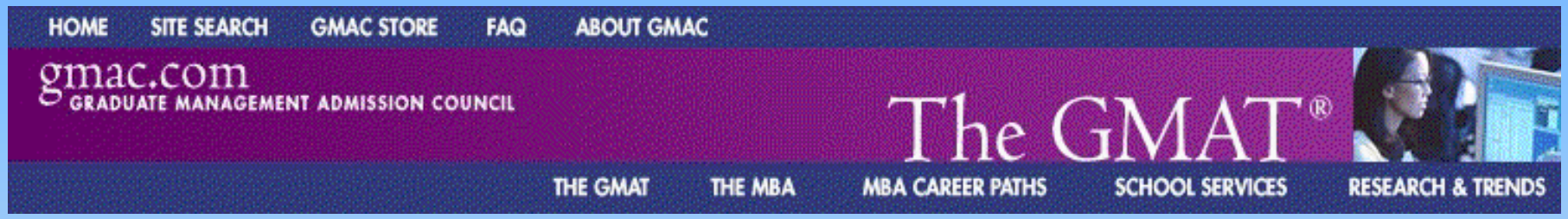
Microsoft®

CAT for Licensure



AMERICAN DIETETIC ASSOCIATION

CAT for College Entrance



The College Board

ACCUPLACER OnLine

toefl.org
Test of English as a Foreign Language

Graduate Record Examinations®

CAT for Education

Northwest Evaluation Association

CAT for Clinical Testing

CAT for Personnel Testing

PSTC

Personnel Systems & Technologies Corporation

Business Language Testing Service | BULATS

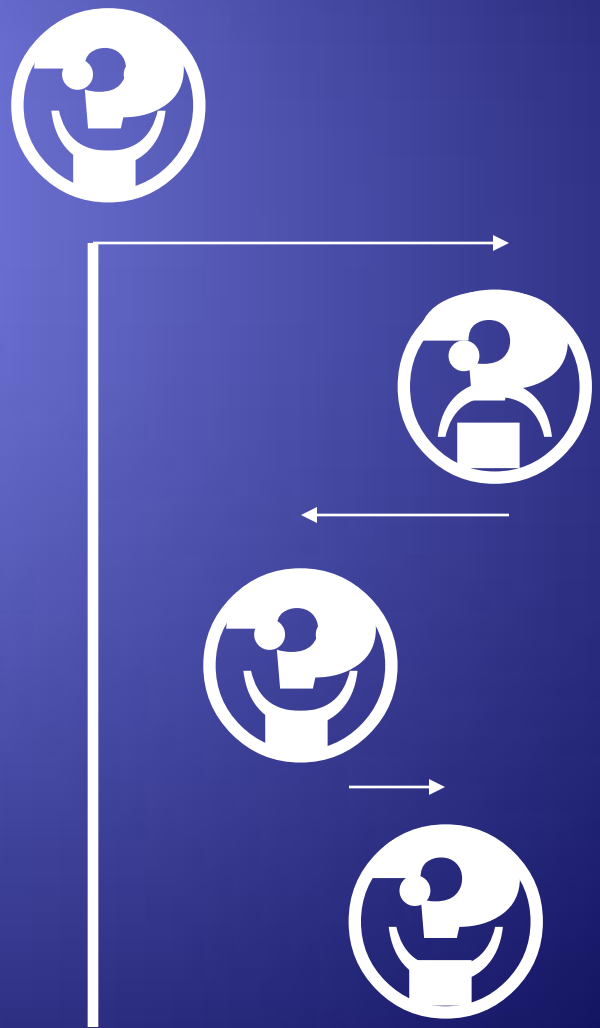
University of Cambridge
Local Examinations Syndicate

ALTE

Low Able

Pass Point

High Able

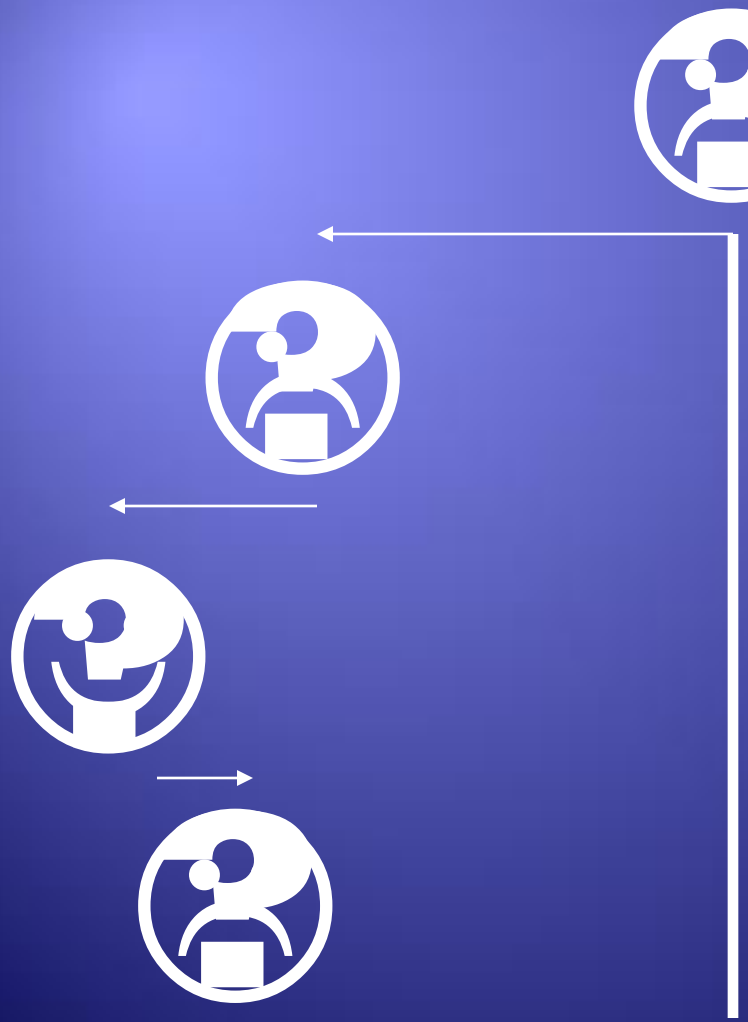


PASS!

Low Able

Pass Point

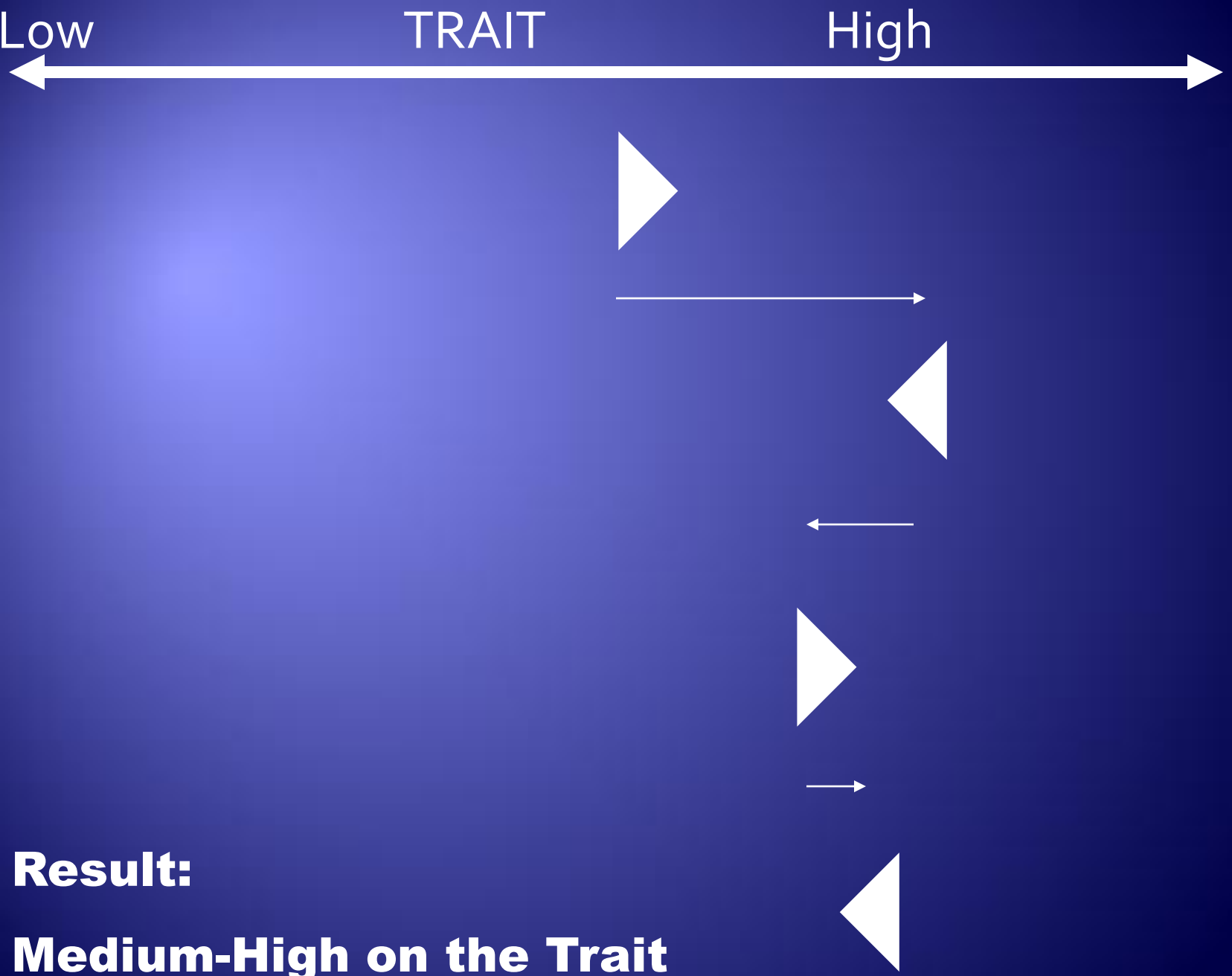
High Able



FAIL

Example – Binary Search

- Binary search



Result:

Medium-High on the Trait

When is CAT appropriate?

- Heterogeneous populations
- Diagnostic tests
- On-demand testing
- Long tests

When is CAT not appropriate?

- Small populations
  - Difficulty in calibrating items
  - Higher administration costs

CAT Requirements

- Calibrated item bank
- Administration software

How do I create a calibrated item bank?

- You probably already have done the hard work!
- It's usually the same as for CBT.

Create an Item Bank
Part 1

- Item sources
  - Previous exams
    - With a CAT bank fewer constraints exist regarding item re-use
  - Write new items

Create an Item Bank

Part 2

- Item quality
  - Statistics relevant to CAT (not necessarily print)
    - Difficulty
    - Other variables relevant to selected IRT model

Calibrating items
Part 1

- Analyze item level data from previously administered items
  - Preferably using raw person-data
  - Alternatively, could create raw estimates from p-values
  - Several software packages exist for this purpose

Calibrating items
Part 2

- Pilot (beta) new test items
  - On paper
  - On computer
- Typically requires a psychometrician to supervise analyses

Test Specifications

- Starting rule
  - With item which provides maximum information
  - At cut point

Test Specifications

- Stopping Rule
  - Fixed length
  - Variable length
  - by Total Test/Subtest
  - Calculated
    - Specified precision of measure
    - Specified confidence in a pass/fail decision
  - Maximum item count
  - Minimum item count

Test Specifications

- Content balancing
  - None
  - Fixed percentage

Test Specifications

- Testing new items (beta testing, field testing, experimental items)

Adaptive Algorithm

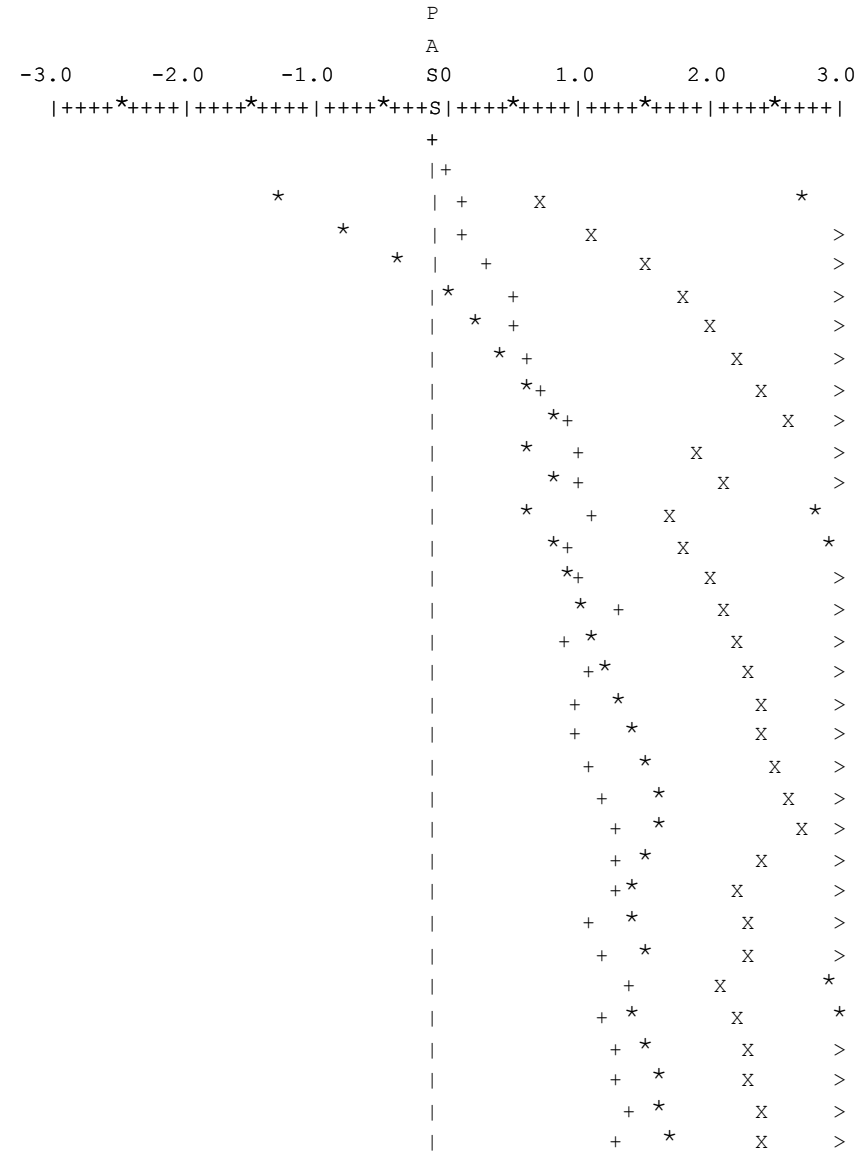
- Person ability algorithm
- Item selection algorithm
  - Test difficulty
  - Maximum jump size
  - Content issues
  - Item exposure control
  - Option to not allow same items to be used during retesting
  - Overlapping items (items that cue other items)

339585909 Entry= 1
MLT Ver: 10/01/01

Tested: 01/28/02
Status: 2

Clear Pass

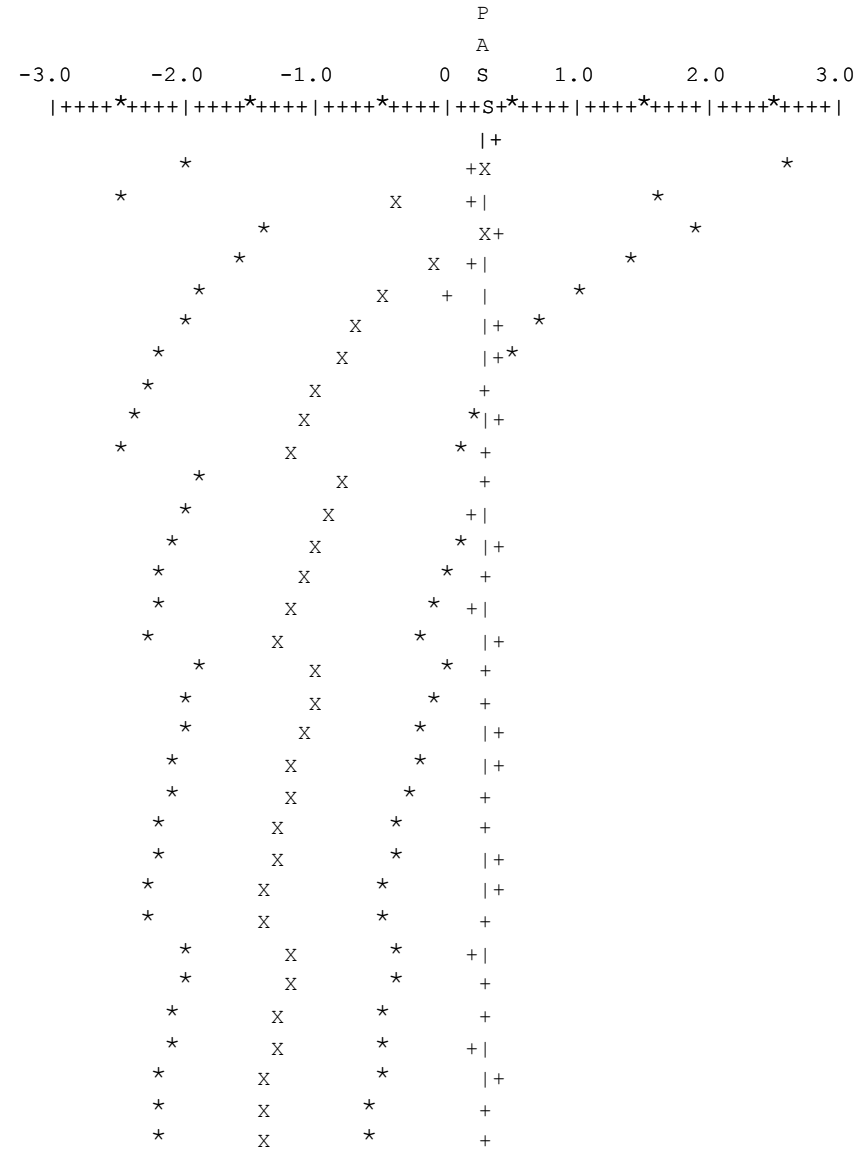
Item	AN	Cont	Diff	Ans	🔔	=	Time	!	Meas	SE
1	21151	BBN	-0.09	1	1	o	2'30		9.99	9.99
2	22805	CHE	0.03	2	1	o	2'56		9.99	9.99
3	22479	HEM	0.13	4	0	o	0'36		0.72	1.22
4	21986	MIC	0.13	3	1	o	0'29		1.15	1.15
5	22397	IMM	0.26	1	1	o	0'10		1.48	1.12
6	21793	UA	0.46	4	1	o	0' 9		1.76	1.10
7	22504	BBN	0.50	3	1	o	0'56		1.99	1.08
8	22083	CHE	0.57	4	1	o	0'22		2.19	1.07
9	22641	HEM	0.74	4	1	o	0'59		2.38	1.06
10	20194	MIC	0.90	2	1	+	3'17		2.56	1.05
11	22032	BBN	1.00	4	0	o	1'26		1.92	0.78
12	20344	CHE	1.00	4	1	o	1' 0		2.08	0.77
13	22261	HEM	1.12	4	0	o	1' 9		1.72	0.66
14	21851	MIC	0.94	4	1	o	1'15		1.85	0.65
15	21511	IMM	1.02	1	1	o	2'14		1.97	0.65
16	21450	UA	1.27	1	1	+	1'17		2.09	0.64
17	20537	BBN	0.93	3	1	o	0'35		2.18	0.64
18	22330	CHE	1.12	2	1	+	2'32		2.28	0.63
19	21218	HEM	1.02	1	1	o	0'37		2.36	0.63
20	21628	MIC	0.96	3	1	o	1' 3		2.44	0.63
21	22748	BBN	1.07	1	1	o	2'10		2.51	0.62
22	22553	CHE	1.22	3	1	o	0'31		2.59	0.62
23	22639	HEM	1.28	1	1	o	0'57		2.66	0.62
24	22646	MIC	1.35	2	0	=	2'44		2.40	0.55
25	22663	IMM	1.27	1	0	o	1'17		2.19	0.50
26	22557	UA	1.06	2	1	o	0'41		2.25	0.50
27	20686	BBN	1.15	1	1	o	0'27		2.31	0.50
28	22634	CHE	1.37	3	0	o	1'19		2.15	0.46
29	21646	HEM	1.16	2	1	o	0'15		2.20	0.46
30	22387	MIC	1.31	4	1	o	0'23		2.26	0.46
31	20018	BBN	1.27	3	1	o	0'34		2.31	0.45
32	22059	CHE	1.40	1	1	o	0'48		2.37	0.45
33	22471	HEM	1.34	1	1	o	0'41		2.42	0.45



434843789 Entry= 1
HT Ver: 01/01/02

Tested: 01/28/02
Status: 1

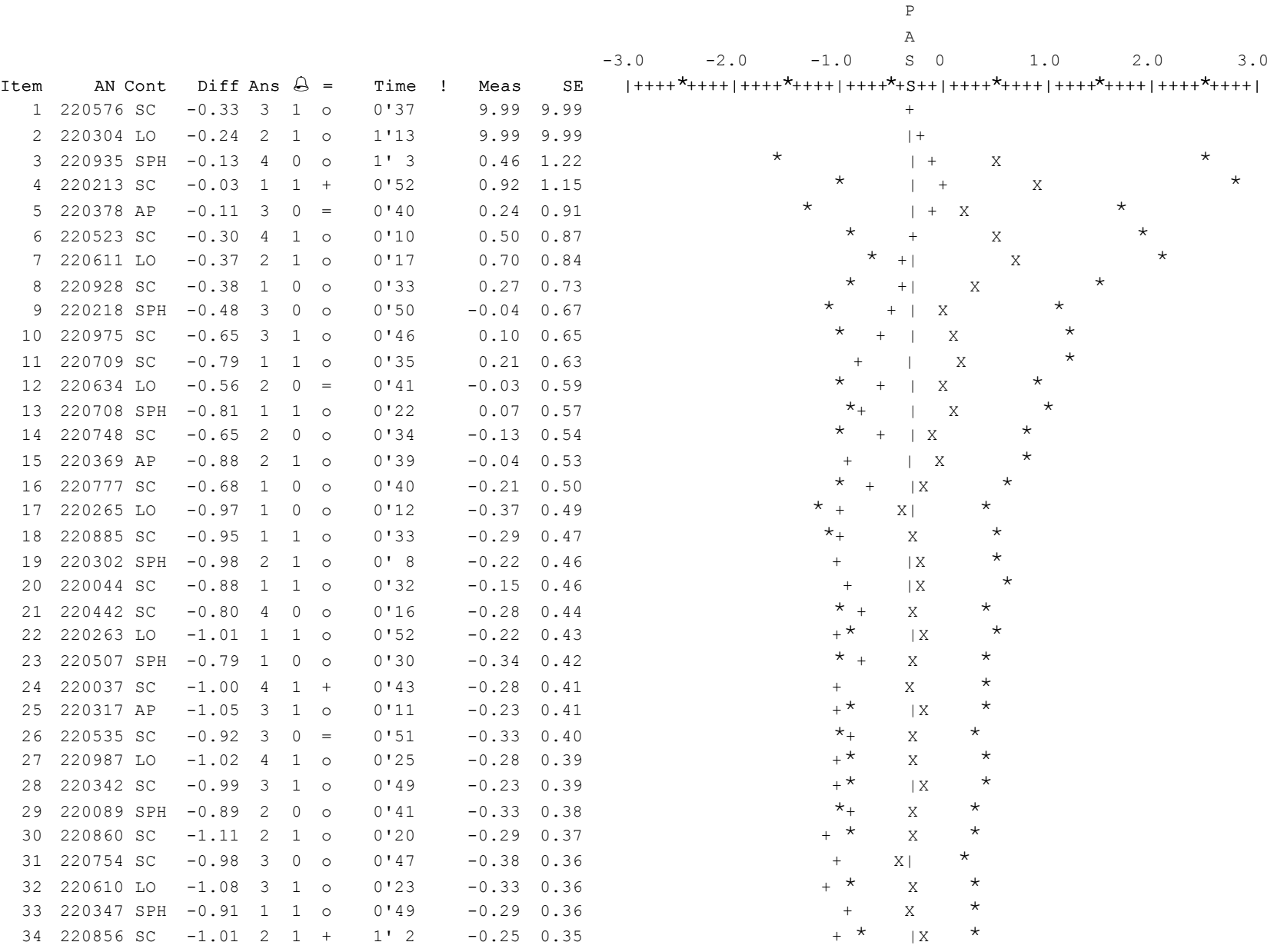
Clear Fail



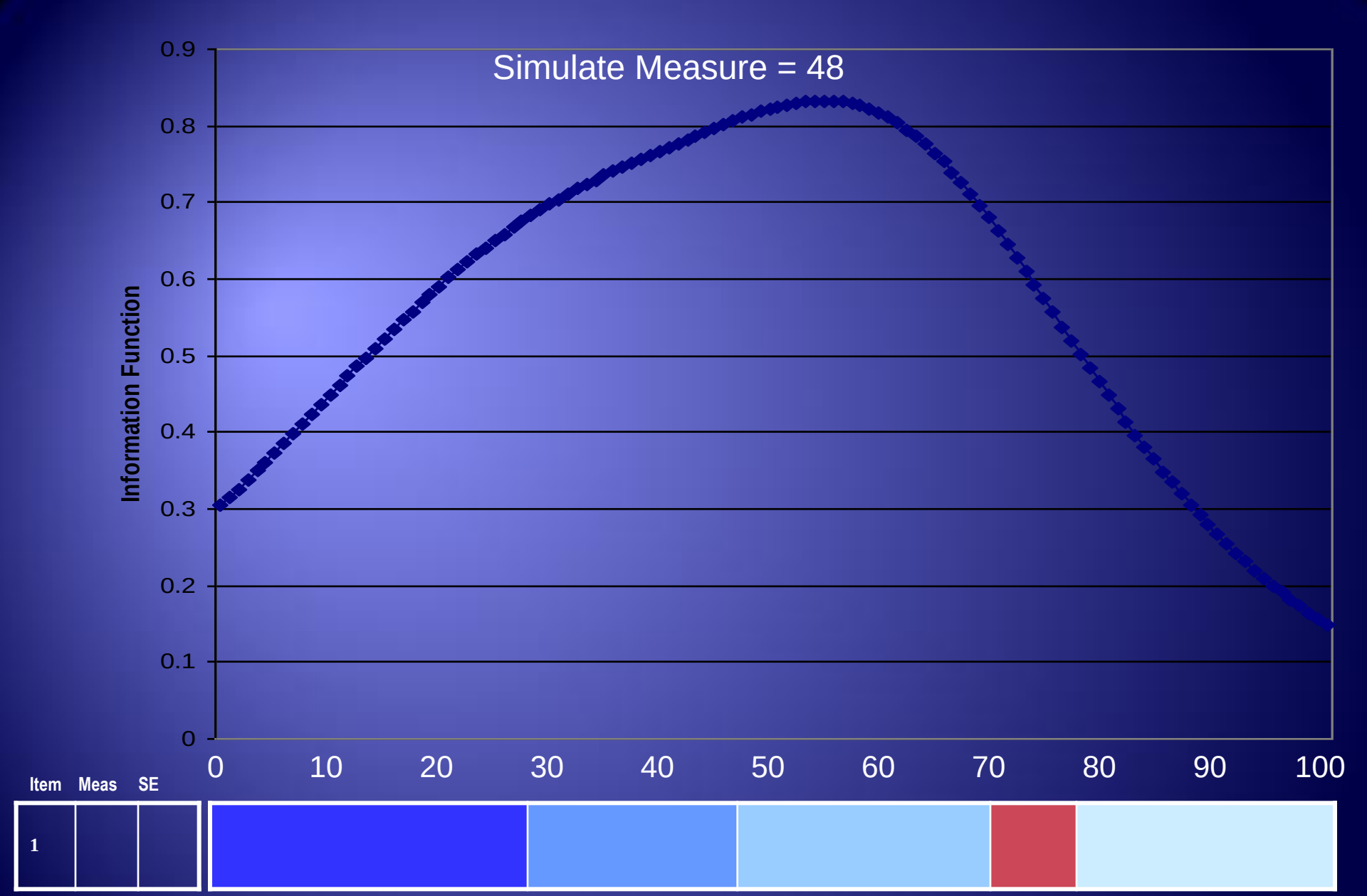
Item	AN	Cont	Diff	Ans	🔔	=	Time	!	Meas	SE
1	31384	ST	0.35	1	0		0'34		-9.99	9.99
2	31009	FIX	0.22	3	1		0'18		0.29	1.41
3	31113	LO	0.18	1	0		0'26		-0.44	1.22
4	30385	MIC	0.36	3	1		0'33		0.28	1.00
5	30873	ST	0.24	3	0		0'31		-0.14	0.91
6	30533	PRO	0.05	2	0		0'30		-0.46	0.87
7	30525	ST	0.35	2	0		0'16		-0.67	0.84
8	31008	FIX	0.37	4	0		0'31		-0.83	0.82
9	30664	ST	0.30	2	0		0'32		-0.98	0.80
10	31086	LO	0.35	4	0		0'12		-1.11	0.79
11	31626	ST	0.34	2	0		0'23		-1.22	0.78
12	31356	MIC	0.32	4	1		0'41		-0.81	0.67
13	31210	PRO	0.21	2	0		0'35		-0.92	0.66
14	31148	ST	0.39	1	0		0'20		-1.01	0.65
15	31620	FIX	0.25	4	0		0'10		-1.10	0.65
16	30224	ST	0.20	4	0		0'25		-1.19	0.64
17	30940	FIX	0.40	2	0		0'32		-1.25	0.64
18	31288	ST	0.25	3	1		1'14		-0.97	0.57
19	31529	LO	0.28	1	0		0'58		-1.04	0.56
20	31120	ST	0.40	2	0		0'11		-1.10	0.56
21	31355	MIC	0.36	2	0		0'59		-1.15	0.56
22	31207	PRO	0.33	2	0		0'34		-1.21	0.55
23	30745	ST	0.33	4	0		0'33		-1.26	0.55
24	31285	FIX	0.40	3	0		0'13		-1.31	0.55
25	30237	ST	0.39	3	0		0'22		-1.35	0.55
26	30179	ST	0.26	1	0		0'24		-1.40	0.54
27	31055	FIX	0.23	4	1		0'24		-1.18	0.50
28	31598	LO	0.29	2	0		0'33		-1.23	0.49
29	30384	MIC	0.27	3	0		0'11		-1.27	0.49
30	30524	ST	0.20	1	0		0'21		-1.31	0.49
31	31470	PRO	0.38	1	0		0'16		-1.35	0.49
32	30188	ST	0.31	3	0		0'21		-1.39	0.49
33	31402	FIX	0.28	3	0		1' 9		-1.42	0.49

411433522 Entry= 1
PBT Ver: 10/01/01

Tested: 01/26/02
Status: 1

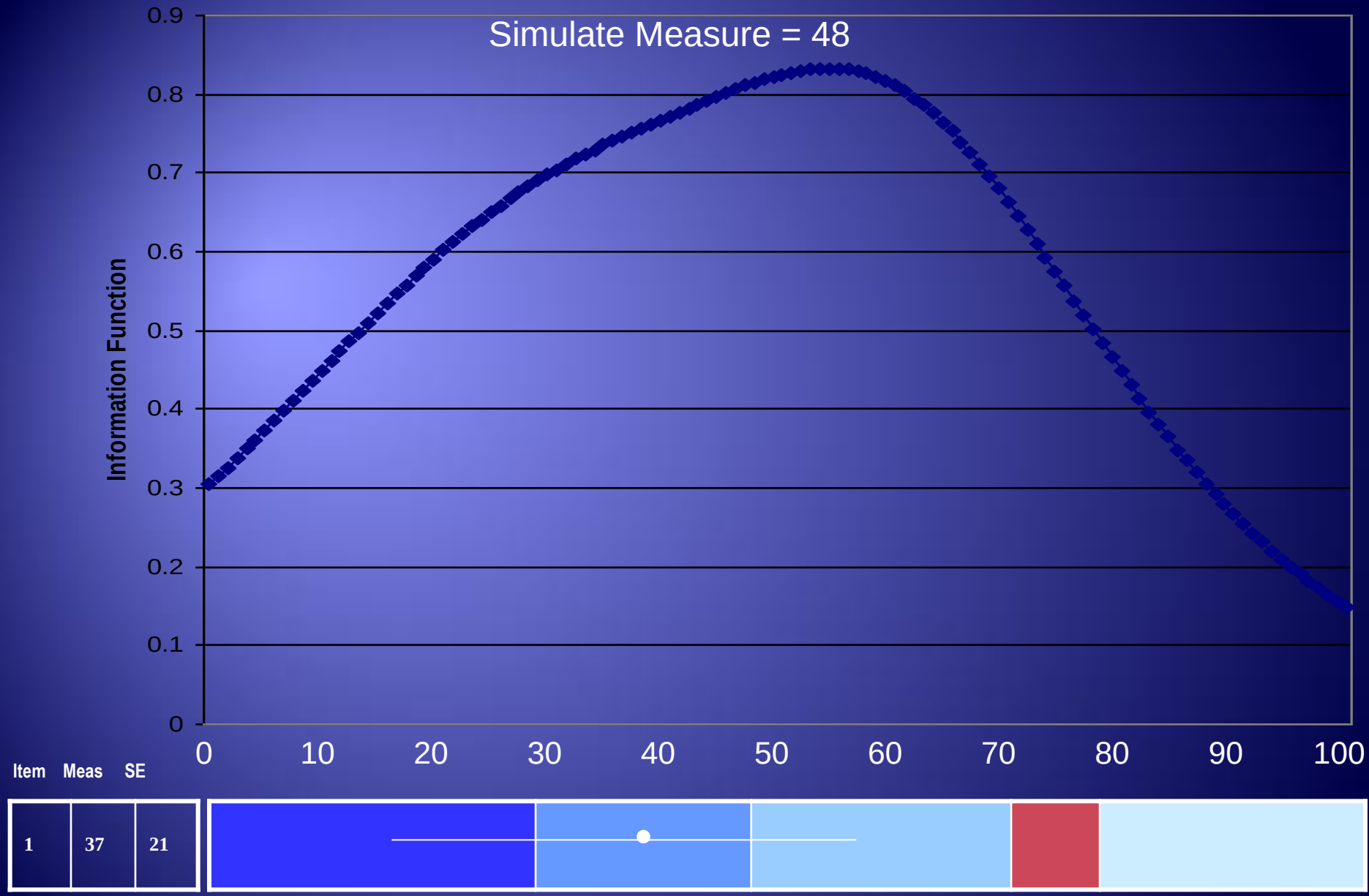


Item	AN	Cont	Diff	Ans	🔔	=	Time	!	Meas	SE
1	220576	SC	-0.33	3	1	o	0'37		9.99	9.99
2	220304	LO	-0.24	2	1	o	1'13		9.99	9.99
3	220935	SPH	-0.13	4	0	o	1' 3		0.46	1.22
4	220213	SC	-0.03	1	1	+	0'52		0.92	1.15
5	220378	AP	-0.11	3	0	=	0'40		0.24	0.91
6	220523	SC	-0.30	4	1	o	0'10		0.50	0.87
7	220611	LO	-0.37	2	1	o	0'17		0.70	0.84
8	220928	SC	-0.38	1	0	o	0'33		0.27	0.73
9	220218	SPH	-0.48	3	0	o	0'50		-0.04	0.67
10	220975	SC	-0.65	3	1	o	0'46		0.10	0.65
11	220709	SC	-0.79	1	1	o	0'35		0.21	0.63
12	220634	LO	-0.56	2	0	=	0'41		-0.03	0.59
13	220708	SPH	-0.81	1	1	o	0'22		0.07	0.57
14	220748	SC	-0.65	2	0	o	0'34		-0.13	0.54
15	220369	AP	-0.88	2	1	o	0'39		-0.04	0.53
16	220777	SC	-0.68	1	0	o	0'40		-0.21	0.50
17	220265	LO	-0.97	1	0	o	0'12		-0.37	0.49
18	220885	SC	-0.95	1	1	o	0'33		-0.29	0.47
19	220302	SPH	-0.98	2	1	o	0' 8		-0.22	0.46
20	220044	SC	-0.88	1	1	o	0'32		-0.15	0.46
21	220442	SC	-0.80	4	0	o	0'16		-0.28	0.44
22	220263	LO	-1.01	1	1	o	0'52		-0.22	0.43
23	220507	SPH	-0.79	1	0	o	0'30		-0.34	0.42
24	220037	SC	-1.00	4	1	+	0'43		-0.28	0.41
25	220317	AP	-1.05	3	1	o	0'11		-0.23	0.41
26	220535	SC	-0.92	3	0	=	0'51		-0.33	0.40
27	220987	LO	-1.02	4	1	o	0'25		-0.28	0.39
28	220342	SC	-0.99	3	1	o	0'49		-0.23	0.39
29	220089	SPH	-0.89	2	0	o	0'41		-0.33	0.38
30	220860	SC	-1.11	2	1	o	0'20		-0.29	0.37
31	220754	SC	-0.98	3	0	o	0'47		-0.38	0.36
32	220610	LO	-1.08	3	1	o	0'23		-0.33	0.36
33	220347	SPH	-0.91	1	1	o	0'49		-0.29	0.36
34	220856	SC	-1.01	2	1	+	1' 2		-0.25	0.35



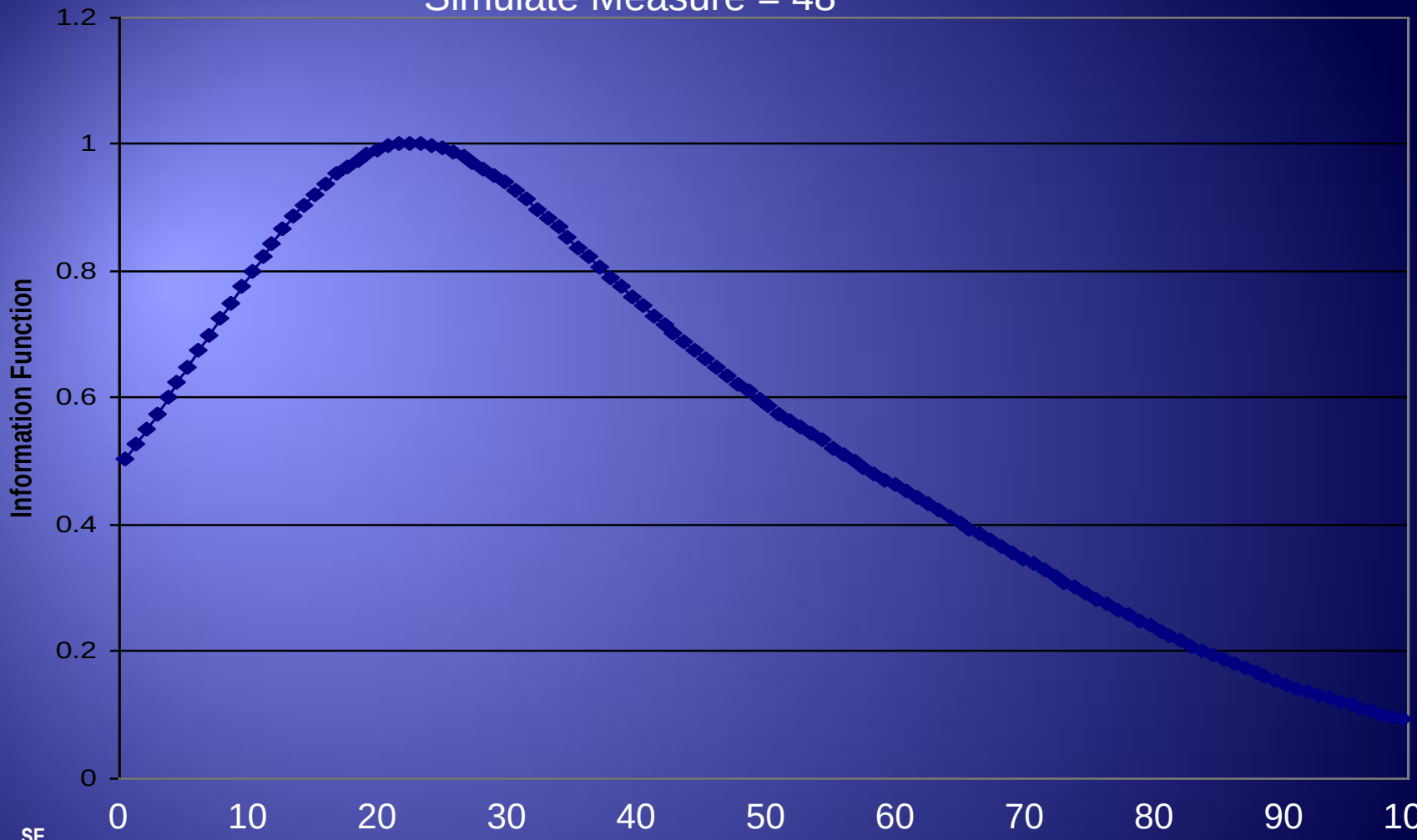
GP1 – I have a lack of energy

0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All

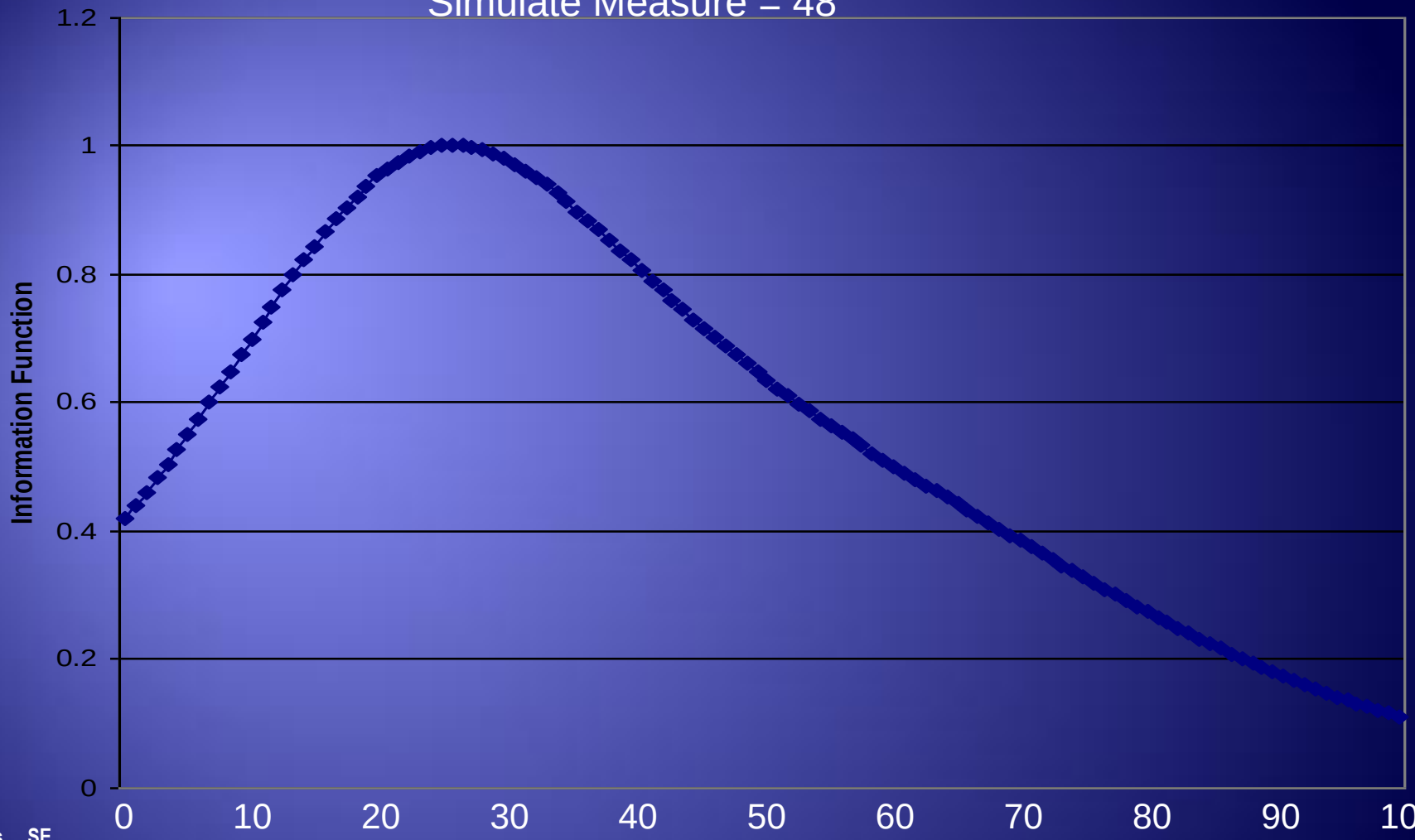


GP1 – I have a lack of energy

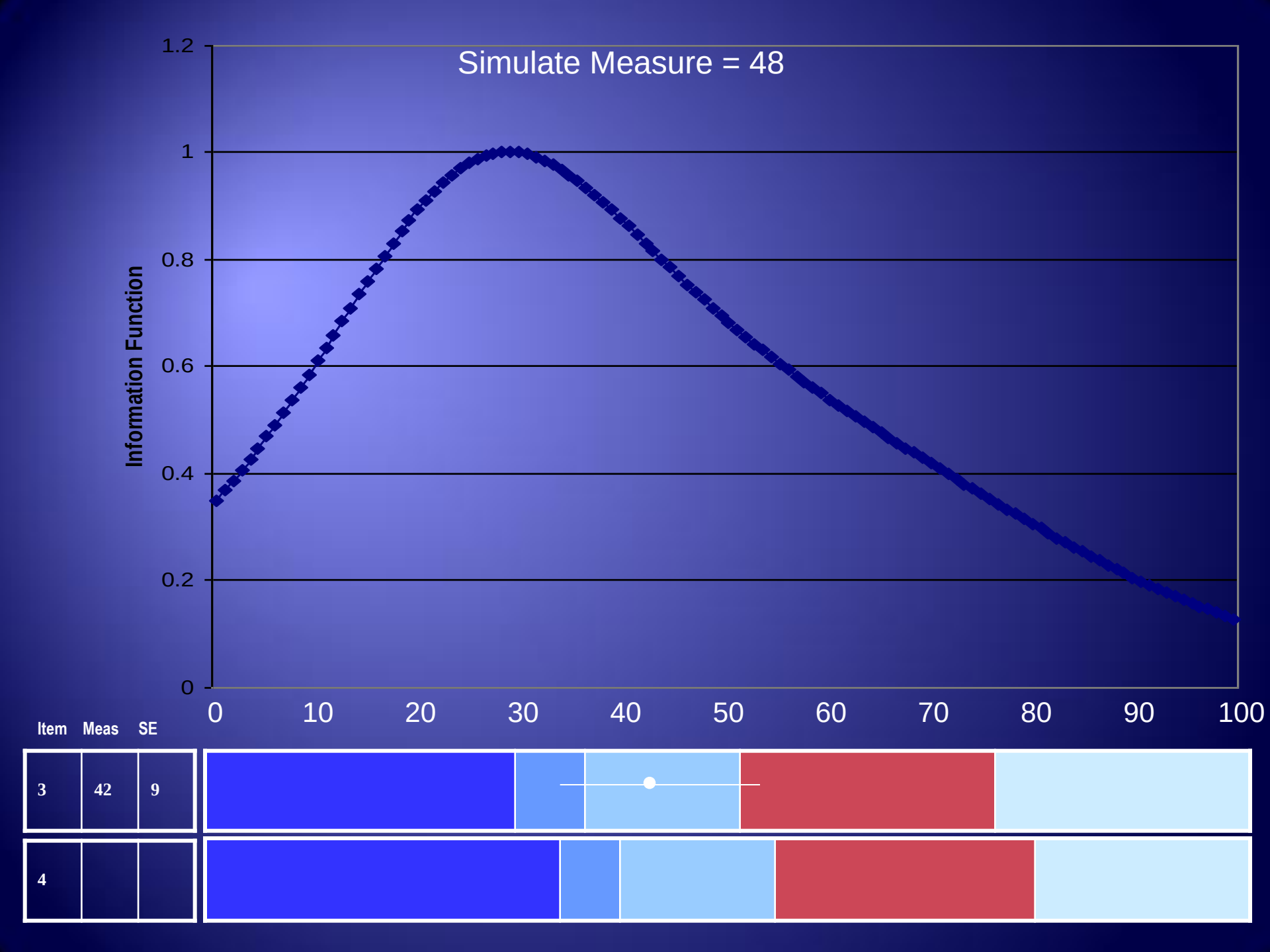
0 = Very Much; 1 = Quite a Bit; 2 = Somewhat; 3 = A Little Bit; 4 = Not at All



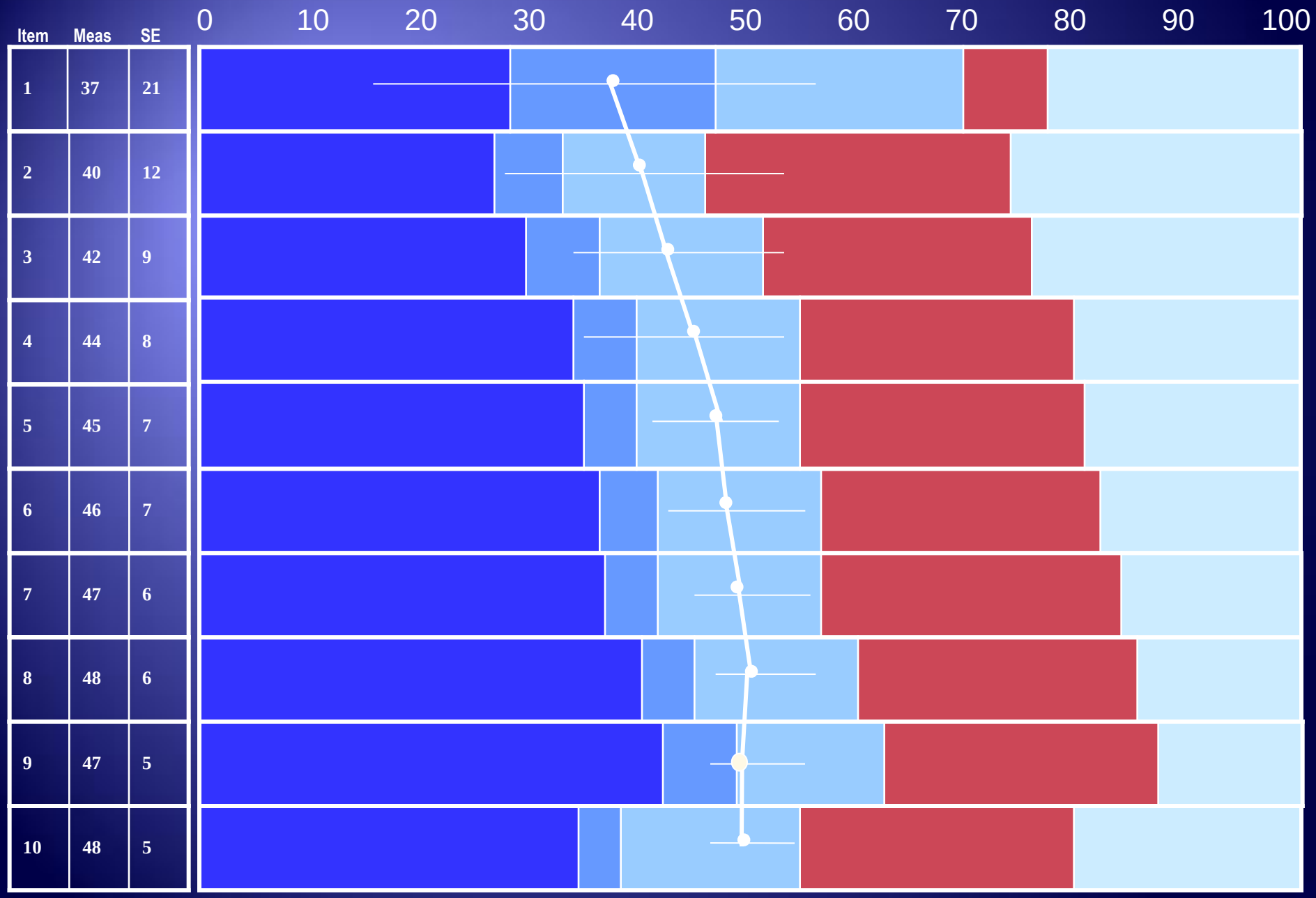
Item	Meas	SE
1	37	21
2		



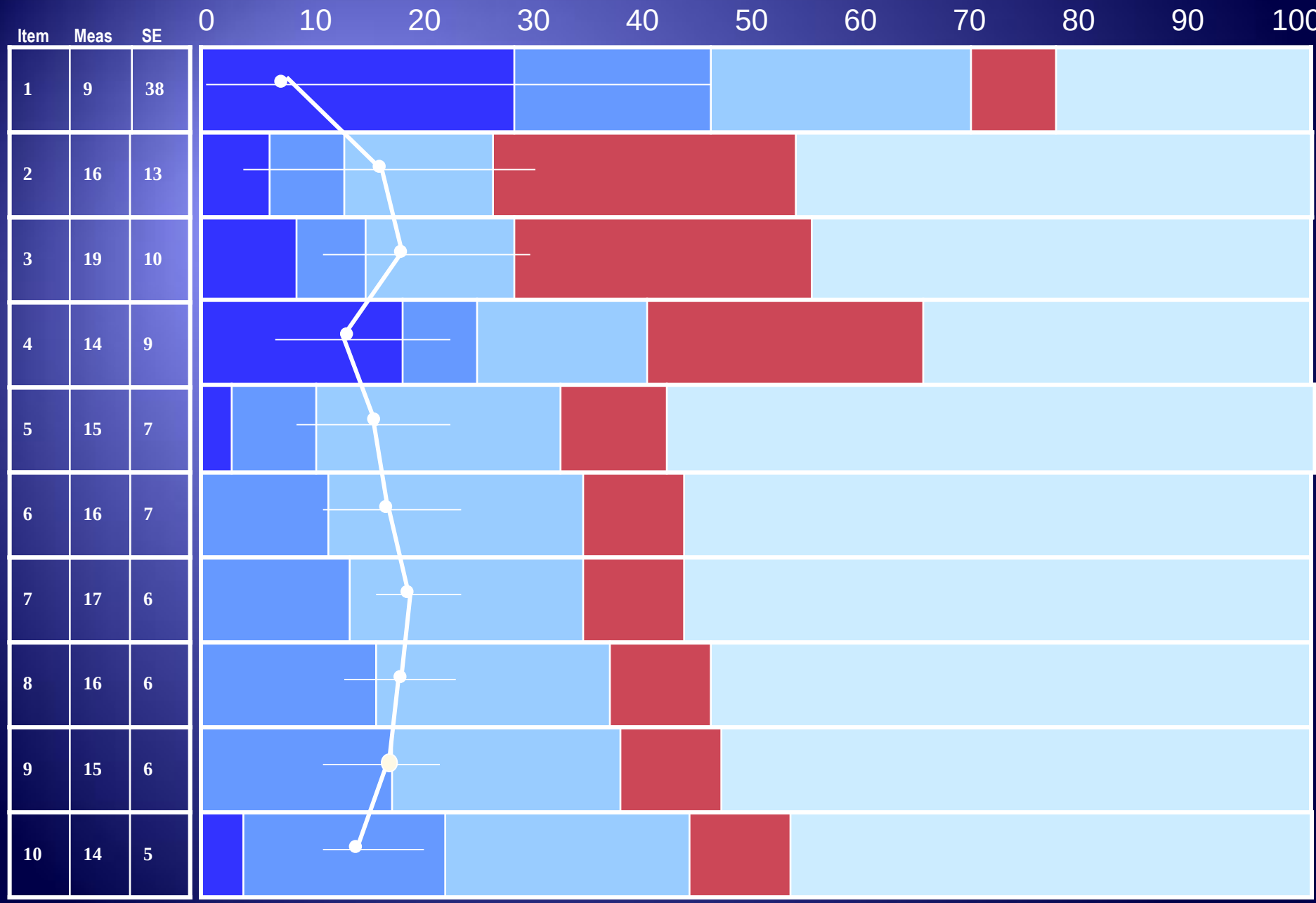
Item	Meas	SE
2	40	12
3		


Simulate Measure = 48
Information Function
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
Item
Meas
SE
3
42
9
4

Simulate Measure = 48

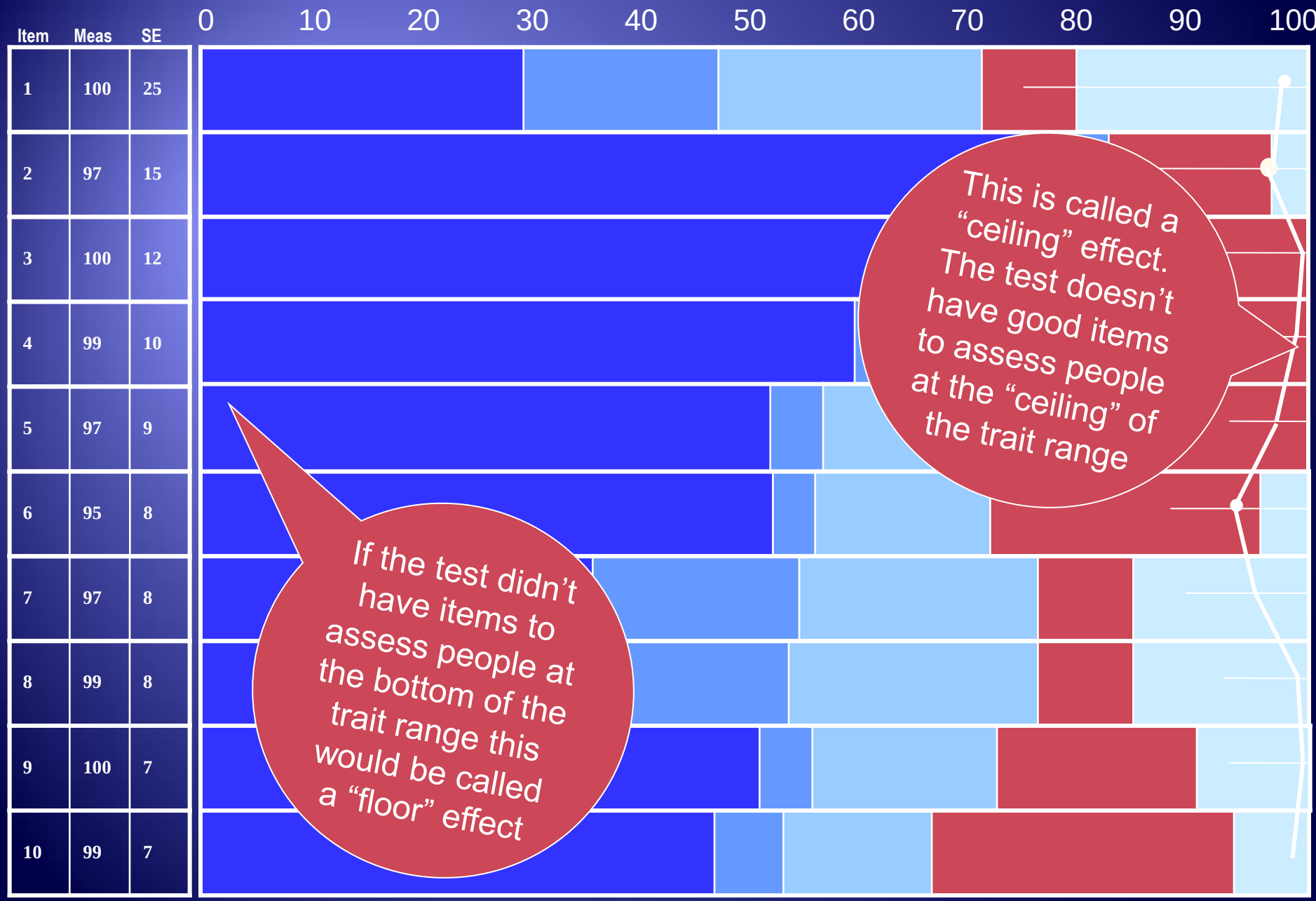


Simulate Measure = 15



Item	Meas	SE
1	9	38
2	16	13
3	19	10
4	14	9
5	15	7
6	16	7
7	17	6
8	16	6
9	15	6
10	14	5

Simulate Measure = 92



Sample PROMIS Fatigue Short Form

	In the past 7 days …	Never	Rarely	Some-times	Often	Always
FATEXP 20	How often did you feel tired?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP 5	How often did you experience extreme exhaustion?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATEXP 18	How often did you run out of energy?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP 33	How often did your fatigue limit you at work (include work at home)?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP 30	How often were you too tired to think clearly?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP 21	How often were you too tired to take a bath or shower?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
FATIMP 40	How often did you have enough energy to exercise strenuously?	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

PROMIS

Patient-Reported Outcomes Measurement Information System

Dynamic Tools to Measure Health Outcomes From the Patient Perspective

Demonstration

- CAT in Assessment Center