Item Response Theory (IRT): Enhancing Health Outcomes Measurement

Bryce B. Reeve, Ph.D.

e-mail: bbreeve@email.unc.edu

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UNC N.C. CANCER HOSPITAL LINEBERGER COMPREHENSIVE CANCER CENTER

UNC Gillings School of Global Public Health

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Presentation Overview

- **IRT Models**
  - Theory of IRT (Reeve)
  - IRT item, scale, and person properties (Reeve)
  - Comparison with Classical Test Theory (Reeve)
  - IRT Assumptions and Model Fit (Orlando Edelen)
  - IRT Scoring (Orlando Edelen)
- **Applying IRT to enhancing health outcomes measurement**
  - Designing and evaluating scales (Siemons; Krishnan)
  - Assessing Differential Item Functioning (DIF) (Orlando Edelen)
  - Linking scales (Glas; Oude Voshaar)
  - Item Banking and Computerized Adaptive Testing (Bjorner; Nikolaus)

Please Note:

- The quality of a health outcomes measure is related to the attention the developer(s) took to use qualitative and quantitative methods integrating multiple perspectives throughout the process.

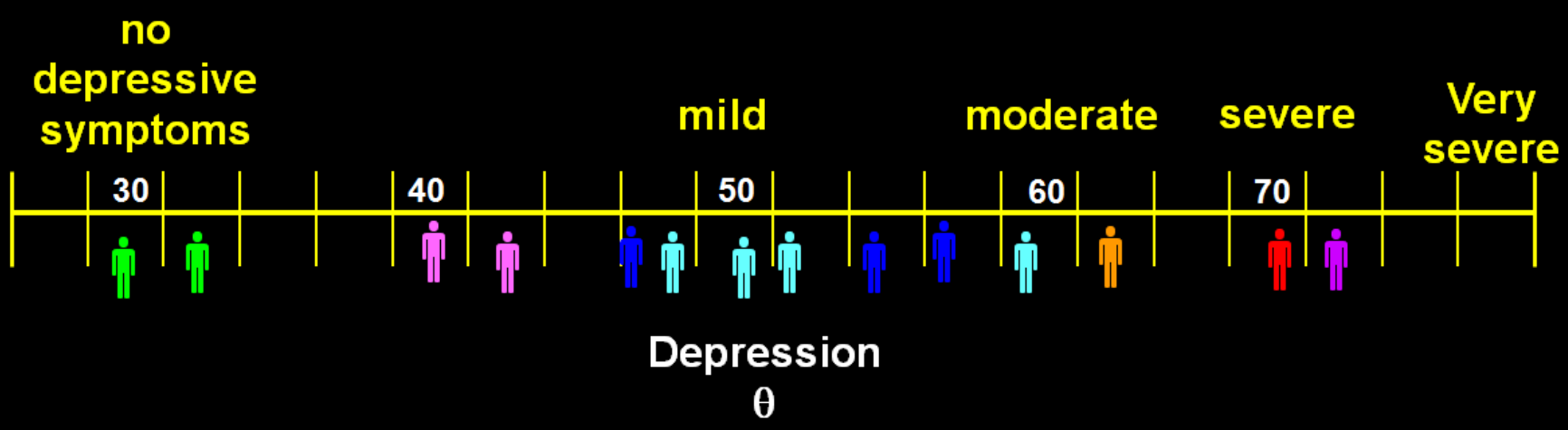
- IRT Methods do not replace the classical/traditional test theory methods for item/scale analysis.

- IRT analysis is not a magic wand!
  - It cannot fix bad data or poorly defined constructs
  - By itself, it does not address all forms of validity and other attributes that evaluate the quality of a questionnaire.

The Need for Better Outcome Measures

Needs	Challenges
Develop measures that are valid, reliable, and sensitive to detect clinically meaningful change	Have a minimum set of questions to reduce respondent burden.
Different forms of an instrument to measure different health levels.	Different forms to be linked on the same metric for group comparisons
Non-biased measurement across groups	Detect differences in group perceptions

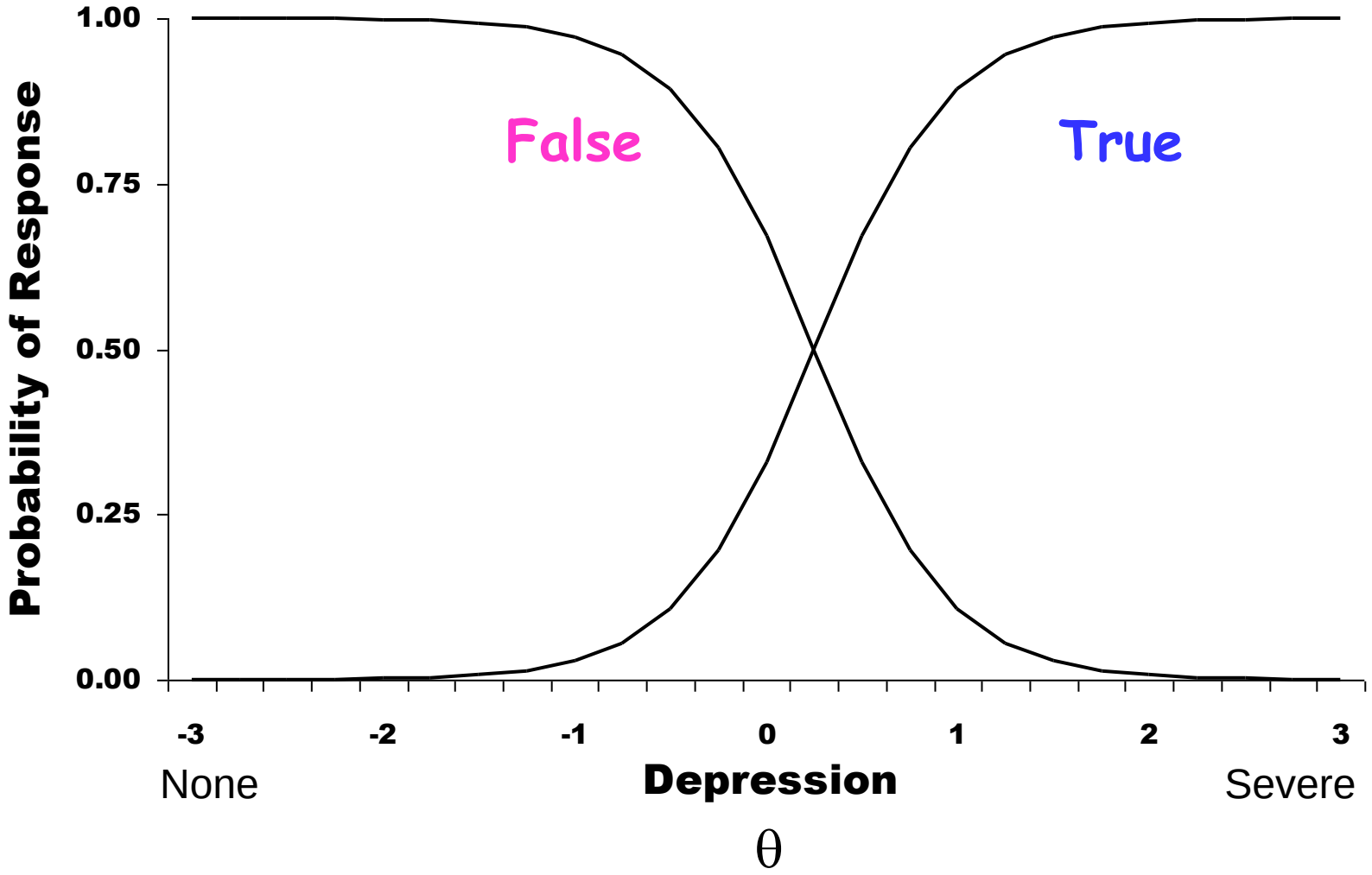
What is Item Response Theory (IRT)?



- **IRT is designed for:**
  - **Modeling latent "unobservable" variables (traits, domains, θ)**
  - **Multi-item Scales/Questionnaires**

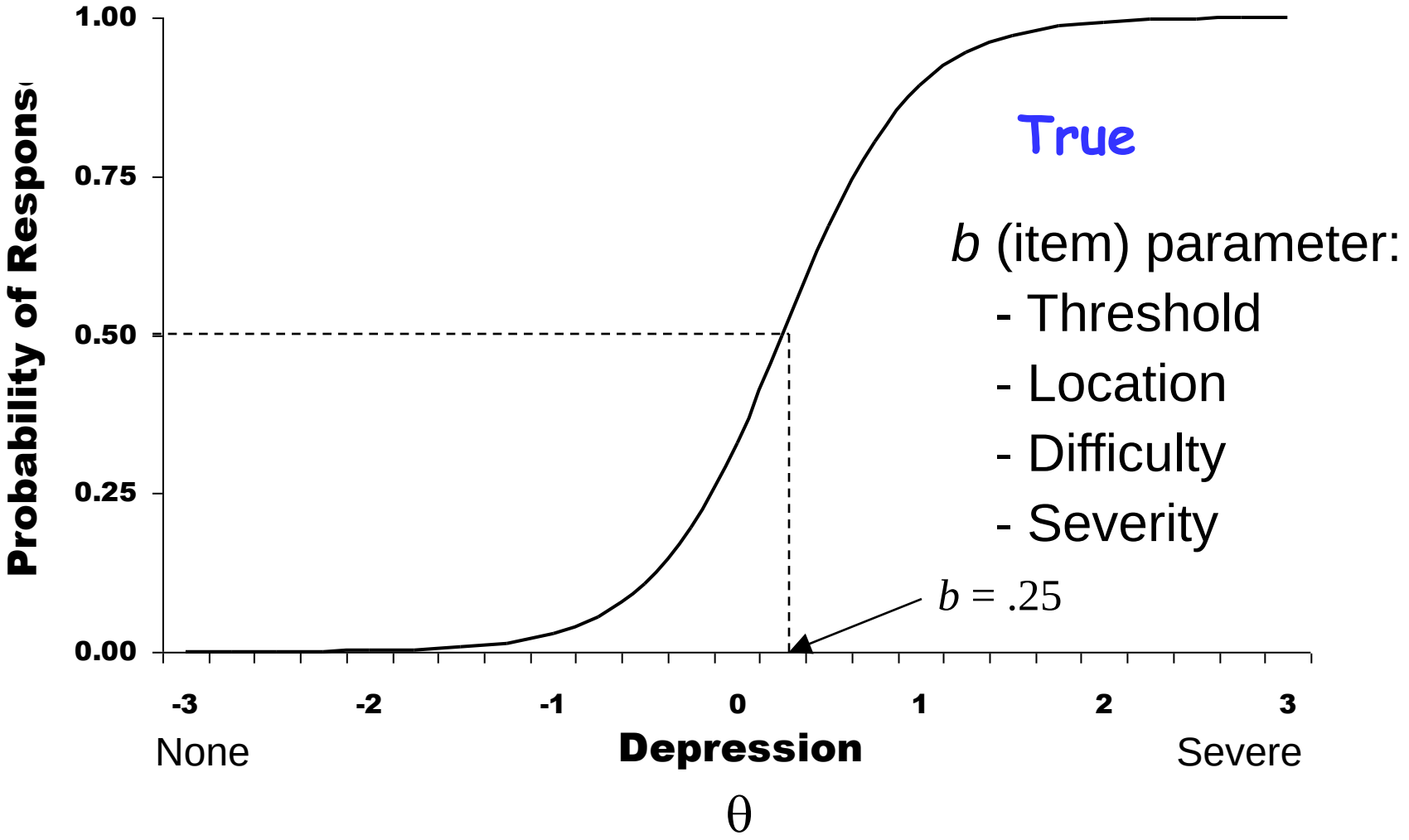
IRT Model: Item Characteristic Curves

I am unhappy some of the time?



IRT Model

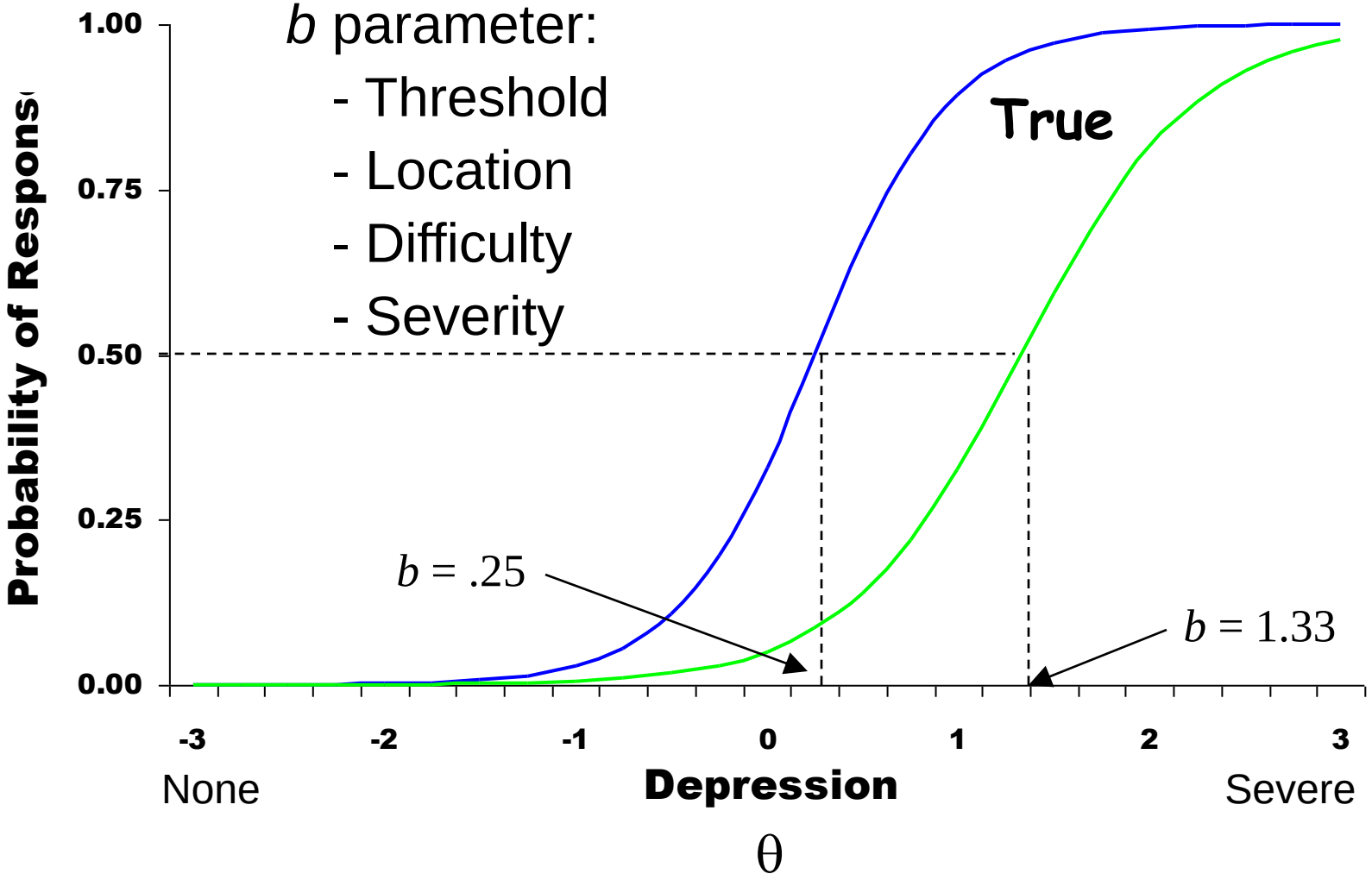
I am unhappy some of the time?



IRT Models

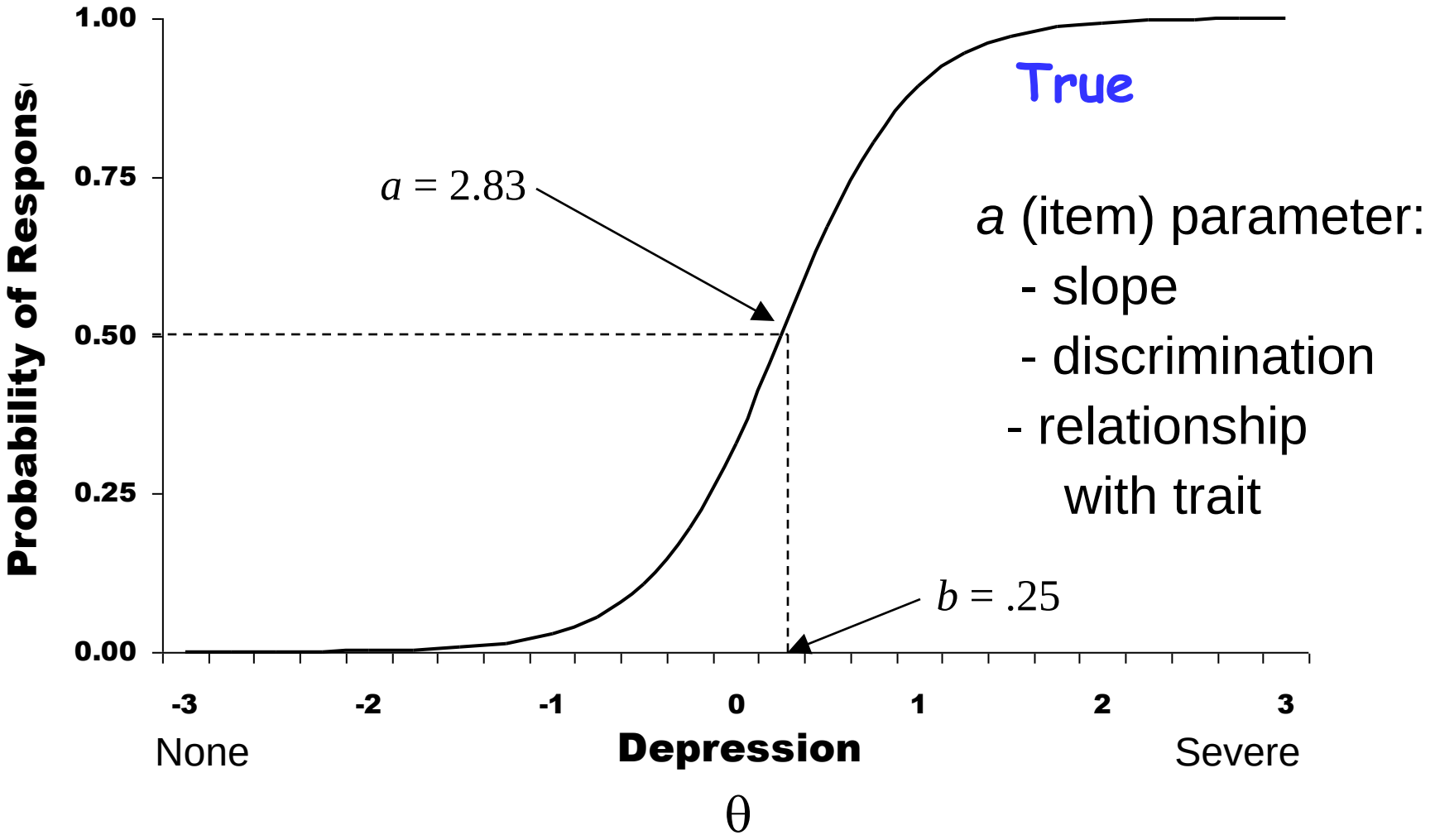
I am unhappy some of the time.

I don't care what happens to me.



IRT Model

I am unhappy some of the time?

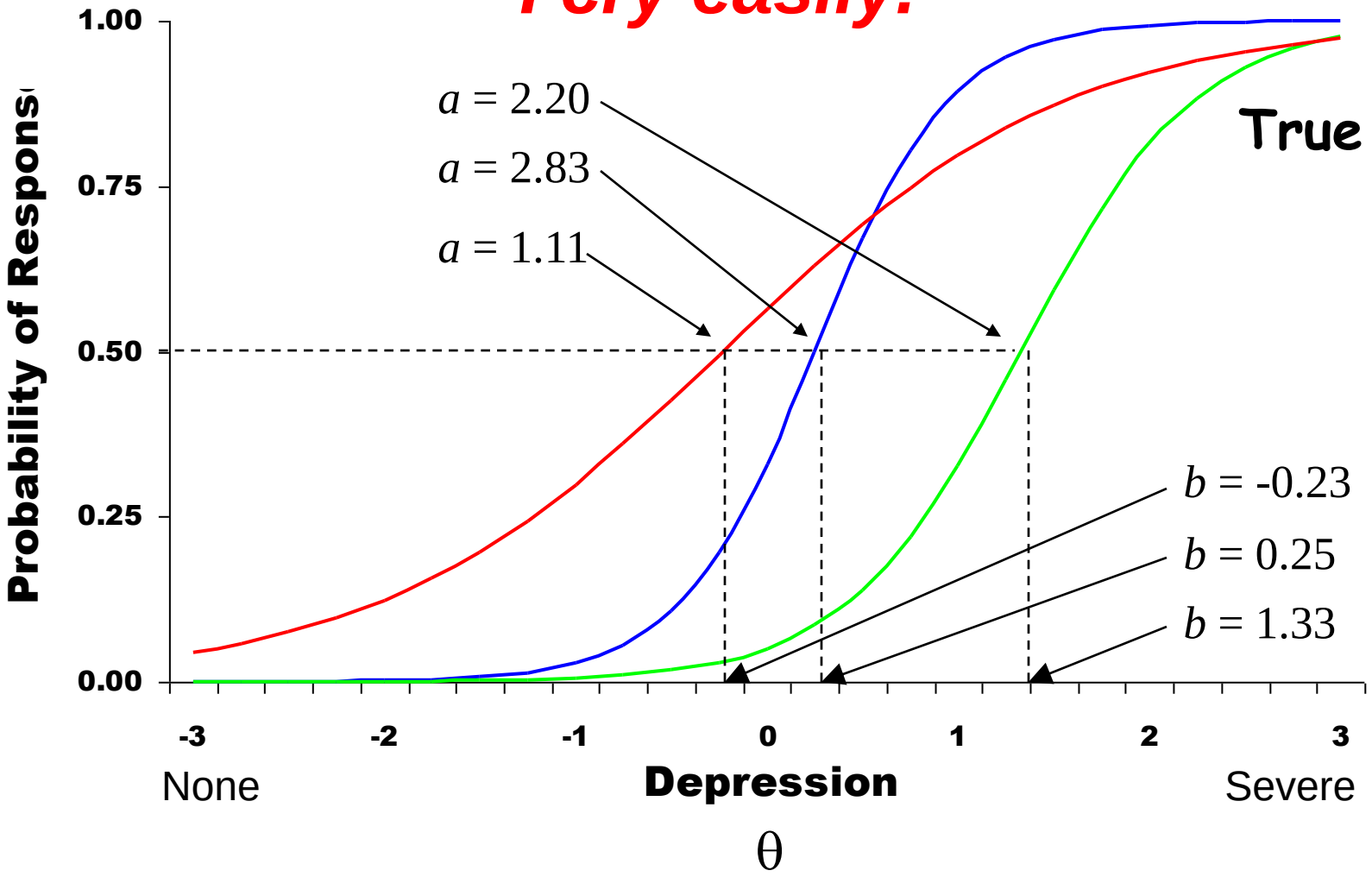


IRT Models

I am unhappy some of the time.

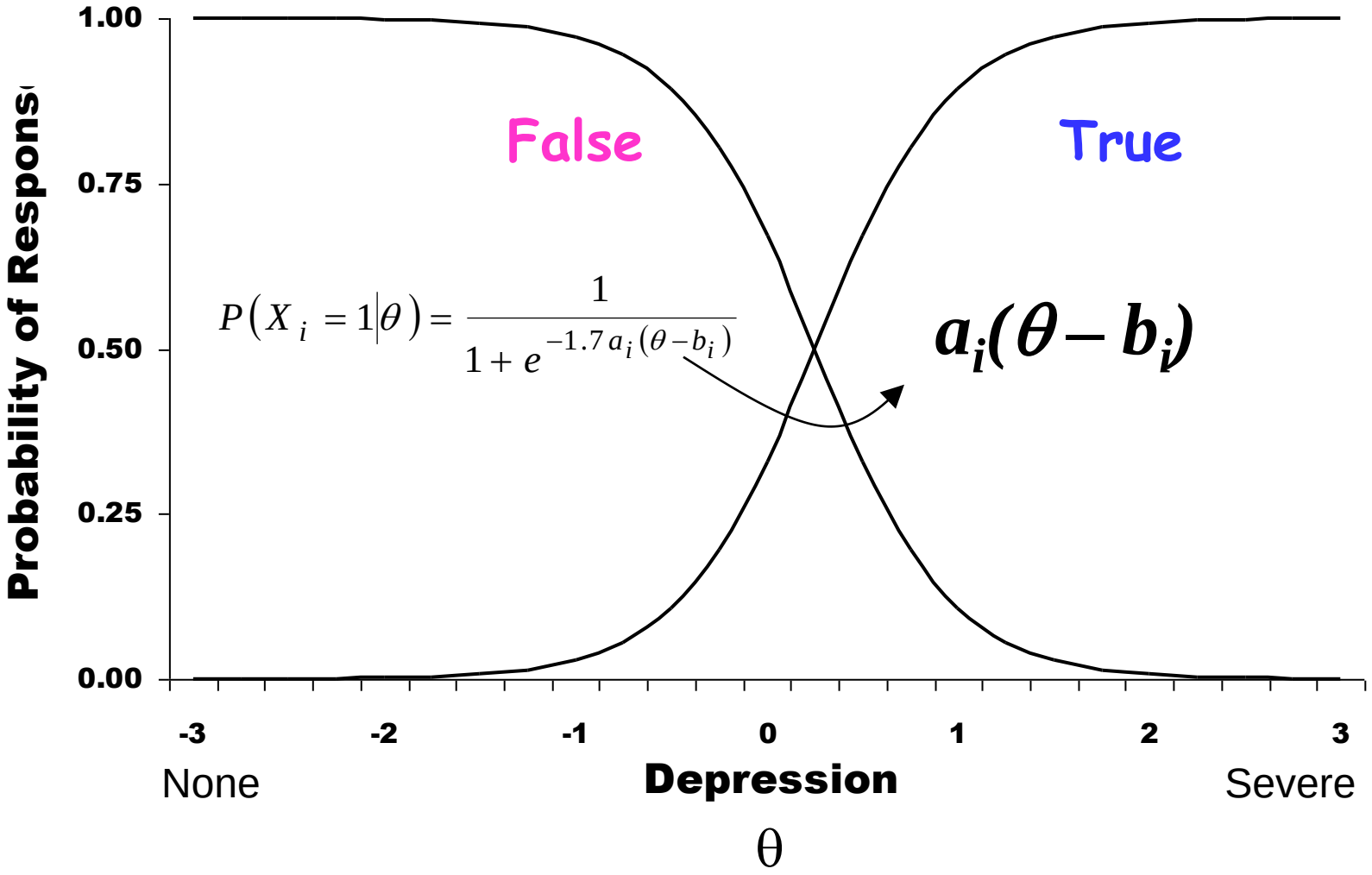
I don't care what happens to me.

I cry easily.



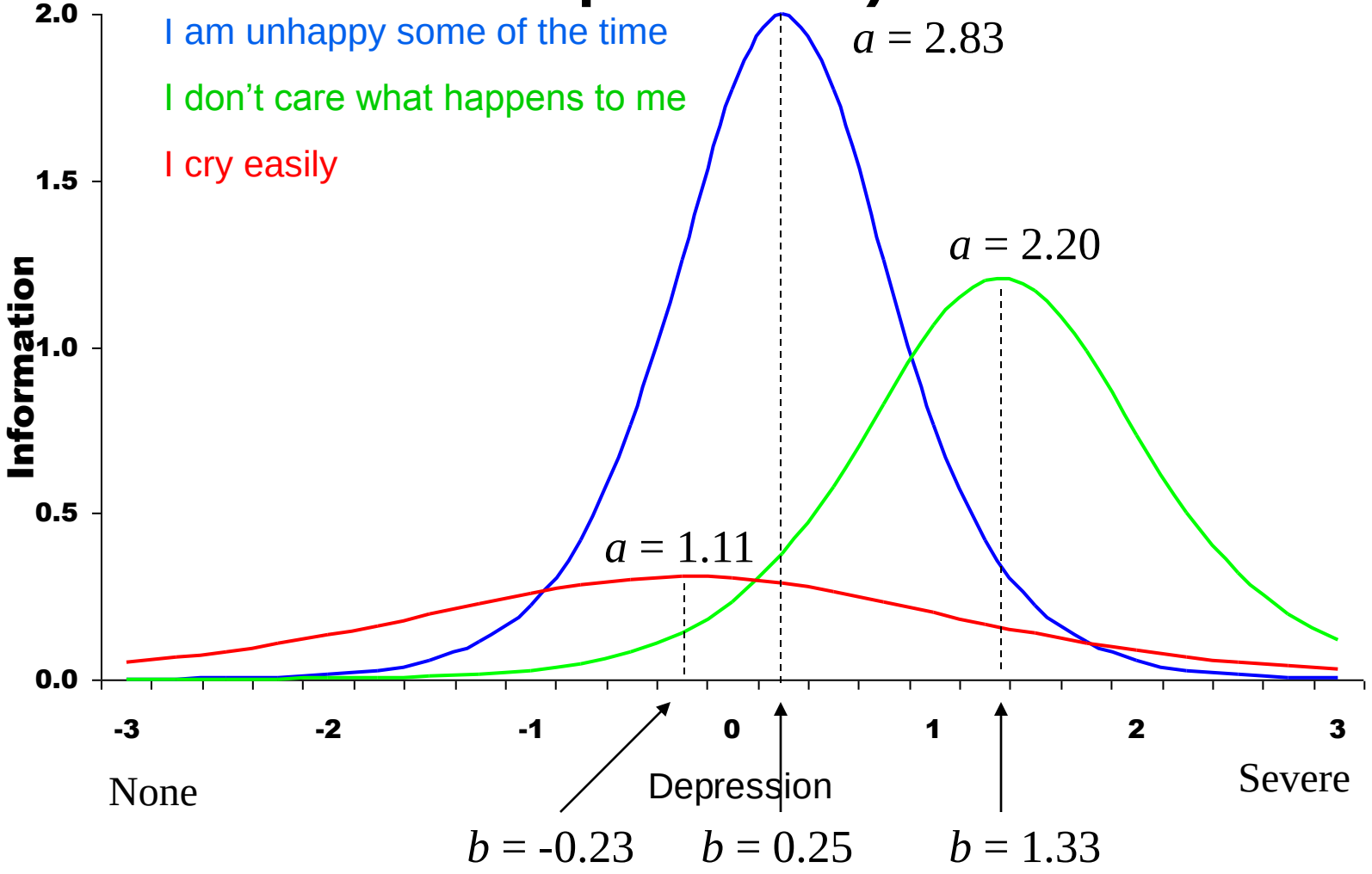
IRT Model: Item Characteristic Curves

I am unhappy some of the time?

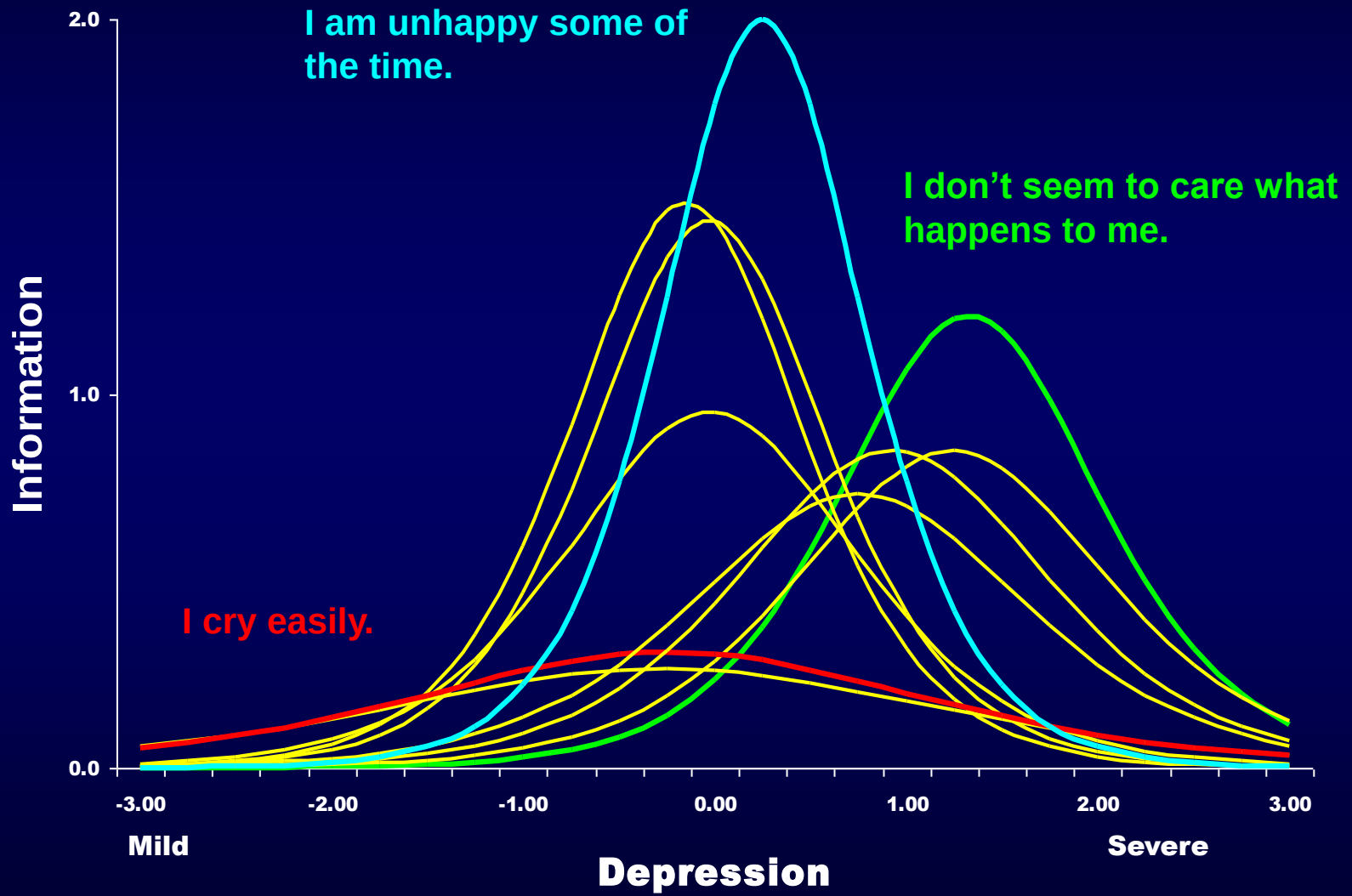


IRT: Item Information Curves

(The range of the latent construct over which an item is most useful for distinguishing among respondents)

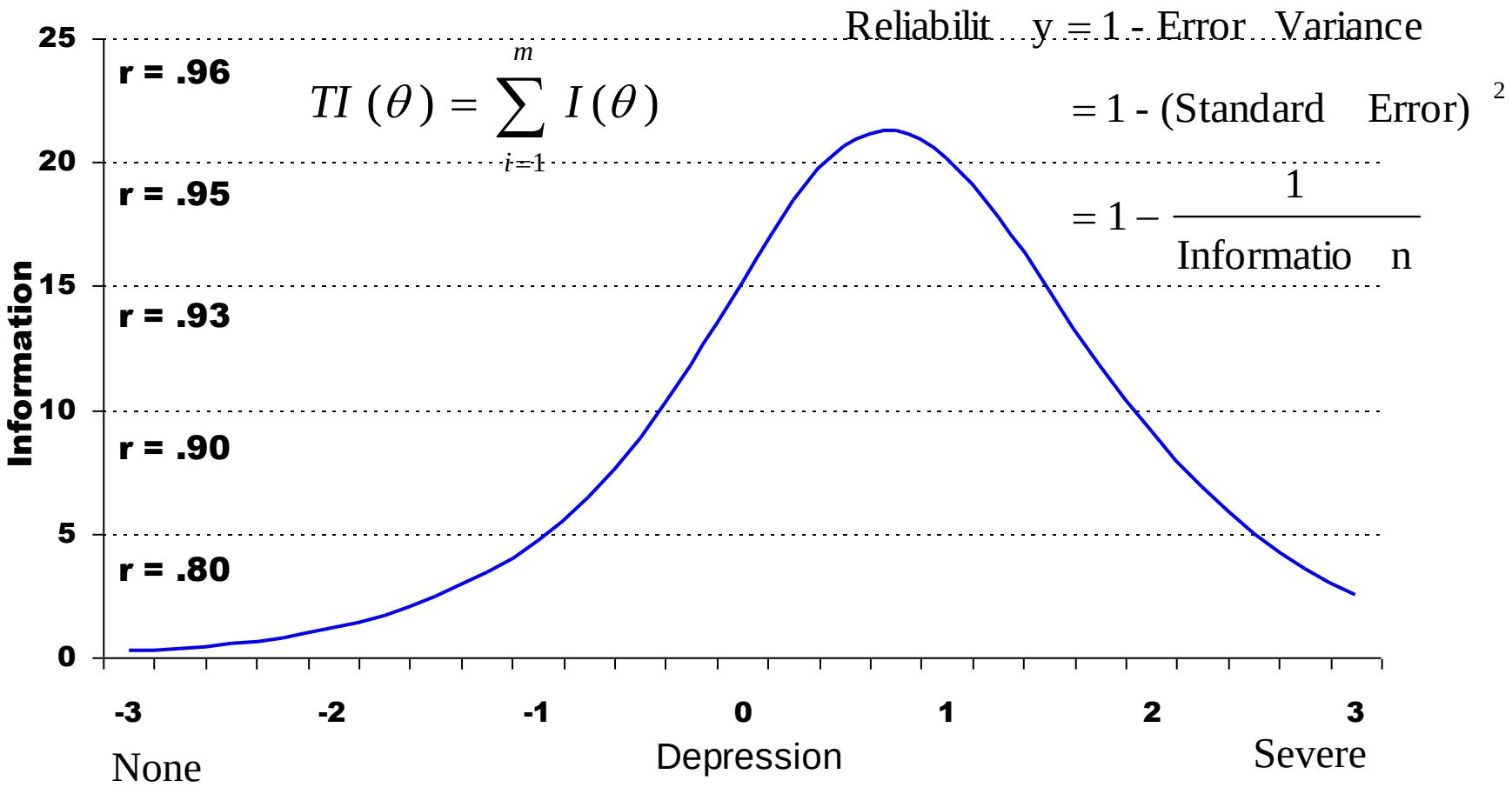


Building reliable and efficient measures…

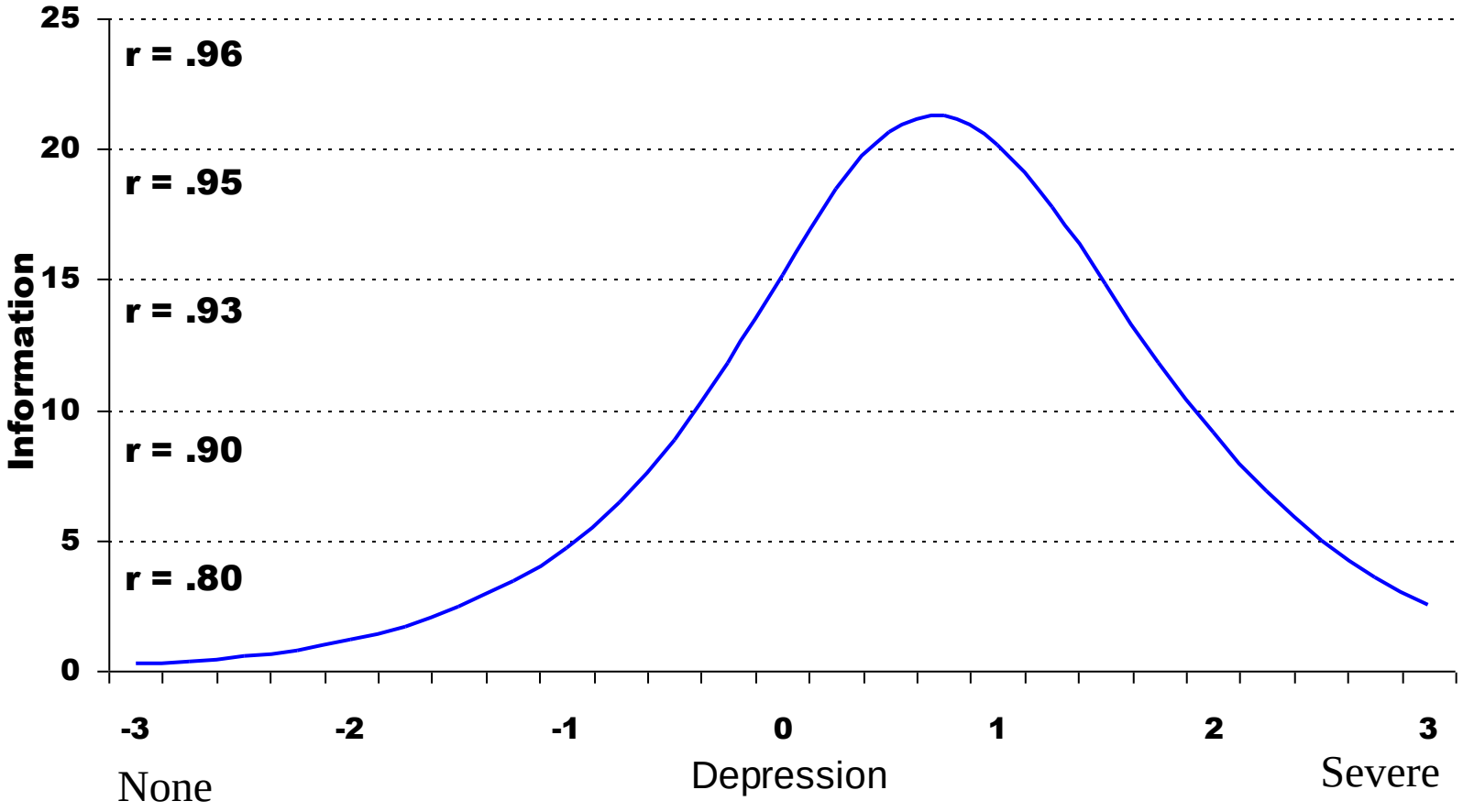


10 Items from the MMPI-2 Depression Scale

Scale (Test) Information Curve
(The range of the latent construct over which a scale is most useful for distinguishing among respondents)



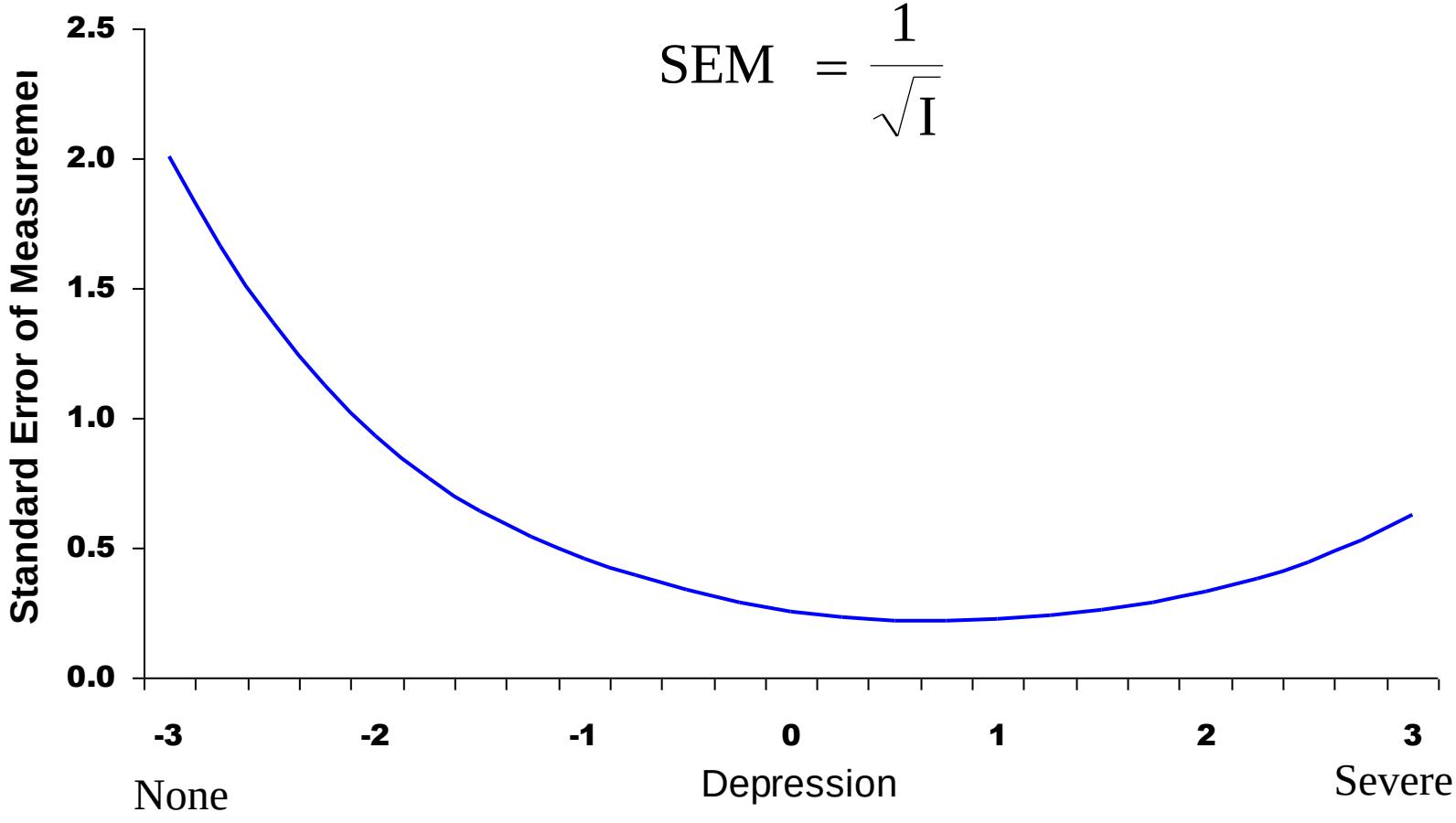
Questions on the MMPI-2 depression scales were chosen because they maximally discriminate a clinically depressed group from a non-clinical group



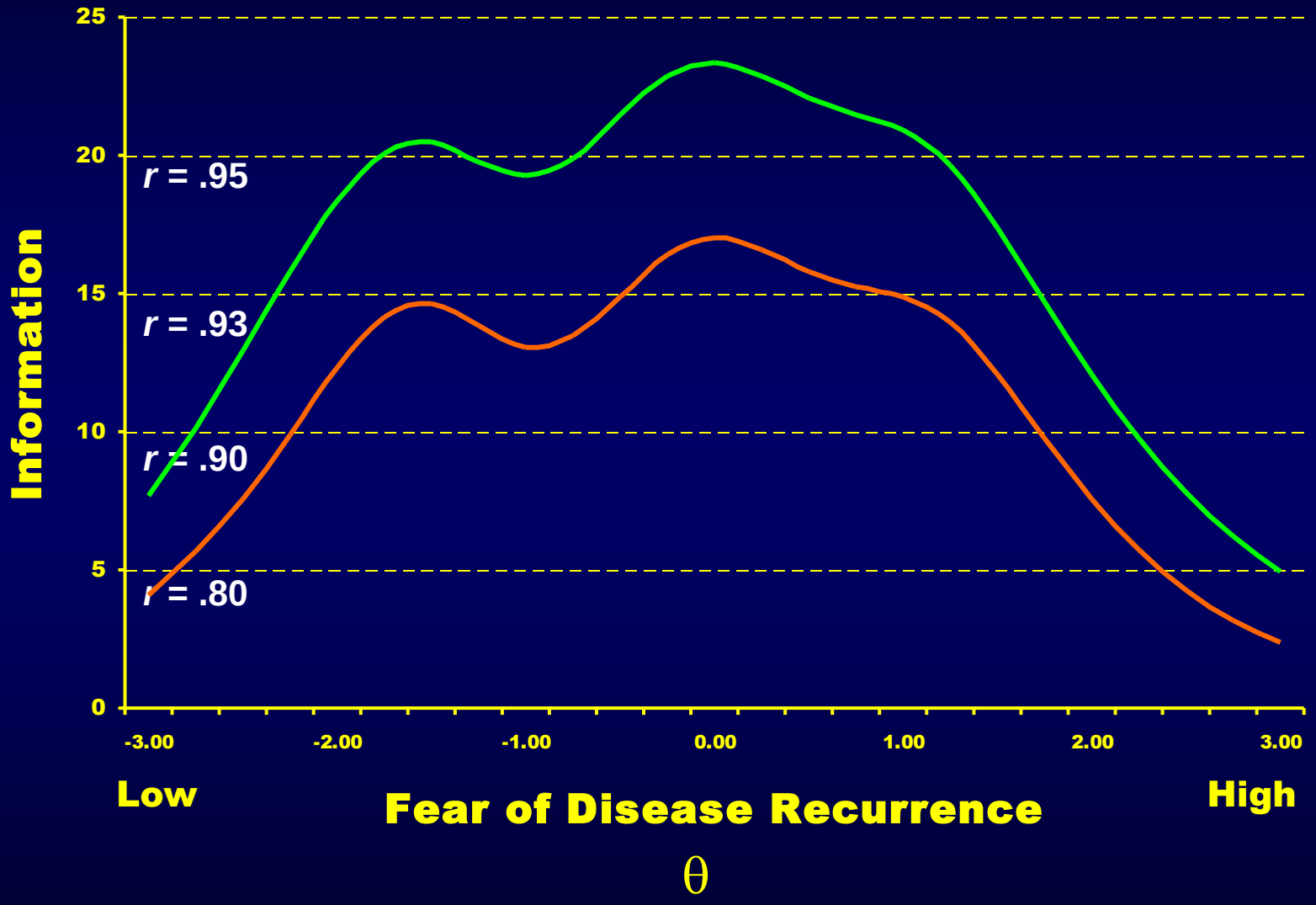
Standard Error of Measurement Curve

(The range of the latent construct over which a scale is most useful for measuring respondent trait levels)

$$SEM = \frac{1}{\sqrt{I}}$$



What is the reduction in information going from a 22 to 12 item scale?



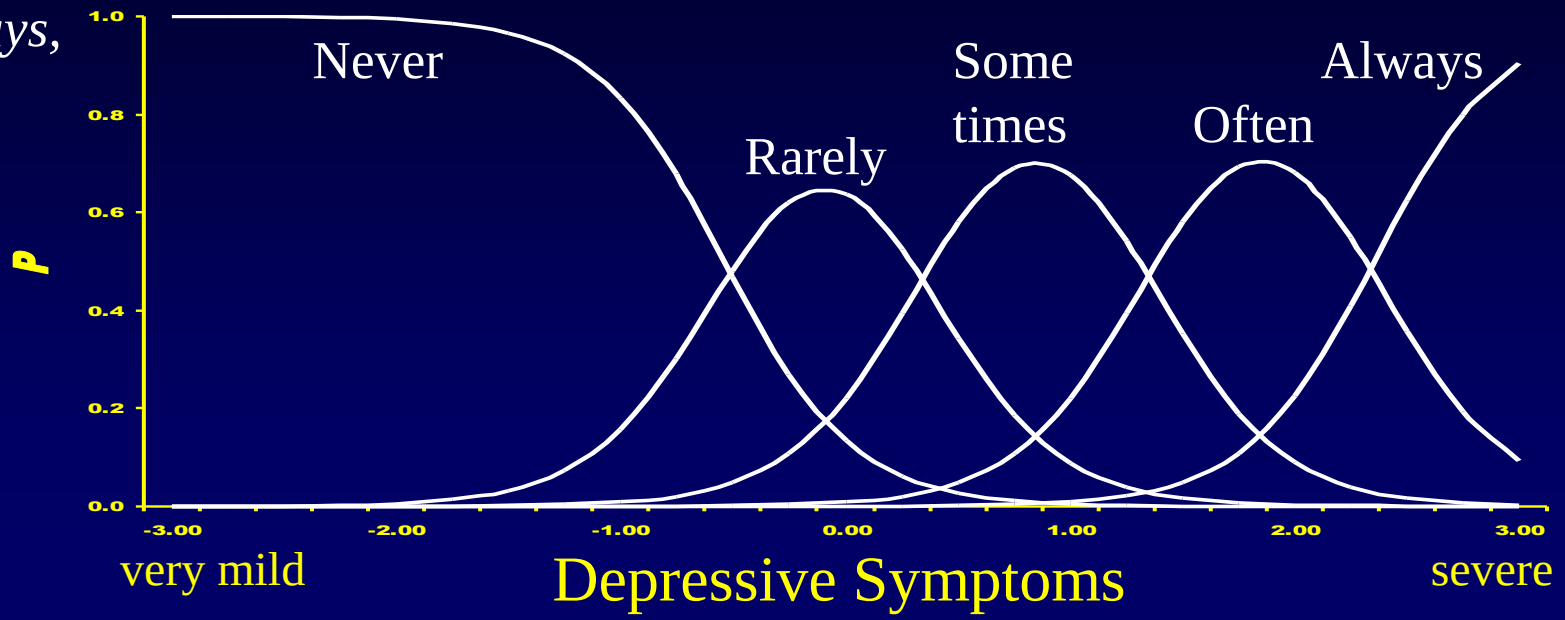
**r* = approximate reliability

What about IRT models for questions with more than two response categories?

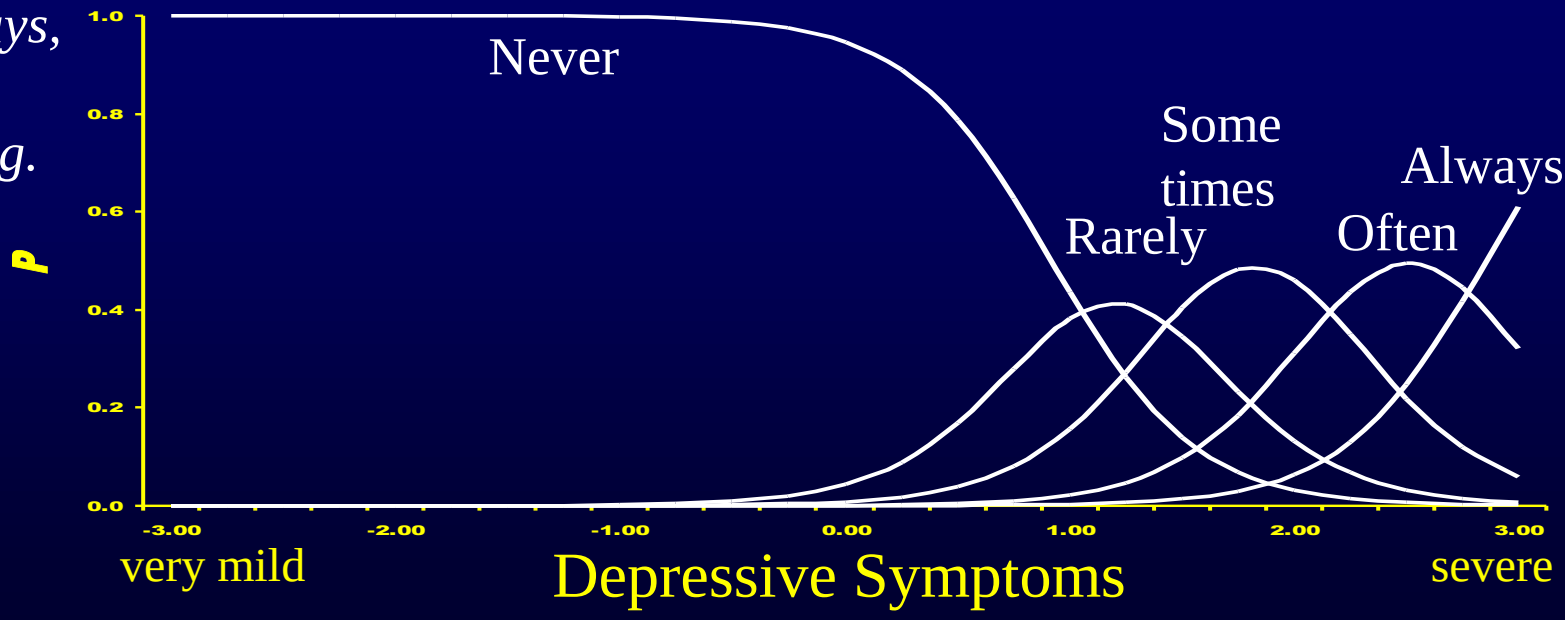
Data from responses to the PROMIS Depression Item Bank.

Item Response Theory (IRT): Category Response Curves

In the past 7 days, I felt unhappy.



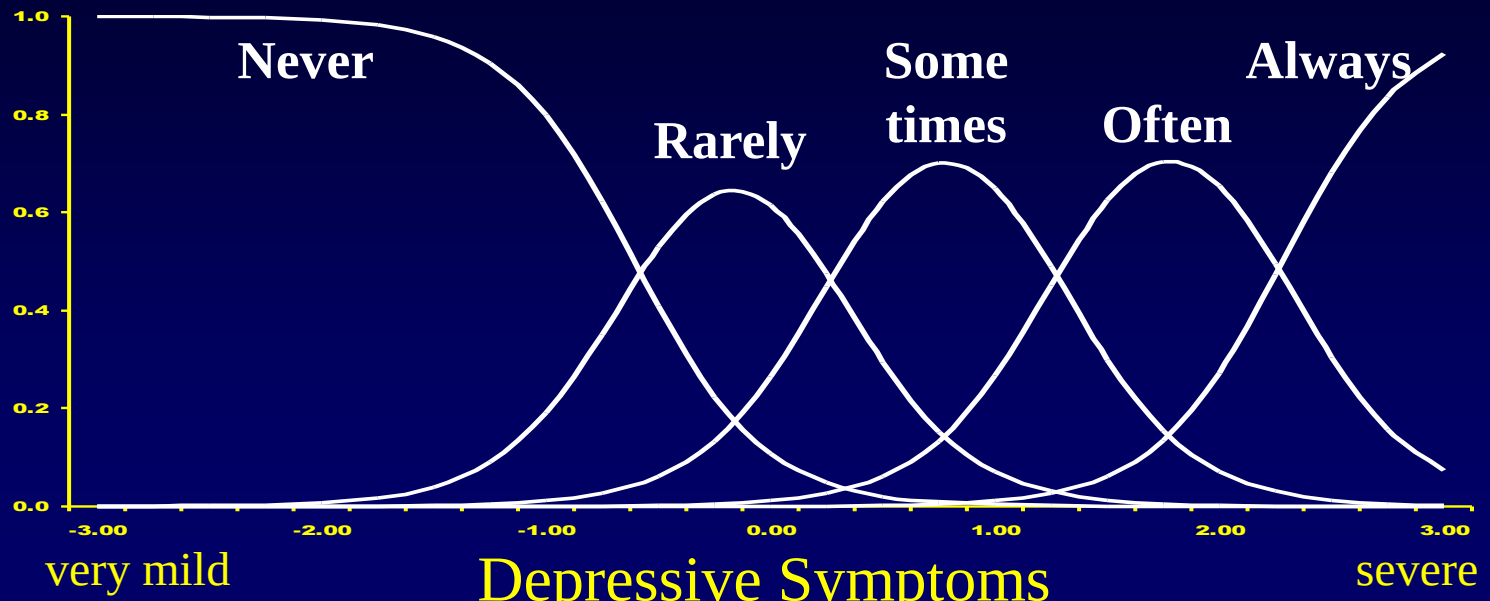
In the past 7 days, I felt I had no reason for living.



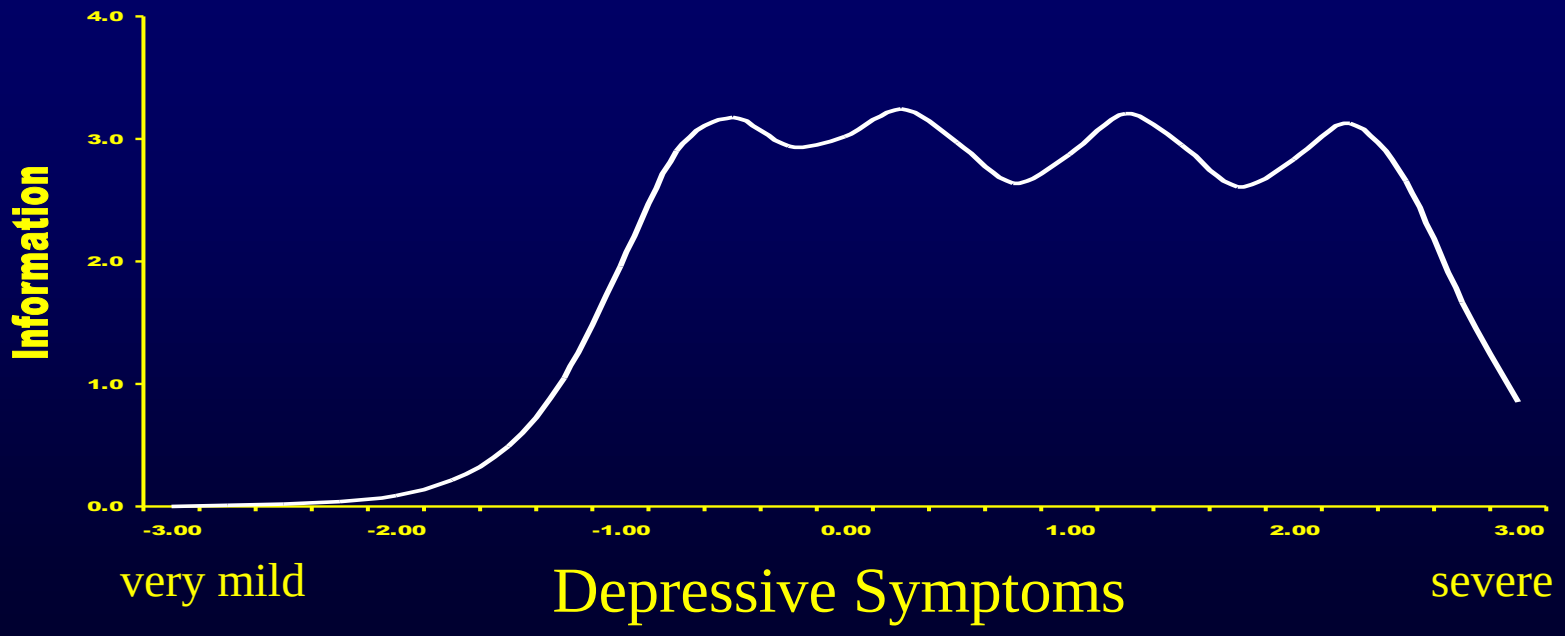
Item Response Theory (IRT)

In the past 7 days, I felt unhappy.

Category Response Curves



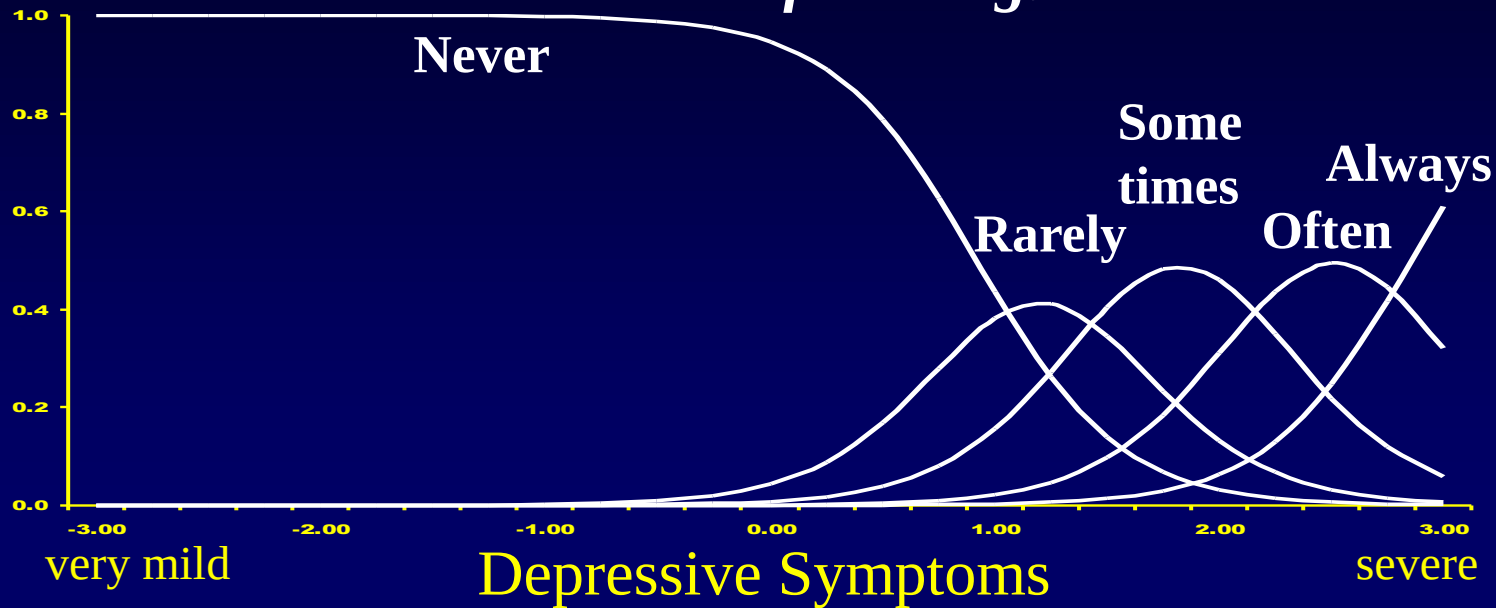
Information Function



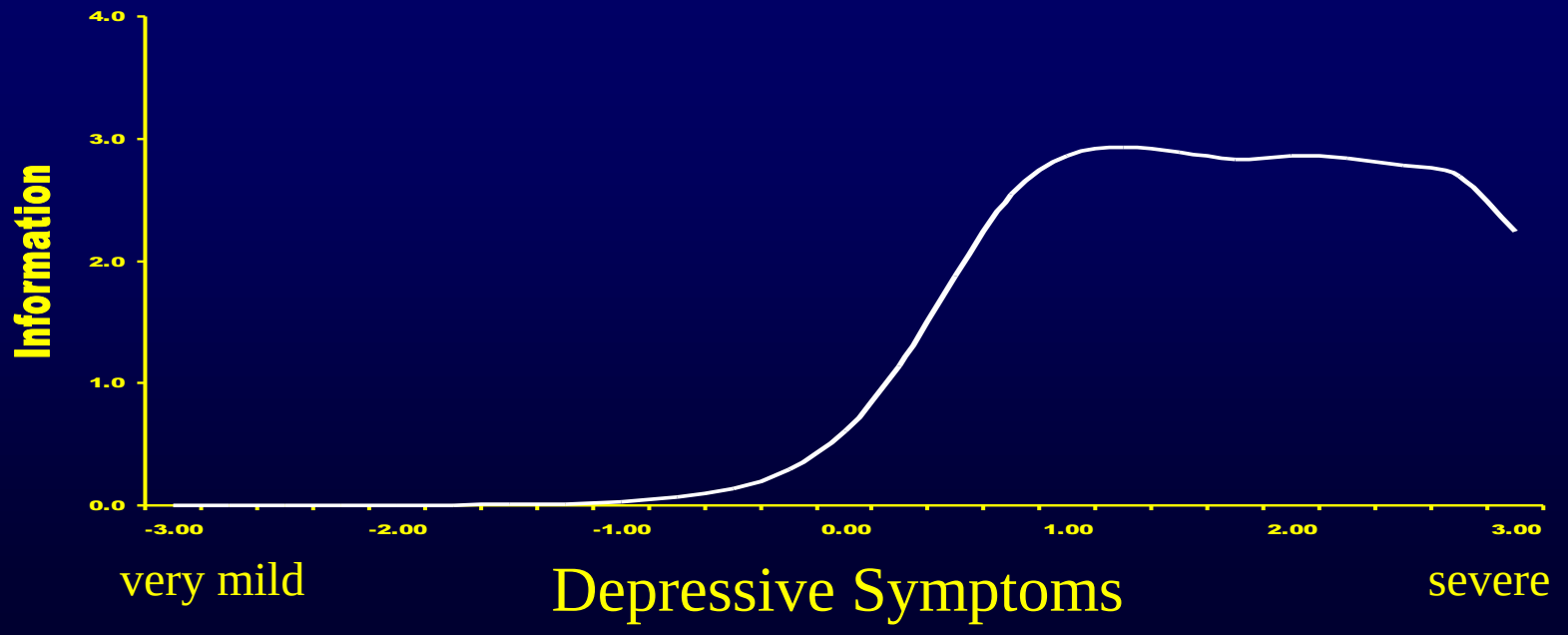
Item Response Theory (IRT)

In the past 7 days, I felt I had no reason for living.

Category Response Curves



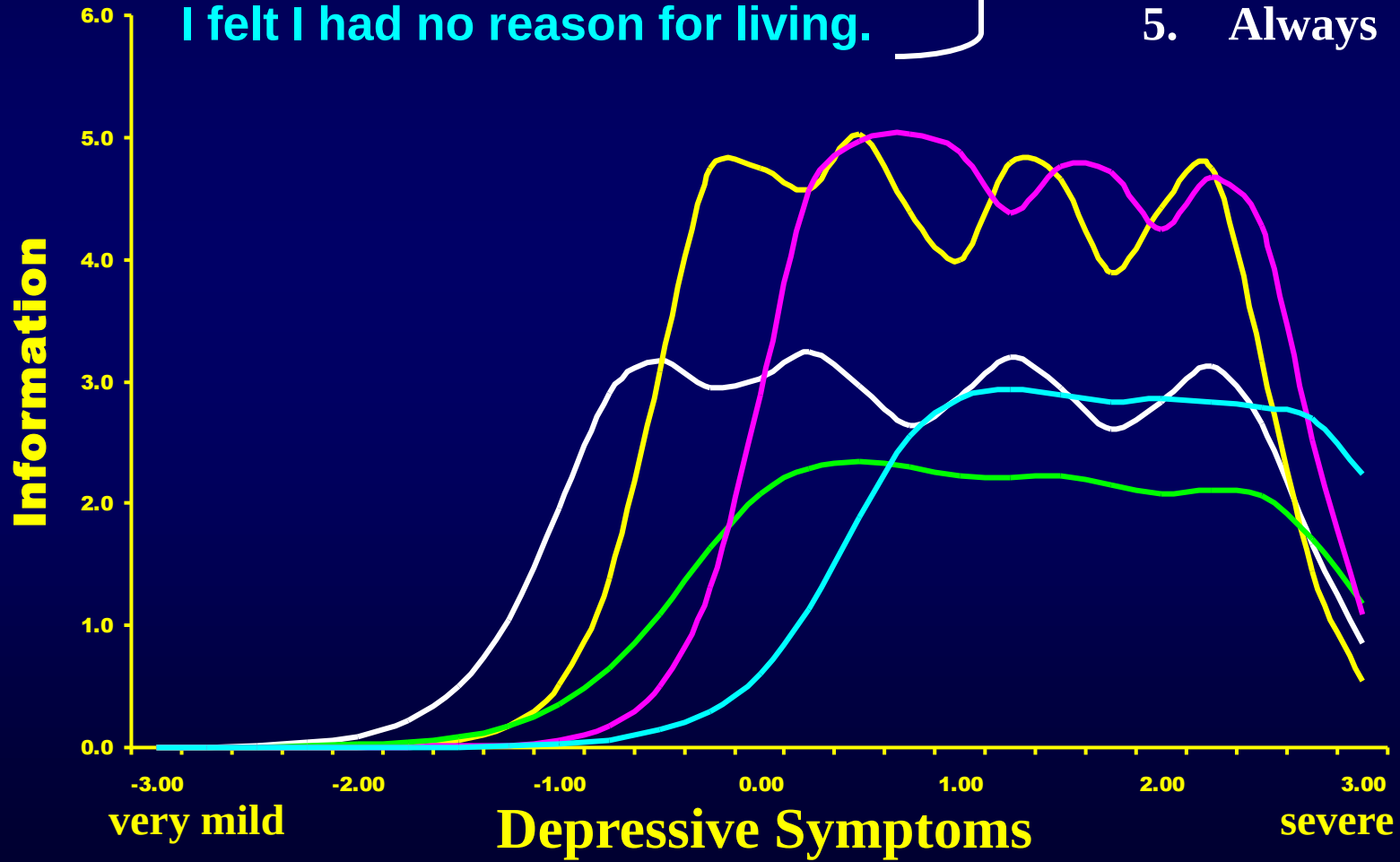
Information Function



Item Response Theory (IRT): Item Information Functions

I felt unhappy.
I felt depressed.
I withdrew from other people.
I felt worthless.
I felt I had no reason for living.

1. Never
2. Rarely
3. Sometimes
4. Often
5. Always



IRT Family of Models

IRT models come in many varieties (over a 100) to handle:

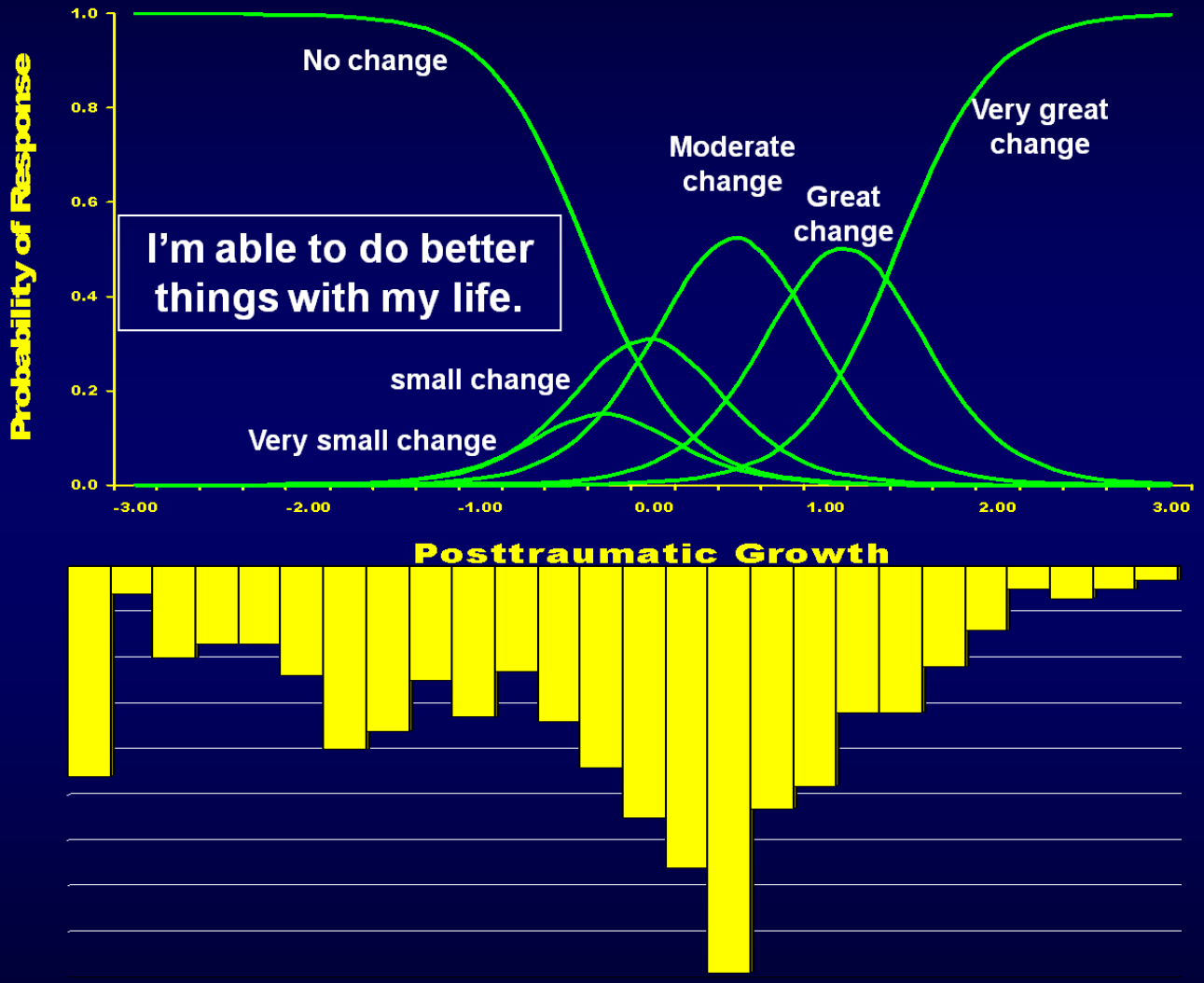
- Unidimensional and multidimensional data
- Binary, polytomous, and continuous response data
- Ordered as well as unordered response data

IRT Models You May See in Outcomes Research

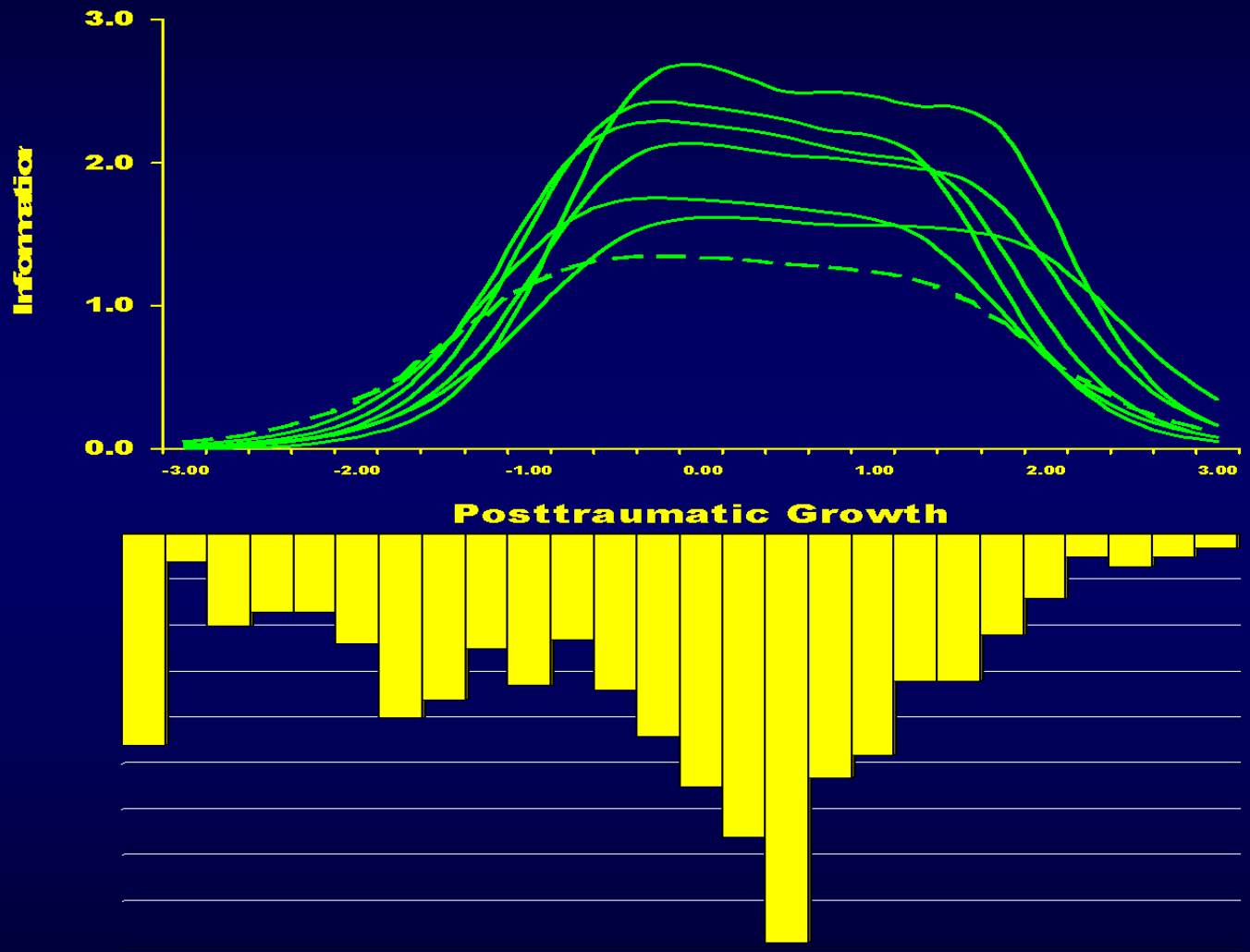
Model	Item Response Format	Model Characteristics
Rasch / 1-Parameter Logistic	Dichotomous	Discrimination power equal across all items. Threshold varies across items.
2-Parameter Logistic	Dichotomous	Discrimination and threshold parameters vary across items.
Graded Response	Polytomous	Ordered responses. Discrimination varies across items.
Nominal	Polytomous	No pre-specified item order. Discrimination varies across items.
Partial Credit (Rasch Model)	Polytomous	Discrimination power constrained to be equal across items.
Rating Scale (Rasch Model)	Polytomous	Discrimination equal across items. Item threshold steps equal across items.
Generalized Partial Credit	Polytomous	Variation of Partial Credit Model with discrimination varying among items.

Applications of IRT models for Health Outcomes Measurement

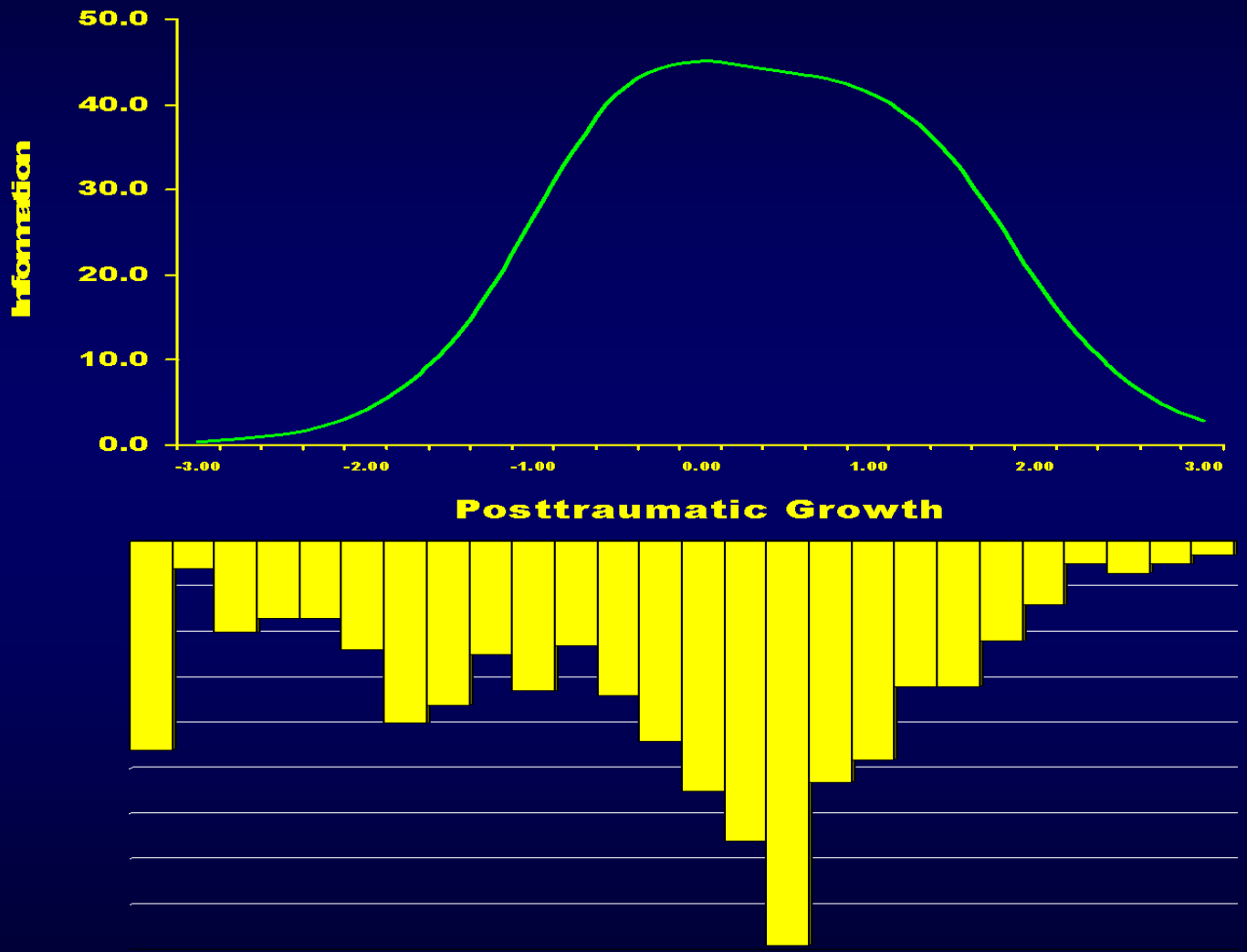
1. Design and Evaluation



1. Design and Evaluation



1. Design and Evaluation



2. Testing for Differential Item Functioning (DIF)

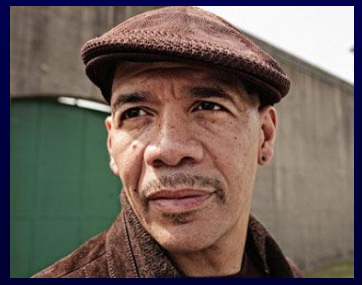
In the past 7 days, I cried

None of the time	A little of the time	Some of the time	Most of the time	All of the time

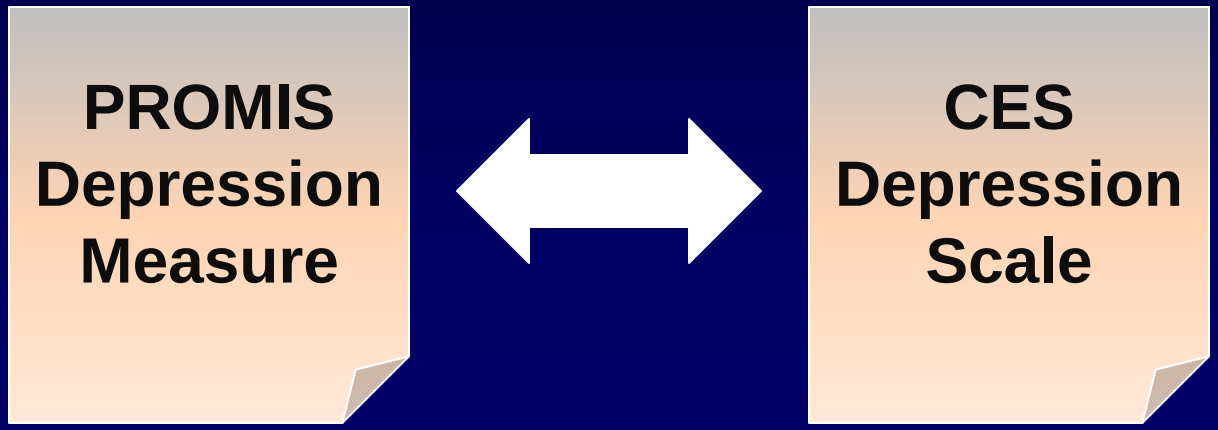
In the past 7 days, I felt blue

None of the time	A little of the time	Some of the time	Most of the time	All of the time

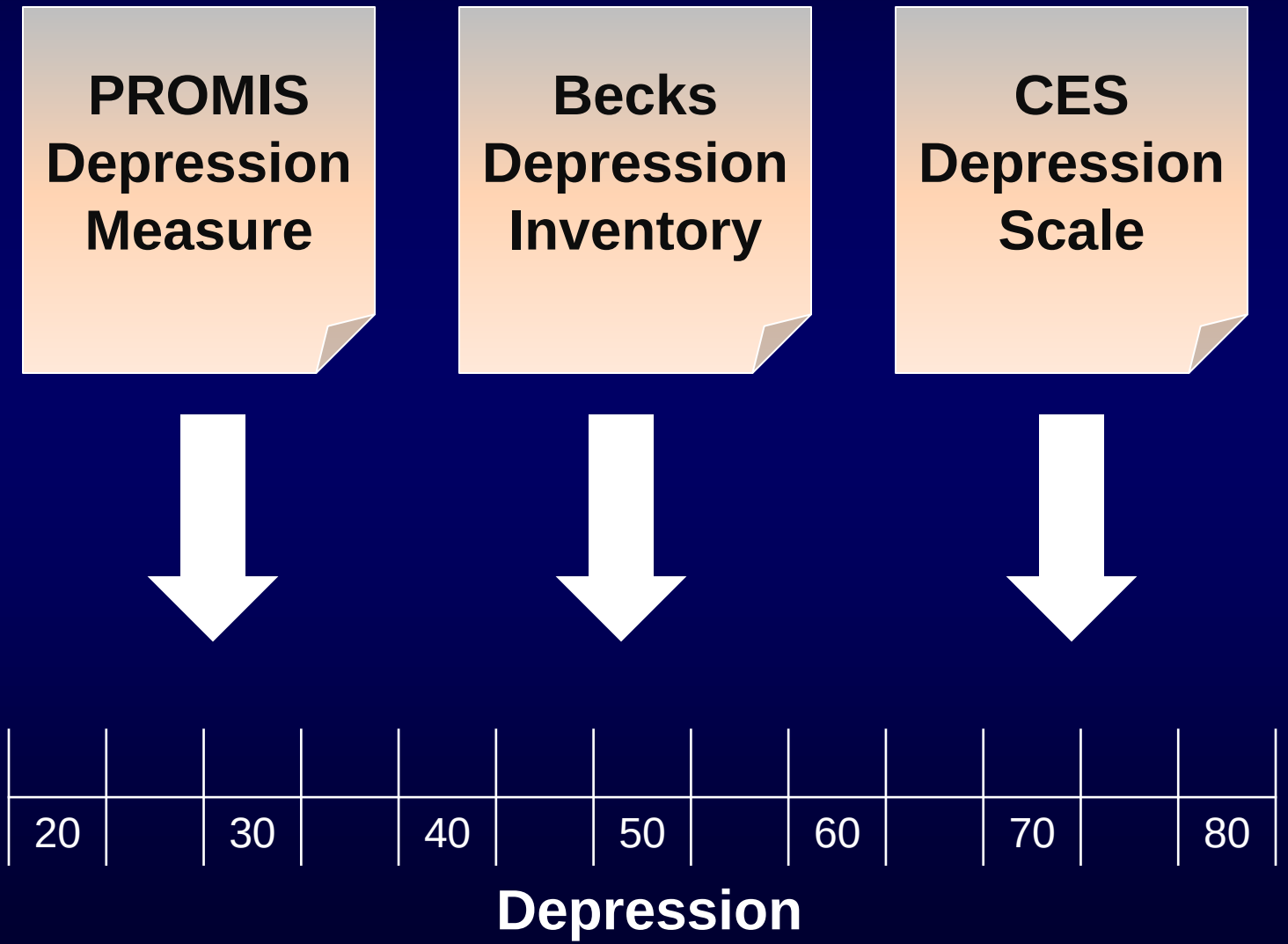
Depression



3. Linking Health Outcome Measures

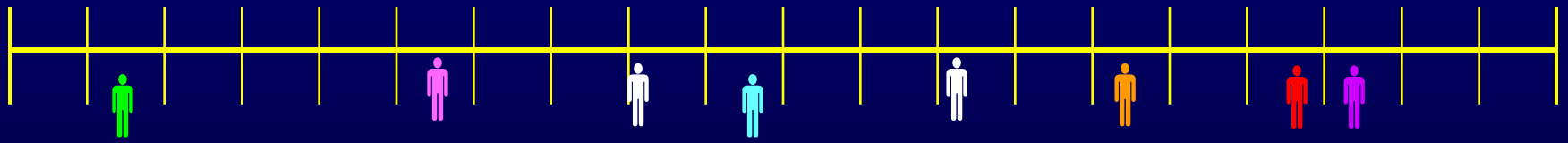


3. Linking Health Outcome Measures



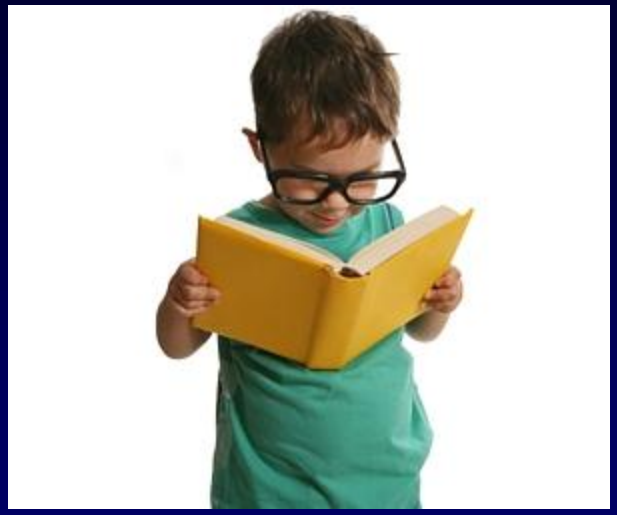
4. Item Banking and Computerized Adaptive Testing (CAT)

no depression | **mild depression** | **moderate depression** | **severe depression** | **extreme depression**



Depression Item Bank

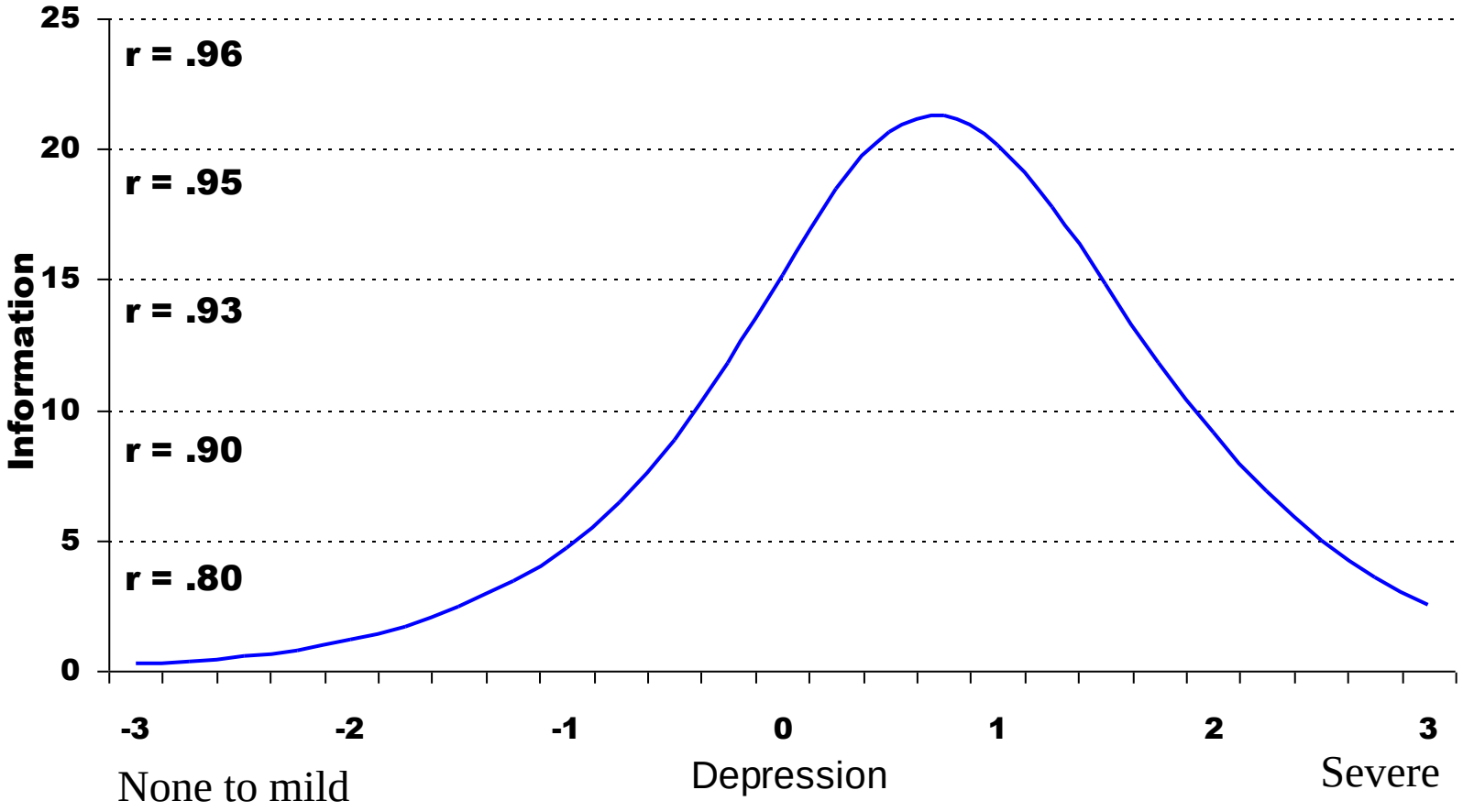
Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 | Item 7 | Item 8 | Item 9 | Item *n*



Traditional Measurement Theory (Classical Test Theory, CTT) versus Modern Measurement Theory

Classical Test Theory	Item Response Theory
Measures of precision fixed for all scores	Precision measures vary across scores
Longer scales increase reliability	Shorter, targeted scales can be equally reliable
Scale properties are sample dependent	Item & scale properties are invariant within a linear transformation
Comparing person scores dependent on item set	Person scores comparable across different item sets
Comparing respondents requires parallel scales	Different scales can be placed on a common metric
Mixed item formats leads to unbalanced impact on total scale scores	Easily handles mixed item formats
Summed scores are on an ordinal scale	Scores on interval scale
	Graphical tools for item and scale analysis

Questions on the MMPI-2 depression scales were chosen because they maximally discriminate a clinically depressed group from a non-clinical group



Conclusions

- **IRT serves as a powerful analytic tool to help design health outcomes measures.**

- **Limitations**
  - **Lack of user-friendliness of software**
  - **Required knowledge of measurement theory.**
  - **Needs large sample sizes**

ISOQOL 20TH ANNUAL CONFERENCE

October 9–12, 2013 **MIAMI, FLORIDA, USA**

21st ISOQOL Conference in Berlin, Germany
October 15-18, 2014

rgizing the Science of
Quality of Life Research:
Where have we been and where can we go?

Important Deadlines
April 12: Oral and Poster Presentation Abstract Submissions Due
May 31: Scholarship Applications Nominations Due
July 1: (...
Au... Ear... Deadline
Septe...mber 16: Advanced Registration Deadline
September 16: ISOQOL Hotel Room Block Closes

isoqol.org/2013conference

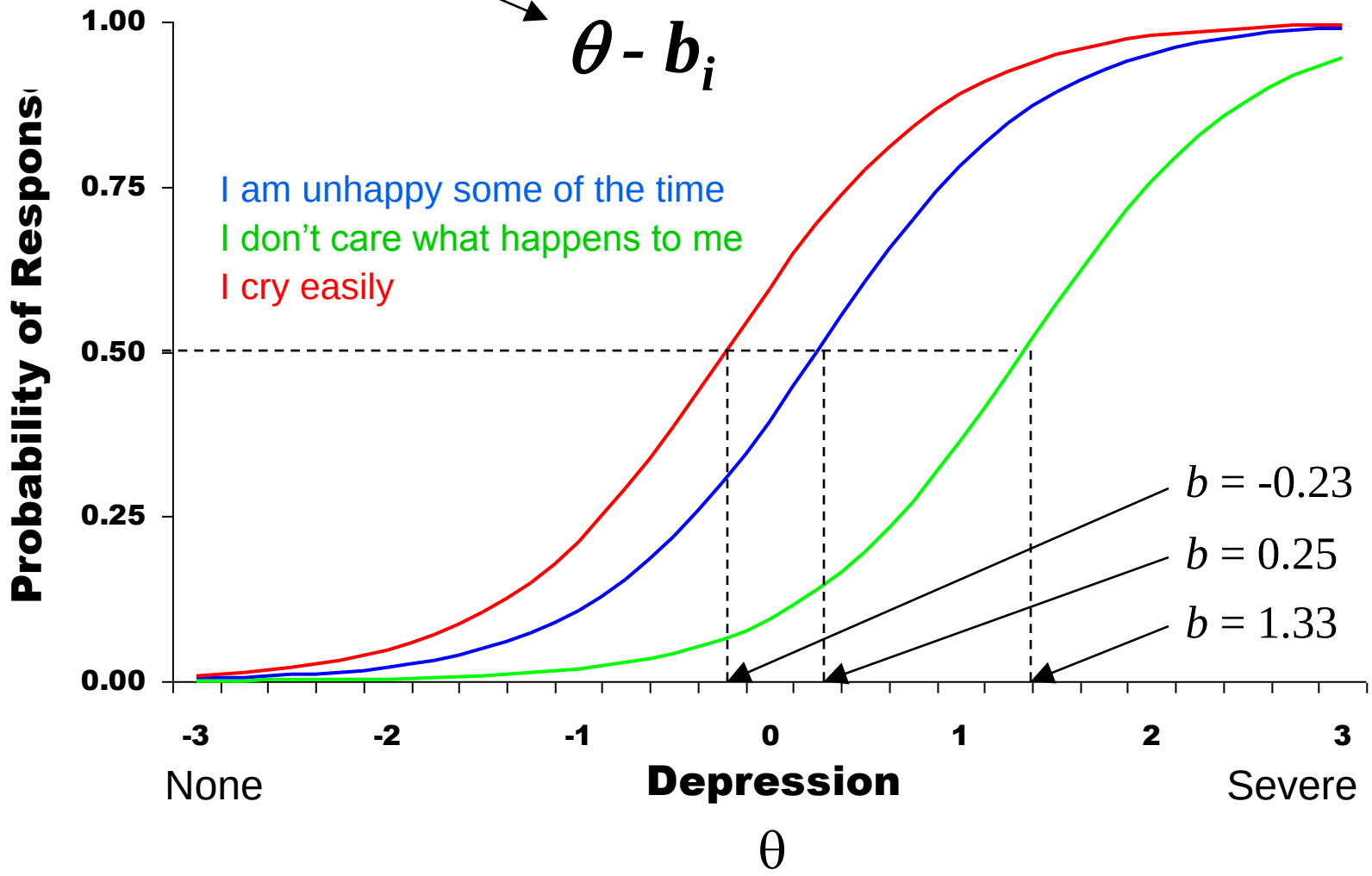
Sample Size Issues

Sample Size Issues

- The IRT model to be estimated
  - Parameters ↑, Sample Size ↑ - Rasch models need less data.
- The number of items or questions.
  - Number of items ↑, Sample Size ↑
- The number of response options.
  - Number of response categories ↑, Sample Size ↑
- Unidimensionality of construct
  - Better the data meet assumption of unidimensionality, sample size ↓
- The item properties
  - Items at the extremes need more data
- Population distribution
  - Distributed across theta continuum, Sample Size ↓
- Purpose of Study
  - Evaluation of an instrument, smaller sample sizes needed
  - Estimate accurate respondent scores, larger sample sizes needed.
  - Calibrating items for an item bank, larger sample sizes

Rasch / 1-Parameter Logistic IRT Model

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \qquad P(X_i = 1|\theta) = \frac{1}{1 + e^{-a(\theta - b_i)}} \qquad P(X_i = 1|\theta) = \frac{1}{1 + e^{-1.7a(\theta - b_i)}}$$



2-Parameter Logistic IRT Model

$$P(X_i = 1|\theta) = \frac{1}{1 + e^{-1.7\,a_i(\theta - b_i)}}$$

$a_i(\theta - b_i)$