Traditional and Modern Approaches to Outcomes Measurement

Ronald K. Hambleton
University of Massachusetts
at Amherst

NCI and DIA Conference, June 24, 2004

Goals of the Presentation

- Strengths and weaknesses of classical test theory
- Item response theory: assumptions, models, features
- Descriptions of IRT applications
- Concluding remarks

Traditional (Classical) Methods to Instrument Development

- Classical test theory has been used in the instrument development field for over 80 years—Basic models, concept of error, p and r values, split-half reliability, coefficient alpha, range restriction corrections, etc.
- HR-literature is full of highly reliable and valid instruments.

Classical Test Theory: Weaknesses

- Dependence of Item Statistics on the Sample of Respondents
- Dependence of Respondent Scores on Choice of Items
- Assumption of Common Errors
- No Modeling of Data at Item Level
- Items and Respondents on Different Scales

What is Item Response Theory?

- A statistical theory that links observable respondent performance to items in an instrument to unobservable trait or traits of interest via statistical models.
- Introduces traits, item parameters, models, etc.

Item Response Theory

- Theory is general—framework for many specific models to be generated:

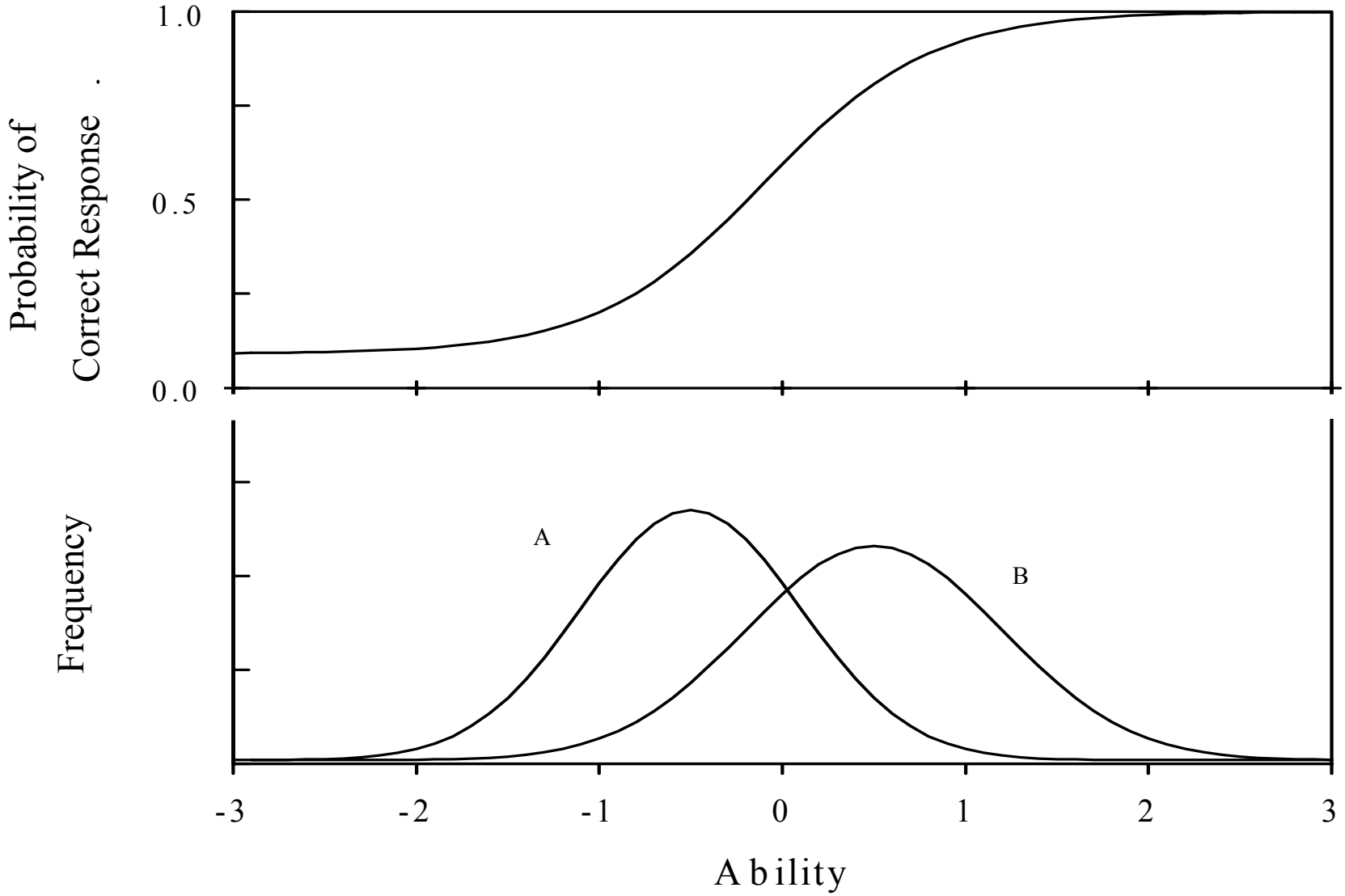
--one or more “abilities” or “traits”

--various assumptions/models

--binary or polytomous data

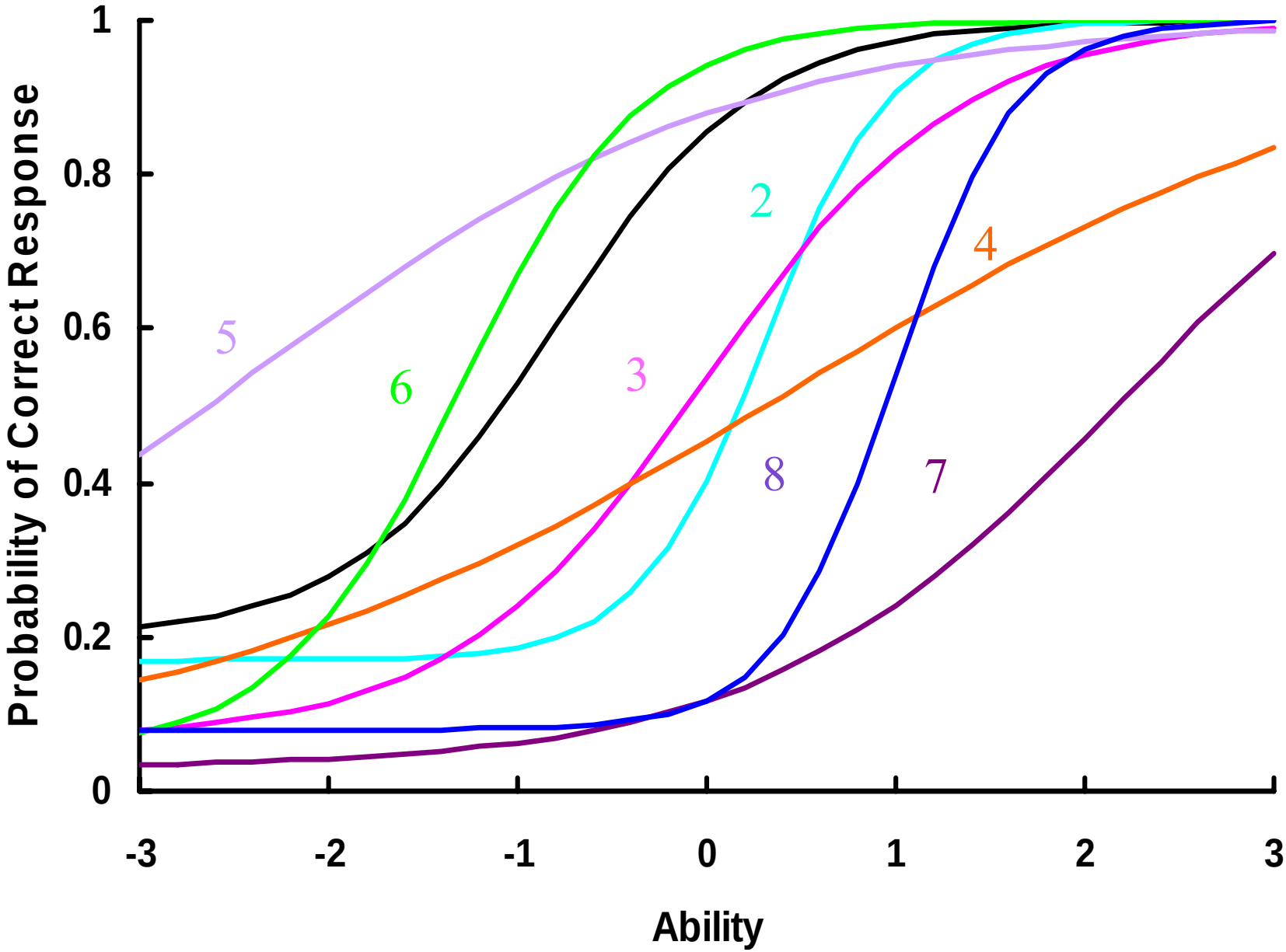
- At the specific model level, fit can be addressed.

Item characteristic Curve (for binary data):

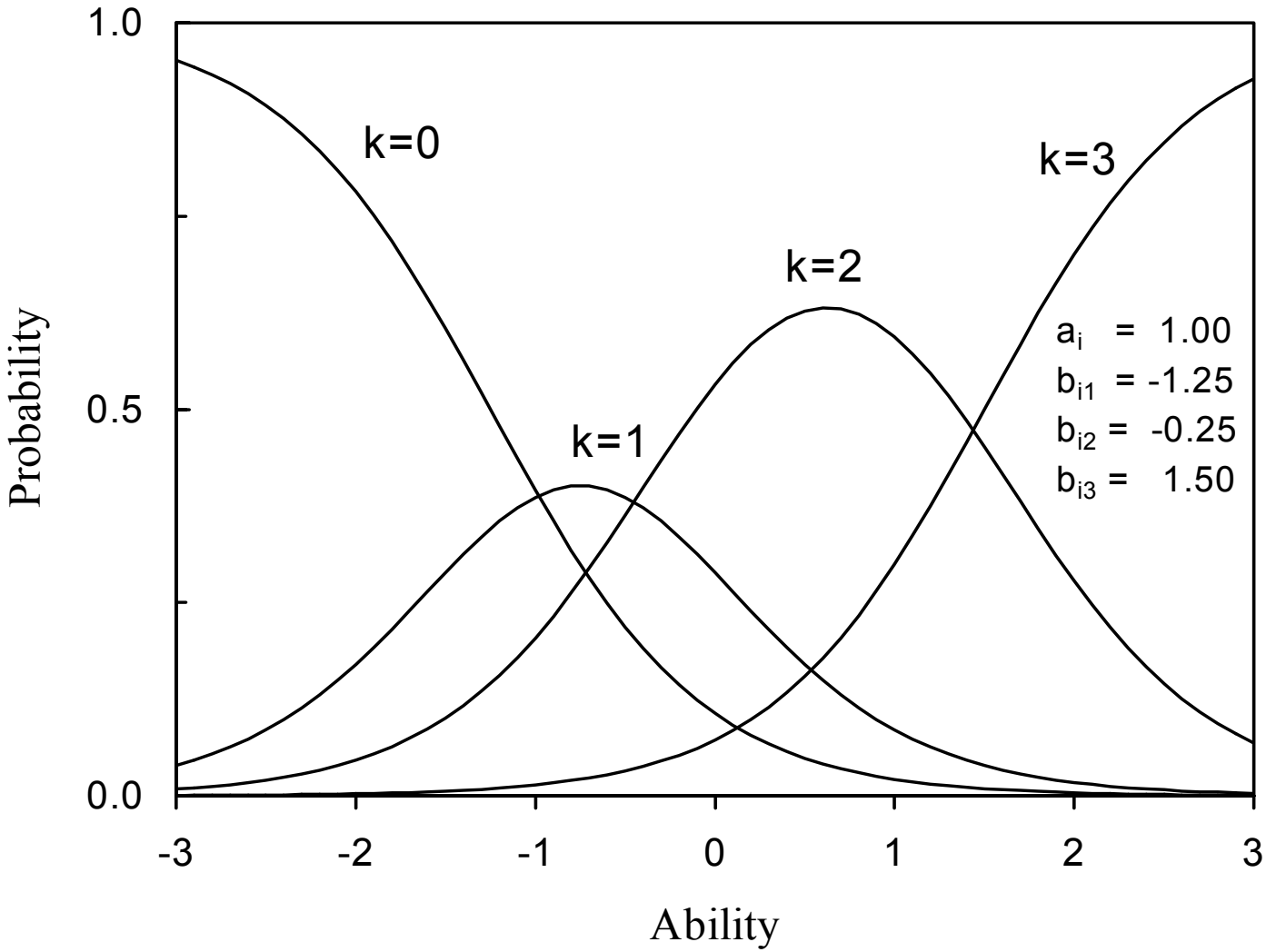


Three-Parameter Logistic Model:

$$P(x_i = 1 | \theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$


Probability of Correct Response
1
0.8
0.6
0.4
0.2
0
-3
-2
-1
0
1
2
3
Ability
5
6
3
2
4
8
7

Polytomous Response Model With Item Score Category Functions



Generalized Partial Credit Model:

$$P(x_i = k \mid \theta) = \frac{\exp[\sum_{s=1}^{k} a_i(\theta - b_{si})]}{1 + \sum_{r=1}^{k} \exp[\sum_{s=1}^{r} a_i(\theta - b_{si})]}$$

What are specific IRT model assumptions?

- The big one for several models: Dominant first factor measured by the items. [Note: multidimensional models do exist too.]
- No dependencies between items.
- Mathematical form of the ICCs linking performance on item and the trait measured by the instrument.

IRT: Many Useful Features

- Item parameter estimation is independent of particular respondent samples
- Trait (ability) estimation is independent of particular choice of items (invaluable property for CAT).

IRT: Many Useful Features, Cont.

- Error of measurement for each respondent.
- Modeling of data at the item level allows for “optimal assessments.”
- Items and respondents calibrated on the same reporting scale (enhances instrument development and interpretation)

IRT Shortcomings/Limitations

- Many practitioners lack the expertise for choosing and applying models.
- Available IRT software is not always straightforward to use.
- Large samples are helpful in item parameter estimation.
- Does not address construct definition/domain coverage.

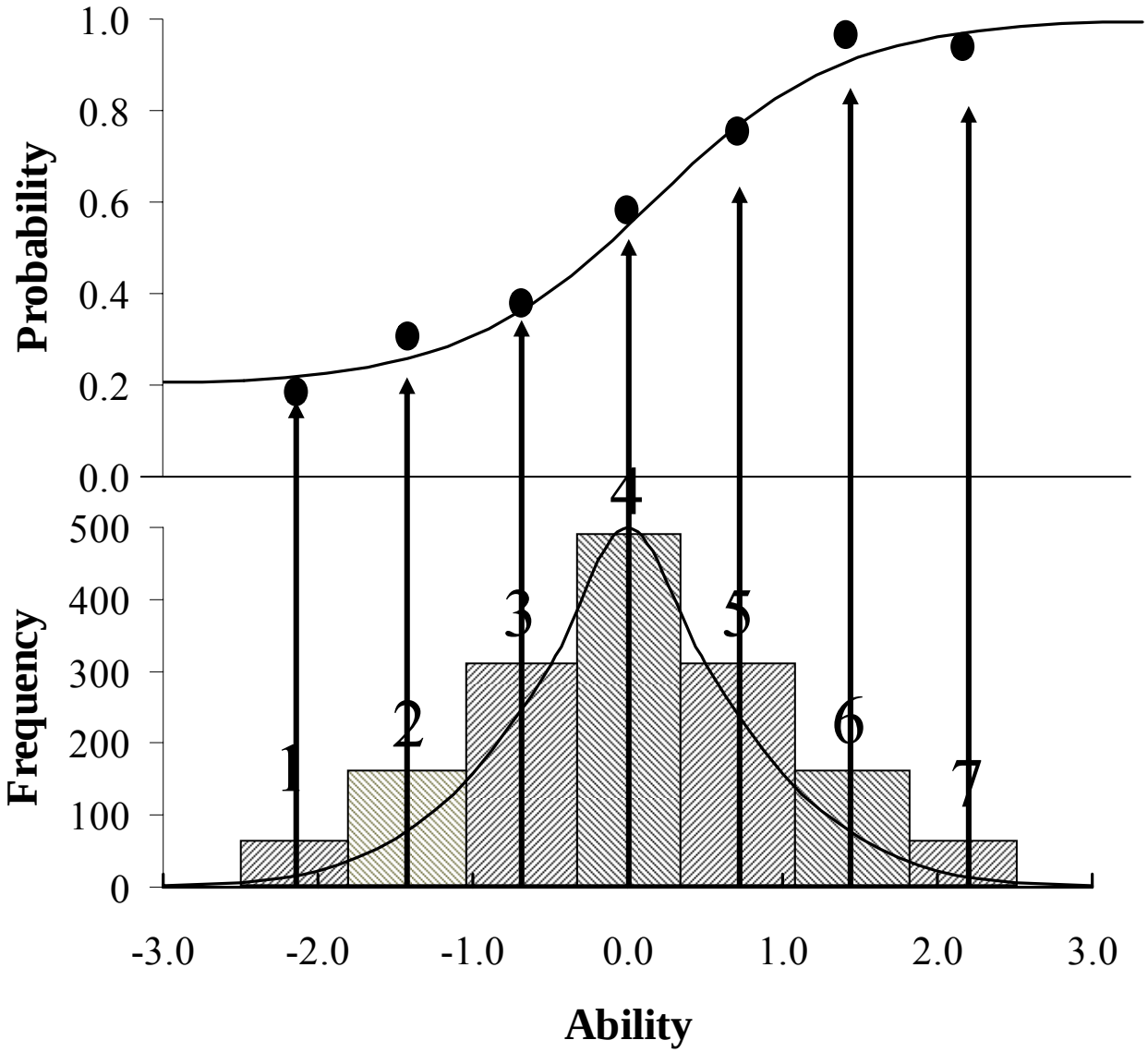
Many Choices of IRT Models

- One (Rasch)-, Two-, and Three-Parameter Logistic and Normal Ogive Models (0-1 Data)
- Partial Credit, Generalized Partial Credit, Graded Response, Nominal Response Models (polytomous response data)
- Multidimensional logistic model (0-1 Data)

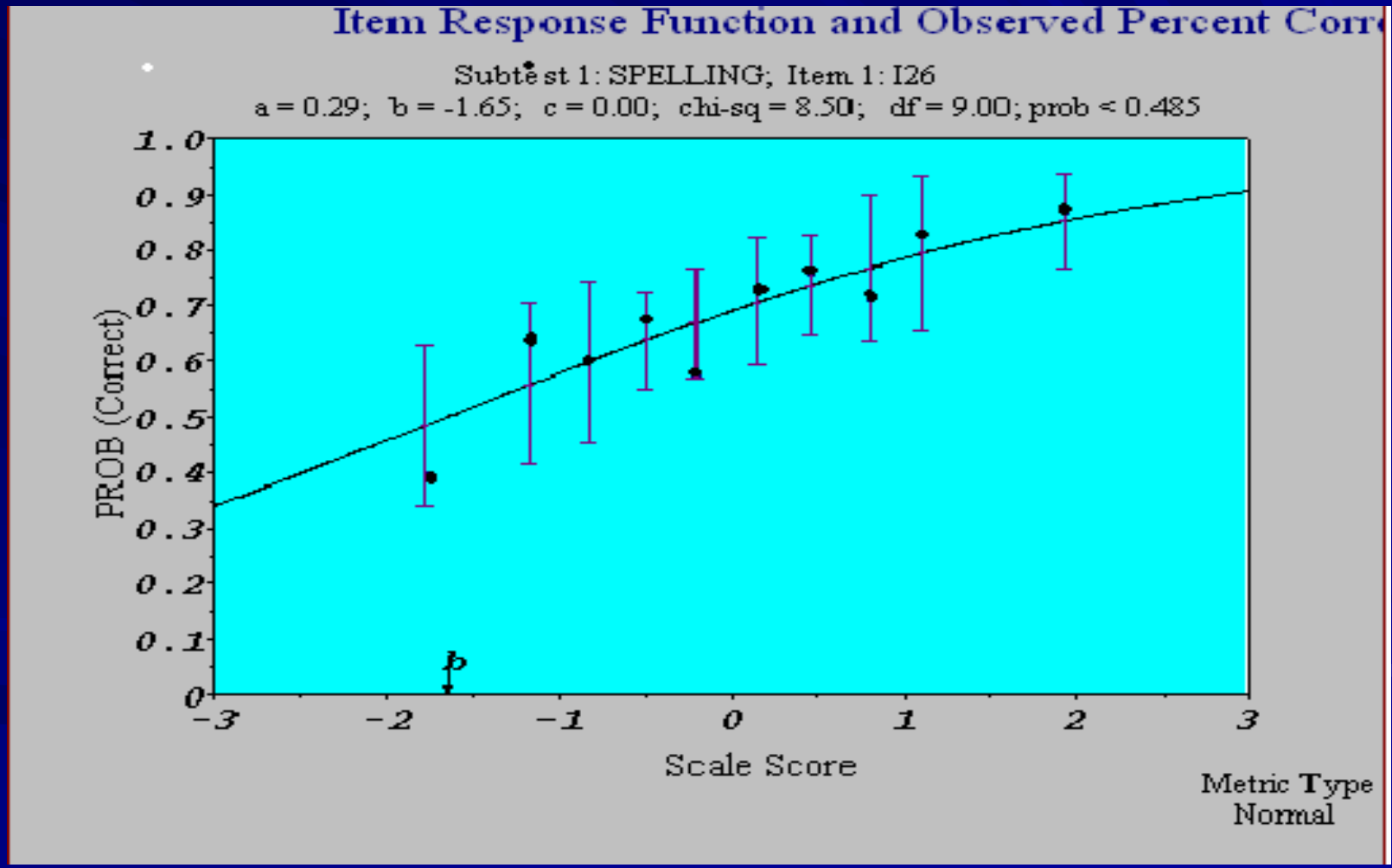
Estimation, Model Fit, Software

- Estimation: Marginal maximum likelihood estimation, Bayesian, and more.
- Software: Bilog-MG, Parscale, Multilog, ConQuest, Winsteps, etc.
- Model Fit: Some statistical tests; but need to link misfit to intended use.

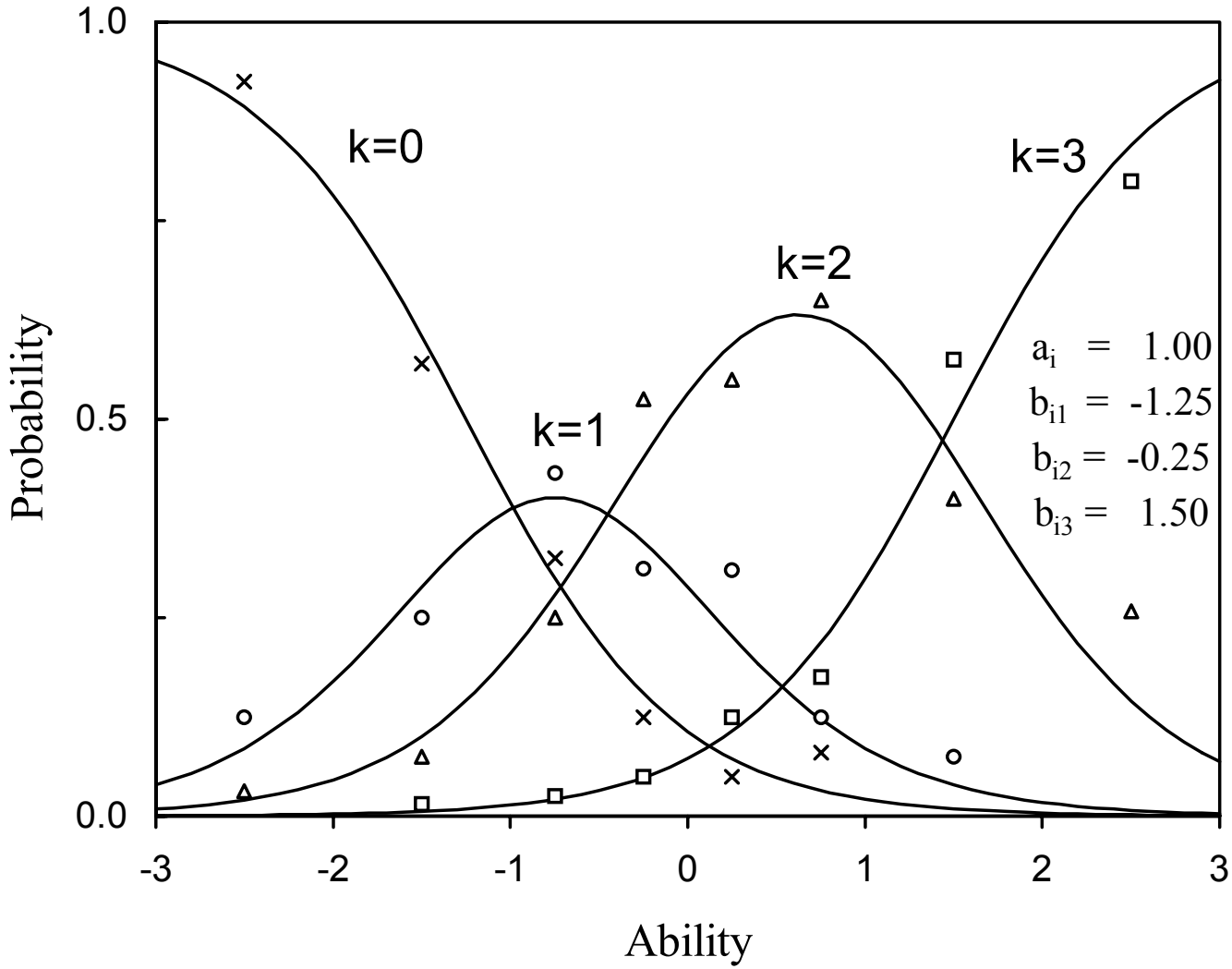
Assessing model fit



BILOG output



GRM Model-Data Fit:



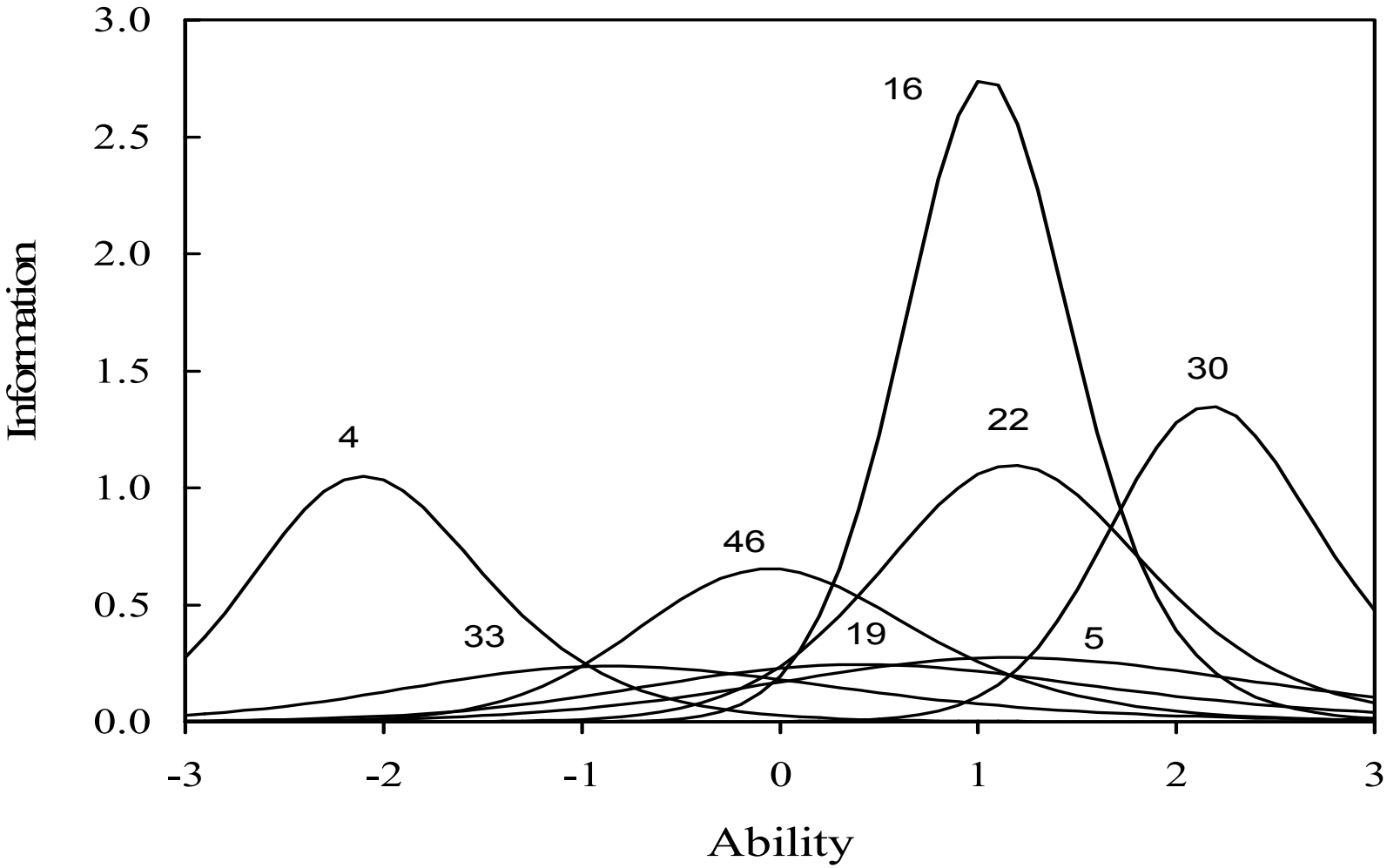
Applications of IRT to HRQOL

- Test Development
- DIF Analysis
- Test Score Linking or Equating
- Computer-Adaptive Testing
- Score Reporting

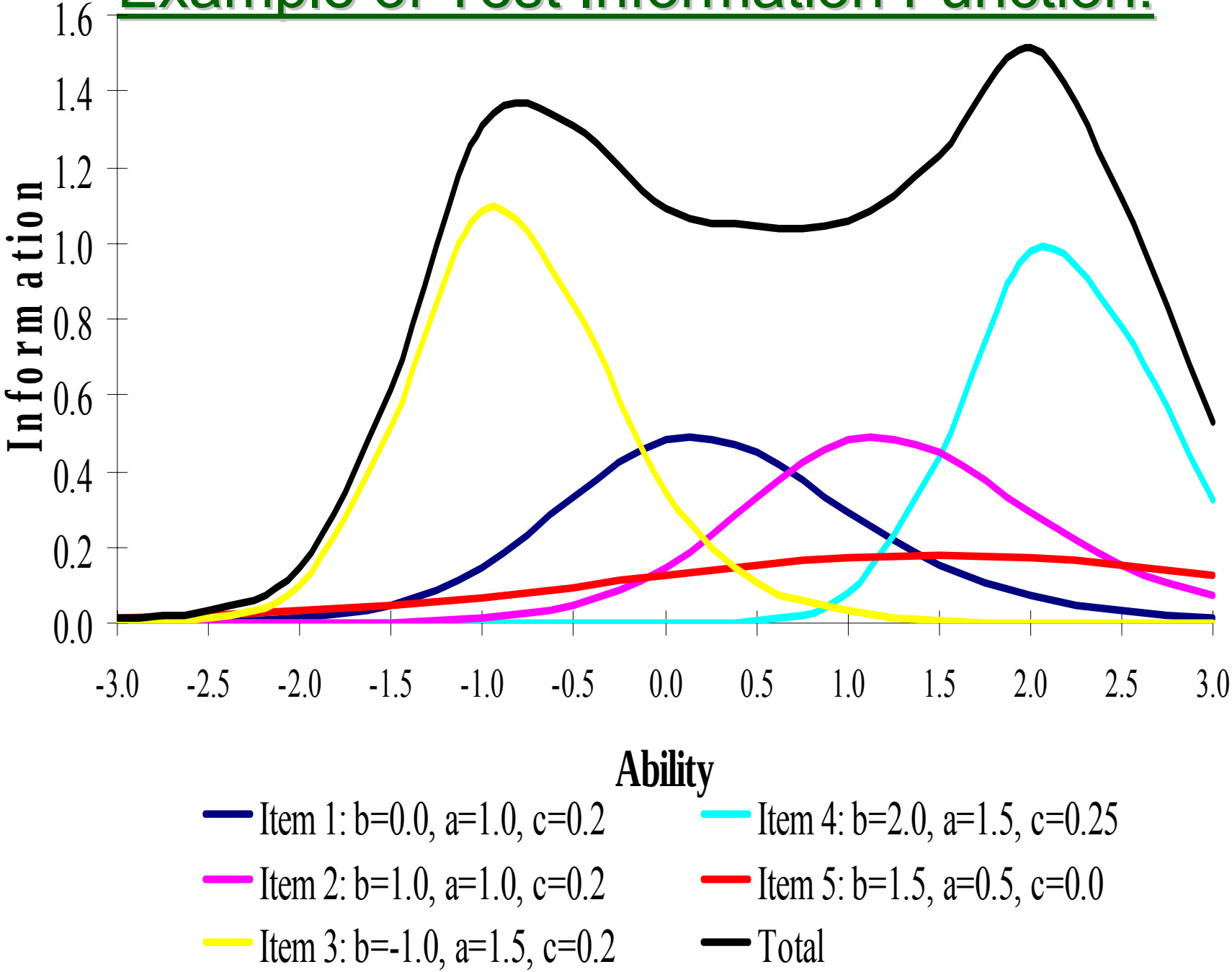
Test Development

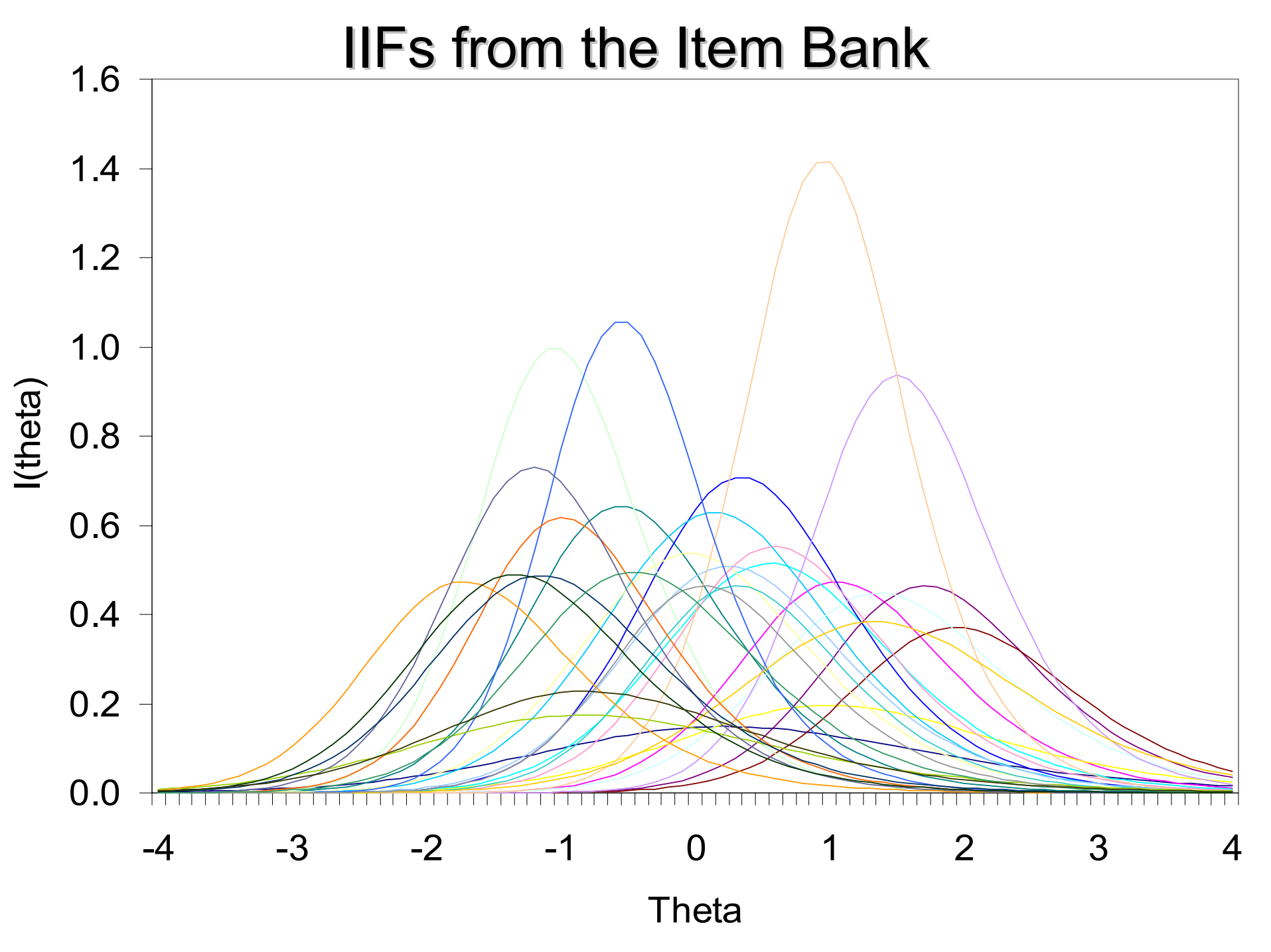
- Item and test information are invaluable in optimally constructing instruments. [These concepts not available in CTT.]
- Test information can either be the target for item selection, or the result. Inversely related to measurement precision.

Examples of Item Information Functions:



Example of Test Information Function:




IIFs from the Item Bank
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
I(theta)
-4
-3
-2
-1
0
1
2
3
4
Theta

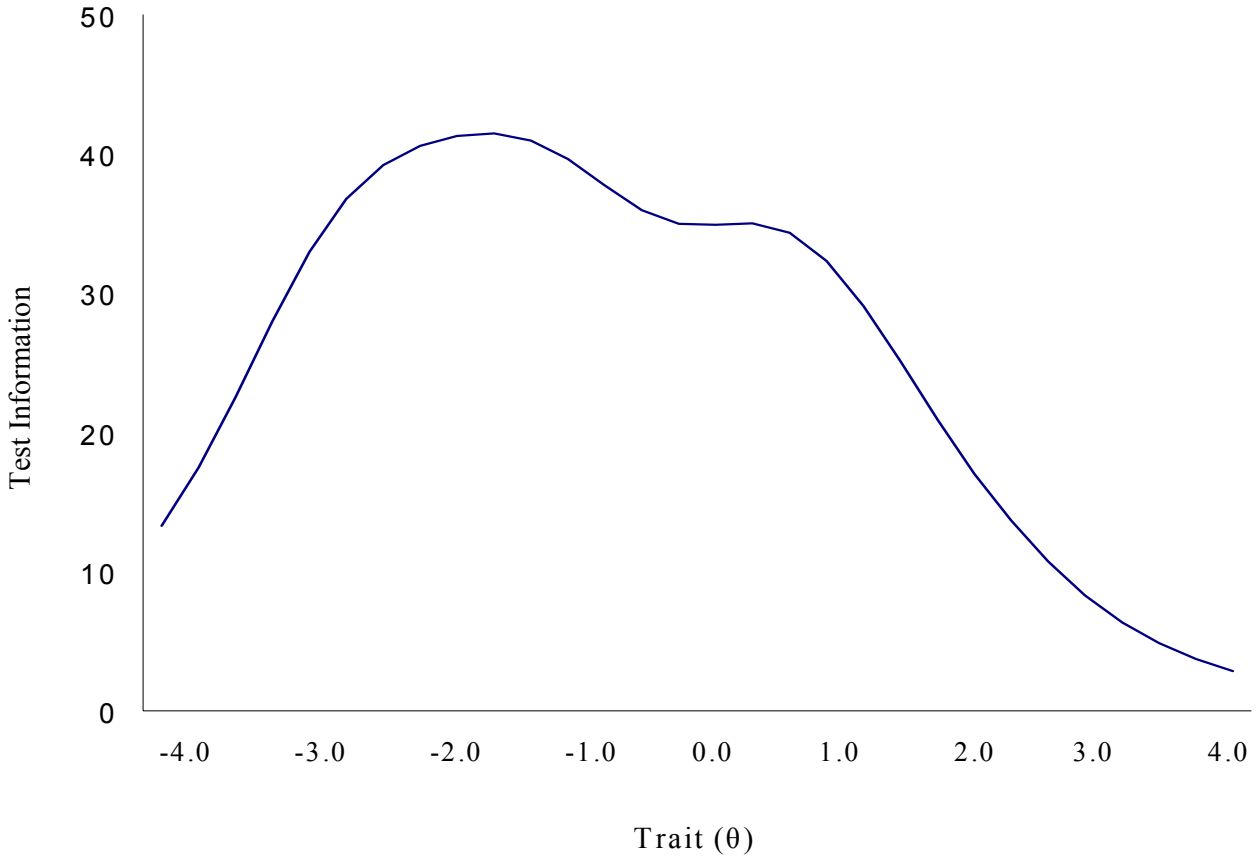
Item Information Function for 3-Parameter Logistic Model:

$$I_i(\theta) = \frac{\left[P'_i(\theta)\right]^2}{P_i(\theta)\left[1 - P_i(\theta)\right]}$$

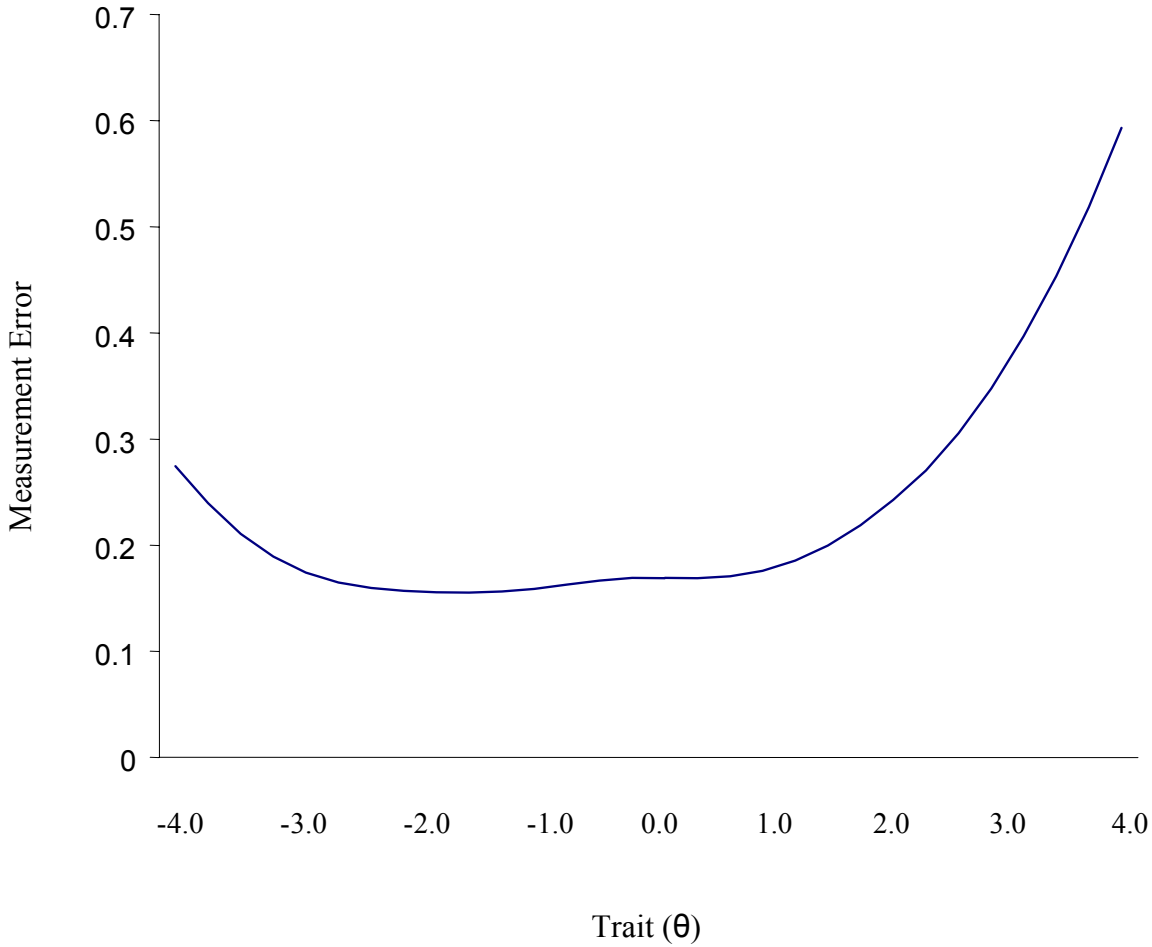
Or (the computational formula):

$$I_i(\theta) = \frac{2.89a_i^2(1 - c_i)}{\left[c_i + e^{1.7a_i(\theta - b_i)}\right]\left[c_i + e^{-1.7a_i(\theta - b_i)}\right]^2}$$

Test Information Function for the 50-item Satisfaction With Medical Care Survey.



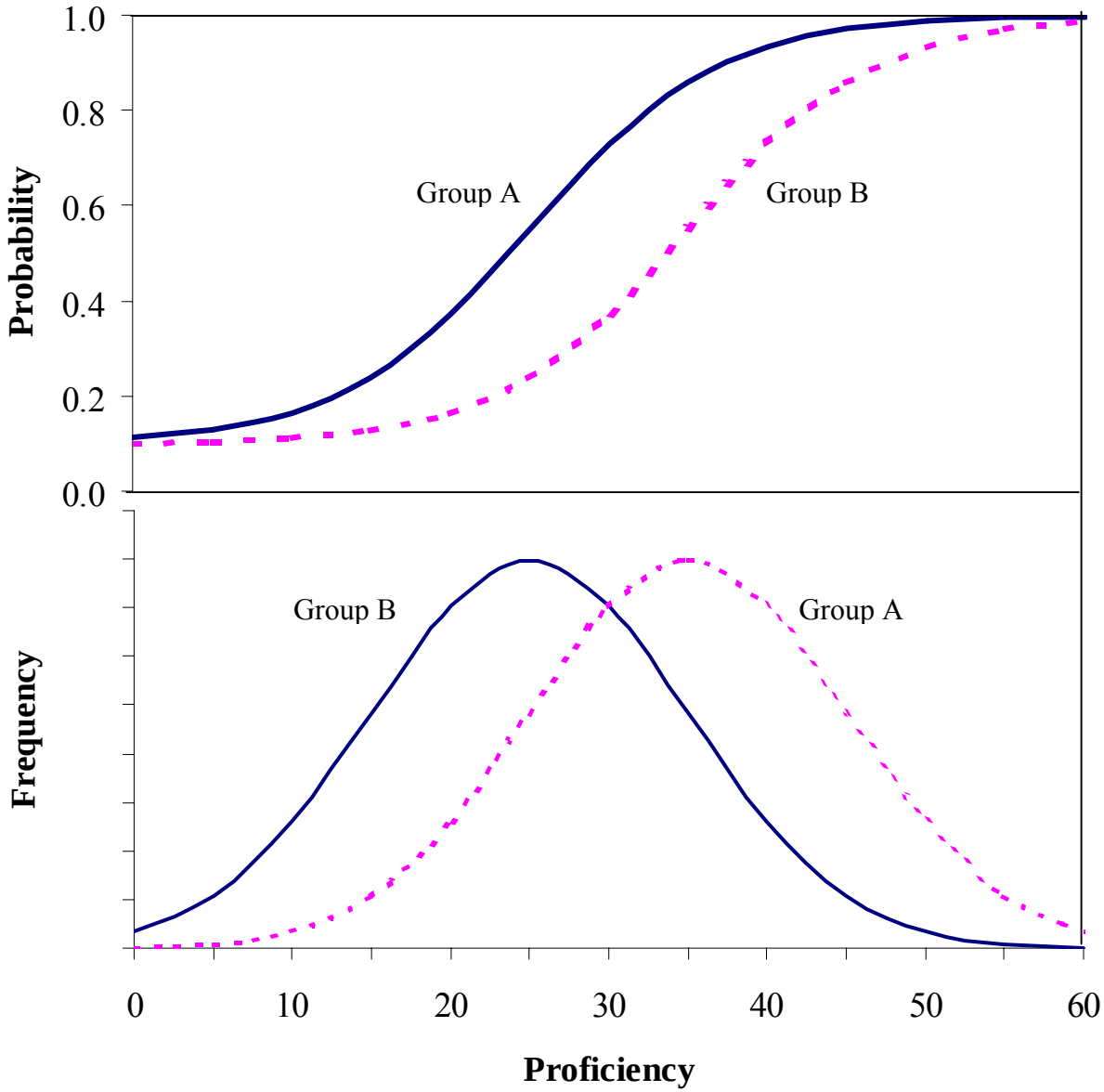
Measurement error at each trait score for the Satisfaction With Medical Care Survey.



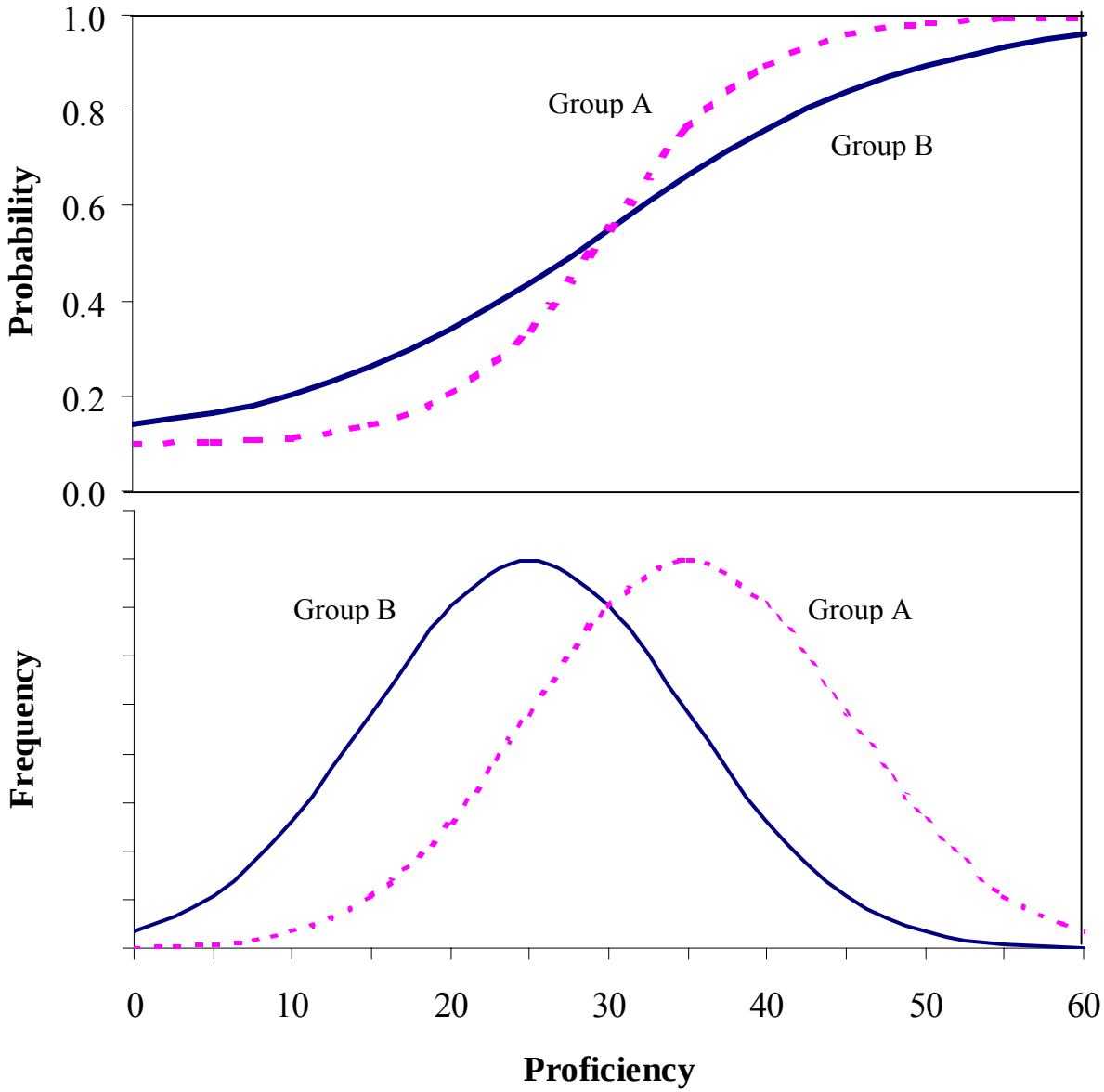
DIF Analysis

- Convenient approach
- Lots of options for loss functions (weights, no weights, sums, sums of squares, signed/unsigned)
- Male/Female, Black/White/Hispanic, One Language Version versus Another, Age Groups, etc.

An example of uniform DIF

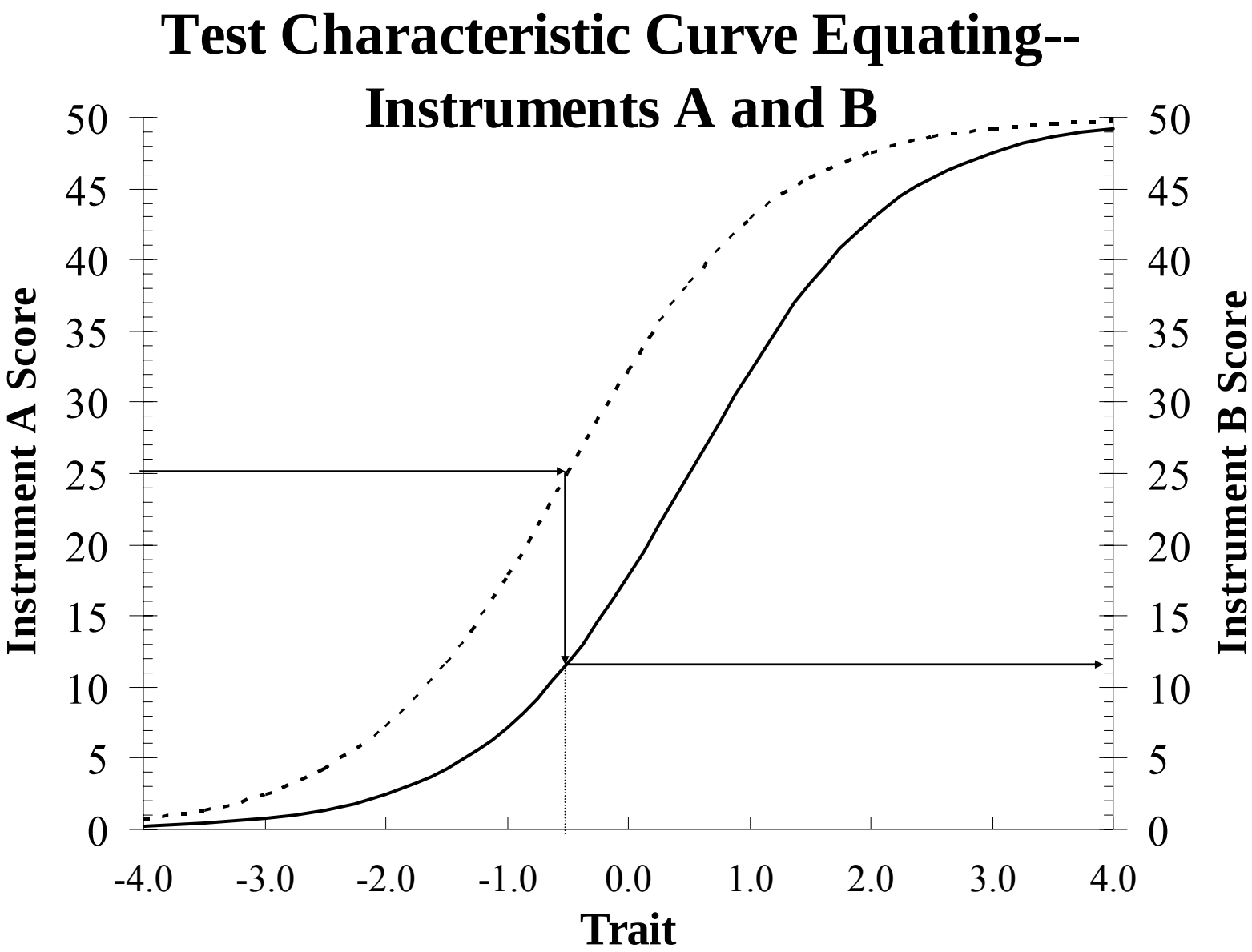


An example of nonuniform DIF



Test Score Linking or Equating

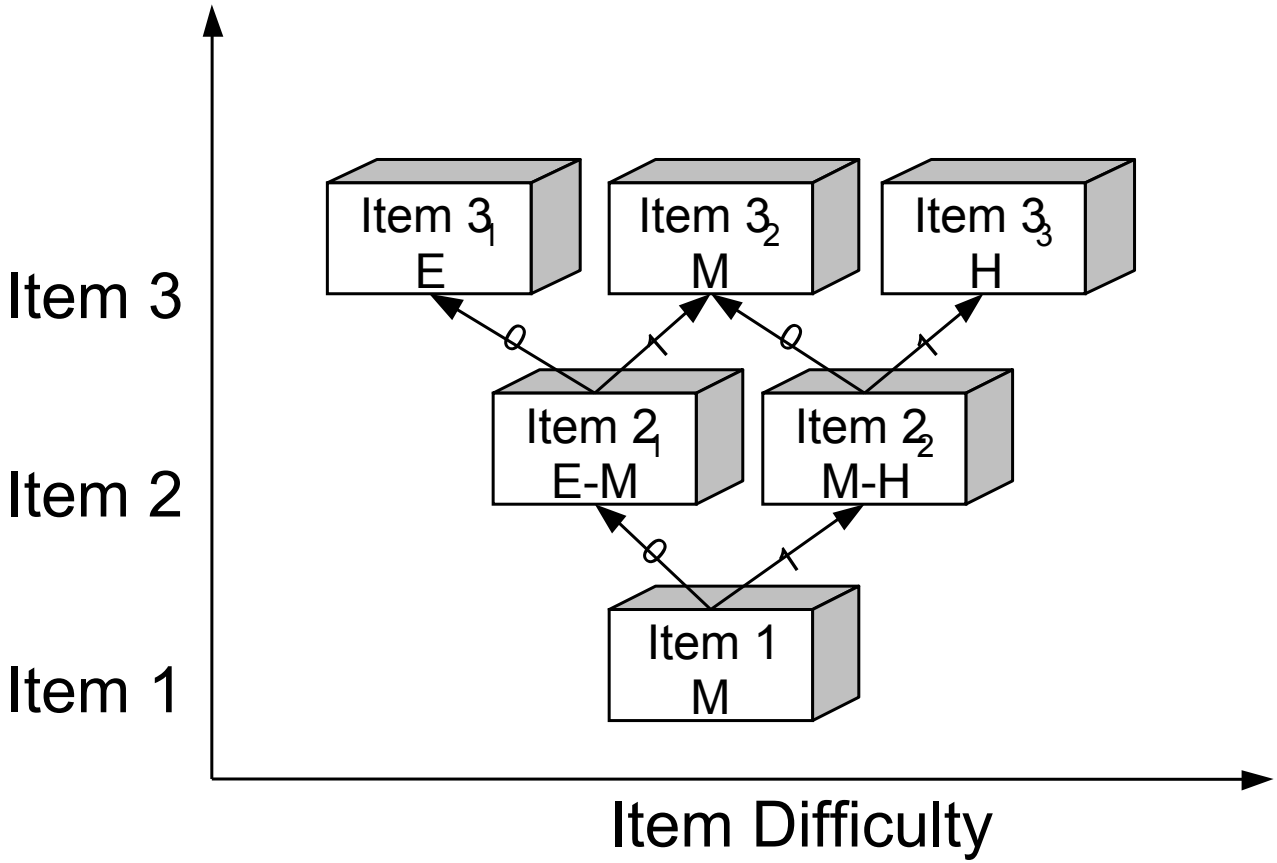
- Very convenient to link new items into a bank.
- Transformations of the data are linear and this is a big advantage
- Especially good approach if instruments differ in “difficult.”
- Common persons can be used to link scores on two instruments.


Test Characteristic Curve Equating--
Instruments A and B
Instrument A Score
Instrument B Score
Trait
50
45
40
35
30
25
20
15
10
5
0
-4.0
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
4.0

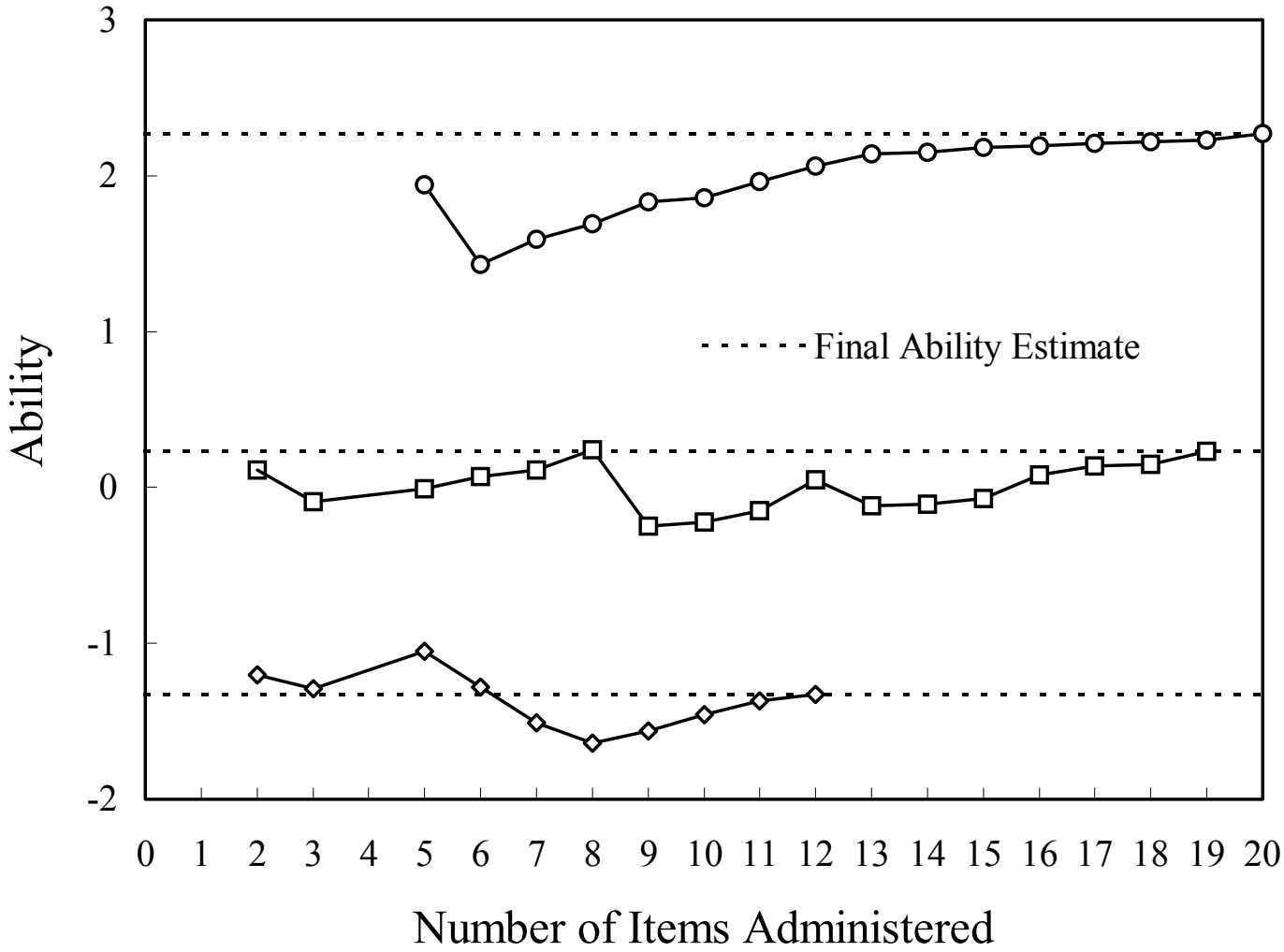
Computer-Adaptive Testing

- Reduce testing time by 50%
- In principle, a different instrument for each respondent
- IRT modeling, can place all trait estimates on a common scale for reporting and analysis.
- Measurement of change, if of interest, can be improved.

Example of a CAT

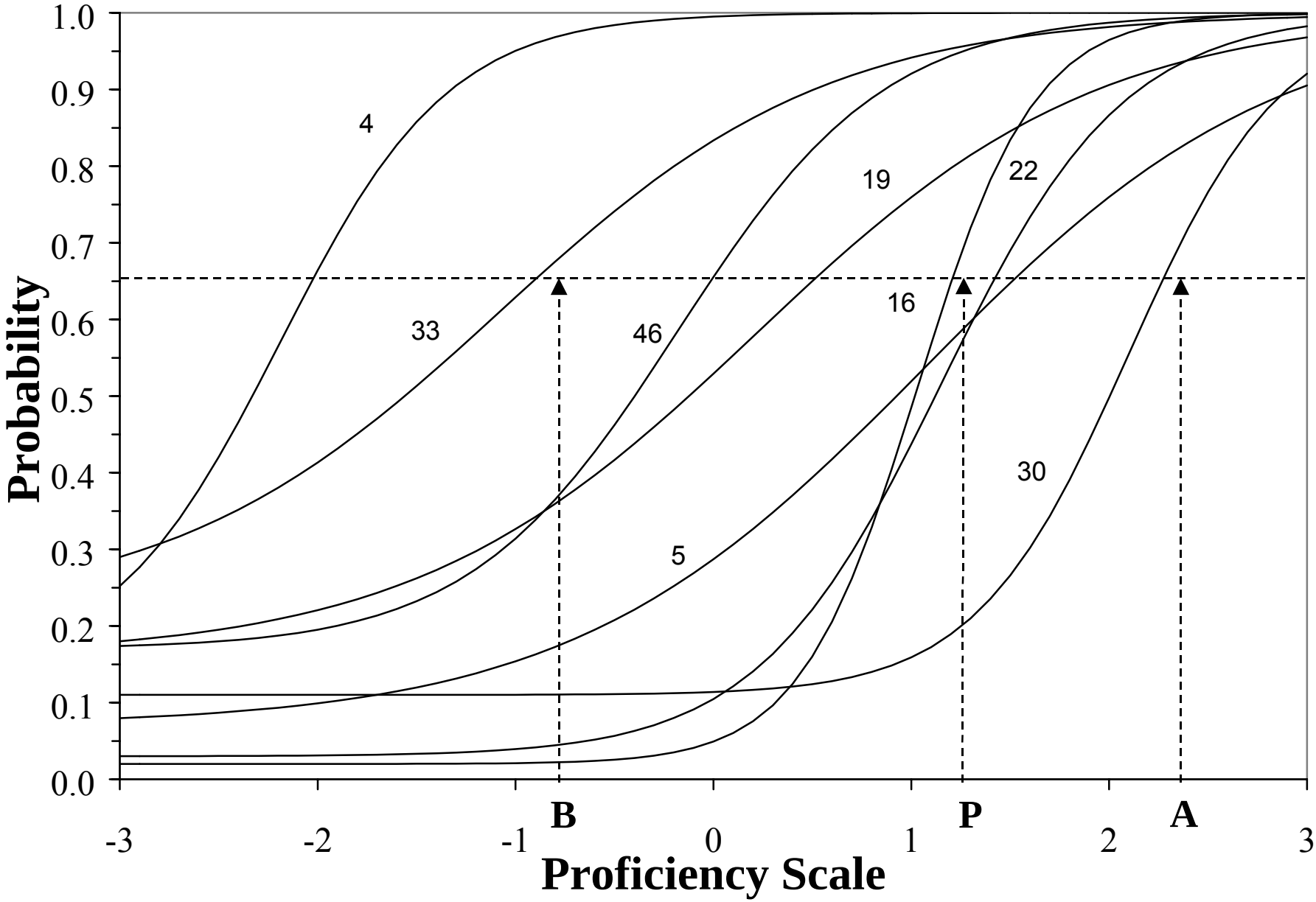


CAT Ability Estimates:



Score Reporting

- Can enhance the meaning of score scales by capitalizing on the fact that items and persons are being reported on the same scale.


Probability
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
4
19
22
33
46
16
30
5
B
P
A
-3
-2
-1
0
1
2
3
Proficiency Scale

Family of Rasch Models vs. Other IRT Models?

- More similar to each other, than IRT models to CTT models.
- Other IRT Models (e.g., 2p, 3p logistic models, graded response model) usually fit the data better. Still, advantage of improved fit needs to be demonstrated.

Family of Rasch Models vs. Other IRT Models? Cont.

- Rasch Model (and extensions) needs smaller respondent samples for effective item parameter estimation. Fewer complications in data analysis.
- Both Rasch and other IRT models can increase psychometric properties of instruments.

Unique Challenges in Health Outcomes Measurement

- Multidimensionality and CAT [could form unidimensional subscales, if necessary, or fit MIRT models.]
- Model fit is important.
- Amount of DIF/lack of item parameter invariance.

Concluding Remarks

- IRT is not a magic want to fix all the mistakes in instrument development!
- And, as will be clear as conference continues, IRT modeling can address many of the challenges of constructing reliable and valid instrumentation.

Concluding Remarks, Cont.

- What's needed next are solid research studies that (1) sort out the models from one another, and address fit, (2) address the handling of multidimensionality in the instruments, if it exists, and (3) provide practical experiences.
- Expect CAT and IRT will be a powerful combination.

- Please contact Ronald Hambleton at rkh@educ.umass.edu for more references.

The Graded Response Model

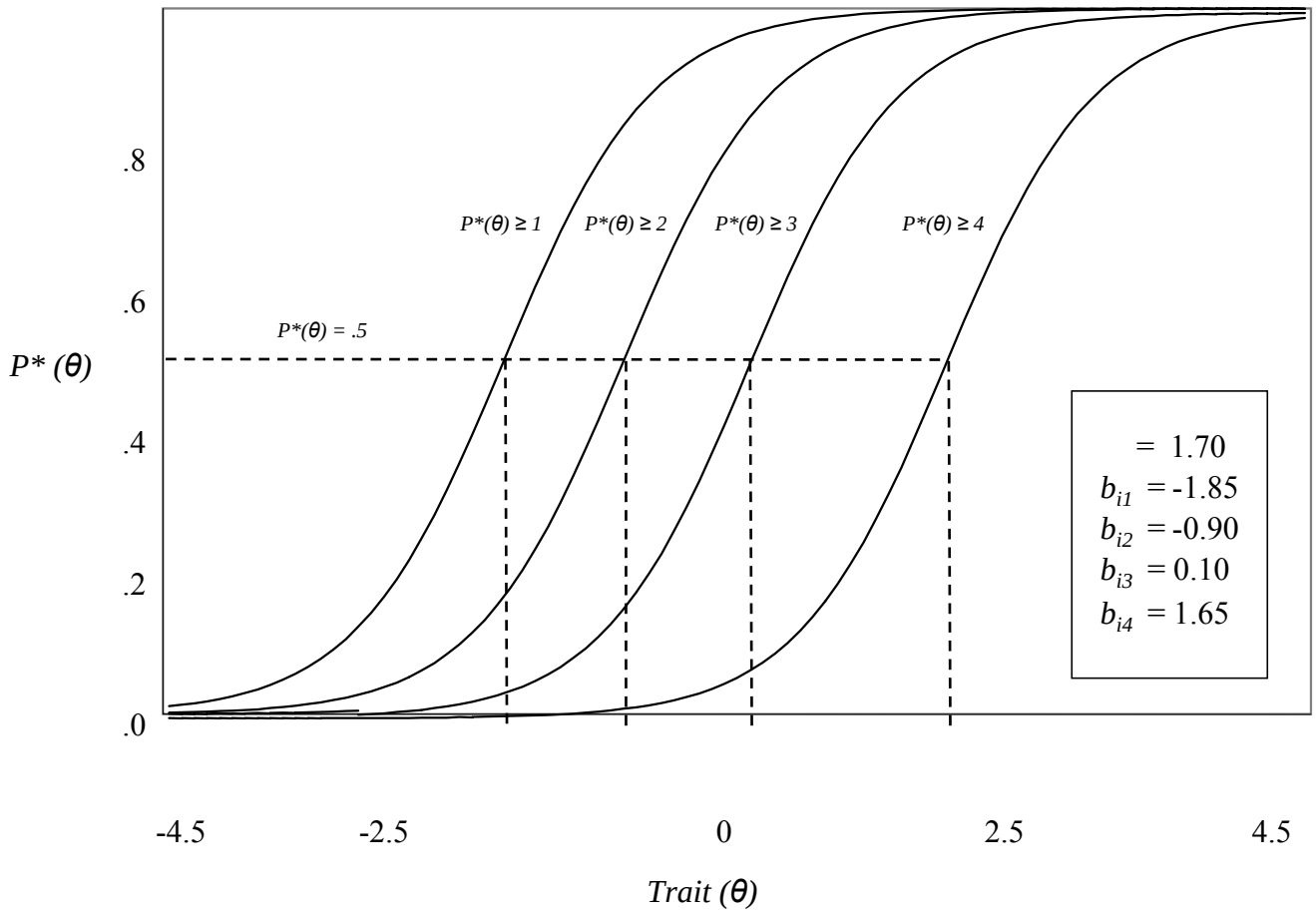
$$P_{ix}^{*}(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}}$$

where $i = 1,2,\ldots,n$ and $x = 0,1,\ldots, m_i$

The probability of a person obtaining a rating of *x* under the Graded Response Model.

$$P_{ix}(\theta) = P^{*}_{ix}(\theta) - P^{*}_{i(x+1)}(\theta)$$

Cumulative score category functions for the graded response model fitted to a five-category item (scored 0 to 4).



Score category functions for the graded response model

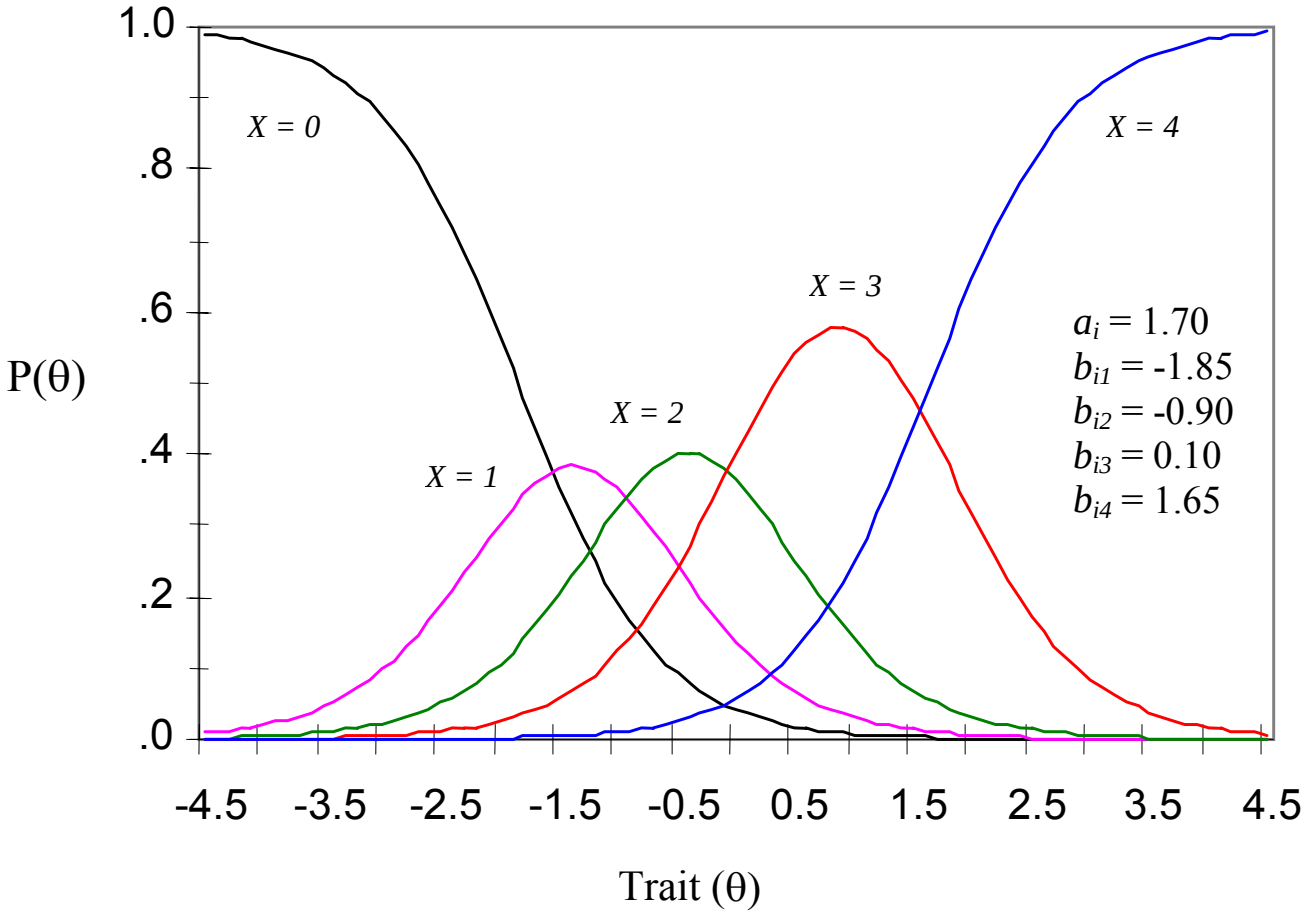


Figure 20. Comparison of male and female thresholds (50 items, 200 thresholds, female scale mapped to the male scale)