

Applications of Item Response Theory to Improve Health Outcomes Assessment

Ronald K. Hambleton

University of Massachusetts, Amherst

Family and Health Services Conference,
University of Massachusetts, October 15, 2006

Introduction

- Methods for developing educational and psychological tests and surveys have been receiving considerable attention.
- Stimulated by (1) federal legislation (NCLB) and educational reform, (2) considerable scrutiny of tests today, (3) computer technology, (4) statistical advances (e.g., SEM) and (5) great interest and research funds to measure new variables in the health sciences such as quality of life.

Presentation: Five Related Topics

- Background/Shortcomings of CTT
- IRT Models, Assumptions, Features
- Technical Matters (Estimation, Model Fit)
- Applications to Health Science Data (Model Fit, Score Reporting, Test Development, Bias, and CAT)
- Next Steps in Research and Final Remarks

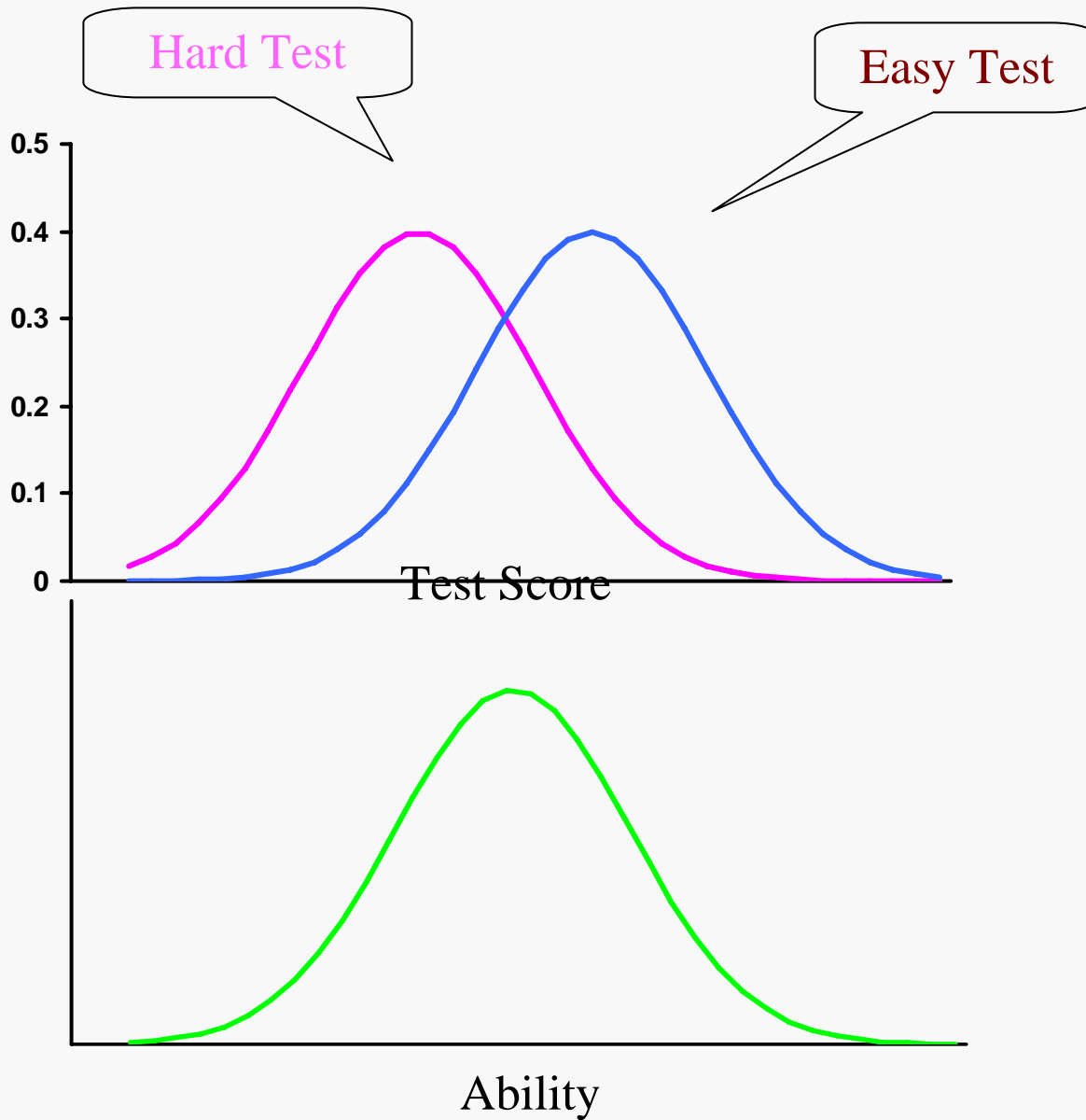
Classical Test Theory

- Influential for 80 years.
- Standard references by Gulliksen (1950) and Lord and Novick (1968).
- Important results--test length and reliability and validity, range restriction formulas, attenuation formulas, estimation of measurement error, etc.

CTT in Practice

- CTT has served the field of instrument development well. Excellent tests **can and have** been produced over the years.
- But, there are CTT shortcomings with the need today for item banks, computer-adaptive tests, improved score reporting (a big new challenge), and more.

Test Score and Ability Distributions for a Group



Limitations of Classical Test Theory

- Descriptors of examinee traits or abilities are item dependent.
- Descriptors of items are examinee sample dependent.
- Popular estimates of error are group-based (standard error of measurement).
- Modeling of data is at the test score level ($X=T+E$) and item level modeling is needed.

What Do Test Developers Want?

- Examinee parameter invariance
- Item parameter invariance
- Estimate of error for each examinee score
- Modeling examinee responses at the item level for flexibility in test item selection
- Examinees and items on a common reporting scale (optimal test design)

Item Response Theory (IRT) can address these five needs, when IRT models can be found to fit the test data.

What is item response theory (IRT)?

- A statistical theory linking examinee traits or abilities (what the test measures, and what is of interest to the test user) and examinee responses to the test items.
- Links between traits or abilities and item responses are made through non-linear statistical models that are based upon assumptions that can be checked for their adequacy with a set of data.

What is item response theory (IRT)?

- The statistical theory is general permitting (1) one or more traits or abilities, (2) various model assumptions, and (3) binary or polytomous response data.
- Over 100 IRT models in the testing field, but about 8 to 10 in wide use. (see van der Linden & Hambleton, 1997)

Applications of IRT in the US in 2006

1. achievement testing
2. selection testing (such as the SAT, GRE, GMAT, TOEFL, ASVAB)
3. psychological tests (e.g., W-J)
4. quality of life measures (great interest)
5. international assessments (such as TIMSS, OECD/PISA)
6. credentialing exams (100s of examples)

Two Important IRT Assumptions

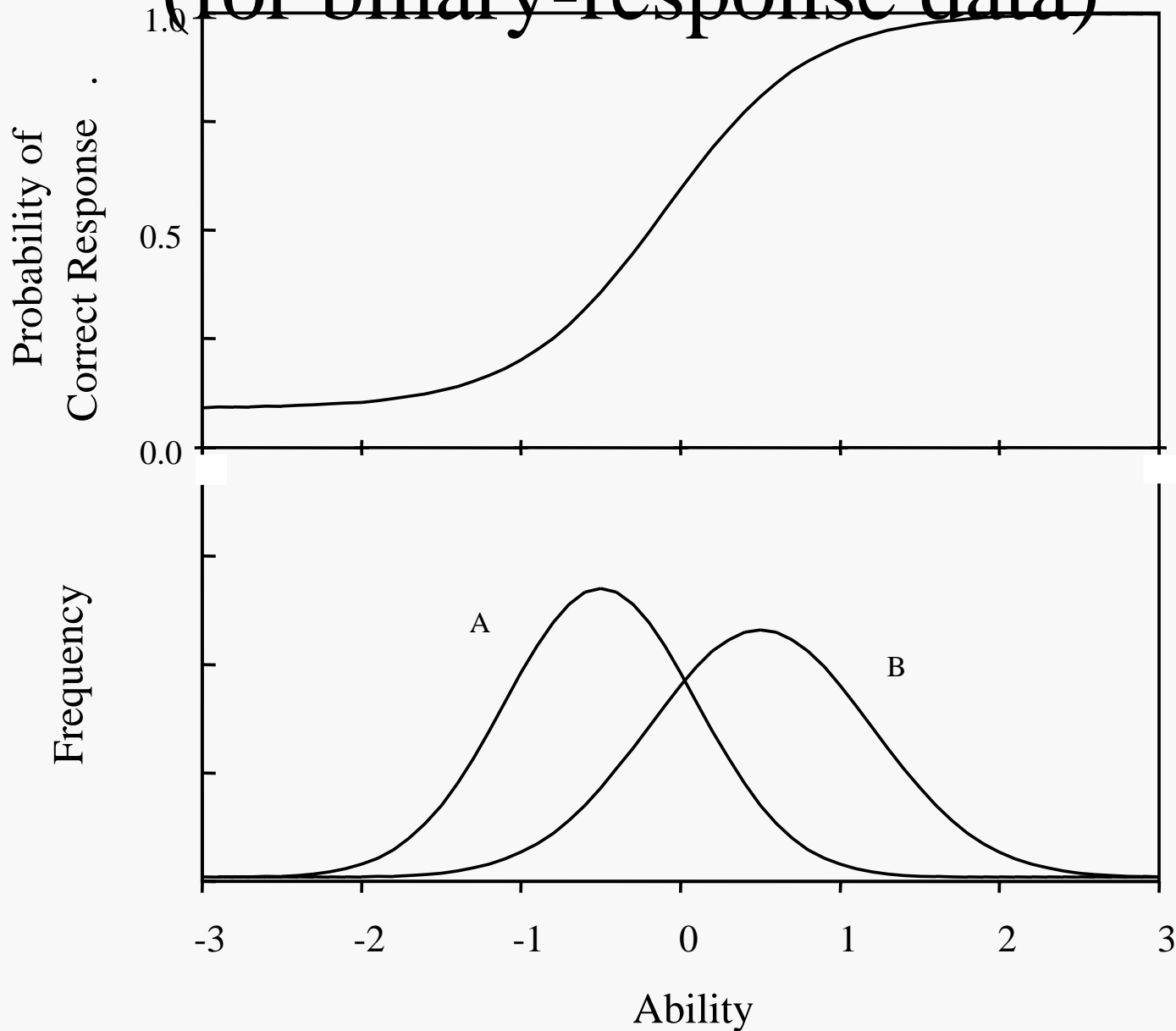
- **Unidimensionality of the Test (or equivalently, The Assumption of Local Independence)**
- **Shape of the Item Characteristic Curve (ICC) or Item Characteristic Function (ICF)**

Assumption of Unidimensionality

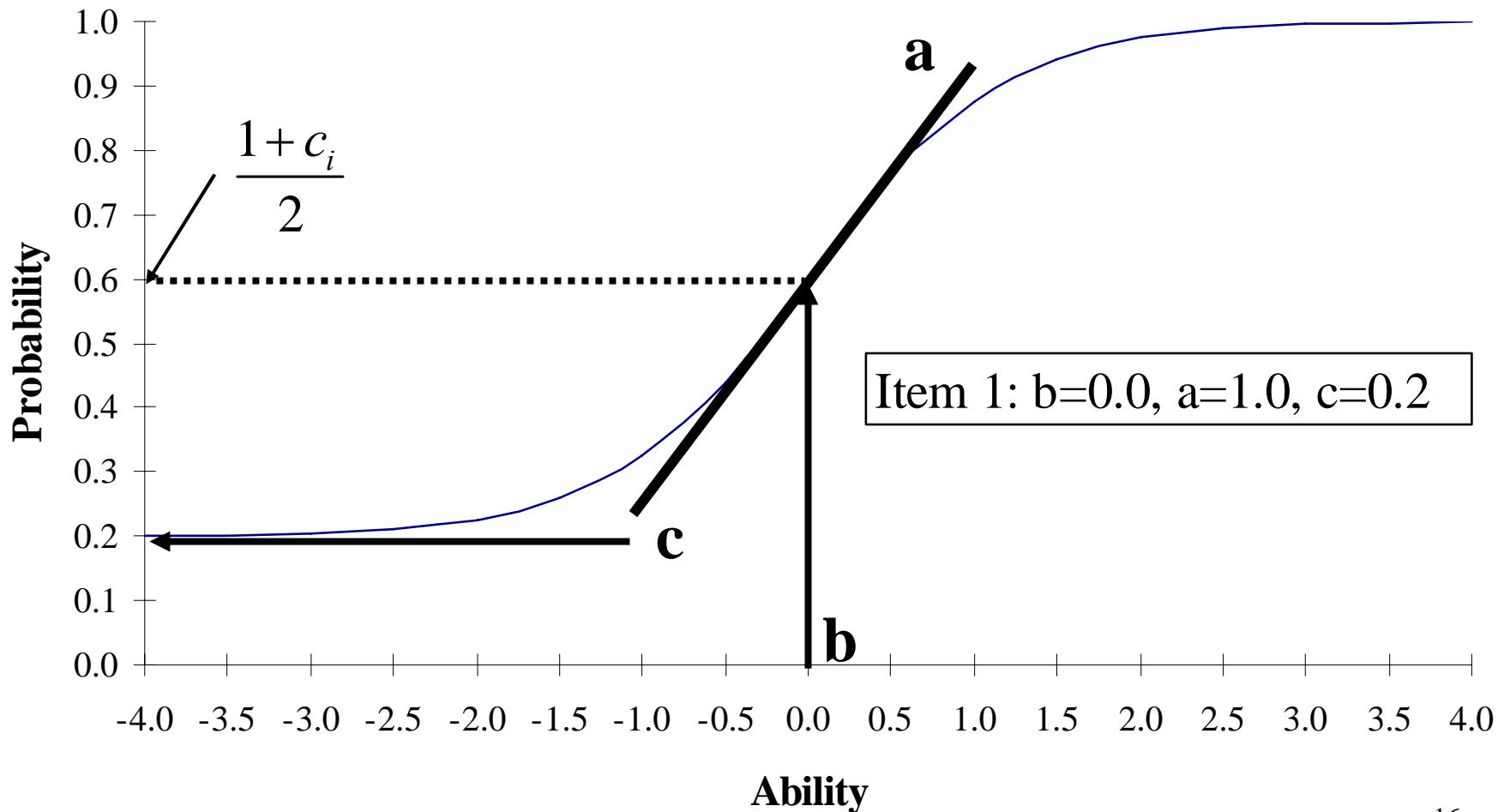
- The main idea is that there is a **single** dominant factor (such as math proficiency, attitudes about [X], physical functioning, quality of life, personality) that is being measured.
- A good check: Is it reasonable to report a single score to describe respondent performance?

Item characteristic Curve (ICC)

(for binary-response data)



Item Parameter Interpretations for the Three-Parameter Logistic Model



Three-Parameter Logistic Model:

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

D is a scaling factor set to a value of 1.7

Two-Parameter Logistic Model:

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

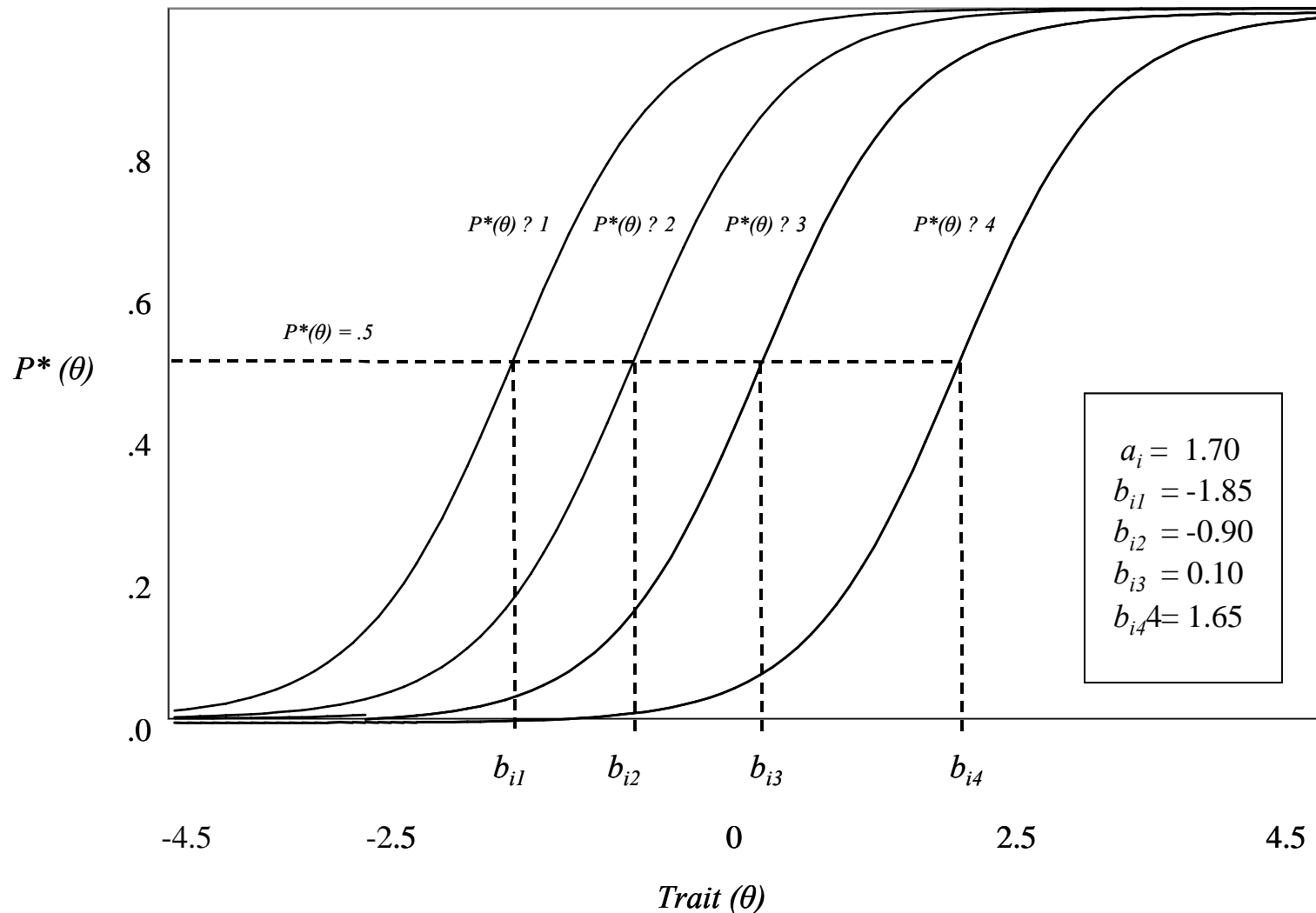
One-Parameter Logistic Model (Rasch Model):

$$P(u_i = 1 | \theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

Big Development in the 1980s: Applications of IRT Response Models to Polytomous Data

- Performance assessments with new item types and polytomous scoring- required new IRT models (polytomous response models).
- Polytomous response already present with many **psychological tests, attitude scales, quality of life instruments, etc.**

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



The Graded Response Model

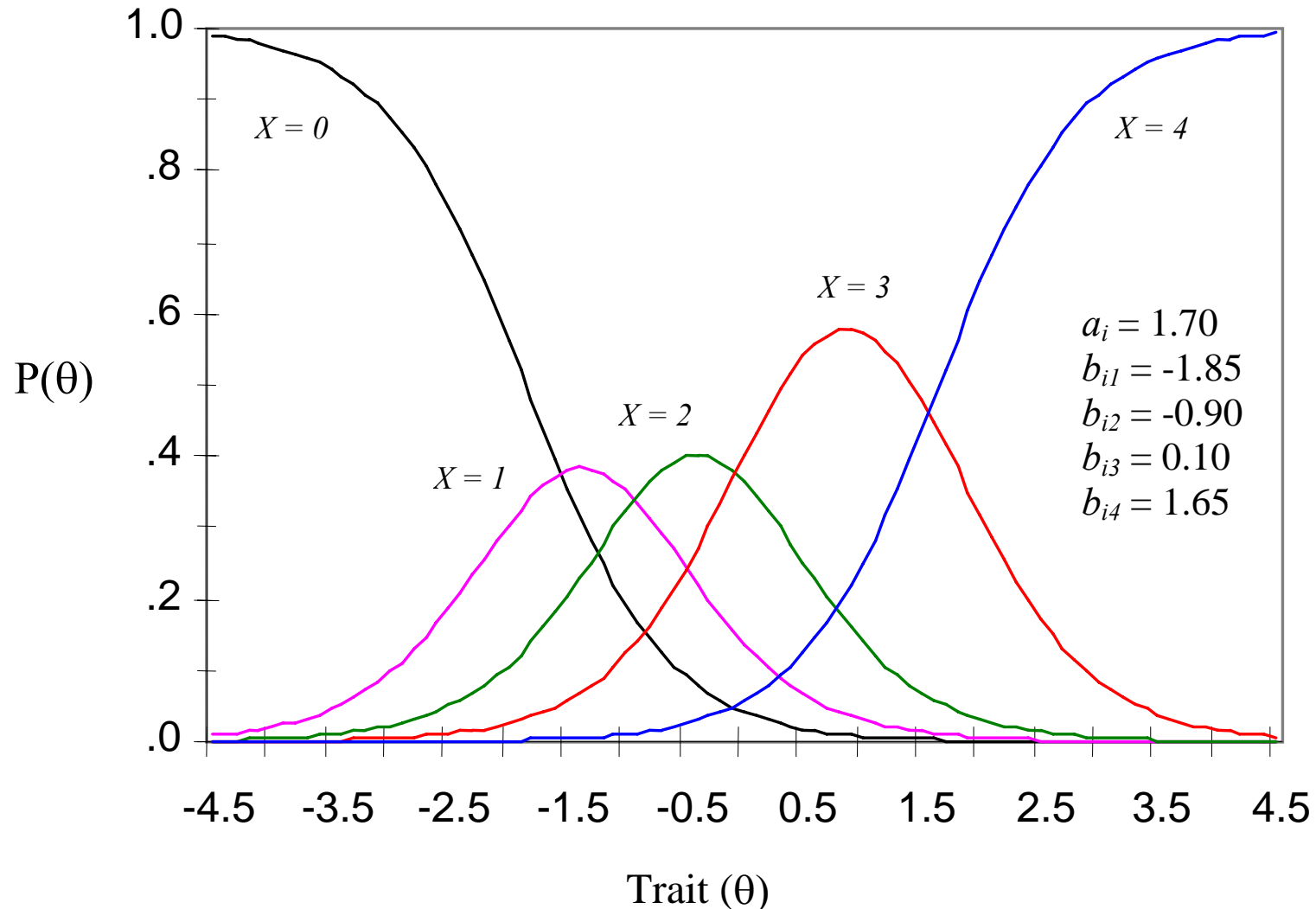
$$P_{ix}^* (\theta) = \frac{e^{Da_i (\theta - b_{ix})}}{1 + e^{Da_i (\theta - b_{ix})}}$$

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

The probability of a respondent giving a rating of x under the graded response model.

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

Score category functions for the GRM fitted to a five-category item ($a_i = 1.70$, $b_i = -1.0$).



Graded Response Model:

$$P_{i0}^*(\theta) = 1.0$$

$$P_{ix}^*(\theta) = \frac{e^{Da_i(\theta - b_{ix})}}{1 + e^{Da_i(\theta - b_{ix})}}, \quad x = 0, 1, \dots, m_i$$

$$P_{i(m_i+1)}^*(\theta) = 0.0$$

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

Several other polytomous response models-to handle Likert scales, and semantic differential, frequency, and other rating scales:

- Partial Credit Model(Masters and Wright)
- Graded Response Model (Samejima)
- Generalized Partial Credit Model (Muraki)
- Nominal Response Model (Bock)

Advantages of IRT Modeling of Data

- *Item Parameters Independent of the Examinees*--useful in field-testing and item banking. Called “item parameter invariance.”
- *Trait or Ability Parameters Independent of the Particular Test Items Used in Estimation*--critically important in computer adaptive testing and multi-stage testing. Called “trait or ability parameter invariance.”
- *Items and Examinees on the Same Scale*--especially helpful in test design and score reporting. [see next slide]

Additional IRT Advantages

- *Estimate of Error for Each Ability Score.*
- *A Convenient Framework for Solving Many Instrument Development Problems (e.g., optimal test design, because items and ability reported on the same scale; bias because ICCs can be compared; CAT because adjustments for difficulty in tests can be made, and item selection is facilitated too).*

Item and Ability Parameter Estimation, Model Fit

- Time permits only brief introduction to model parameter estimation.
- Researcher starts with item response data for a group of examinees (examinee responses to the items in the test).
- A software package is needed, and a choice of one of the IRT models.

Matrix of Data: Items (0 or 1, or 0 to m_i , $i= 1, n$)

		1	2	3	n
		Examinees	1	2	0	1
2	2		3	4	2
3	4		3	3	4
⋮	⋮		⋮	⋮						⋮
⋮	⋮		⋮	⋮						⋮
⋮	⋮		⋮	⋮						⋮
⋮	⋮		⋮	⋮						⋮
⋮	⋮		⋮	⋮						⋮
N	3		2	3	4

Ability and Item Parameter Estimation Procedures

- Maximum Likelihood Estimation (MLE)
- Joint MLE
- Conditional MLE
- Marginal MLE*
- Bayesian
- Markov-Chain Monte Carlo (MC-MC)

IRT Software Availability

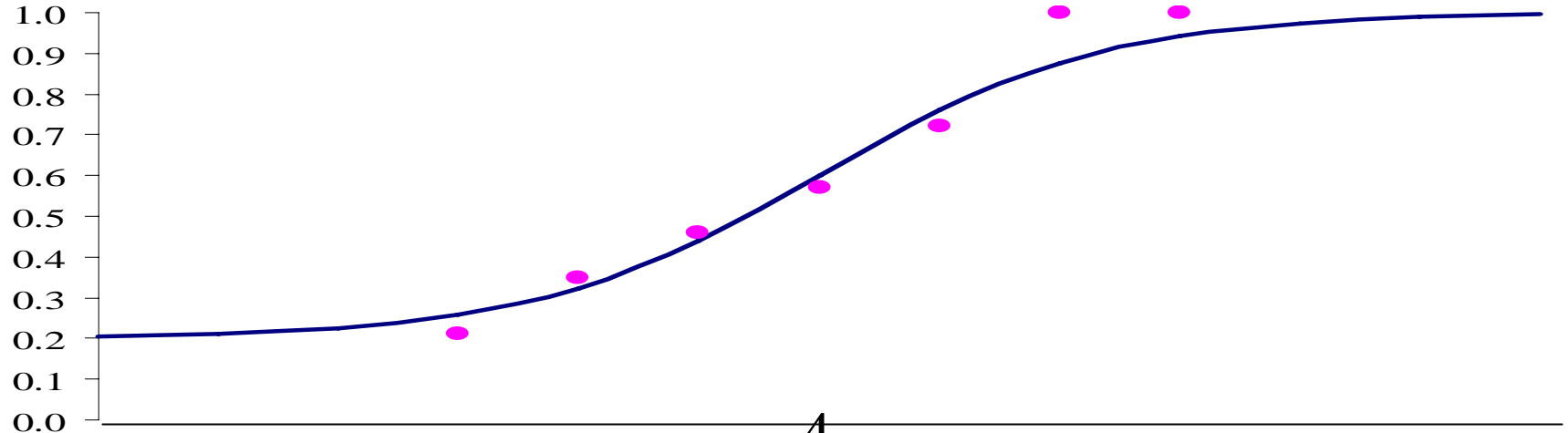
- <http://www.assess.com>
- <http://www.ssicentral.com>
- <http://www.winsteps.com>
- <http://www.scienceplus.nl>

IRT Model-Examinee Data Fit

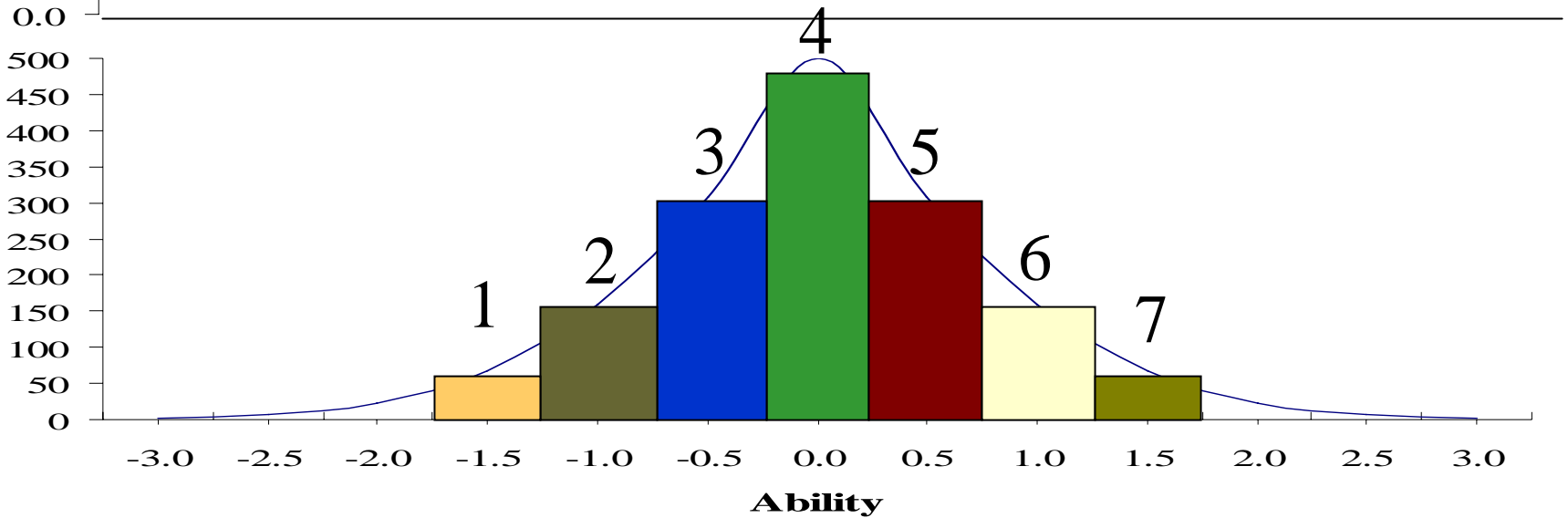
- Assess model assumptions such as unidimensionality
- Assess residuals and standardized residuals and examine consequences of model misfit (e.g., predicting score distributions)
- Check invariance properties (e.g., item bias)

Example of Fitting An ICC to Binary Data

Probability

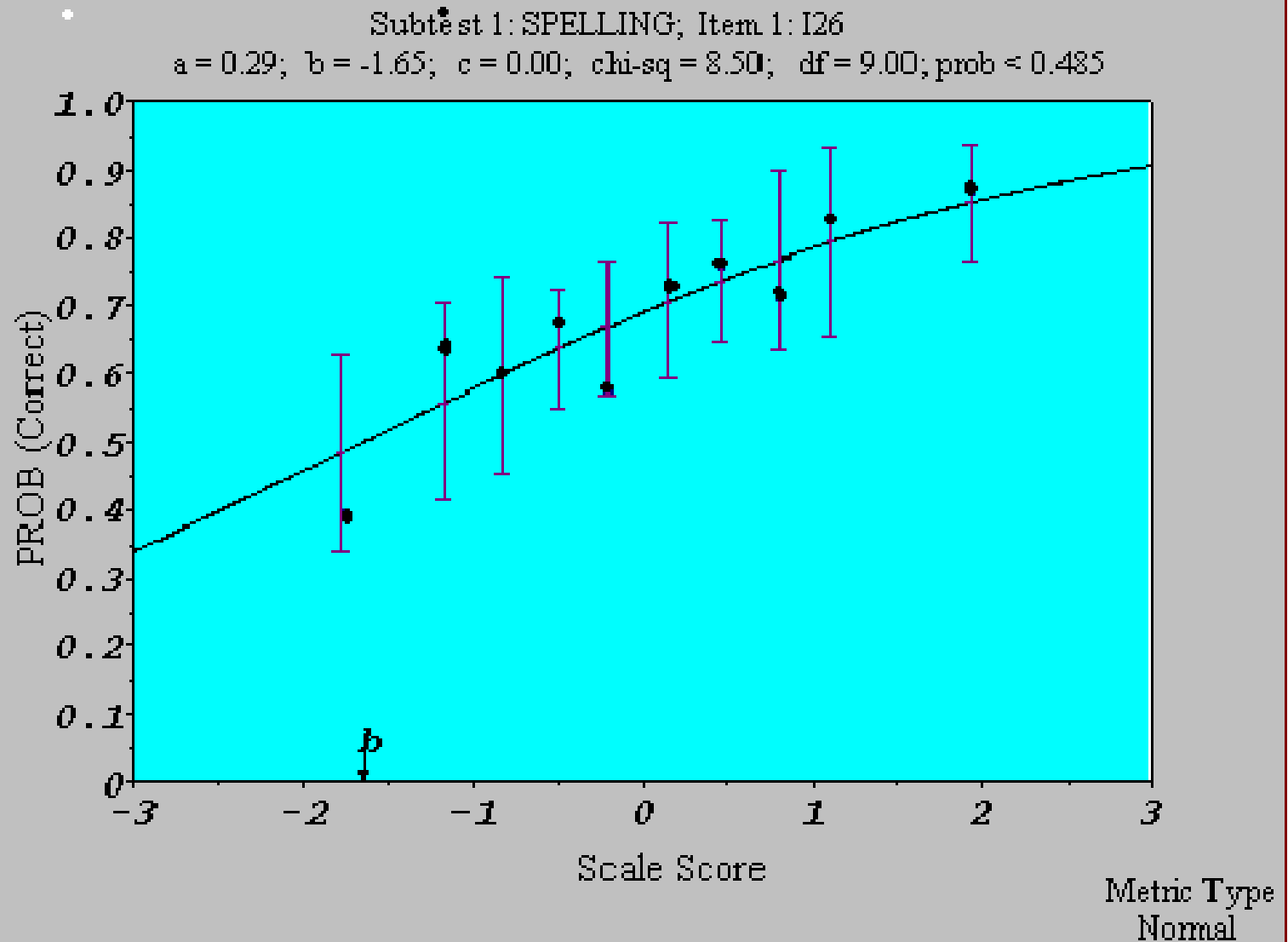


Frequency



BILOG-MG Output

Item Response Function and Observed Percent Correct

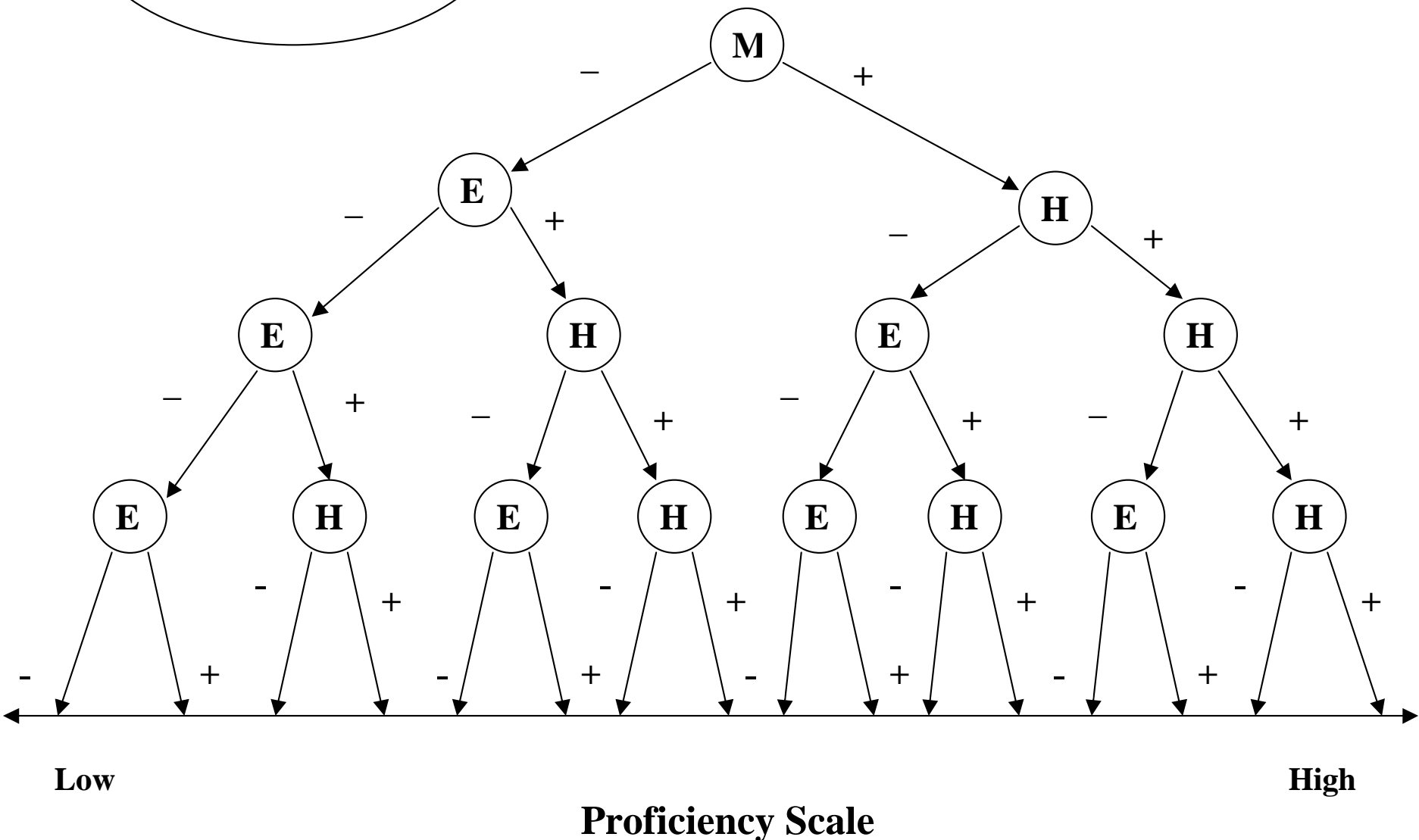


Applying the Graded Response IRT
Model to Health Sciences Polytomous
Response Data: An
Example [ConferenceExample.ppt](#)

Application of IRT Models to Computer-Adaptive Testing (CAT)

- In CAT, the items administered to an examinee are dependent on previous performance
 - an examinee doing well (or responding positively) sees harder questions, and if he/she is doing poorly (or responding negatively), the examinee sees easier questions.

Item Bank



Application of IRT Models to Computer-Adaptive Testing (CAT)

- Testing stops when (1) content specifications are met, (2) some minimum number of items have been administered, or (3) desired degree of measurement precision has been reached.

Why is IRT critical to CAT?

- Examinees see different test items, and tests will be of different difficulty and length. Test scores are not suitable then for comparing examinees, or examinees to cut-scores.
- IRT provides (1) scores, independent of difficulty of test items, (2) items and scores on a common scale (so optimal item selection is possible), and (3) a popular stopping rule-- measurement precision at the examinee level.

Next Steps in IRT Research

- Cognitively-Based IRT Models for Test Development and Analysis
- Automated Test Assembly
- Multidimensional IRT Modeling
- Model Suitability Based on Consequences of Misfit

Final Remarks

IRT is sufficiently well-developed that it provides a measurement theory for test development, scale construction, CAT, equating, study of bias (or DIF), and score reporting.

--But, IRT is **NOT** a magic wand to wave over vague test specifications and poorly written items or statements!!

Follow-Up Study

- Hambleton, Swaminathan, and Rogers (Sage, 1991). (Good Introduction)
- Van der Linden and Hambleton (1997) (Springer, Introduction to All Models)
- Wright and Stone (1979) (MESA, Rasch Model)
- Wainer et al. (2000) (Erlbaum, CAT)