Tutorial

Classical Test Theory
Item Response Theory
Computer Adaptive Testing
Automatic Item Generation
Automatic Essay Grading

Jozef Tvarožek

Faculty of Informatics and Information Technologies
Institute of Informatics and Software Engineering
Slovak University of Technology in Bratislava

Basics

- Test = collection of questions (items)
- Examinee = person taking the test
- Ability = examinee's level of attainment of a skill

CTT – Classical Test Theory

- Examinee and test characteristics not separable
  - Ability = true score (expected value of performance on test)
  - Item difficulty = proportion of examinees in a group of interest who answer the item correctly
  - Taking a “hard” test, examinee will appear to have low ability
  - Taking an “easy” test, examinee will appear to have higher ability

CTT – Classical Test Theory (2)

- **Item characteristics are group-dependent**
  - Preparing test for a “different” population is hard
- **Examinee scores are test-dependent**
  - Contain different amount of error
- **Reliability = correlation between test scores on parallel forms of test**
  - Parallel forms – do they exist?
- **Theory is test oriented, not item oriented**
  - No predictions can be made about item perf.

Requirements for a new theory

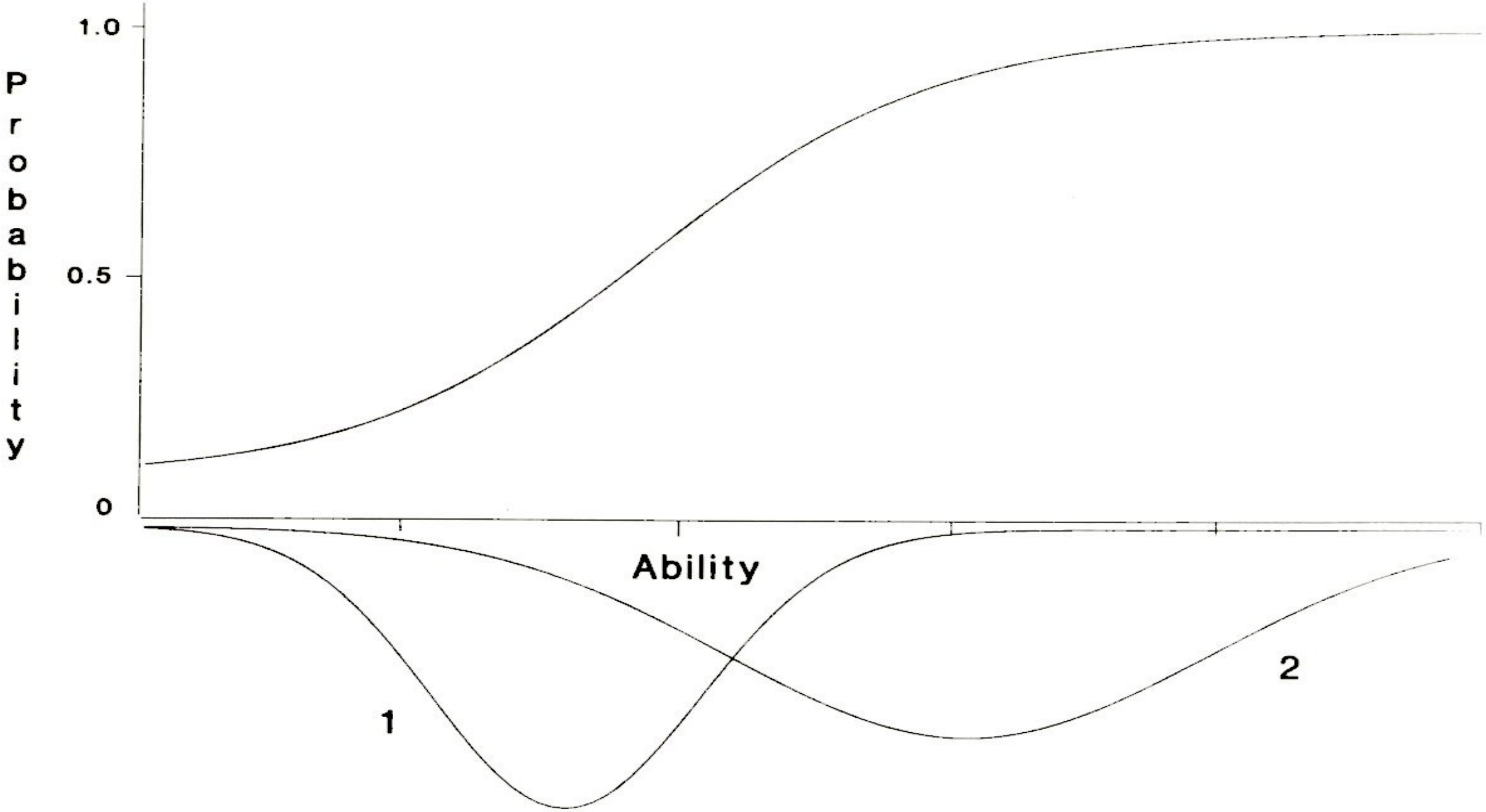
- Item characteristics that are not group-dependent
- Scores that are not test-dependent
- Model of items, not test
- Reliability not defined by parallel forms
- Measure of precision for each ability score

IRT – Item Response Theory

- Postulates:
  - Performance of an examinee can be predicted by a sef of factors (abilities)
  - Relationship between examinees' item performance can be described by monotonically increasing function (Item characteristic curve – ICC)
- IRT models are falsifiable
  - Need to assess the fit of the model to the data.

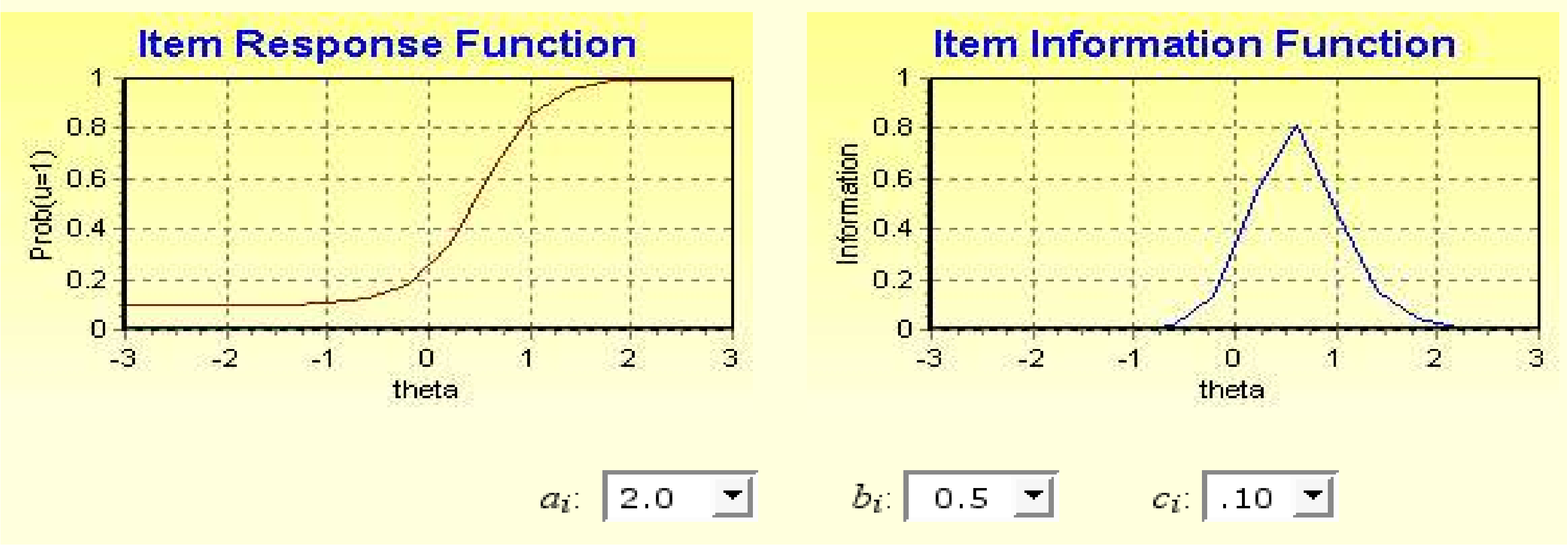
IRT – Item Response Theory (2)

- Item and ability parameters are **invariant**



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

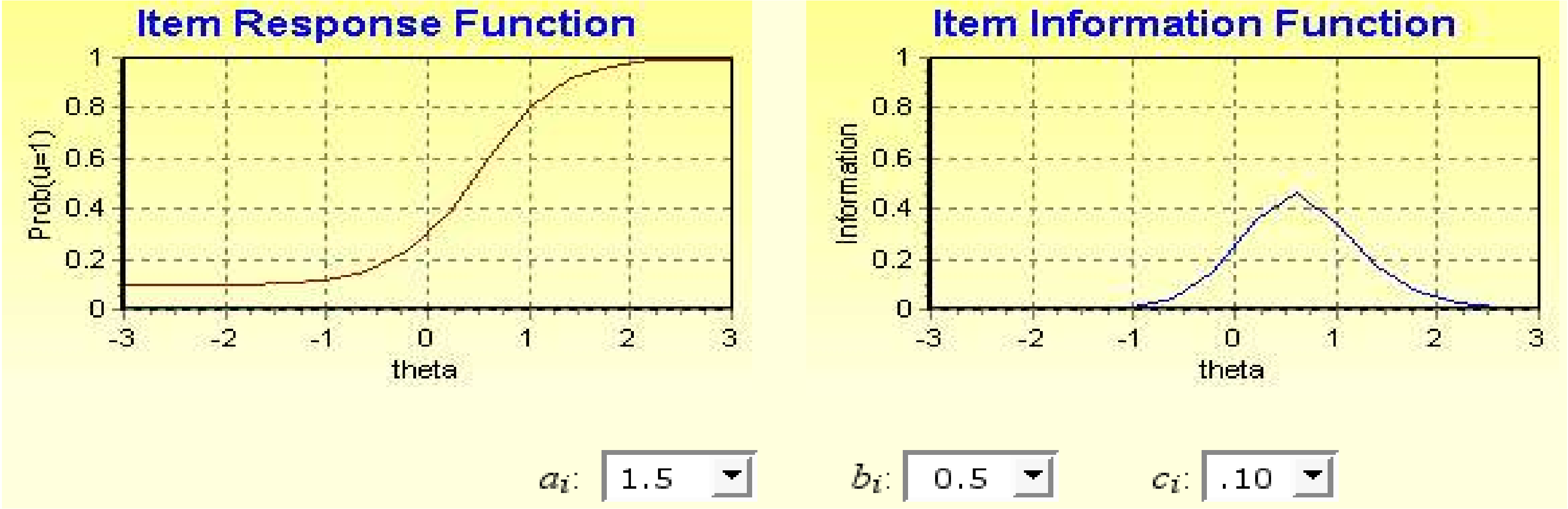
$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)\,(1 - P_i(\theta))}$$



Item Response Function

Item Information Function

a_i: 2.0 b_i: 0.5 c_i: .10

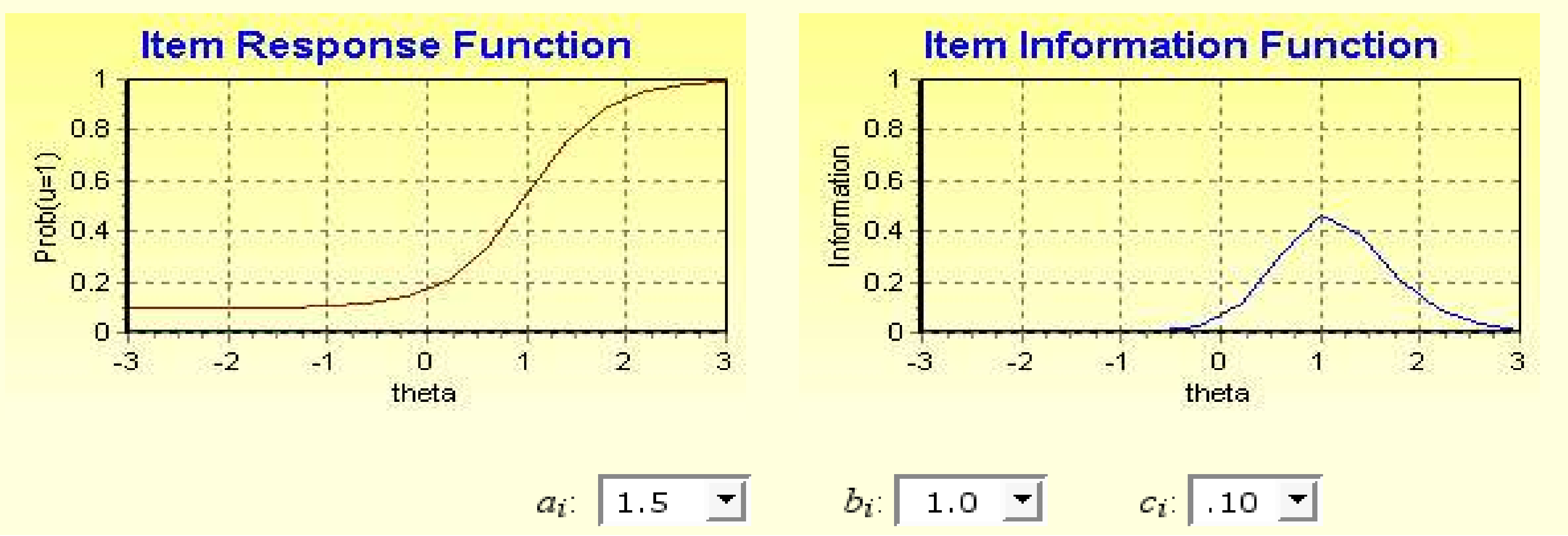


Item Response Function

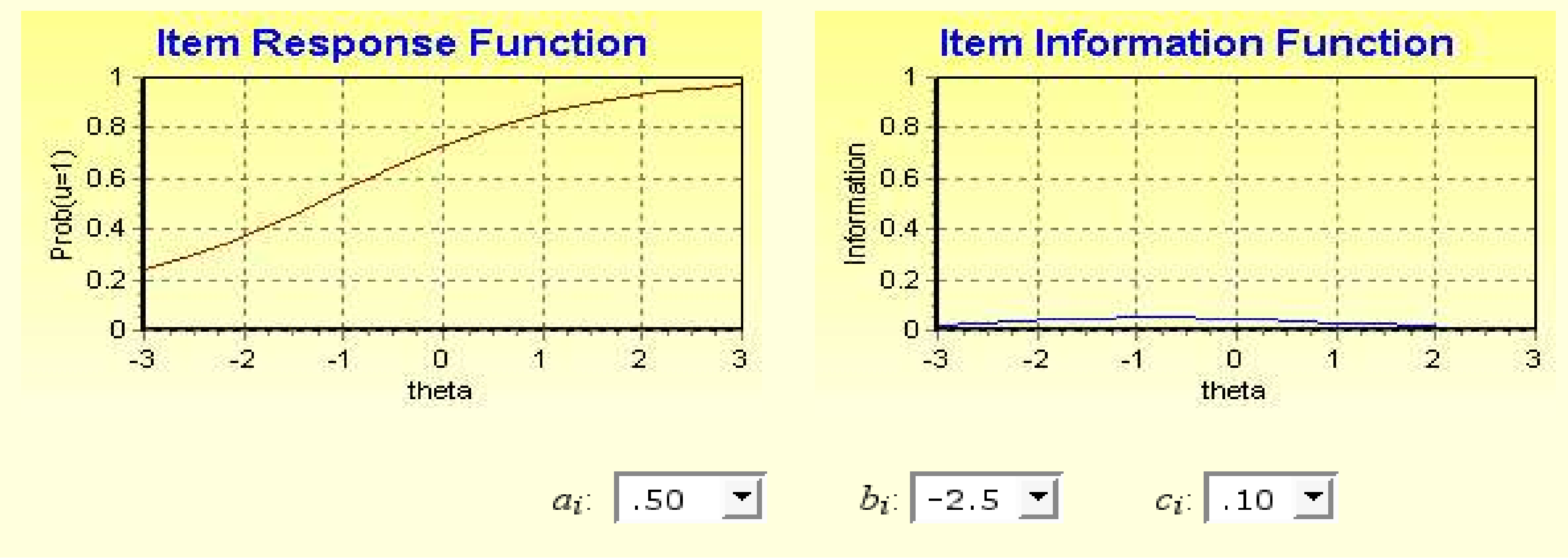
Item Information Function

a_i: 1.5 b_i: 0.5 c_i: .10

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

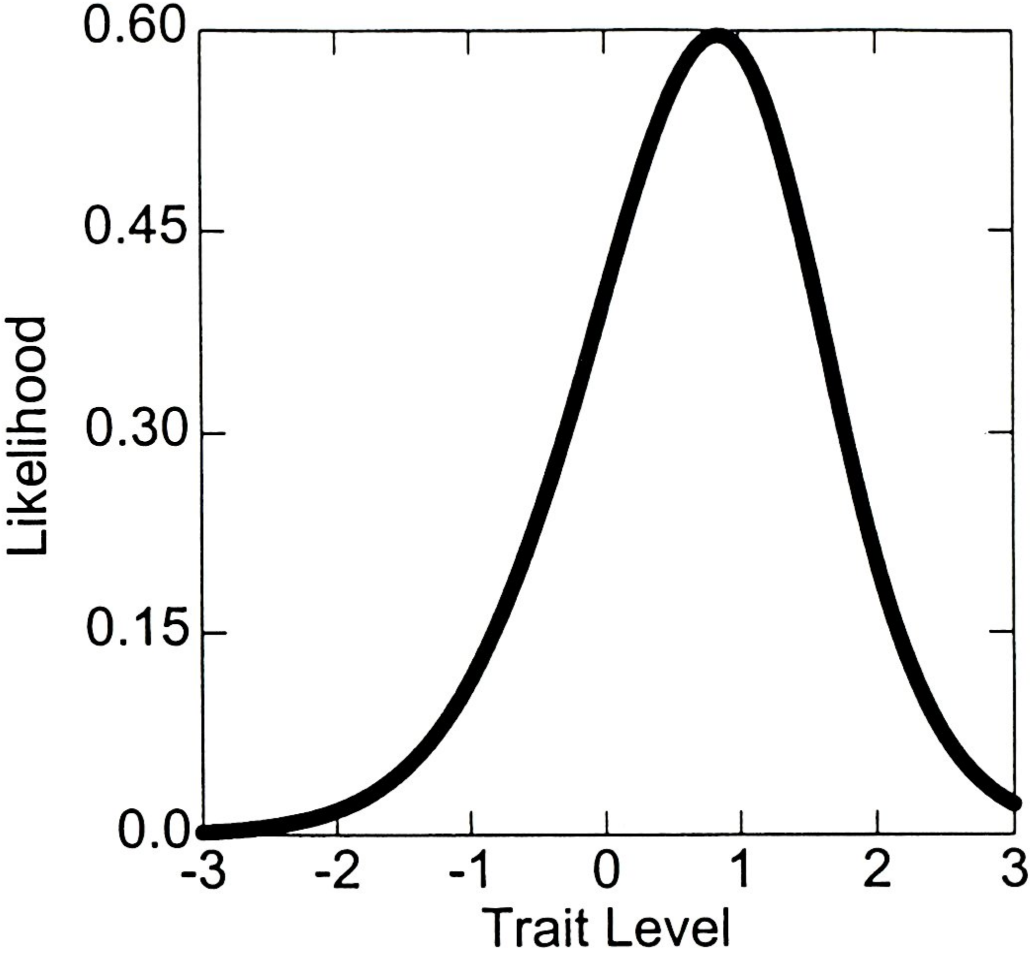


a_i: 1.5 ▾ b_i: 1.0 ▾ c_i: .10 ▾



a_i: .50 ▾ b_i: -2.5 ▾ c_i: .10 ▾

$$L(u_1, u_2, \ldots, u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$



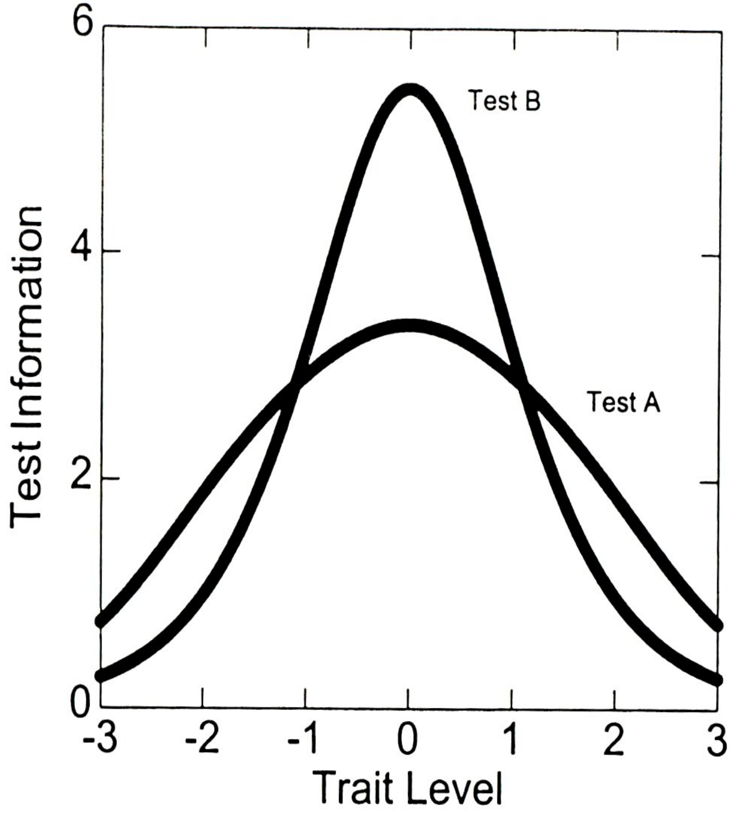


FIG. 7.3. Test information curves for example tests A and B.

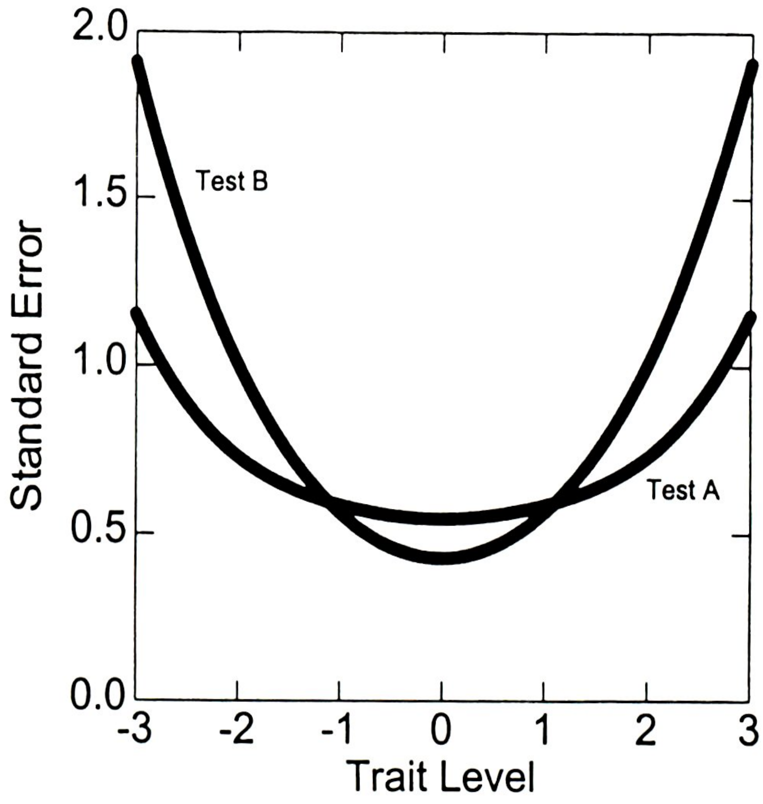


FIG. 7.4. Standard error curves for example tests A and B.

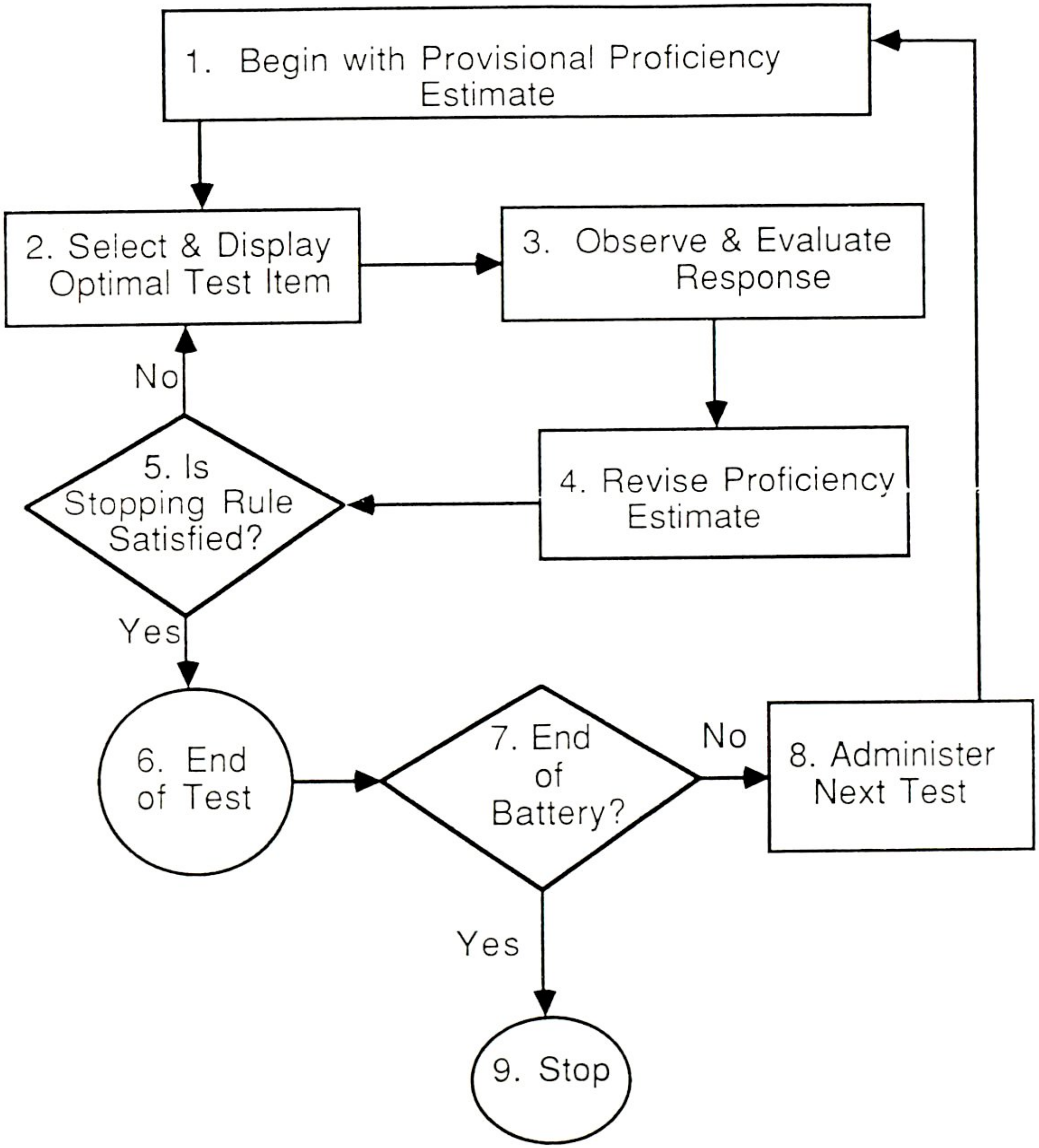
IRT – References

- **Hambleton, R. K., Swaminathan, H., Rogers, H. J., (1991).**
  - Fundamentals of Item Response Theory
- **Embretson, S. E., Reise, S. P. (2000).**
  - *Item Response Theory for Psychologists*.
- **Baker, F. B., Kim, S.-H. (2004).**
  - *Item response theory: Parameter estimation techniques. Second Edition, Revised and Expanded*.

CAT – Computer Adaptive Testing

- Individual vs. Group testing
- Improving entire measurement process:
  - Improved test security
  - Each indiviual stays busy productively
  - The test can be scored immediately
  - Unobtrusive pretesting

Adaptive Test Logic


1. Begin with Provisional Proficiency Estimate
2. Select & Display Optimal Test Item
3. Observe & Evaluate Response
4. Revise Proficiency Estimate
5. Is Stopping Rule Satisfied?
No
Yes
6. End of Test
7. End of Battery?
No
8. Administer Next Test
Yes
9. Stop

CAT – Key questions

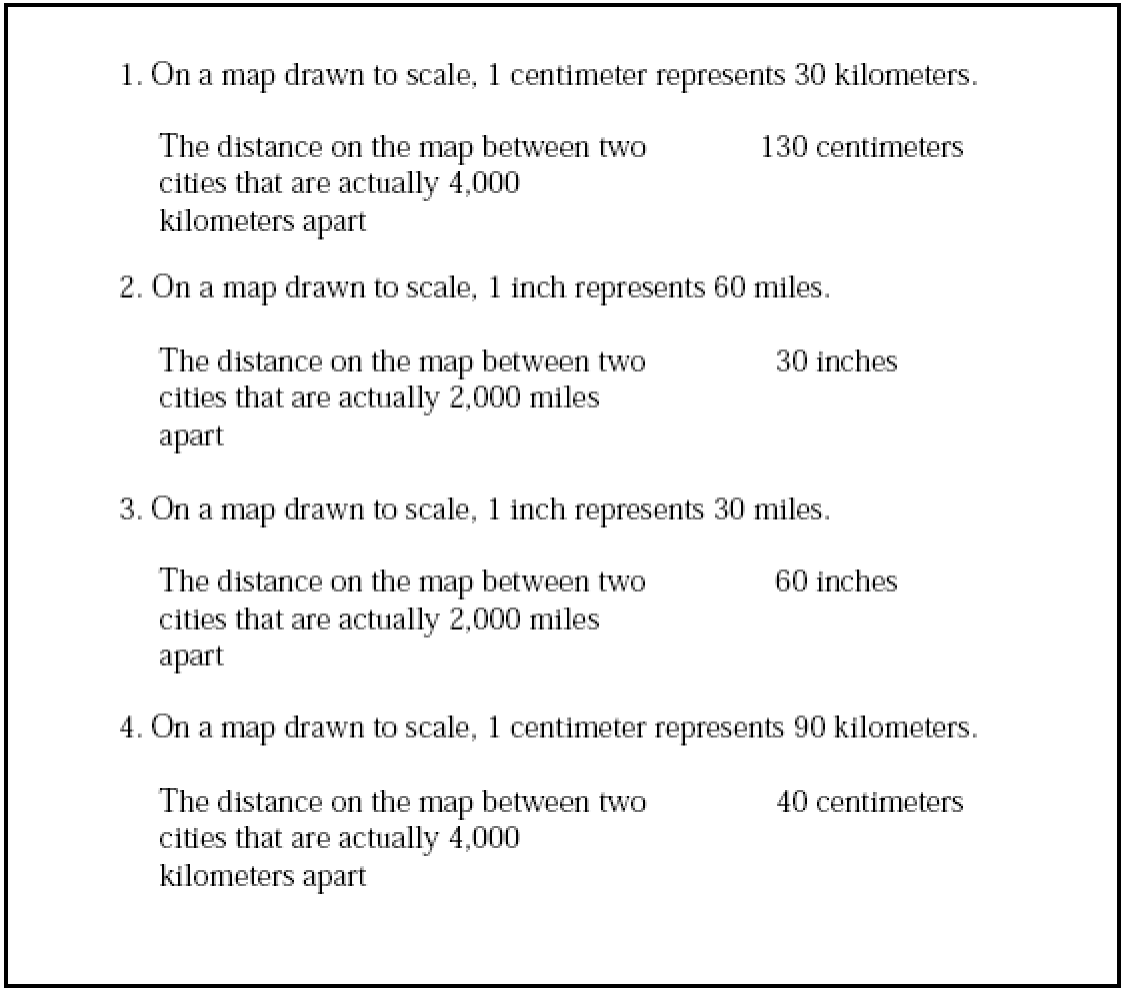
- How to START
  - Medium difficulty item?
- How to CONTINUE
  - Item exposure control
  - Stratification
- How to STOP

CAT – References

- **Wainer, H., (Ed.) (2000).**
  - *Computerized adaptive testing: A primer (2nd Edition).*
- **Sands, W. A., Waters, B. K., McBride, J. R., (Eds.). (1997).**
  - *Computerized adaptive testing: From inquiry to operation.*
- **van der Linden, W. J., Glas, C. A. W., (Eds.). (2000)**
  - *Computerized Adaptive Testing – Theory and Practice*

AIG – Automatic Item Generation

- Item models used to generate new items:



1. On a map drawn to scale, 1 centimeter represents 30 kilometers.

   The distance on the map between two cities that are actually 4,000 kilometers apart — 130 centimeters

2. On a map drawn to scale, 1 inch represents 60 miles.

   The distance on the map between two cities that are actually 2,000 miles apart — 30 inches

3. On a map drawn to scale, 1 inch represents 30 miles.

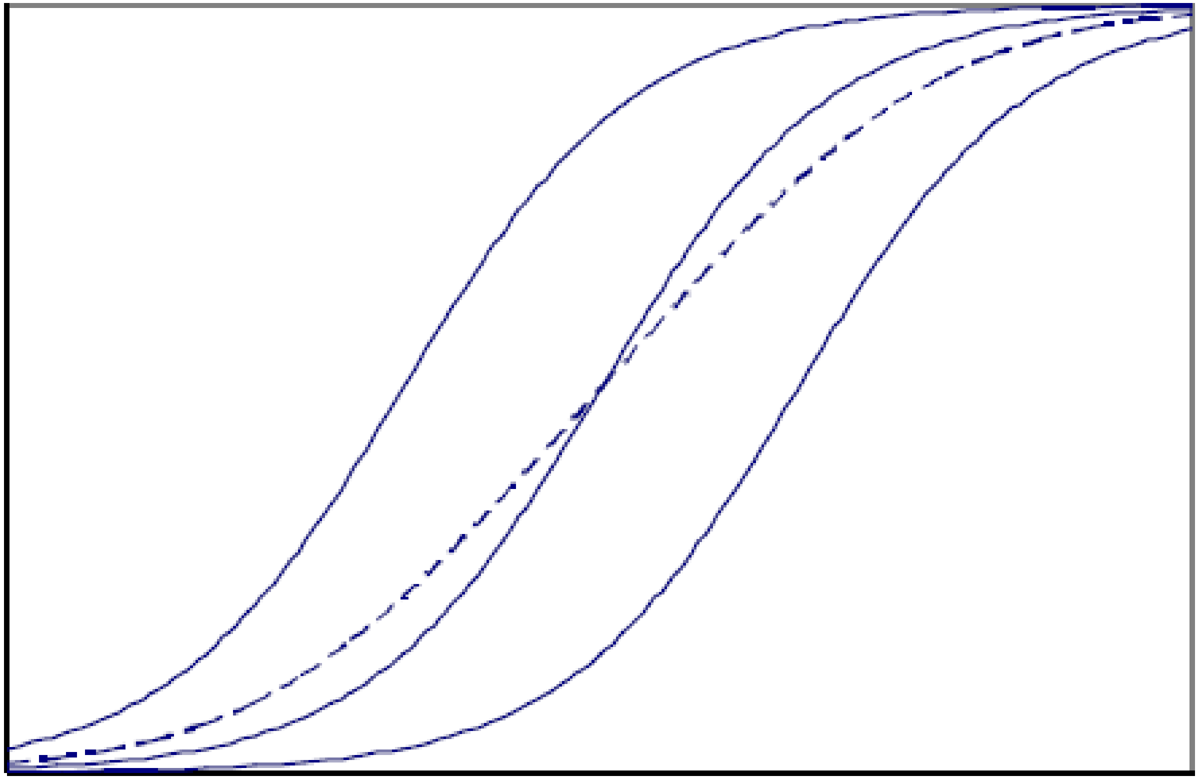
   The distance on the map between two cities that are actually 2,000 miles apart — 60 inches

4. On a map drawn to scale, 1 centimeter represents 90 kilometers.

   The distance on the map between two cities that are actually 4,000 kilometers apart — 40 centimeters

AIG – Item model calibration

- Expected response function:



Graph of expected response function (dashed curve) against three item characteristic curves at three levels of difficulty.

AIG – References

- **Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., Revuelta, J. (2003).**
  - *A feasibility study of on-the-fly item generation in adaptive testing.*
- **Deane, P., Sheehan, K. (2003).**
  - *Automatic item generation via frame semantics: Natural language generation of math word problems.*

AEG – Automatic Essay Grading

- Essay / short free-text response
- Statistical and NLP techniques
- **Electronic Essay Rater (E-Rater)**
  - Syntactic structure, vocabulary use
  - Grades writing skills on six-point scale (performance: 87 - 94 %)
- **Conceptual Rater (C-Rater)**
  - Assessment of short-answer to content-based questions (performance: 80%)

AEG – References

- **Valenti, S., Neri, F., Cucchiarelli, A. (2003).**
  - *An overview of current research on automated essay grading.*
- **Burstein, J. C., Kaplan, R. M., Wolff, S., Lu, C. (1996).**
  - *Using Lexical Semantic Techniques to Classify Free Responses.*