CTB McGraw-Hill

Adaptive Testing: A Simple Idea Finally Come True

Wim J. van der Linden

Outline

- Brief history of IRT
- Adaptive testing
- A few new developments

Brief History of IRT

- 18-19th century tradition of anthropometrics
  - Gauss, Galton, etc.
- 19th century tradition of psychophysics
  - Fechner, Ebbinghaus, Wundt's laboratory
- Early 20th century: measurement of purely psychological properties
  - Binet, Thurstone

Brief History of IRT

- Binet-Simon intelligence test (1905)
  - Placement in special education in Paris
  - How to test intelligence?

> *There is no difficulty in measurement as long as it is a question of experiments on tactile, visual, and auditory sensations. But if it is a question of the keenness of intelligence, where is the method to be found to measure the richness of intelligence, the sureness of judgment, the subtlety of mind?*
>
> A. Binet, 1898

Brief History of IRT

- Binet's innovations:
  1. Multiple tasks/test items
     - Memory
     - Reasoning
     - Judgment
     - Abstraction
     - Etcetera
  2. Rigorous standardization of testing procedures

Alfred Binet
(1857-1911)

Brief History of IRT

- Binet's innovations (*cont'd*)
  3. Pretesting and scaling
     - Pretesting of items for in groups 3-11
     - Use of *chronological age* as scale for intelligence
       - Scale value of items: age at 75th percentile
     - Use of scale values of the items to score the students
       - Definition of *mental age*
- Stanford-Binet intelligence test (1916)
  - Now in its 5th edition

Brief History of IRT

- Thurstone's (1925) ideas of scaling
  - Problem: use of age as a scale assumes implies decline of intelligence while aging
  - Introduction of a latent scale
- Model-based measurement
  - Assumption of response function for each item
  - Statistical estimation of scale values and scores

Louis L. Thurstone (1887-1955)

Brief History of IRT

- Thurstone actually reversed Binet's idea of scaling
  - Binet
    - Defined a scale (age)
    - Looked for resulting response functions
  - Thurstone
    - Defined mathematical form of response functions
    - Looked for resulting scale

Brief History of IRT

- Thurstone's idea of scaling was the beginning of item response theory (IRT)
  - Persistent confusion between (i) normal cumulative distribution function and (ii) normal-ogive response function
  - Appropriate statistical methods and computational power still lacking
  - Classical test theory omnipresent
  - Some distrust ("latent" variables???)

Brief History of IRT

- 1950s: Lazersfeld, Lord, Rasch
- 1960s: Alan Birnbaum
  - Aware of earlier developments in bio-assay (dose-response curves)
  - Berkson ("Why I prefer logits over probits")
  - 3-parameter logistic model
- 1970-1990s: Major statistical developments
  - Bayesian methods
  - Software development

Brief History of IRT

- From 1990s: New models and applications
  - Polytomous, rating scale, multidimensional, hierarchical, etc. models
  - Large-scale educational testing programs
  - (Inter)national assessments
  - Adaptive testing
  - Health measurement
  - Marketing research

Adaptive Testing

- Binet's intelligence test was already adaptive!
  - Everything standardized except item selection
- Requirements for computerized adaptive testing
  - Response model with separate parameters for examinees and item properties
  - Calibrated item pool
  - Real-time ability estimation

Adaptive Testing

- Requirements *(Cont'd)*
  - Selection of initial item
  - Item selection rule
    - Rules should realize test specifications
  - Stopping rule

A Few New Developments

- Other response models
  - Polytomous responses
  - Multidimensional person parameters
- Developments in educational testing
  - E.g., cheating detection
- Hierarchical modeling
- Adaptive item calibration

Empirical Response Functions for Binet's Items (Thurstone, 1925)

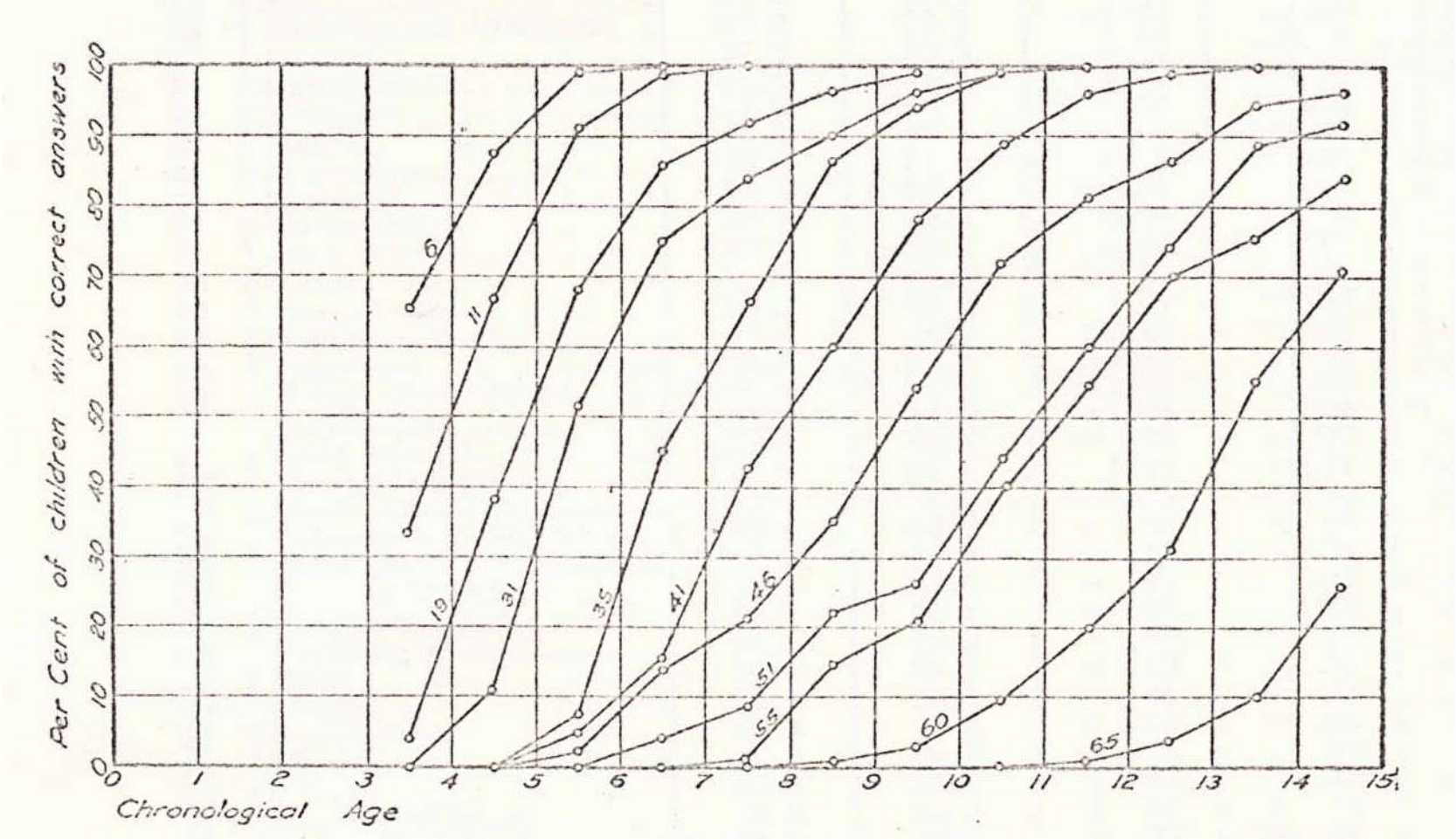
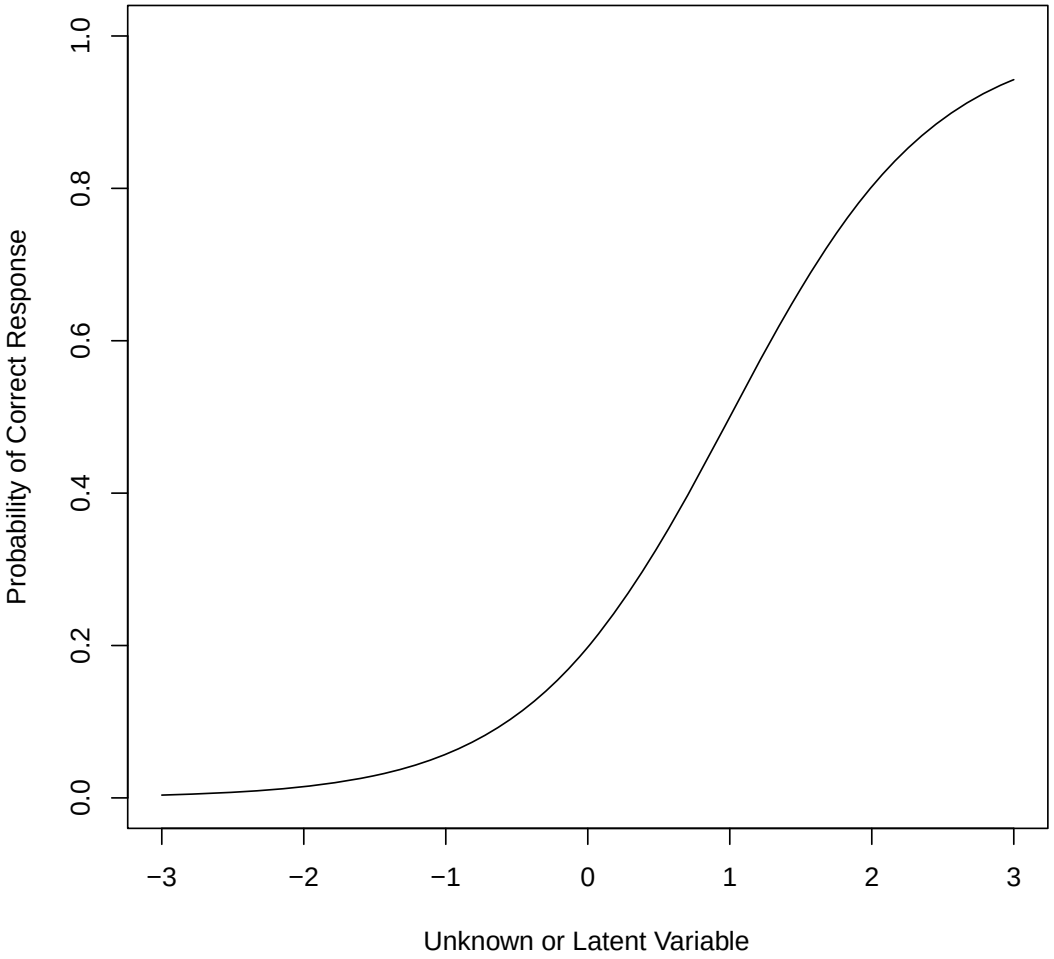
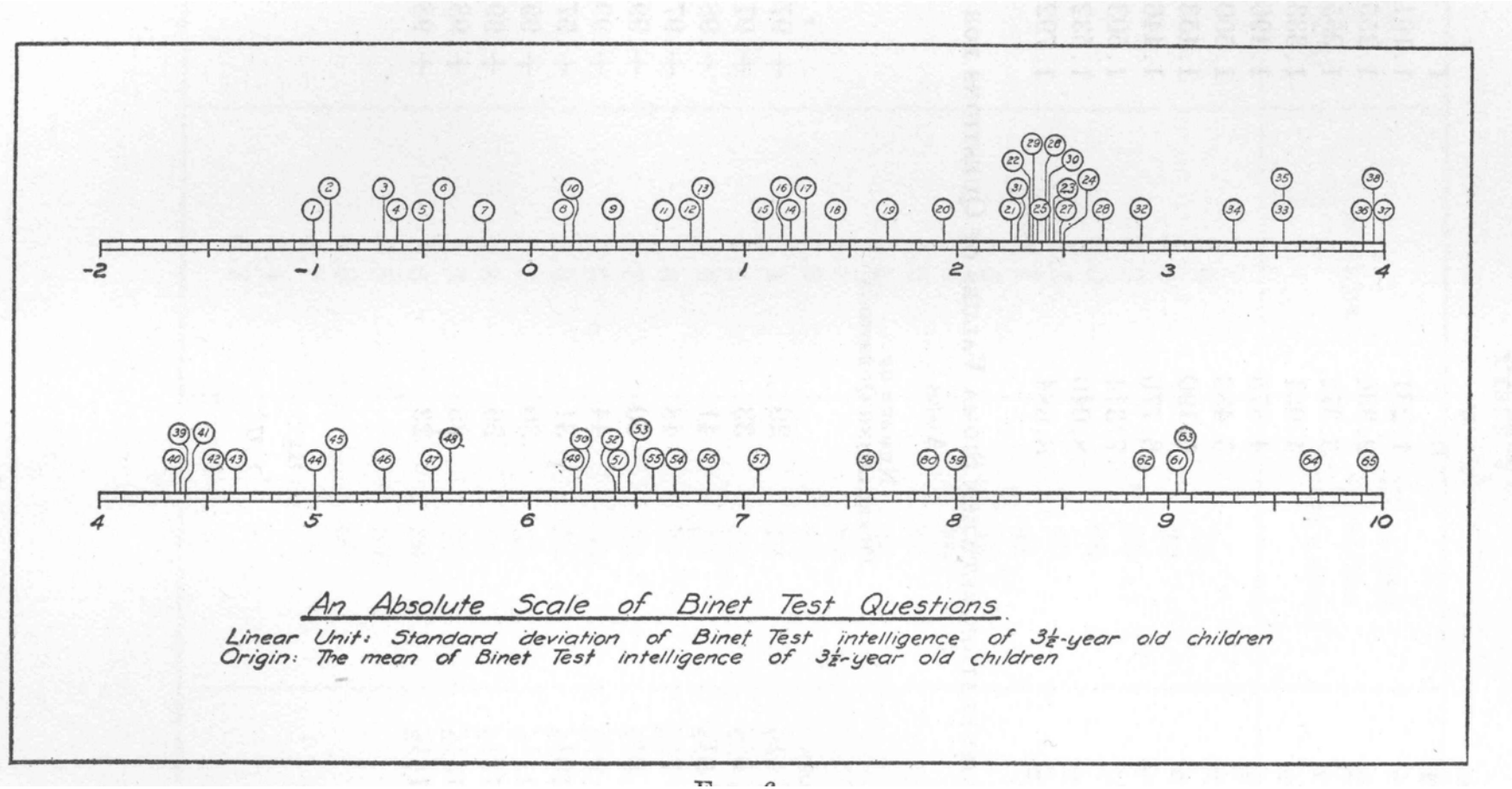


FIG. 5.

Thurstone's Normal-Ogive Response Function



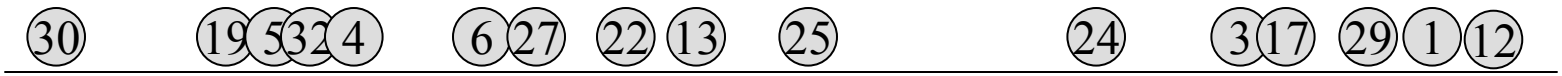
IRT Scale for Binet's Items (Thurstone, 1925)



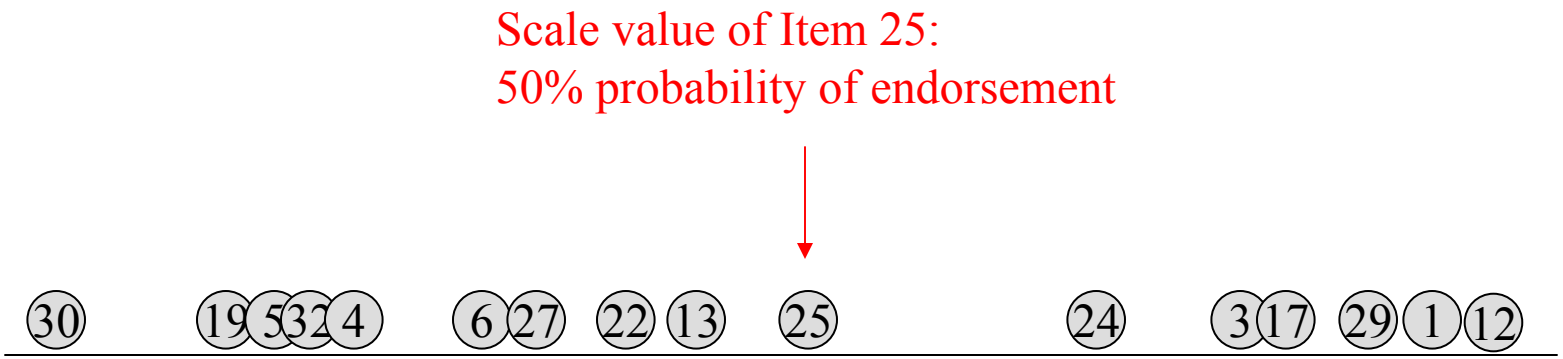
Test of Body Height

1. I bump my head quite often
2. For school pictures I was always asked to stand in the first row
3. In bed, I often suffer from cold feet
4. When walking down the stairs, I take two steps at a time
5. I think I would do well on a basket ball team
6. As a police officer, I would not make much of an impression
7. In most cars, I sit uncomfortably
8. I literally look up to most of my friends
9. Etc.

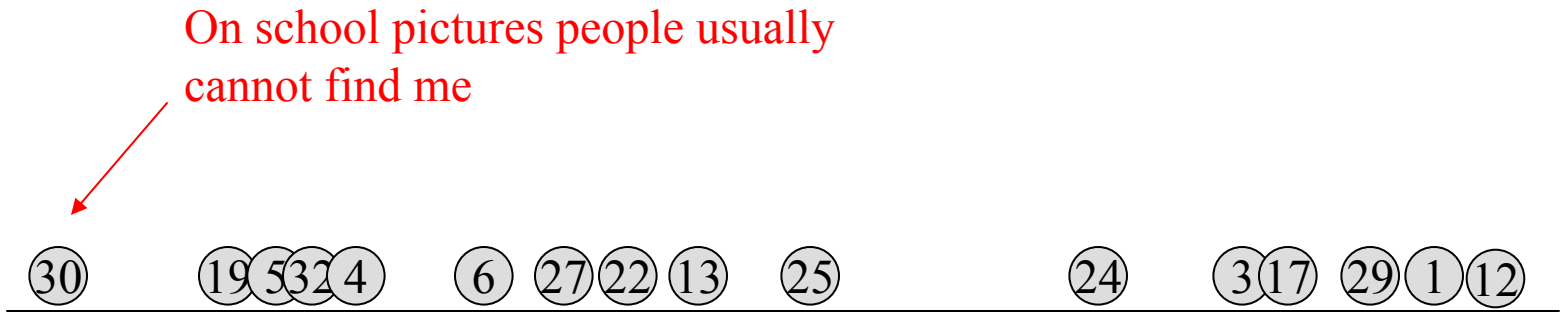
Test of Body Height



Test of Body Height



Test of Body Height



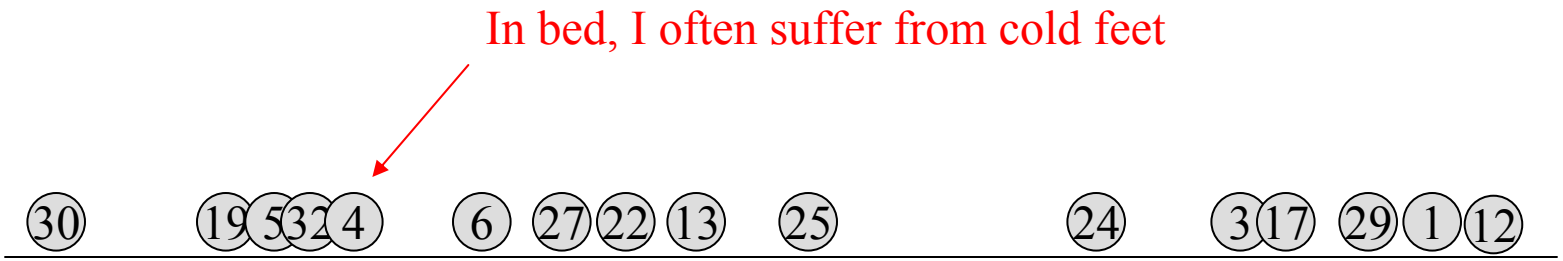
Test of Body Height

I often have to stand on my toes
to see my face in the mirror

Test of Body Height

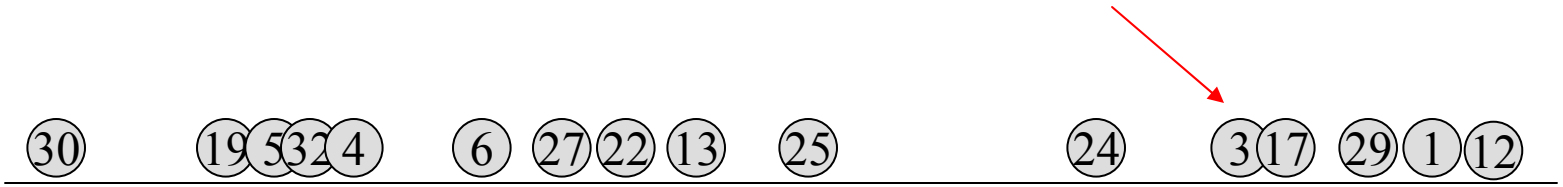
In libraries, I often have to use a ladder to reach the books

Test of Body Height



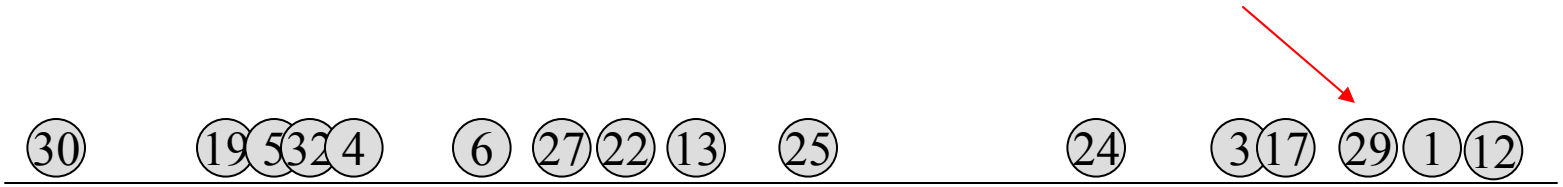
Test of Body Height

For most people my shoes would be too large



Test of Body Height

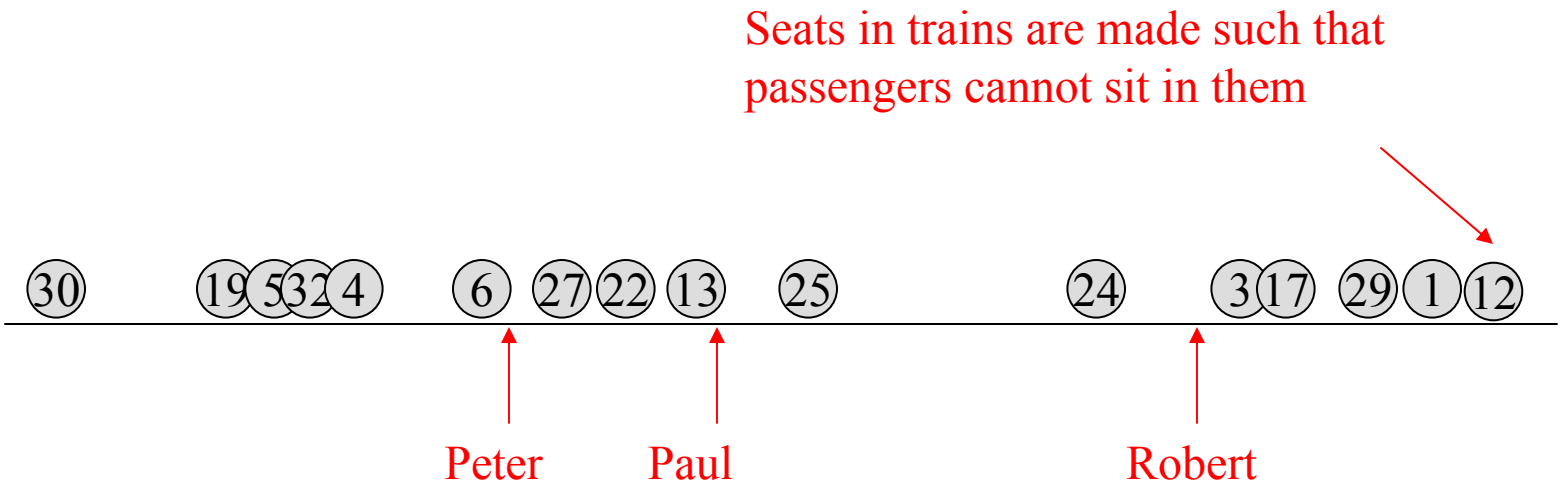
In trains, I have to watch out not to bump my head against the doorposts



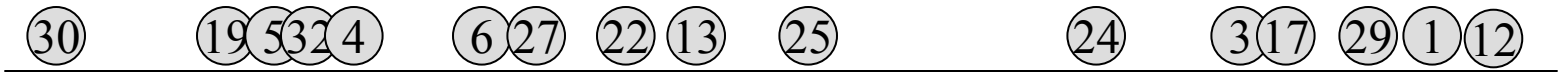
Test of Body Height

Seats in trains are made such that passengers cannot sit in them

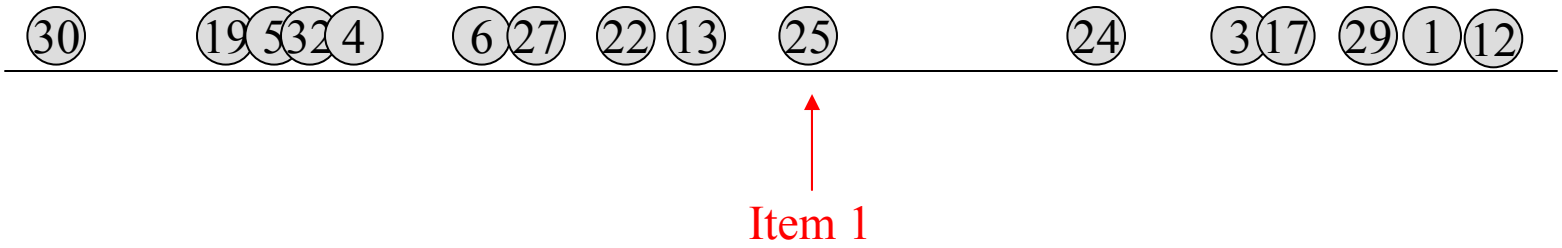
Test of Body Height



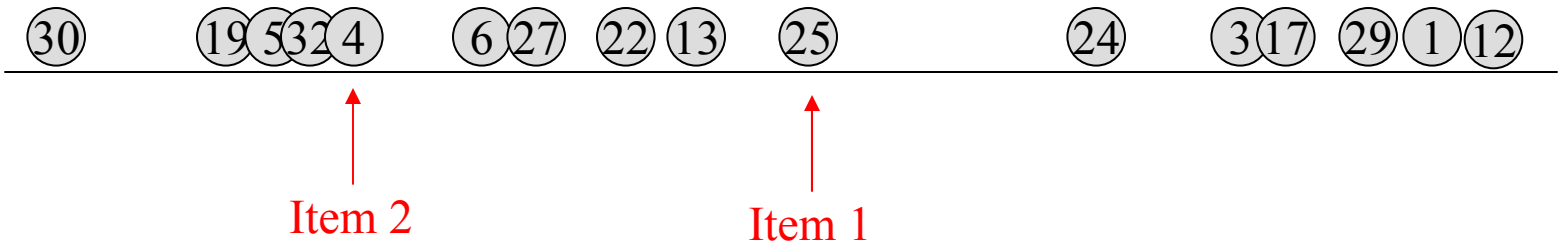
Adaptive Testing



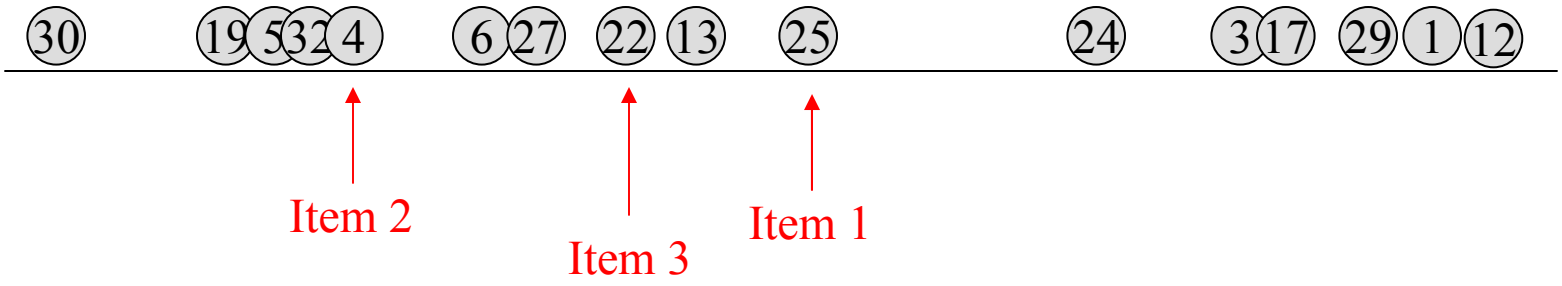
Adaptive Testing



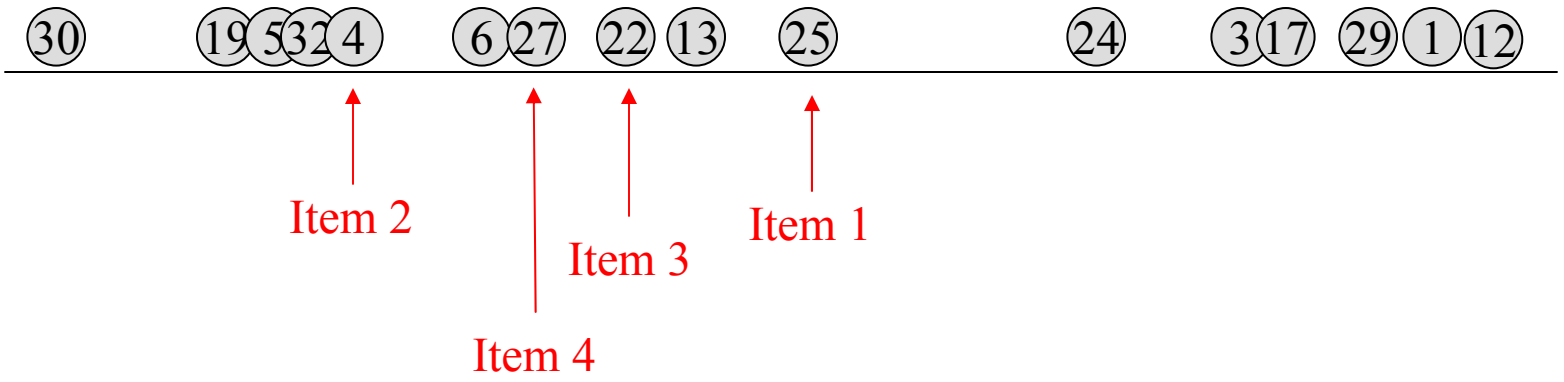
Adaptive Testing



Adaptive Testing



Adaptive Testing



Response Model

- Example of 3PL response function

$$P_i(+|\theta) = c_i + (1 - c_i)\,\Psi[a_i(\theta - b_i)]$$

Item Item Person Item

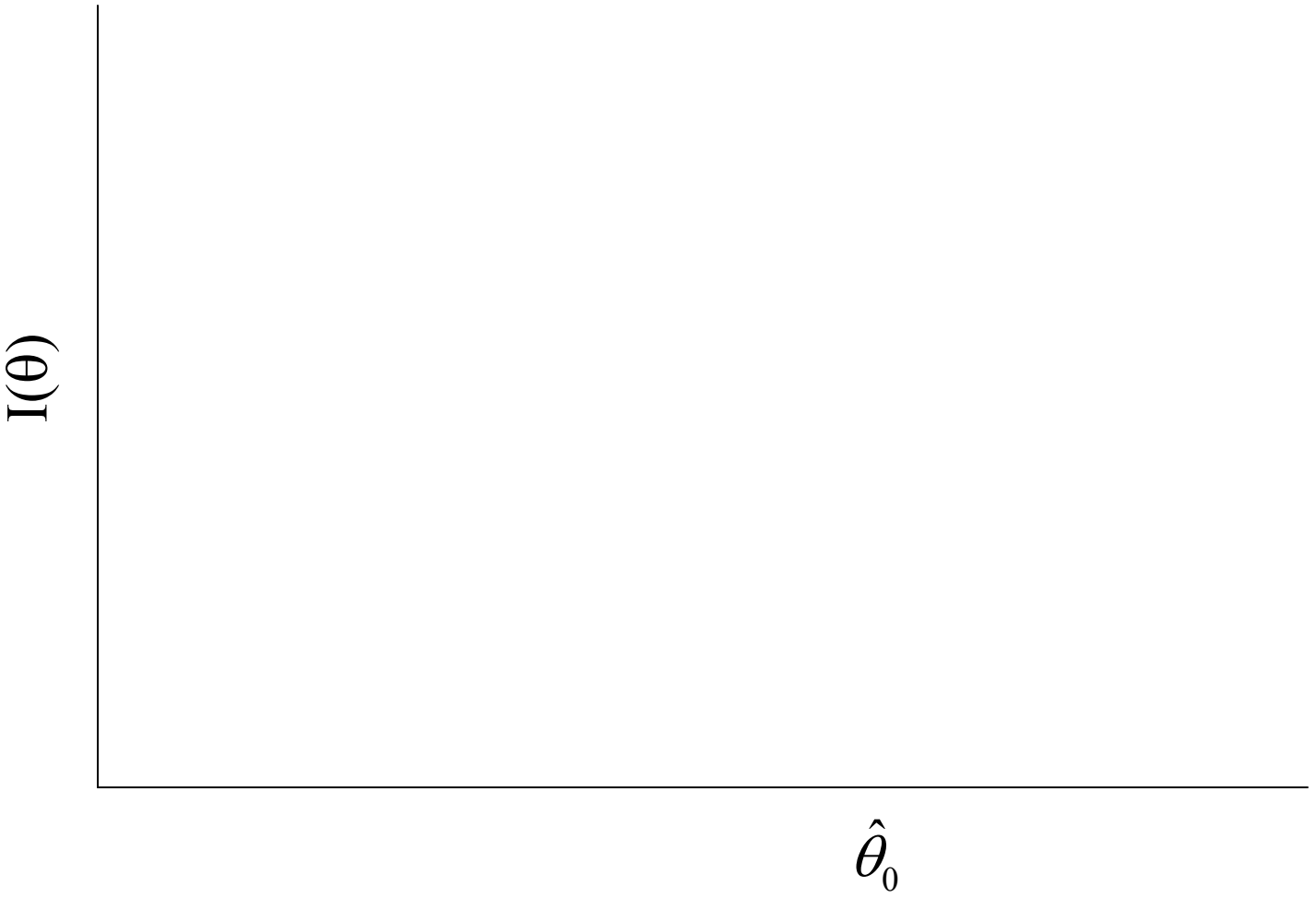
Response Model

- No iid response distributions across persons or items
  - Model with incidental parameters (*not*: nuisance parameters
- Great optimal design opportunities

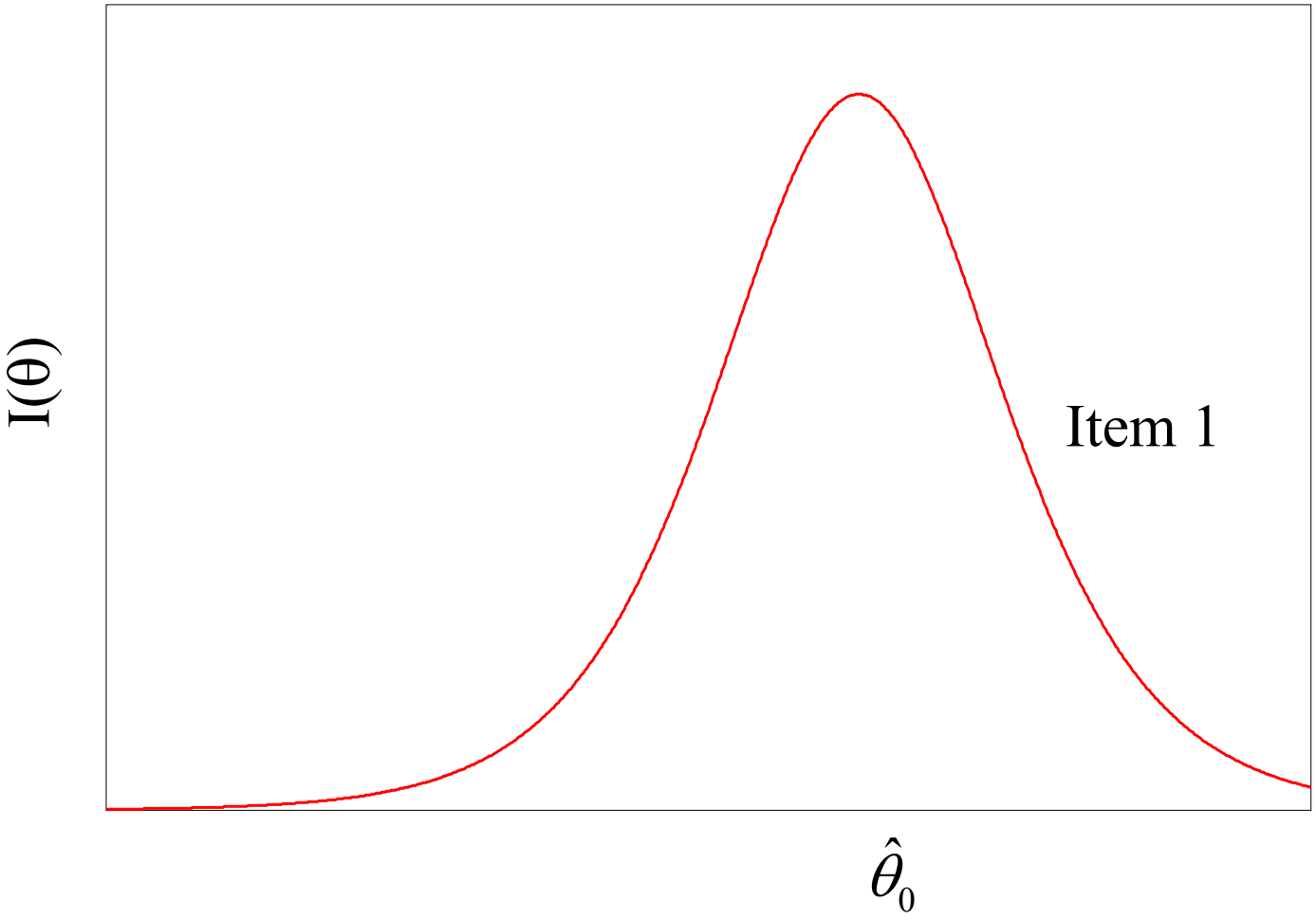
$$(\theta, a_i, b_i, c_i) \rightarrow \text{Efficiency}(\hat{\theta})$$

$$(\theta, a_i, b_i, c_i) \rightarrow \text{Efficiency}(\hat{a}_i, \hat{b}_i, \hat{c}_i)$$

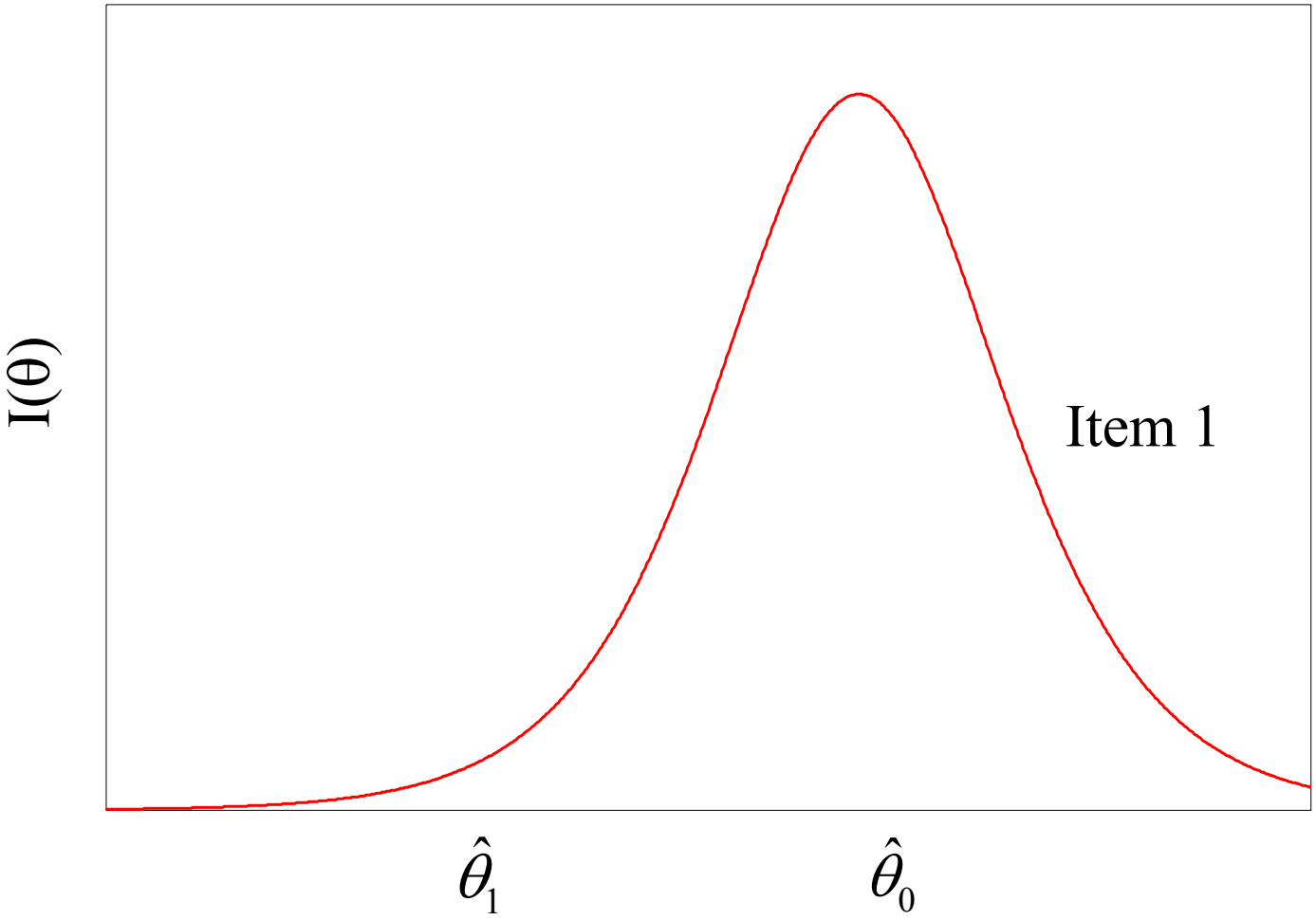
Adaptive Testing



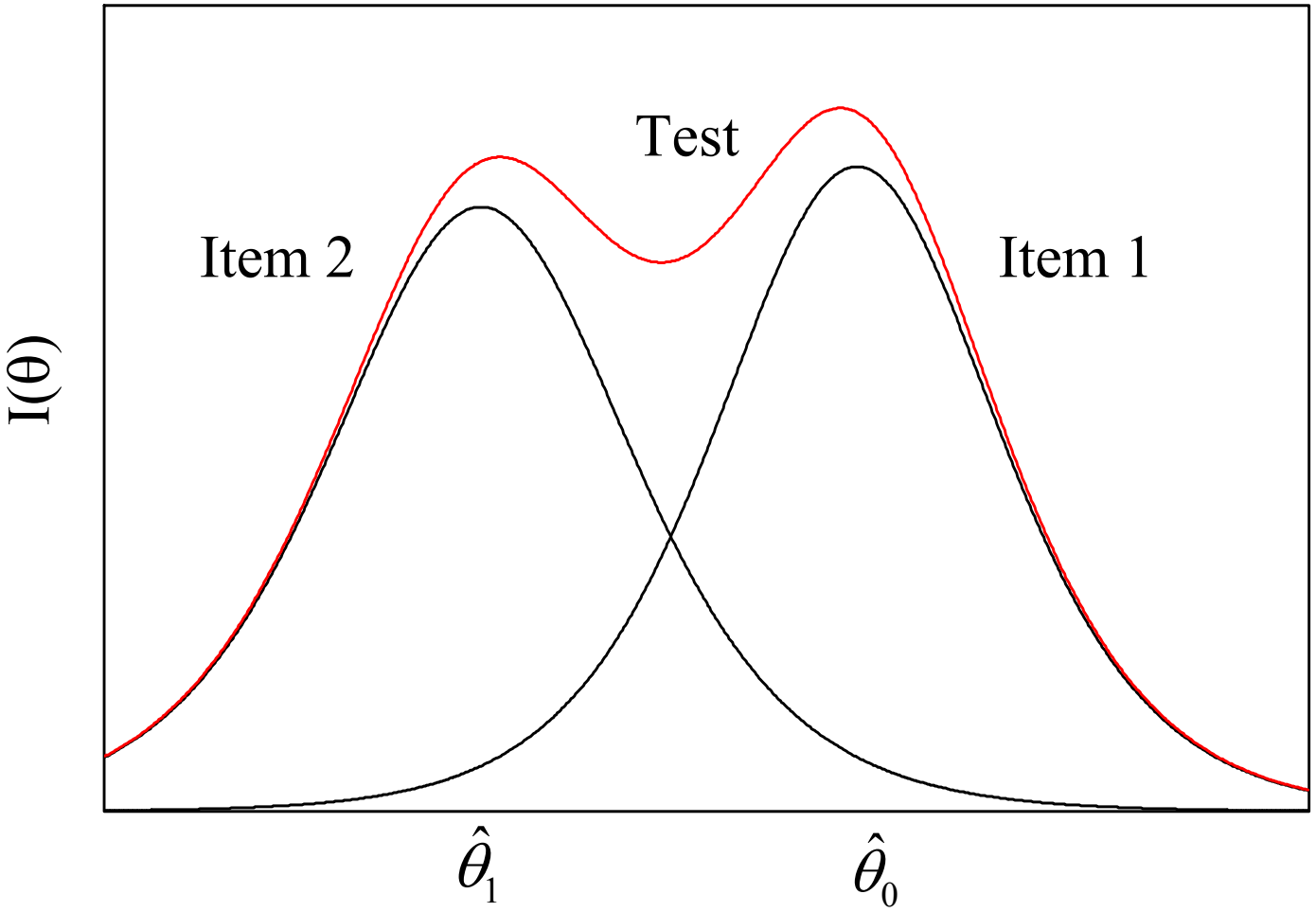
Adaptive Testing



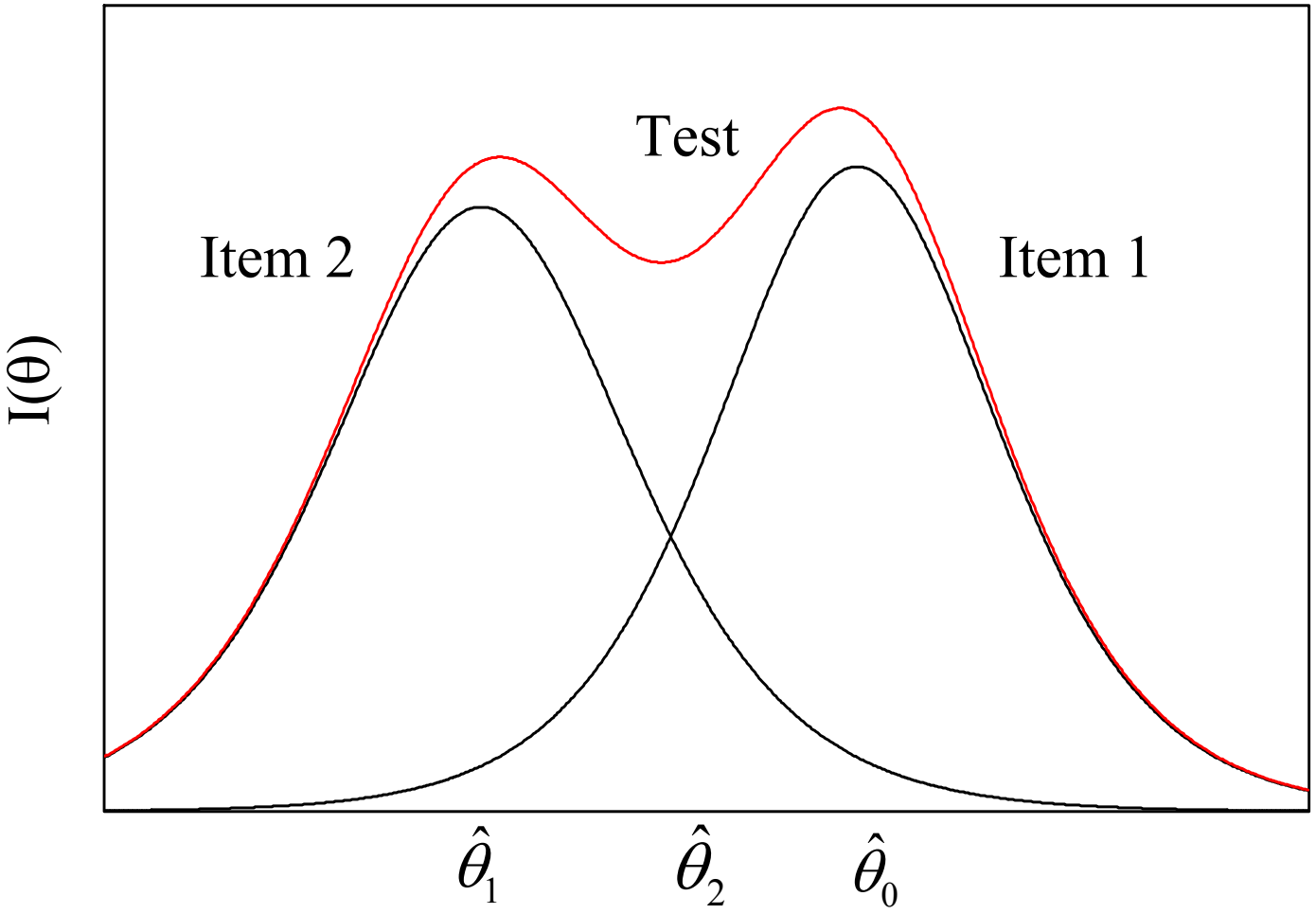
Adaptive Testing



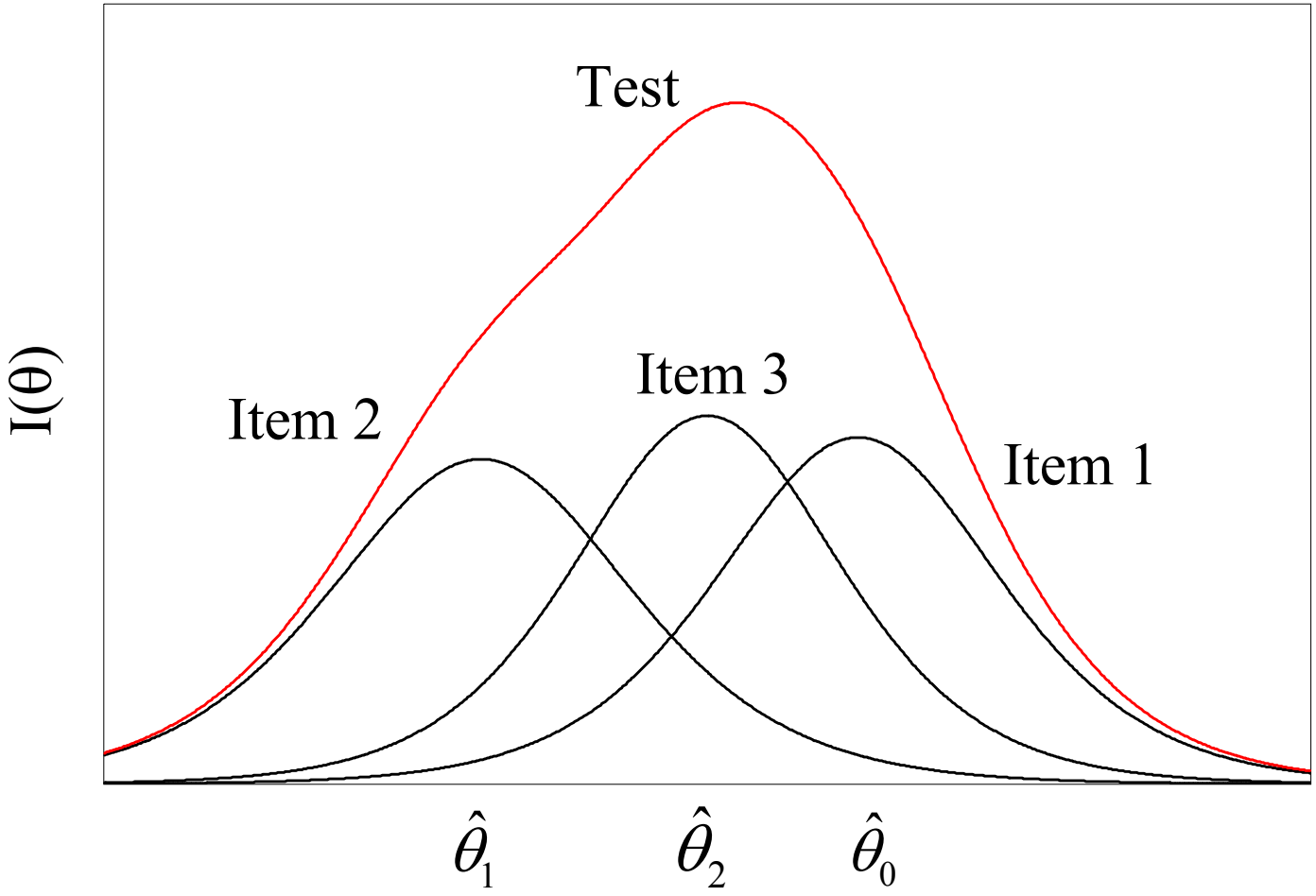
Adaptive Testing



Adaptive Testing



Adaptive Testing



Hierarchical Modeling

- Main advantages
  - Use of collateral information ("borrowing information") to improve test design and/or accuracy of scores
- Two basic types of applications
  - Random person parameters
  - Random item parameters

Adaptive Test Battery

- Test batteries used in diagnosis, guidance, etc.
  - Different subtests for distinct but strongly related abilities
- Use of two types of adaptation
  - within each subtest
  - sequencing of subtests
- Empirical Bayes approach

Adaptive Test Battery

- 3PNO response model for θ_h, $h=1,\ldots,H$

$$P_i(+|\theta_h) = c_i + (1-c_i)\,\Phi[a_i(\theta_h - b_i)]$$

- Ability distribution

$$\theta_1,\ldots,\theta_H \sim MVN(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$$

- Item and distribution parameters estimated during item calibration

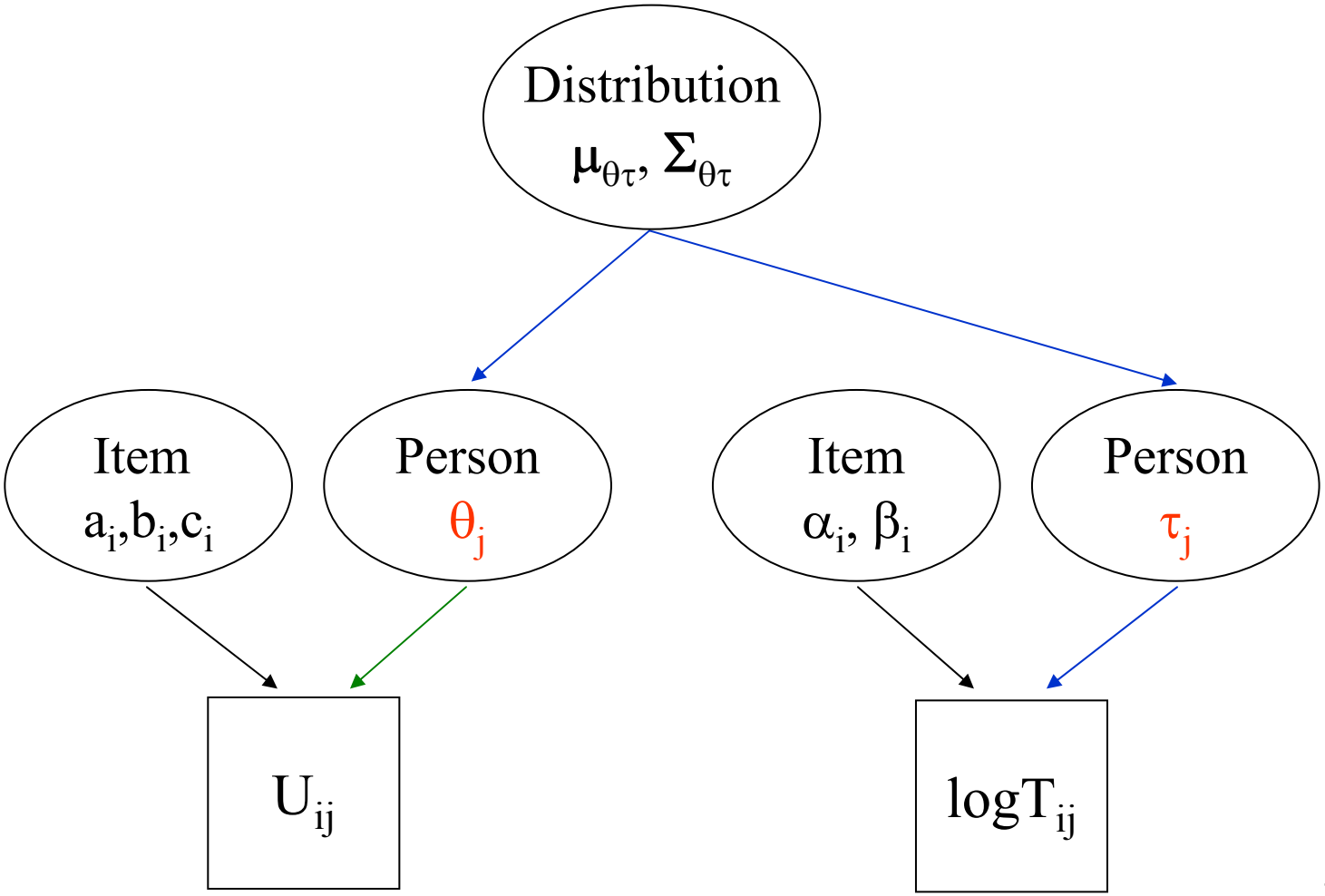
Adaptive Test Battery

- Adaptive sequencing
  - Select first subtest using initial guess of all abilities
  - Select second subtest using posterior distribution of first ability
  - Use same information to initialize second subtest
  - Etc.

Use of Response Times

- Test taker's speed typically correlates with his/her ability
- Use response times as collateral information to update ability estimates and select items
- Benefits
  - Shorter tests (for moderate correlation between speed and ability up to 50% reduction of test length)
  - Compensate ability regions where the item pool is less rich

Hierarchical Model *Cont'd*



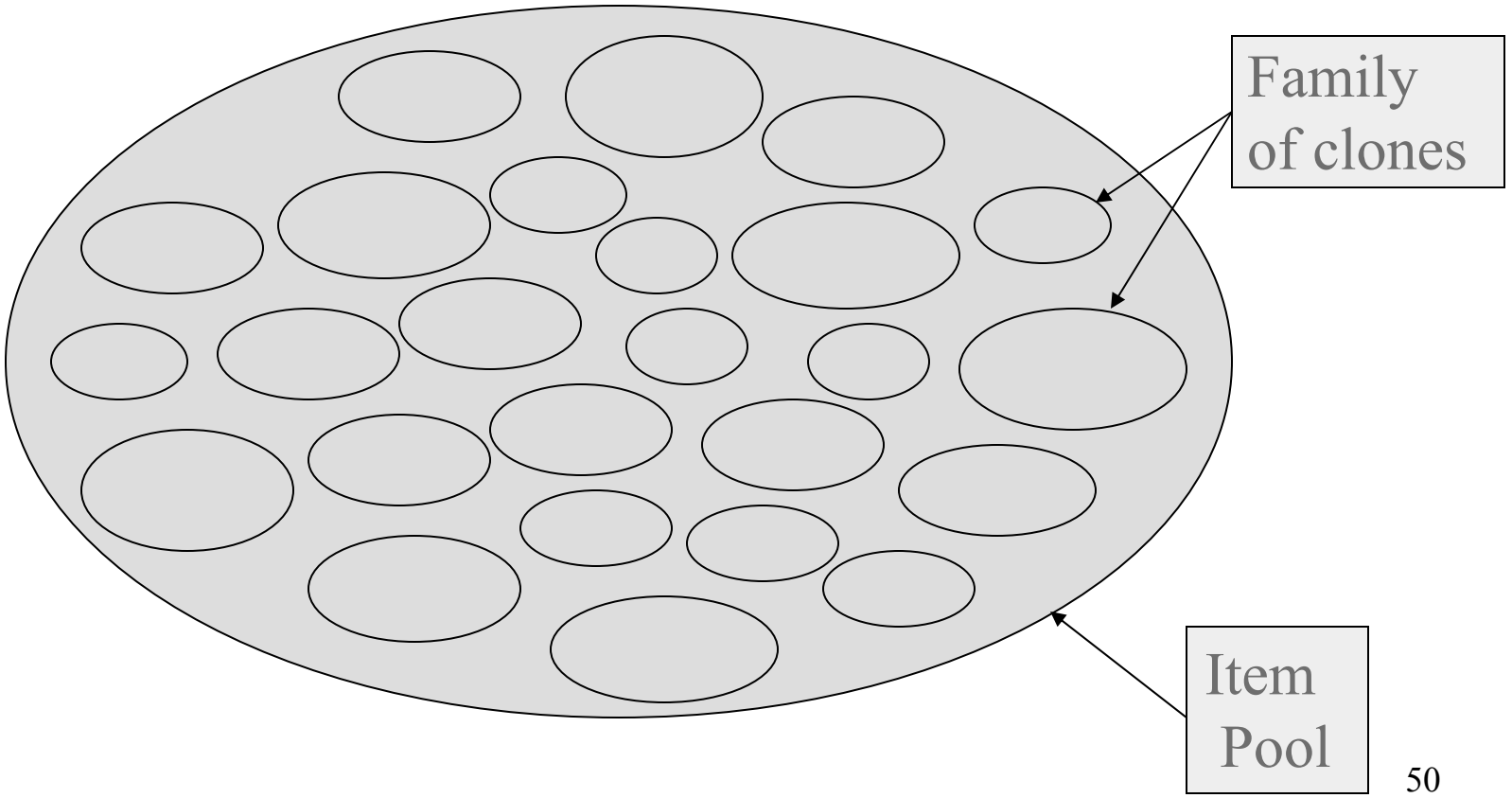
Use of Response Times

- Same approach works with any other source of collateral information, e.g., physiological measures, confidence marking, etc.

Item Cloning

- Use computer to generate new items (on the fly)
- Calibrate item families of clones rather than each individual item
- Use hierarchical IRT to allow for (small) random variation in item parameter values

Item Cloning



Item Cloning

- 3PNO model for each individual item *i* in family *f*

$$P_{i_f}(+|\theta) = c_{i_f} + (1 - c_{i_f})\,\Phi[a_{i_f}(\theta - b_{i_f})]$$

- For each family *f*,

$$(\ln a_{i_f}, b_{i_f}, \text{logit } c_{i_f}) \sim MVN(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$$

Item Cloning

- Use in adaptive testing
  - Optimal selection of family based on family parameters
  - Random generation of item from family for administration to examinee

Adaptive Item Calibration

- Efficiency of item parameter estimates depends on distribution of test takers' θs
- Adapt selection of pretest items to test takers
  - Run test for *n*-5 items or so
  - Select pretest items (3, say) using
    - current posterior distribution of test taker's θ
    - current posterior distributions of the pretest items
  - Insert selected pretest items randomly into 3 of the last 5 slots

Adaptive Item Calibration

- Benefits
  - Pretesting under fully operational conditions
  - Minimal sample size
  - No separate linking study (all pretest items automatically on the operational scale)
  - Possibility of a "self-replenishing" item pool

2011 Conference of the International Association for Computerized Adaptive Testing (IACAT)
October 3-5, Pacific Grove, CA

ctb.com/iacat