UNIVERSITY OF CAMBRIDGE
The Psychometrics Centre

Item Generation

University of Cambridge
The Psychometrics Centre
Aiden Loe
bsl28@cam.ac.uk

Traditional Item Writing

- Items are normally generated by professional item writers
- Training and supervising item writers can be expensive ($1,500 to $2,000 per item: Rudner, 2010) and time consuming.

Traditional Item Writing

- 40% of expertly generated items fail to perform (Haladyna 2012)
- Smartphones and the Internet make it easier to compromise test security
- Tech improvements allow widespread adoption of Computer Adaptive Testing
  - More efficient (reduce length by 50%-90% Gibbons et al. 2008).
  - Item banks bigger (four to five times larger than traditional).

Item Generation Approaches

- Crowd Sourcing Item Generation
- Automatic Item Generation
  - Rule-based Generator
  - Semantic Frame Generator
  - Machine Learning Generator

AIG

- Rule based
  - Item Models (Gierl, Zhou & Alves, 2008)
  - Cognitive Models (Embretson & Yang, 2007)
  - Schema Theory (Singley & Bennett, 2002)
  - Automatic min-max (Arendasy & Sommer, 2011)
- Natural Language Processing
  - Semantic Frame Work analysis (Deane & Sheehan, 2003)
- Machine Learning
  - Uses machine learning algorithms to extract data. (Gütl et al., 2011)

Benefits

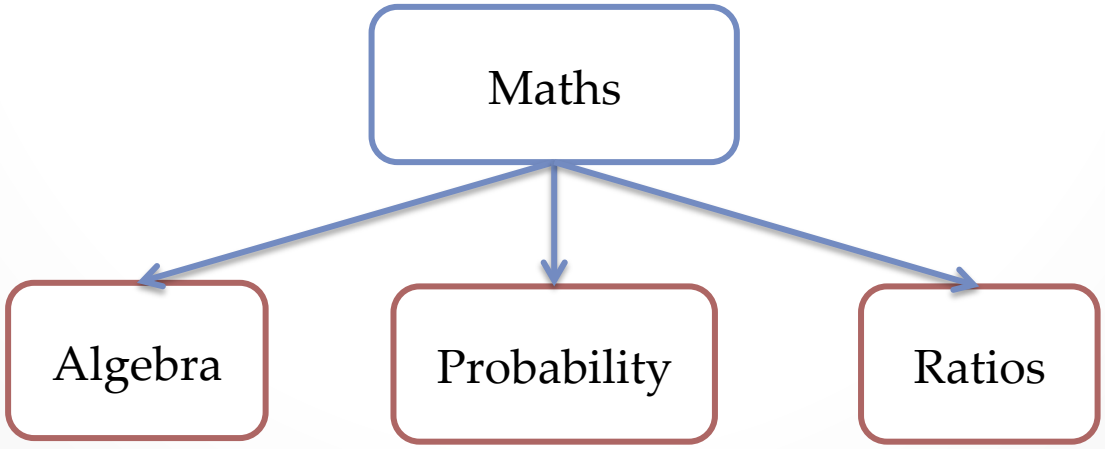
- Items calibrated ahead of time
- Difficulty levels are predicted*
- Save time and improve efficiency
- Reduce time exposure
- Clones of each other or variants

Item Features

- Radicals – Features that influence item difficulty
- Incidentals – Features that no do influence difficulty but changes the content of the question.
- E.g. John has 15 sweets. He shared 5 with Mary, and 2 with Paul. How many sweets does he have left?

Domain Map

- Decompose broad domain by identifying subclasses of items.
- These subclasses are defined by the different knowledge and solution process.
- Once identified, we can model difficulty within subclasses of items and use these models for item generation.



Domain map

- Domain map is useful because it allows us to ask questions about how we should integrate the item-generation capabilities into existing assessment.

Strategy

- Make test specification more explicit about the skills, processes, and strategies that are being assessed and to link these specifications to item difficulty.

- The more restrictive the item content, the more Item Design Rules contribute to Item Difficulty

Isomorph / Clones

- Derived from cognitive theory (*Simon & hayes, 1976)
- Generate or produce items that are in all respects equivalent, isomorphic, to all other items produced by a given model. (Difficulty is the same)
- CAT selects the item models rather than the item difficulty.
- Items are different but with psychometric properties identical to the original item

Item Variant

- Refer to instances of an item model that range in difficulty or some other psychometric characterization of the items.
- Hold constant the psychometric attributes of the generated items, but producing those items with a range of difficulty (Bejar, 1990).
- Items with a wide range of psychometric properties, specifically with a variation in item difficulty.
- Appear quite different to the examinees yet measure the same underlying construct
- Important in the role of test security.

Item models

- Selection of one or more items from existing tests that exhibit unique combination of processing-relevant item stimulus features (radicals).
- Usually narrowly defined in terms of the surface features that are allowed to be substituted to generate new items.
- Allow within-family variation of the psychometric characteristics of the items generated on their basis.

Example 1 – Item Clone

- Alex is coloring a paper mural using 80 crayons shared with 9 of his friends. Each of his friends has the same number of crayons, x . There were 8 crayons left over after Alex handed them out to his friends.

Parent Item

- Which of the following equations represents this situation?
  - 80=9x +8*
  - 80=8x +9
  - 80=9(8)+ x
  - 8=9x –80
  - 80=8x –9

Example Taken from : Gierl, M.J and colleagues (2015)

1 Layer Item Template

Alex is coloring a paper mural using *<Product.Material>* crayons shared with 9 of his friends. Each of his friends has the same number of crayons, x . There were 8 crayons left over after Alex handed them out to his friends.

- Which of the following equations represents this situation?

  - *<Product.Material>* =9x +8*
  - *<Product.Material>* =8x +9
  - *<Product.Material>* =9(8)+ x
  - 8=9x —*<Product.Material>*
  - *<Product.Material>* =8x —9

N-Layer Item Template

<Name> is coloring a <Product.Name> using <Product.Material> shared with <Gender.number> of <Gender> friends. Each of <Gender> friends has the same number of <Product. Material>, <Product.number>. There were <Product. Material> left over after <Name> handed them out to <Gender> friends. Which of the following equations represents this situation?

- Which of the following equations represents this situation?

1. *<Product.Material> =<Gender.number> <Product.number>+<Product. Material>* *
2. *<Product.Material> =<Product. Material> <Product.number> +<Gender.number>*
3. *<Product.Material> =<Gender.number> (<Product. Material>)+ <Product.number>*
4. *<Product.Material> =<Gender.number> <Product.number> — <Product. Material>*
5. *<Product.Material> =<Product. Material> <Product.number> — <Gender.number>*

N-layer item model

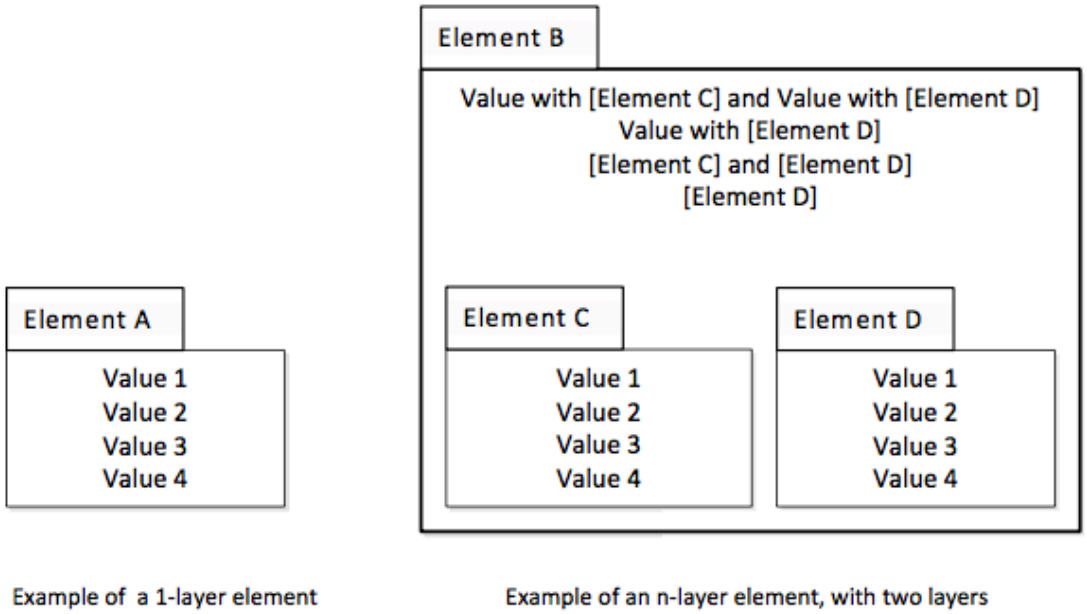


Figure 1. A comparison of the elements in a 1-layer and n-layer item model.

Gierl, M.J and colleagues (2015)

Example 2 – Medical Item

- A 10-year-old, previously healthy boy has a 10-day history of progressive cough, low grade fever and slight dyspnea on exertion. Physical examination shows diffuse rales bilaterally. A chest roentgenogram shows diffuse perihilar infiltrate. The most likely diagnosis is

A. pneumonia due to respiratory syncytial virus

B. pneumonia due to Streptococcus pneumoniae (Pneumococcus)

C. pneumonia due to *Mycoplasma pneumoniae (correct)*

D. congestive heart failure

E. tuberculosis

Example Taken from : LaDuca & colleagues (1986)

Item content

1. Description of the patient and the past medical history
    - A 10-year-old, previously healthy boy has a 10-day history of progressive cough, low grade fever and slight dyspnea on exertion.
2. Symptoms and their duration
    - Physical examination shows diffuse rales bilaterally.
3. Physical findings
    - A chest roentgenogram shows diffuse perihilar infiltrate.
4. Results of diagnostic studies
    - Response options

Response Distractors Rules

- One correct option

1. Any four other options. (Random)
2. Include all other pneumonias. (Constrained)
3. Include only 2 other pneumonias. (Mixed)
4. Key option and distracters do not change across generated items (Fixed)

Table

Type A stem				Type A Choices		
Patient/ PMH	Symptoms/ duration	Physical examination	Diagnostic aids	Response options	Key	Distractors
(0) 10-year-old boy previously healthy	10-day progressive cough, low-grade fever, dyspnea on exertion	Diffuse rales, bilaterally	Chest x-ray: perihilar infiltrate	(A) Pneumonia due to respiratory syncytial virus (B) Pneumonia due to *Streptococcus pneumoniae* (Pneumococcus) (C) Pneumonia due to *Mycoplasma pneumoniae* (D) Congestive heart failure (E) Tuberculosis (F) Pneumonia due to staphylococcus	C	(1) Any four others (2) Include all other pneumonias (3) Include only 2 other pneumonias
(1) Same as 0	10-day progressive cough, 24-hours spiking fever and dyspnea on exertion	Diffuse rales, bilaterally, egophony on R	Chest x-ray: 2 small air fluid levels on R	Same as 0	F	2
(2) Same as 0	Same as 1	Fine crackling rales in R posterior base, with impaired resonance	Chest x-ray: infiltrate in R lower lobe, fluid in R fissure; WBC= 38,000; 96% PMN	Same as 0	B	3

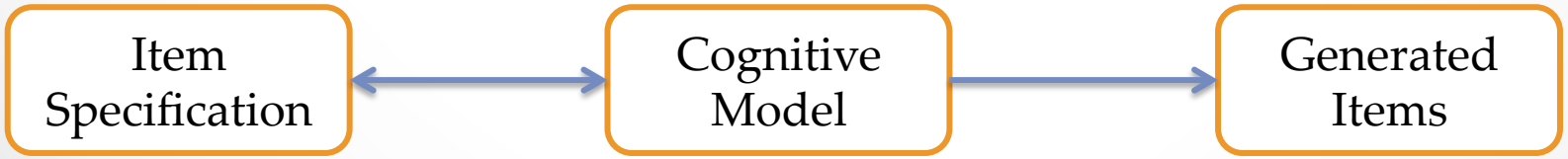
LaDuca & colleagues (1986)

Cognitive Design System

1. Specify goals of measurement.
2. Identify design features in the task domain
3. Develop Cognitive model.
4. Generate items
5. Evaluate models for generated tests.
6. Bank items by cognitive complexity
7. Validation:
    1. Construct Representation
    2. Nomothetic Span

CDS approach

- Item construction process starts with a specification of item-stimulus features (radicals) that can be systematically varied.
- Item stimulus features (radicals) are selected on the basis of a cognitive model that outlines the cognitive processes involved in solving a particular item type.
- Main difference
  - Selection of item stimulus feature is based on a cognitive processing model
  - Integrates cognitive science research and individual difference research



Example

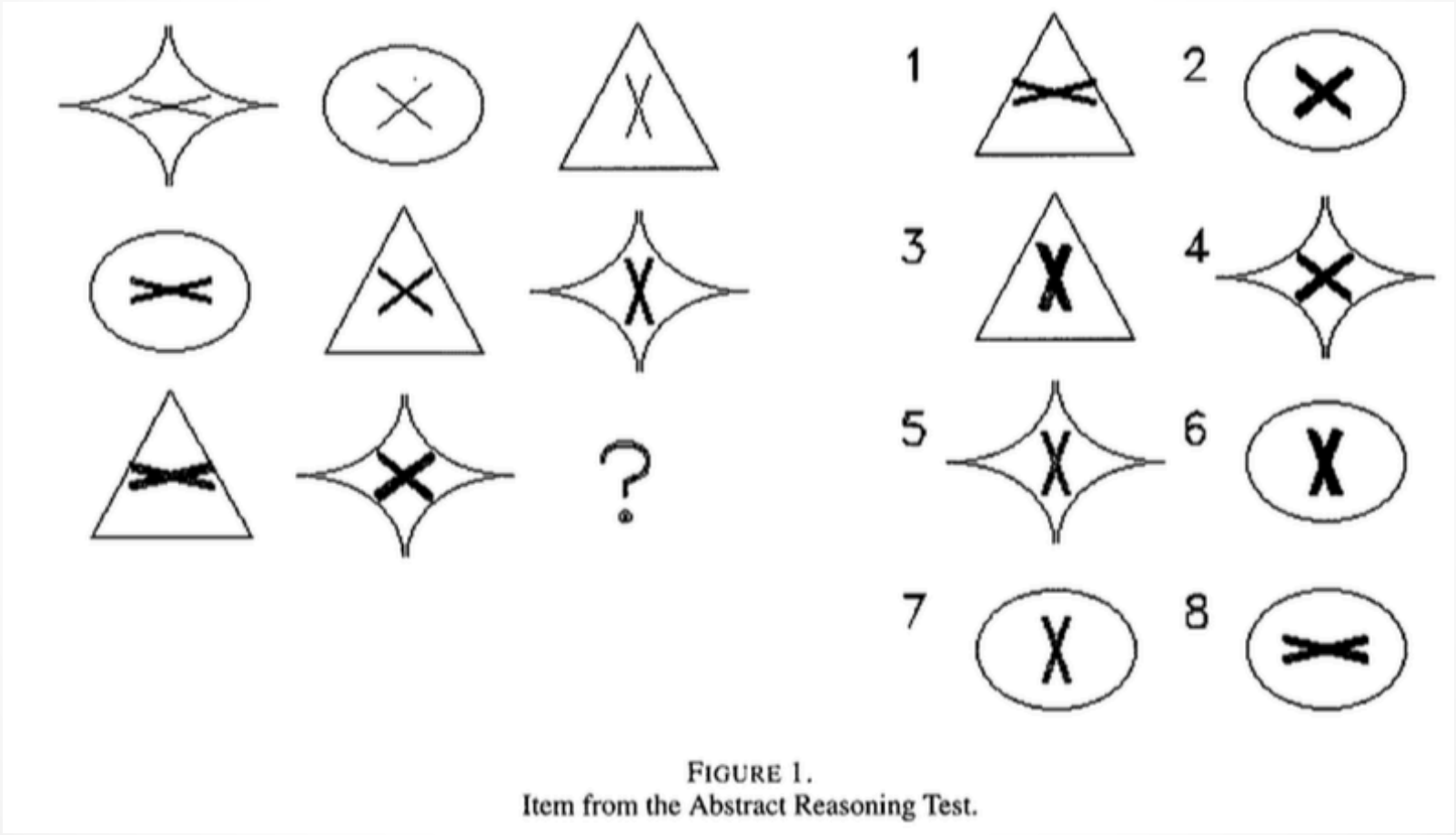
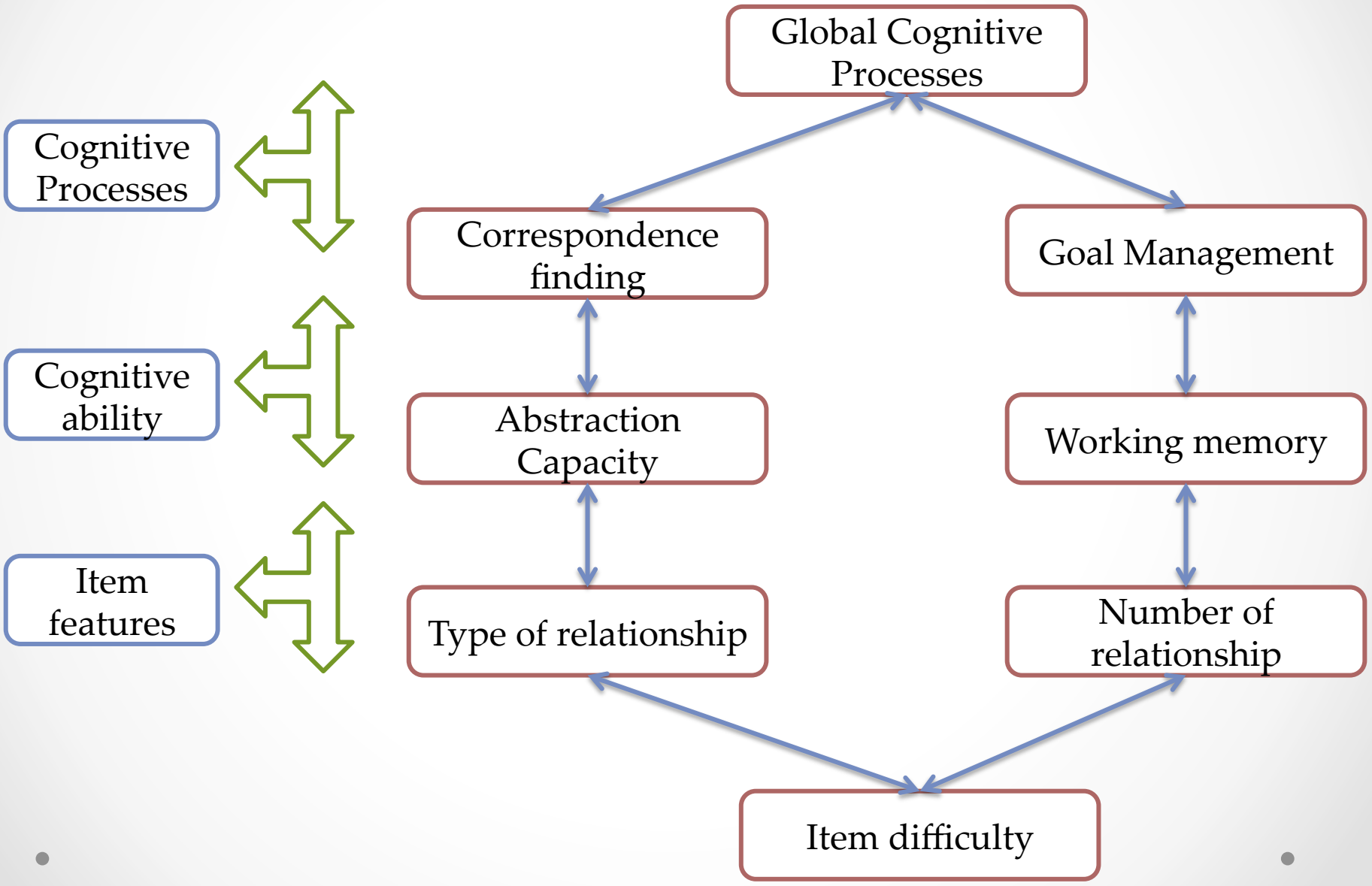


FIGURE 1.
Item from the Abstract Reasoning Test.

Carpenter, Just and Shell's (1990)

Abstract Reasoning items



Cognitive Model

- Cognitive
  - Number of rules
  - Abstract Correspondence
- Perceptual variables
  - Overlay - if objects are overlaid in an array
  - Fusion - if 2 separate objects appear as a single object
  - Distortion - if corresponding objects are distorted

AIGs

- Mental rotation items (Bejar, 1993; Embretson, 1994)
- Progressive matrix problems (Embretson, 1999; Hornke & Habon, 1986
- Hidden Figures (Bejar & Yocom, 1991)
- Mathematical items (Hively, Patternson, & Page, 1968).
- Number Series items (Arendasy & Sommer, 2012)

Construct Validity

Purpose

- AIG or automated scoring is grounded on the constructs we aim to measure rather than the technology per se.

Construct representation

- Concerns the processes, strategies, and knowledge structures that are involved in item solving.
- Aspects of the stimuli are manipulated to vary cognitive demands in a task.
- Mathematical modeling of item difficulty is a major method for such research.

Item difficulty

- Utilized as a method of construct representation (Embtretson, 1983)
- "Construct representation is concerned with identifying the theoretical mechanism that underlie item responses, such as information processes, strategies, and knowledge stores.
- The ability to explain item difficulty underlies the ability to generate items with known difficulty.
- Difficulty modeling can be used to develop and evaluate alternative classification systems.

Nomothetic Span

- Concerns the relationship of test scores to other measures.
- Consists of individual differences correlations across variables.
- A strong system of hypotheses generated from construct representation research should guide nomothetic span research.

Empirical Methods

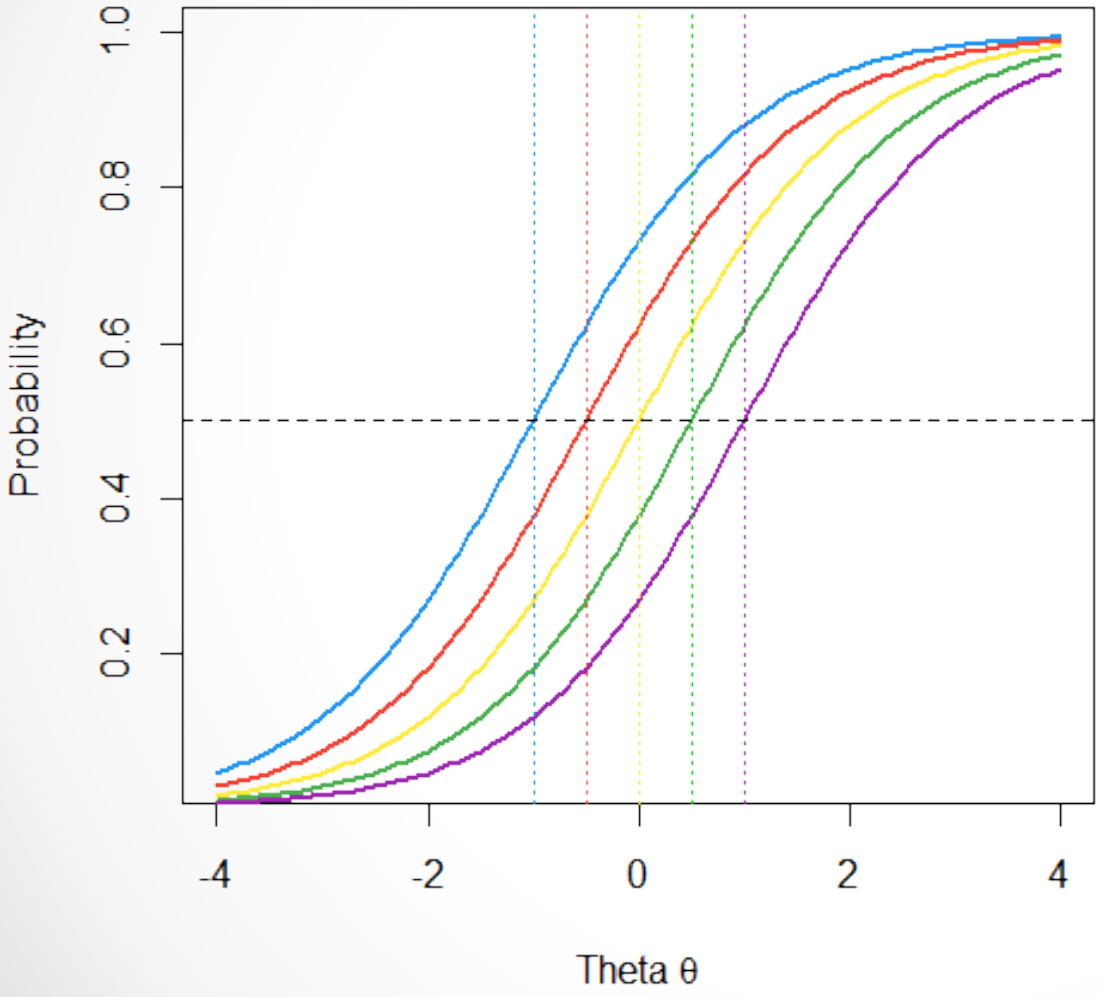
- Regression Models
- IRT (Rasch Model)
- Logistic Linear Test Model
- Multicomponent Latent Trait Modelling
- HIRT (Hierarchical IRT)
- Cognitive Diagnostic Assessment

Rasch Model

$$P(x_{ij} = 1) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}}$$

Θ = theta
b = item difficulty

1 Parameter Logistics Model (1PL)
Item 'difficulty' (b) parameter



a	b
1	-1
1	-0.05
1	0
1	0.05
1	1

CONCERTO

Easy item

$$P(x_{ij} = 1) = \frac{e^{0-0}}{1 + e^{0-0}} = .50$$

Difficulty Item

$$P(x_{ij} = 1) = \frac{e^{2-2}}{1 + e^{2-2}} = .50$$

Linear Logistic Test Model (LLTM)

LLTM

$$P(x_{ij} = 1) = \frac{e^{\theta_j - b_i}}{1 + e^{\theta_j - b_i}}$$

Linear Constraint on item difficulty

$$b_i = \sum_m \eta_m q_{im} + a,$$

$\eta\downarrow m$ = difficulty of the processing operation m
$q\downarrow im$ = number of operators of type m
a = normalisation constant

Example

Q) 90, 90 , 90 , 90 , 58 , 58 , 58 , __ , __

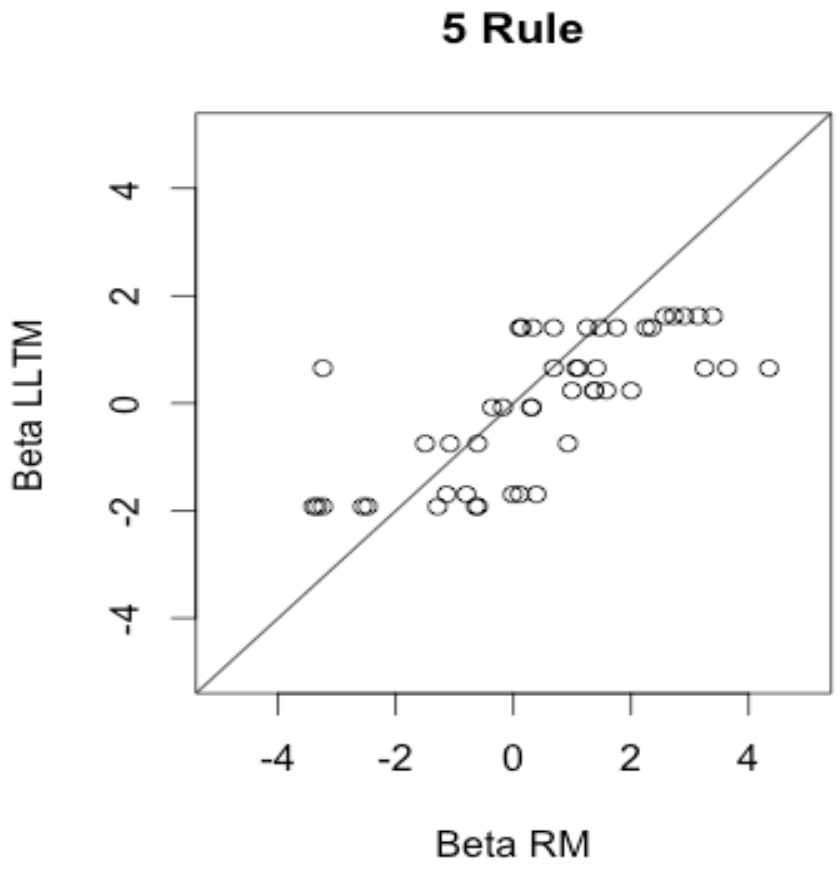
Q) 22, 280, 23, 290, 24, 300, 25, 310,__ , __

Q) 5, 67, 108, 36, 126, 167, 67, 102, __, __

Q-Matrix

Table 4. Q-Matrix of the cognitive operators.					
Item	Apprehension of succession	Parallel sequences	Categorisation	Non-progressive coefficient patterns	Progressive coefficient patterns
1	1	0	0	1	0
2	0	1	0	0	1
3	0	1	0	0	1
4	1	0	0	0	1
5	1	0	0	0	1
6	0	0	1	1	0
7	1	0	1	0	0
8	0	1	0	1	1
9	0	0	1	1	1
10	0	1	1	1	0
11	0	1	1	1	0

Rasch vs LLTM



Correlational relationship

Multi-Component Latent Trait model

- The model assumes that information to solve the item is derived from several component events.
  - Independent component events
  - Sequential dependent events
- Components are identified from subtasks that represent the processing components in solving the item.
- MLTM uses subtask data to identify the components.
- Participants respond to the total item, as well as the series of subtasks that represents the processing component.

Whitely (1981)

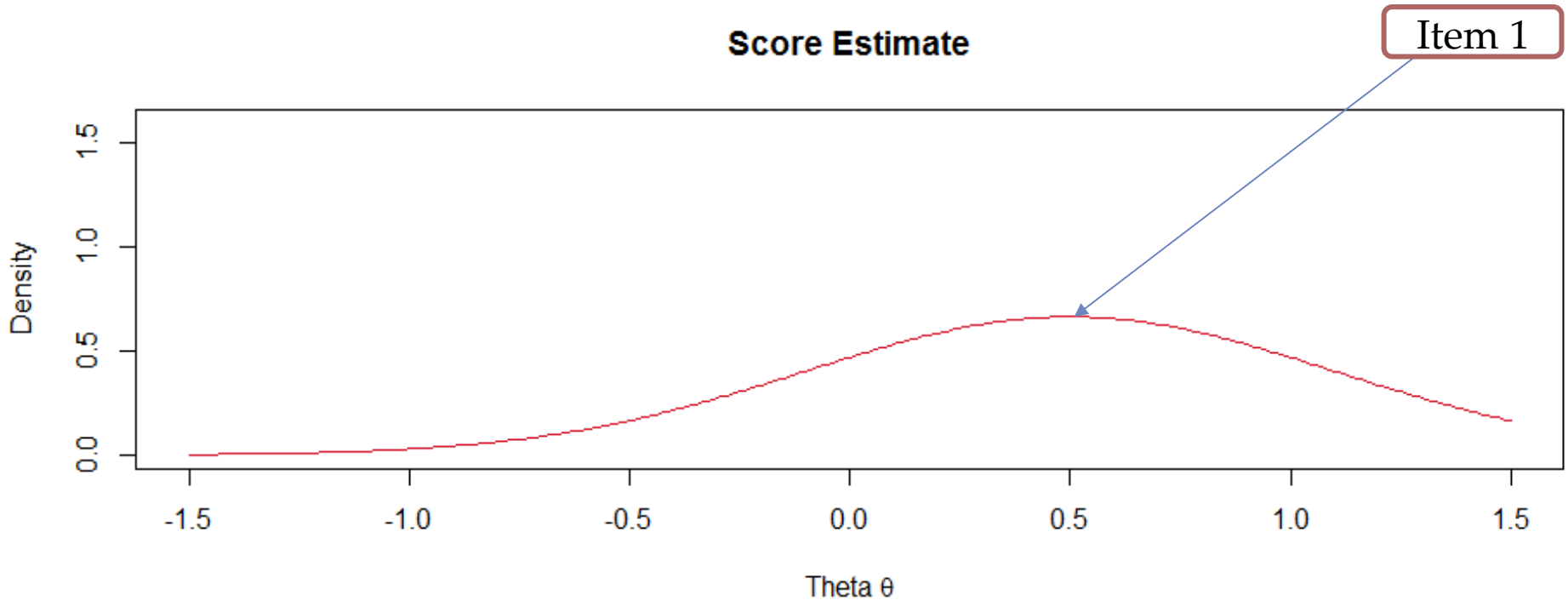
LLTM vs MLTM

- LLTM and MLTM are different.
- LLTM estimates difficulty of complexity factors that are related to item difficulty.
- MLTM estimates item and person parameters for component outcomes.

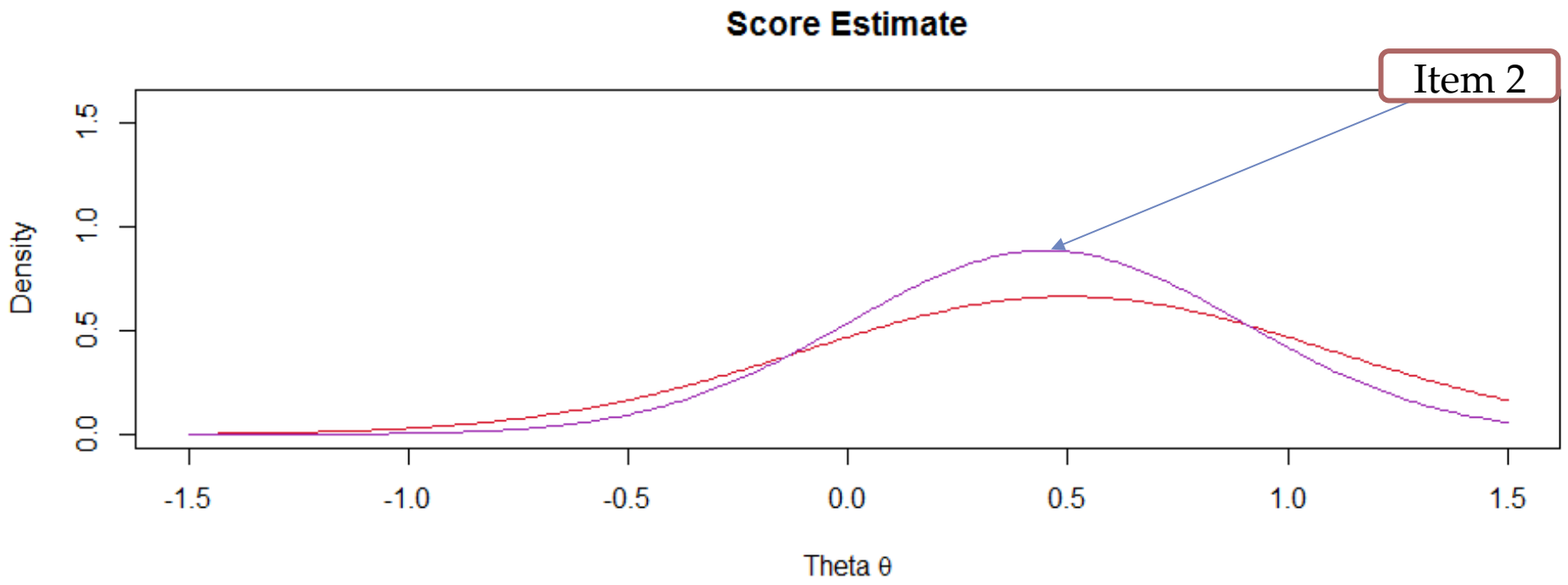
Computer Adaptive Testing

- Intelligence testing
  - Generate optimally informative item for the examinee during the test
- Optimally informative item
  - Based on the previous pattern of the examinee's response
- Psychometric methods for adaptive testing
  - Intelligence measurement
  - Adaptive item selection leads to shorter and more reliable tests

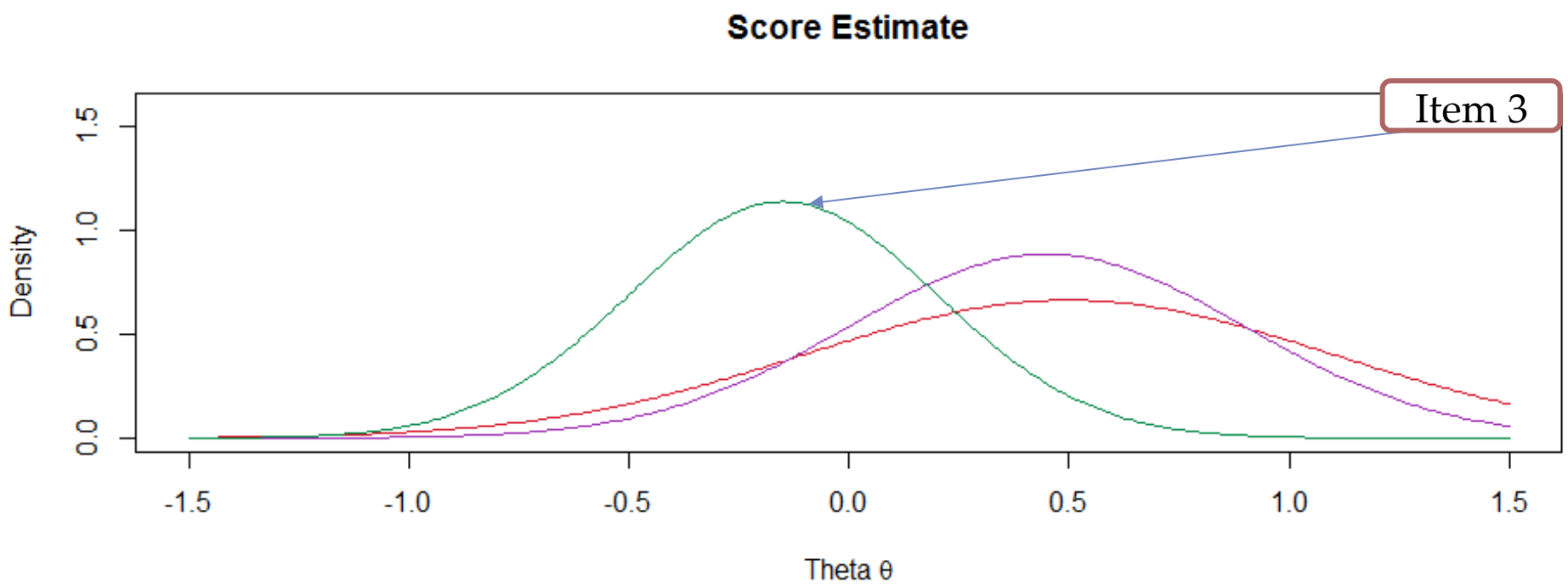
Score estimate during CAT (Liklihood function)



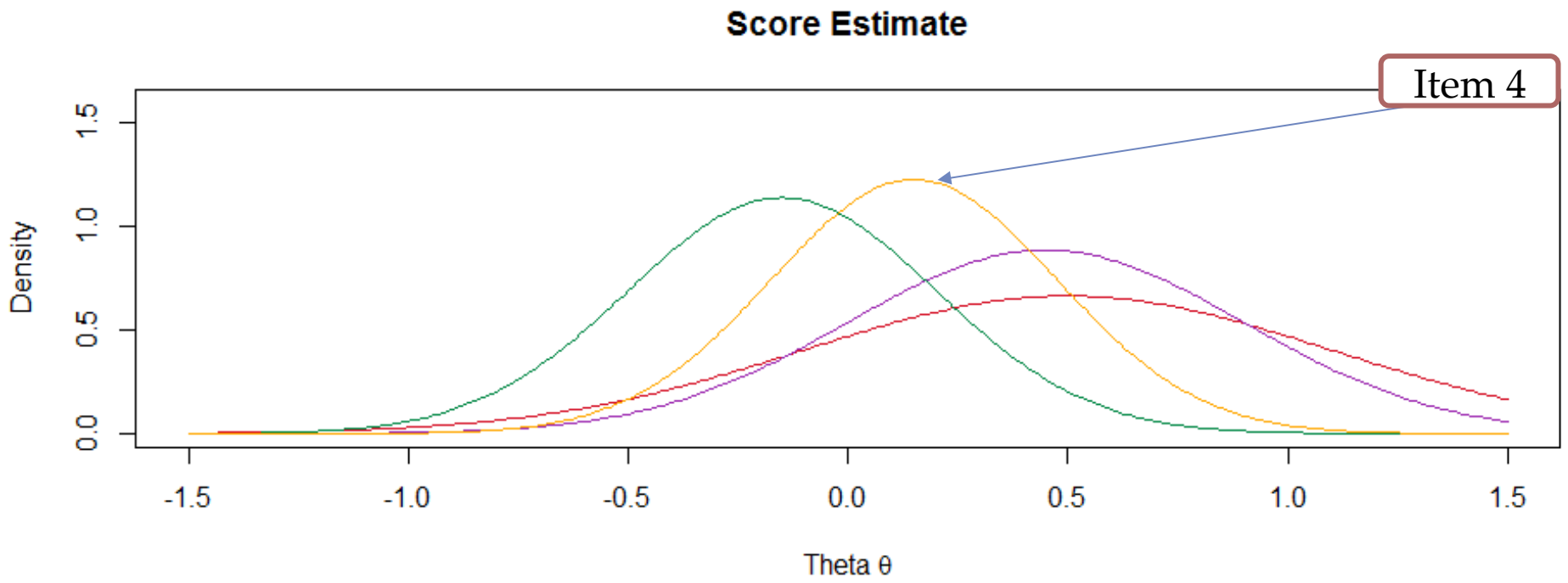
Score estimate during CAT



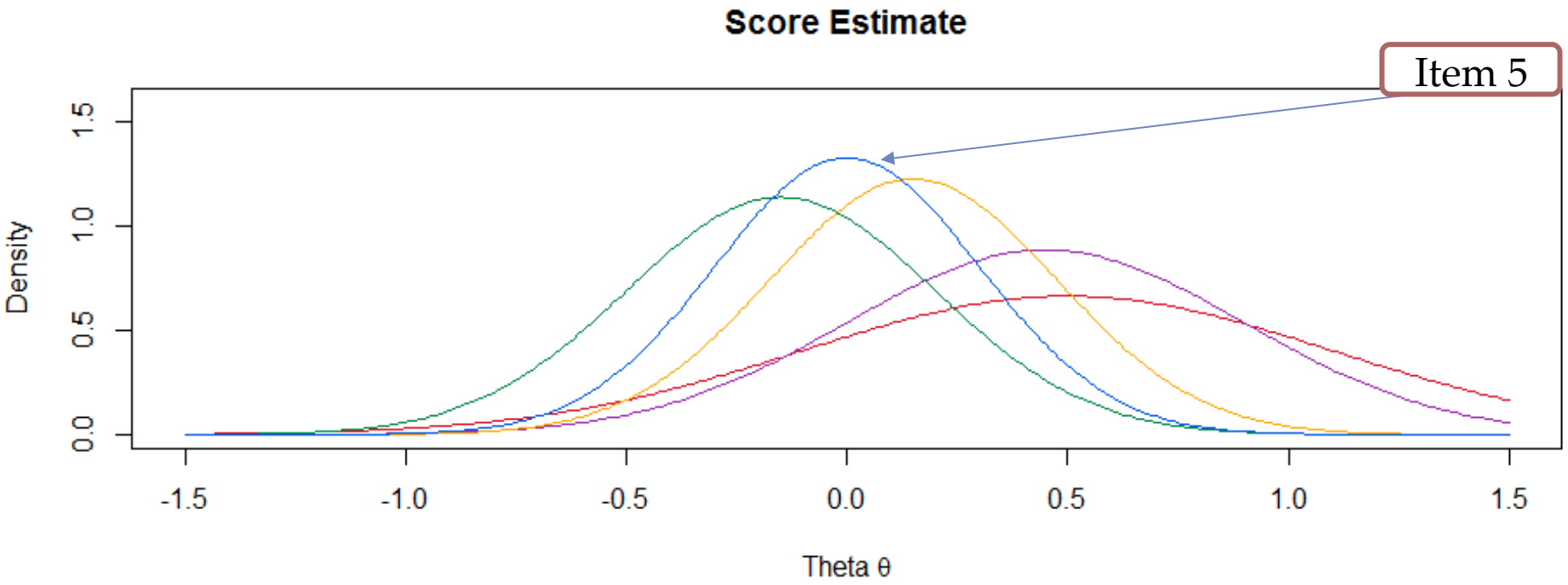
Score estimate during CAT



Score estimate during CAT



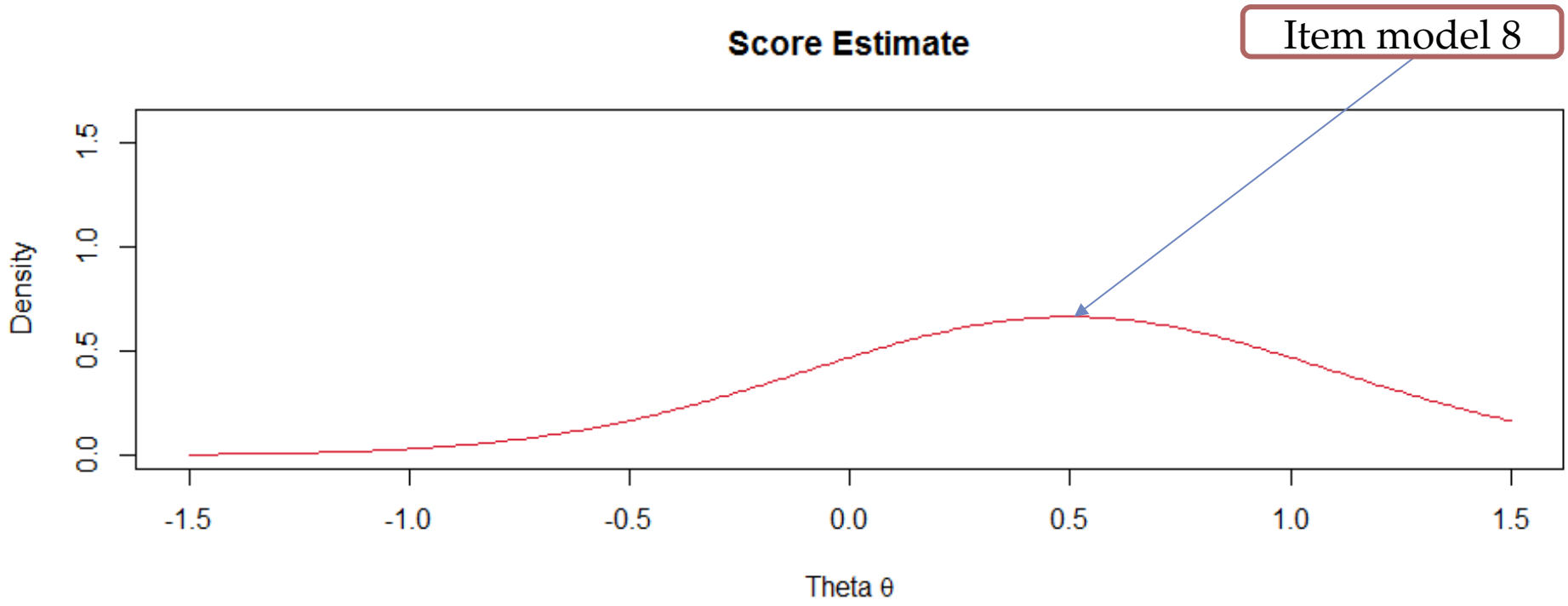
Score estimate during CAT



Adaptive Item Generation

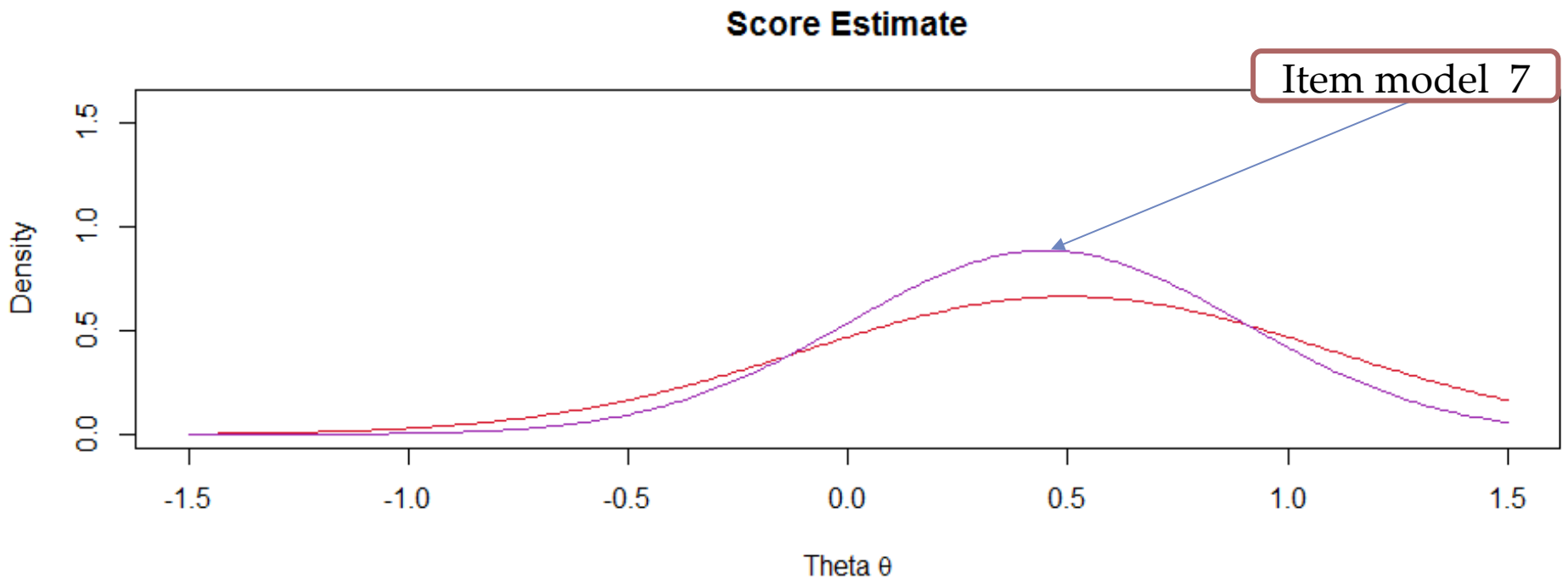
- A cognitive analysis of items
  - Knowledge is required of how stimulus features in specific items impact the ability construct
- CDS Approach to Adaptive Item Generation
  1. Theoretical Foundations of Item models
  2. Supporting Developments
     - Psychometric models : Construct validity
- Supporting Data for CDS (Publication)
  1. Initial Cognitive Model for Matrix Items
  2. Algorithmic Item Generation and Reversed Cognitive Model
  3. Item Generation by Artificial Intelligence
  4. Empirical Tryout of Item Generation

Score estimate during CAT (Liklihood function)

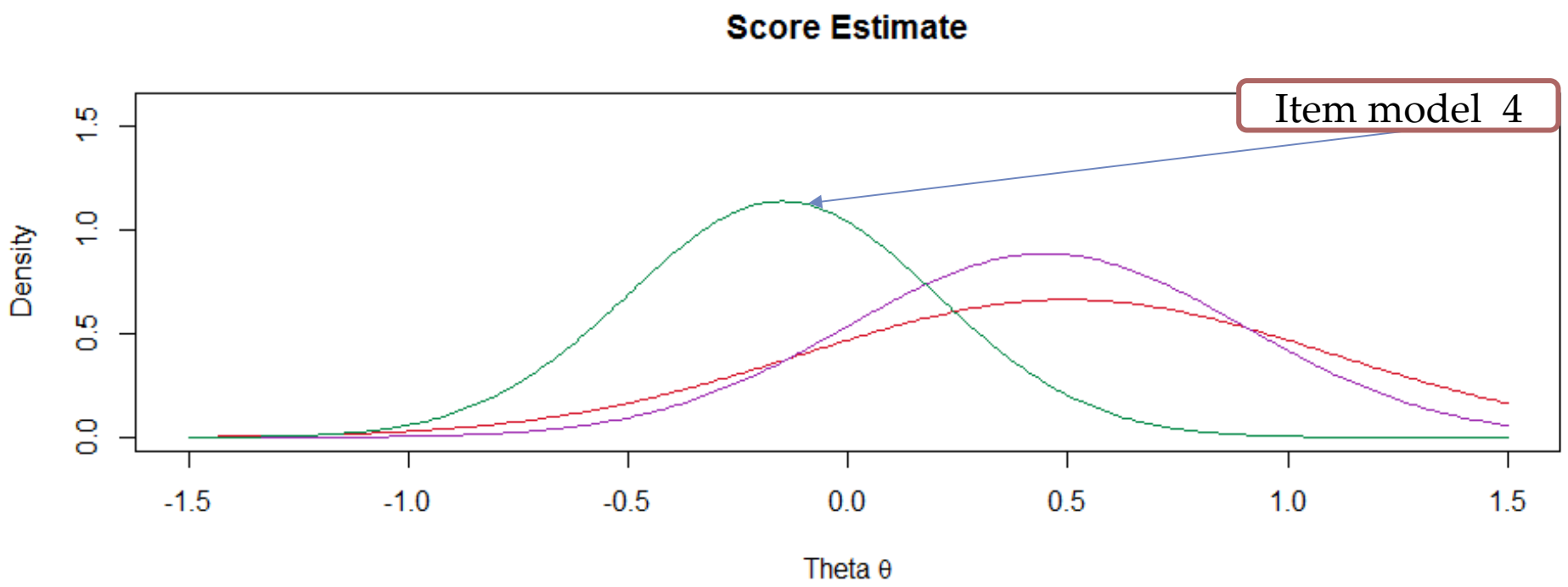


An item is randomly selected from the item model

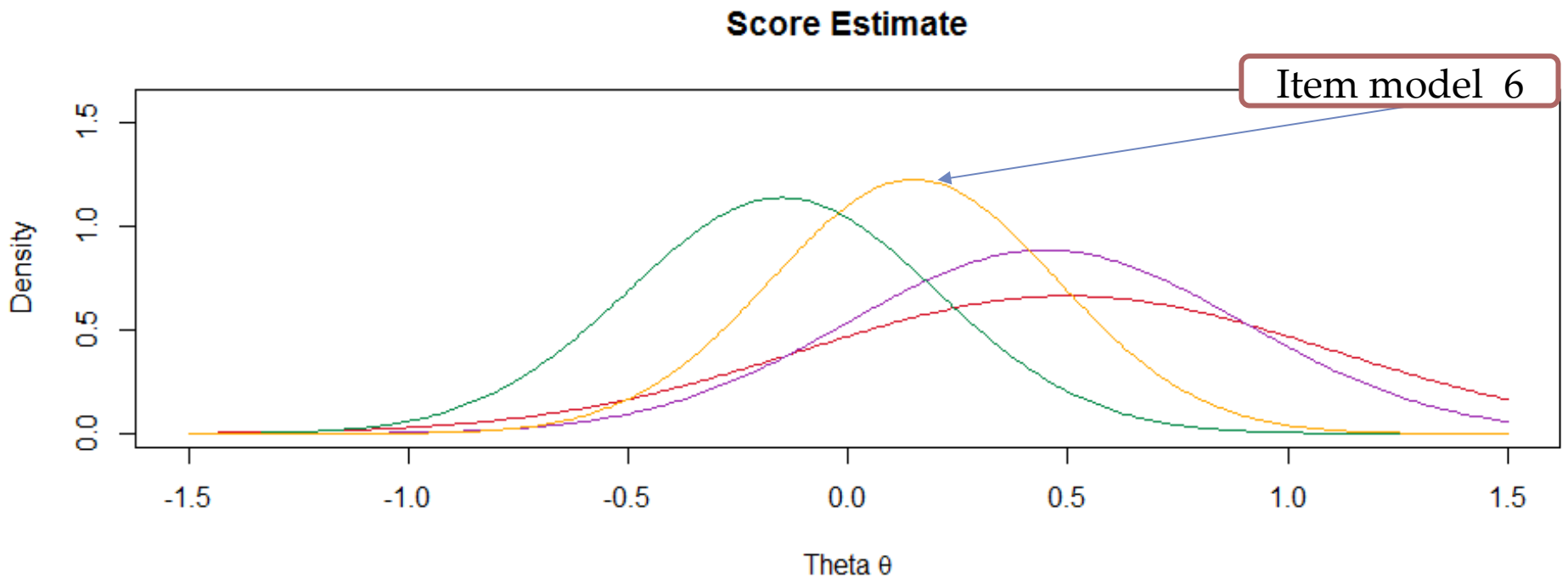
Score estimate during CAT



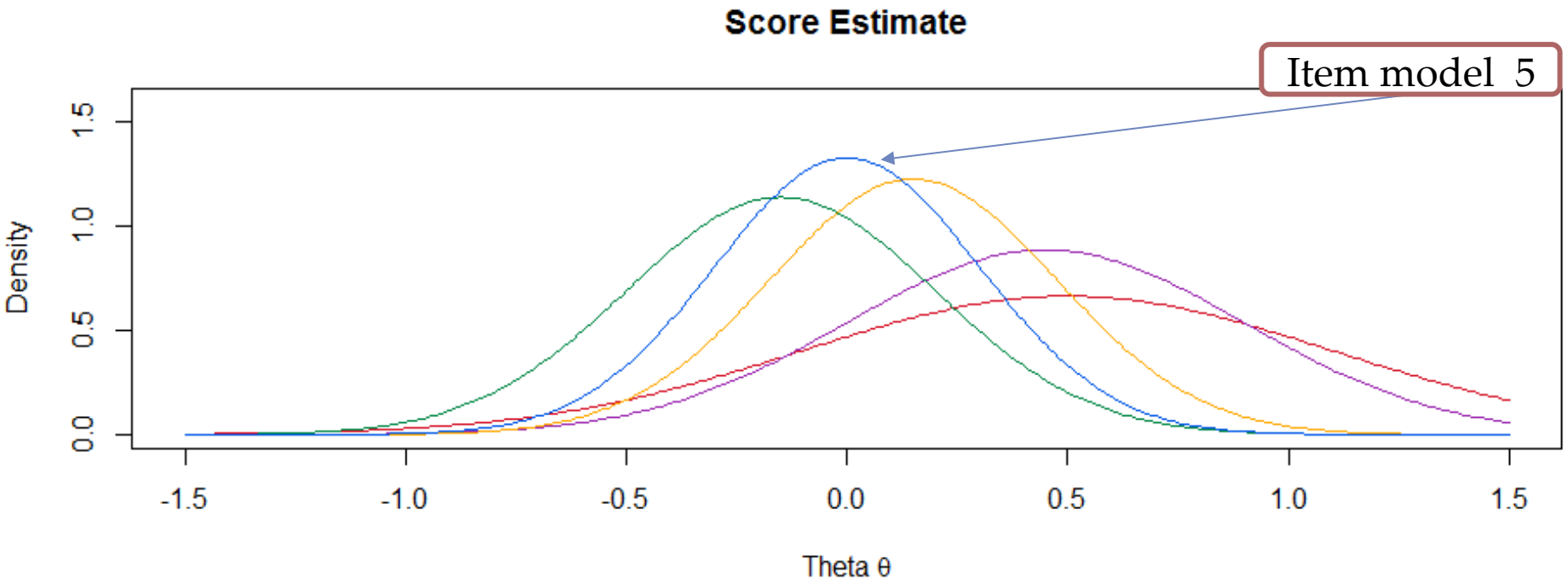
Score estimate during CAT



Score estimate during CAT



Score estimate during CAT



Adaptive Item Generation

- An item generator program (R-package)
  - Produces Item model blue print.
  - Select item models of targeted cognitive complexity.
  - Create items with predicted difficulty.
- An adaptive testing program that can be interfaced with the generator. (Concerto)
  - Display items adaptively.
  - Record responses.

References Upon Request

Thank you