Item Response Theory: Basic Notions

Norman Verhelst
Eurometrics, The Netherlands

EALTA Webinar
January 13 2017

Overview

- Some problems with Classical Theory
- The structure of a theory
- The prototype of IRT: Scalogram analysis
- The logistic function (a bit of mathematics)
- The Rasch model
- Our tasks
- What to do if our hypothesis fails
- The Partial Credit Model
- Closing the circle

An interesting, yet often ignored aspect of classical theory

- An old test, X, is replaced by a new test, Y
- An interesting question: do X and Y measure the same construct?
- Possible answer: if they do, then the correlation between both tests $\rho(X,Y)$ must equal one.
- Objection: What we observe (scores) is blurred (polluted) by measurement error, and these errors will suppress (‘attenuate’) the correlation

Correction for attenuation

$$\rho(T_X, T_Y) = \frac{\rho(X,Y)}{\sqrt{\rho(X,X')\times\rho(Y,Y')}}$$

Example: $\rho(X,Y) = 0.70$

but $\rho(X,X') = 0.85$ and $\rho(Y,Y') = 0.60$

So, we find: $\rho(\mathrm{T}_X, T_Y) = \dfrac{0.70}{\sqrt{0.85\times 0.60}} = 0.98$

Some weak points of classical theory

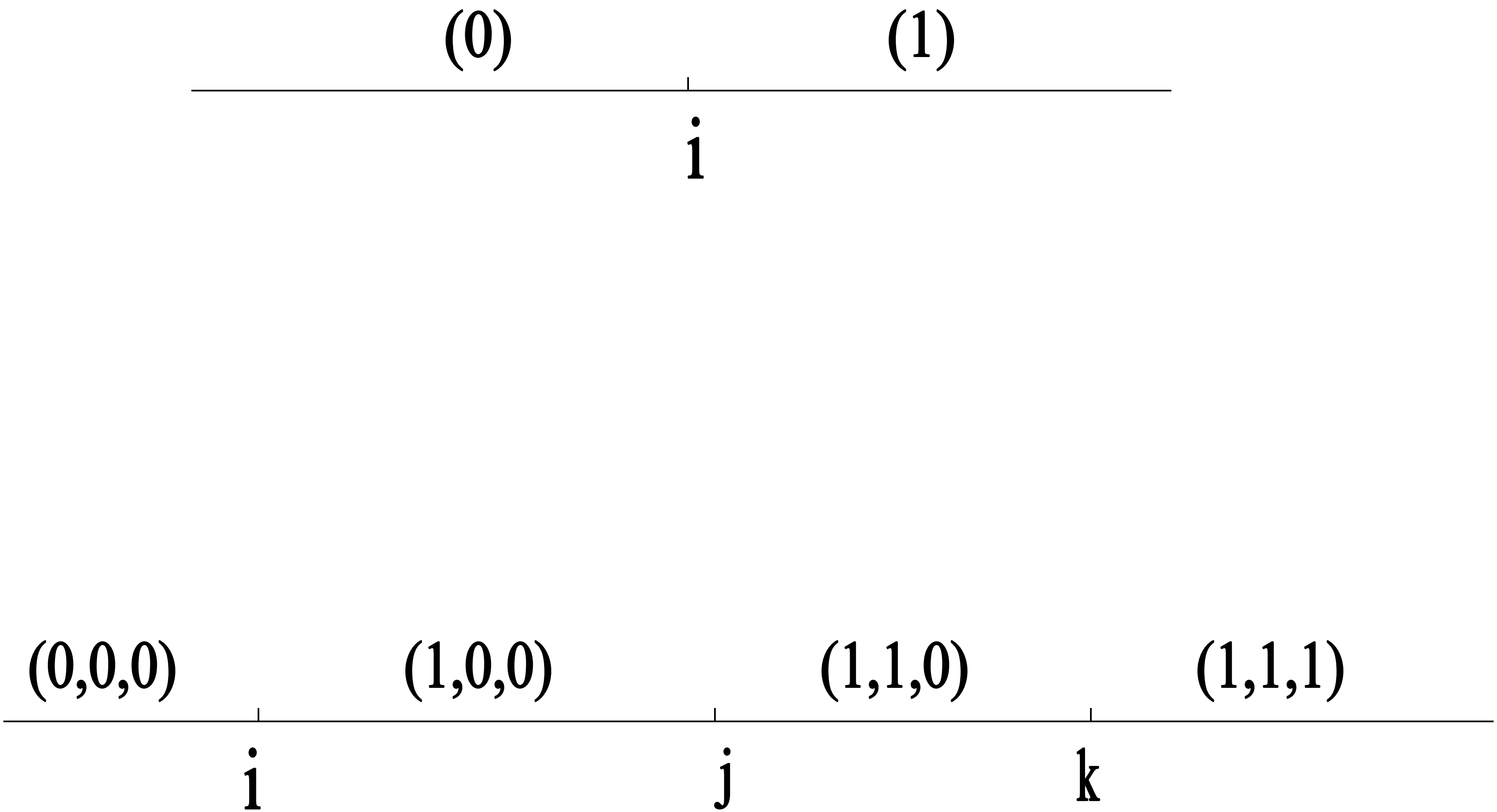
- The construct plays a minor role; central concept is the score
- Indices of difficulty and discrimination are population dependent
- Comparison of scores on different tests is very difficult and often impossible
  - Example: at university P students of faculties *a, b* and *c* take a test of English in March. In May students of faculties *d, e* and *f* take another exam of English
  - How can we compare the performance of faculties *a* and *d*?

The structure of a theory

- A theory is a narrative about the observable world
  - Using concepts and relations between concepts
  - Clarifying the relation between concepts and observable phenomena
  - Imposing restrictions on ‘possible’ phenomena
  - Falsifiable
- Is Classical Test Theory a theory in this sense?

Guttman’s scalogram

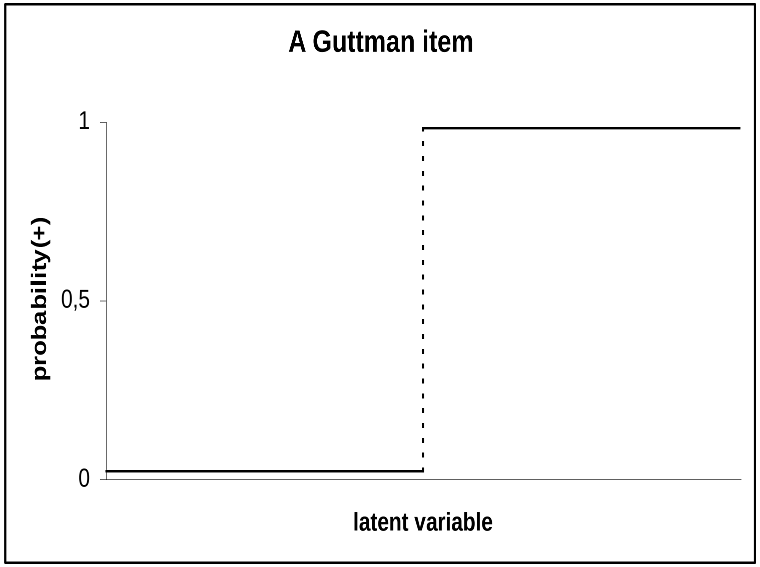
- The targeted concept is represented by a line (continuous variable) (for example, an attitude, a competence)
- Persons and items are represented by points on the line
- The observable responses reflect the ordinal relationship between the points ‘person’ and ‘item’ on the underlying line. This ‘line’ is non-observable or latent


(0)
(1)
i
(0,0,0)
(1,0,0)
(1,1,0)
(1,1,1)
i
j
k

i	j	k	m
0	0	0	0
1	0	0	0
1	1	0	0
1	1	1	0
1	1	1	1

#items	possible	allowed
4	16	5
n	2^n	n+1

Guttman items



- Not-decreasing
- Not continuous (‘step function’)
- The model is deterministic

A bit of algebra: exponentiation

$$3^5 \times 3^7 = 3^{5+7} = 3^{12}$$

$$\text{Definition: } 3^{-5} = \frac{1}{3^5}$$

$$1 = 3^5 \times \frac{1}{3^5} = 3^5 \times 3^{-5} = 3^{5-5} = 3^0 = 1$$

$$\boxed{\begin{array}{l}\text{For any positive number } c, \\ \text{it holds that } c^0 = 1\end{array}}$$

$$3^5 : \begin{cases} 3 \text{ is the basis} & \text{(positive)} \\ 5 \text{ is the exponent} & \text{(arbitrary number)} \end{cases}$$

The number *e*

- A famous number which deserves its own name.
- The basis of the natural logarithms
- $e = 2.7182818\ldots$
- The exponentiel function (see ‘webinar2.xls’, page ‘exp-function’):

$$y = e^x$$

$$y = \exp(x)$$

The logistic function

$$f(x) = \frac{e^x}{1+e^x}$$

$$f(x) = \frac{\exp(x)}{1+\exp(x)}$$

"the exp of something (x) divided by one plus the exp of the same something"

The Rasch model (1)

- Definition of the item response function:

$$f_i(\theta) = P(X_i = 1 \mid \theta)$$

- X_i stands for 'the score on item *i* '
- β_i is a non-specified number (parameter)

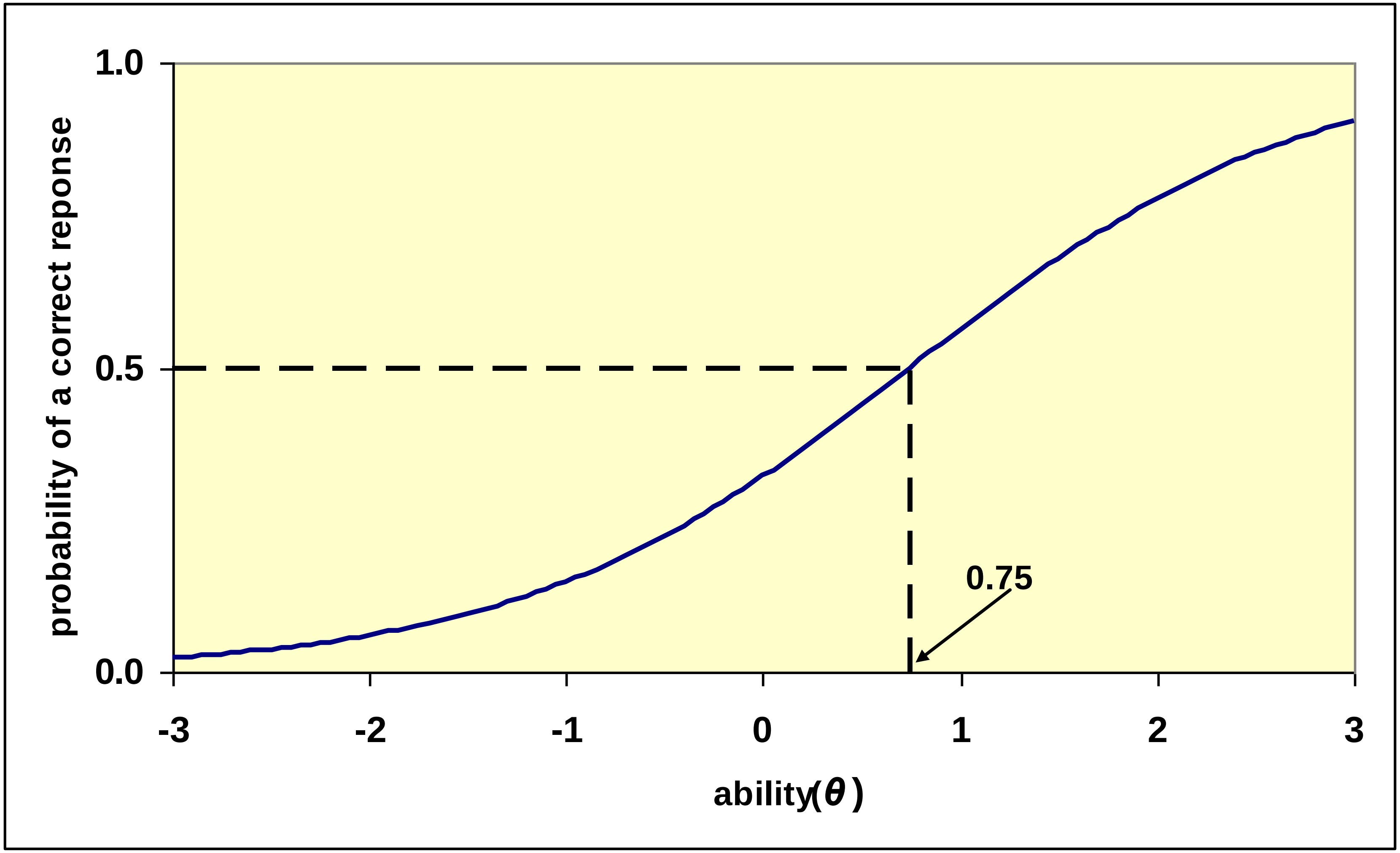
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

The Rasch model (2)

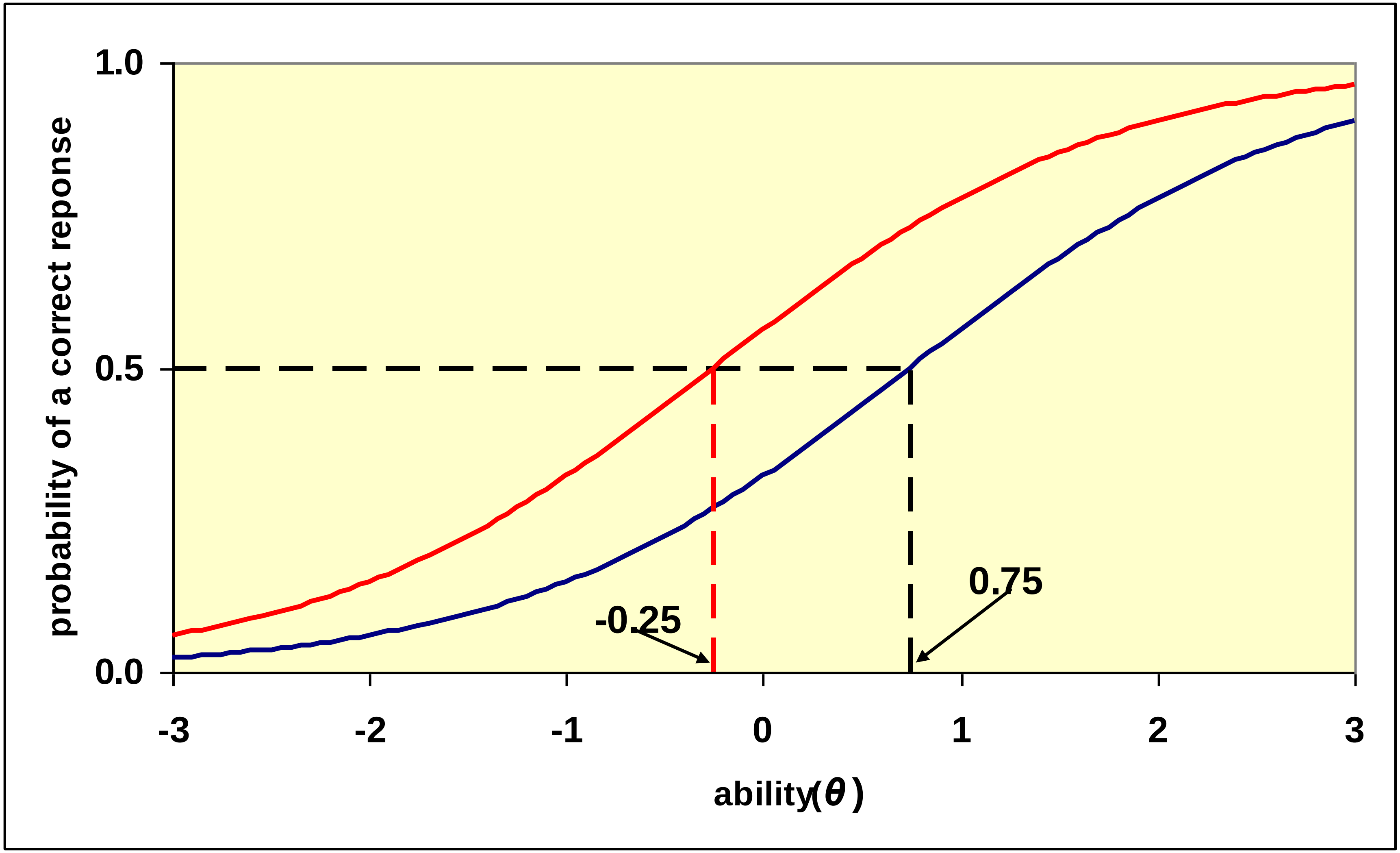
$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

- θ is the symbol for the latent variable
- Each item has its own function; hence f_i
- β_i is a non-specified number (parameter)
- If $\theta = \beta_i$, or $\theta - \beta_i = 0$ (and remember exp(0) =1), then $f_i(\theta) = ½$
- β_i is 'the amount of ability' needed to grant a probability of exactly ½ for a correct response.
- The more ability needed, the more difficult the item.
- β_i is called the **difficulty parameter**

The item response function for an item with difficulty parameter 0.75



The item response functions for two Rasch items



Our tasks

- What do we have?
  - A narrative (hypothesis)
  - A binary table with realisations of X_{vi} (0 or 1)
- What do we have to do?
  - Estimate the parameters
  - Check the narrative
  - Accept or reject the narrative
  - Use the test: go and measure

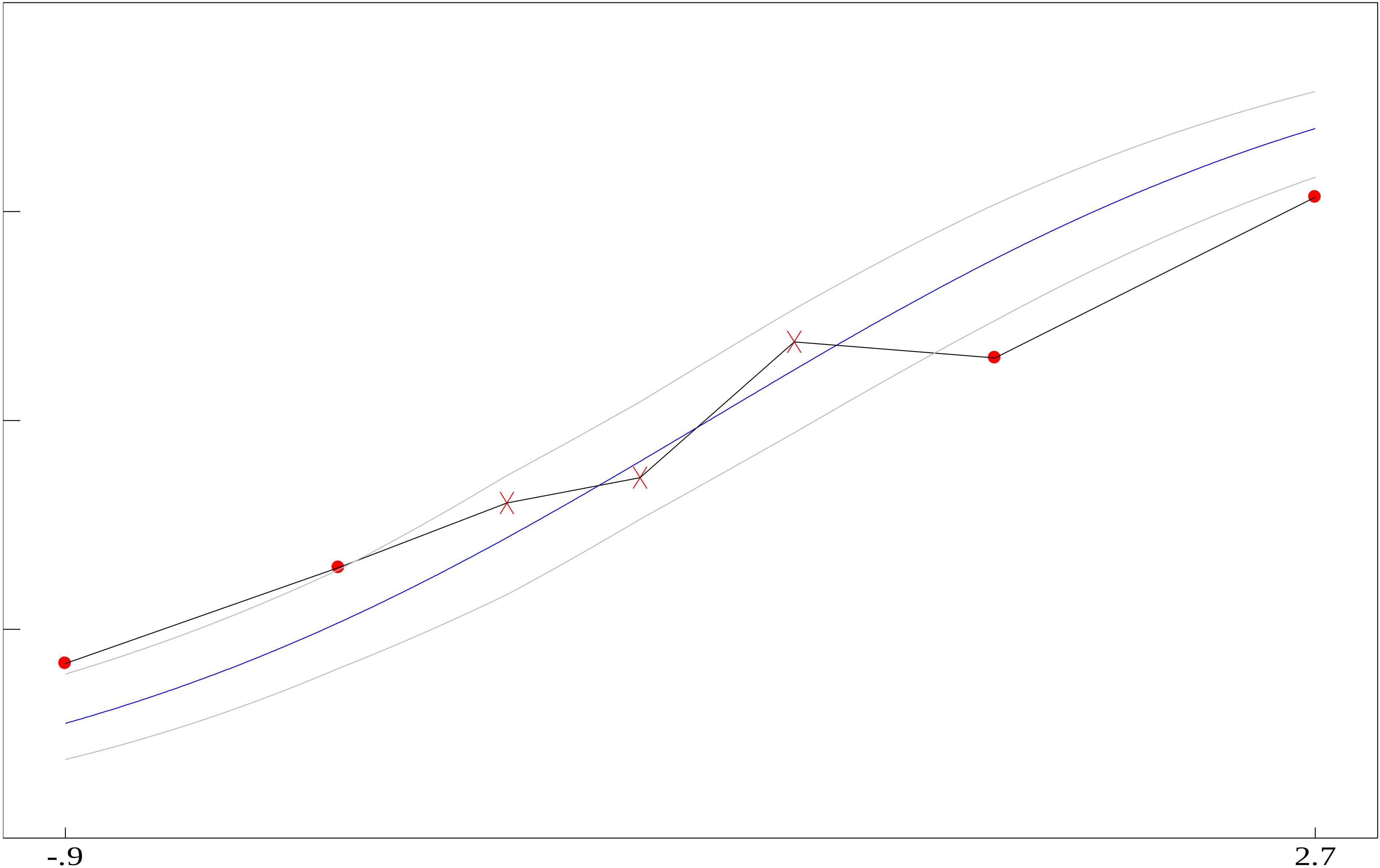
Parameter estimation

- Is a technically difficult problem
- Still a controversy about some methods
- The non-technical user can use public software, but beware…
- The matter is too complex to be discussed in a webinar
- This shows the necessity of inter-disciplinary cooperation

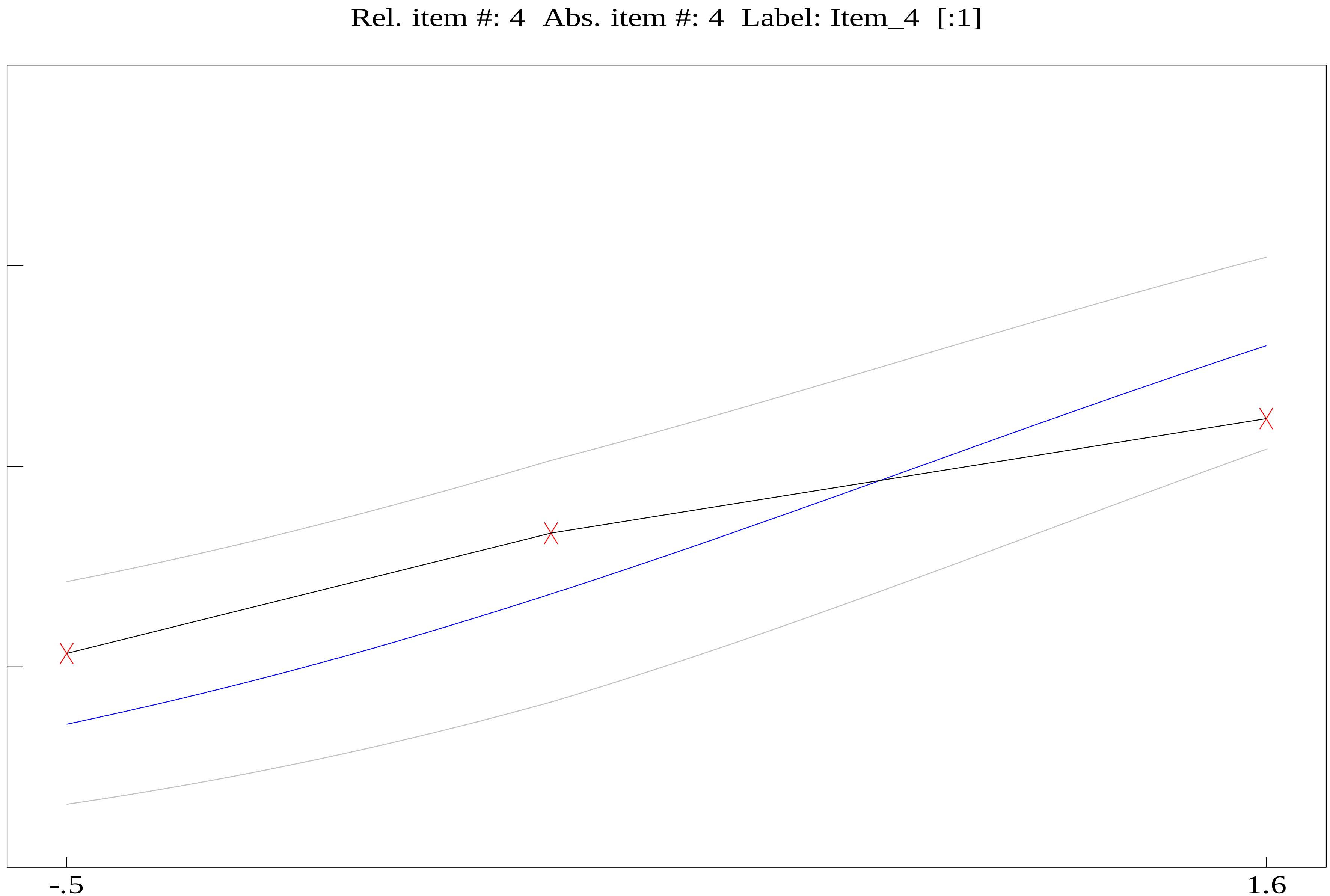
Check the narrative

- As the model is probabilistic, testing is not straightforward.
- Often statistical tests are used.
  - We need to understand the logic of statistical tests
- Graphical aids are helpful too

Rel. item #: 4 Abs. item #: 4 Label: Item_4 [:1]



with n = 133 instead of n = 1332



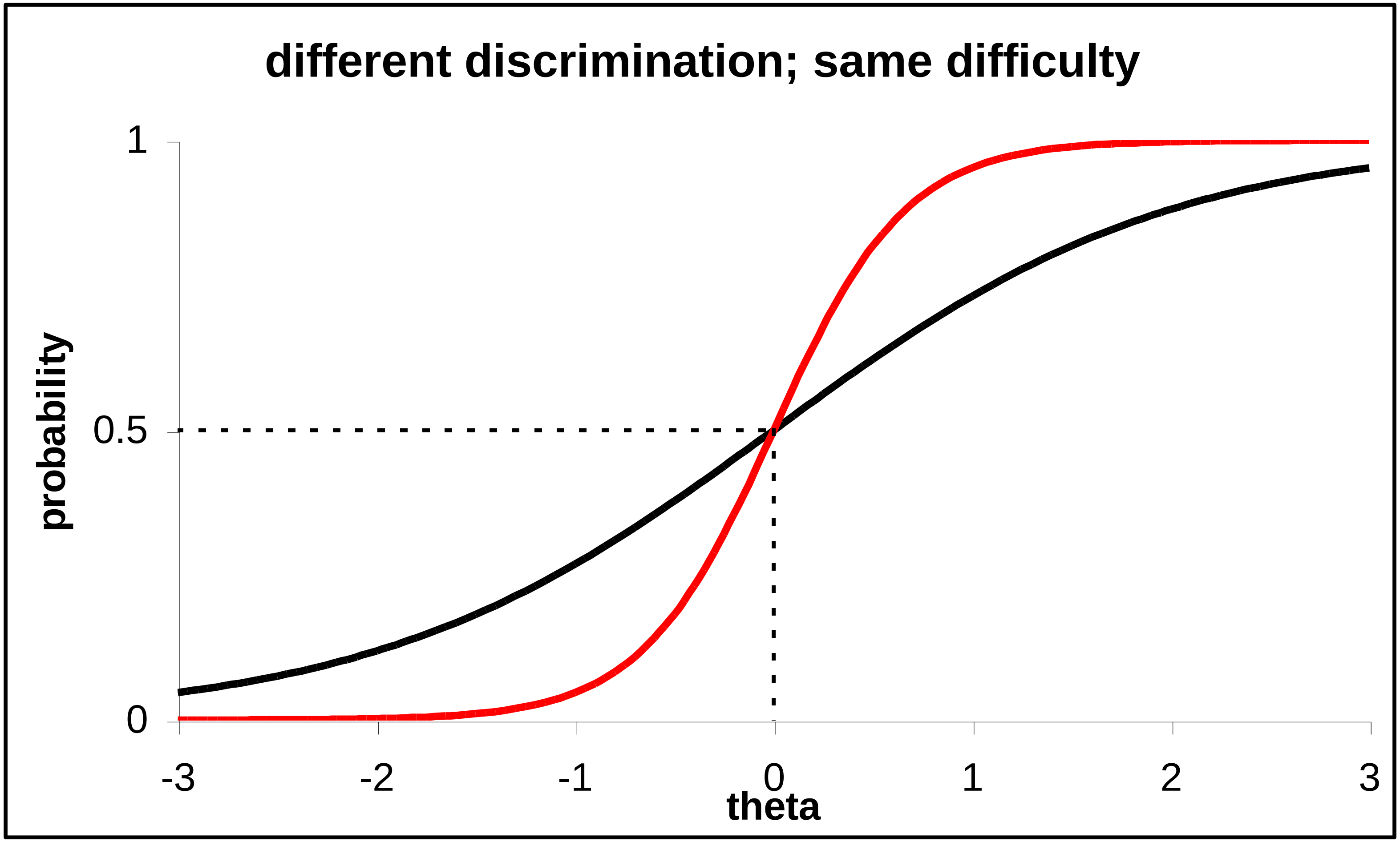
Power of a (statistical) test

- The probability that a deviation from the narrative will be discovered is dependent
  - On the seriousness of the deviation (which we do not know)
  - On the sample size (which is under our control)
- This probability is called the statistical power
- Discovered: yield a significant result
  - In every-day language: will ring the alarm
- The deviation from the narrative in the example has to do with discrimination
  - Item 4 discriminates less well than assumed by the narrative

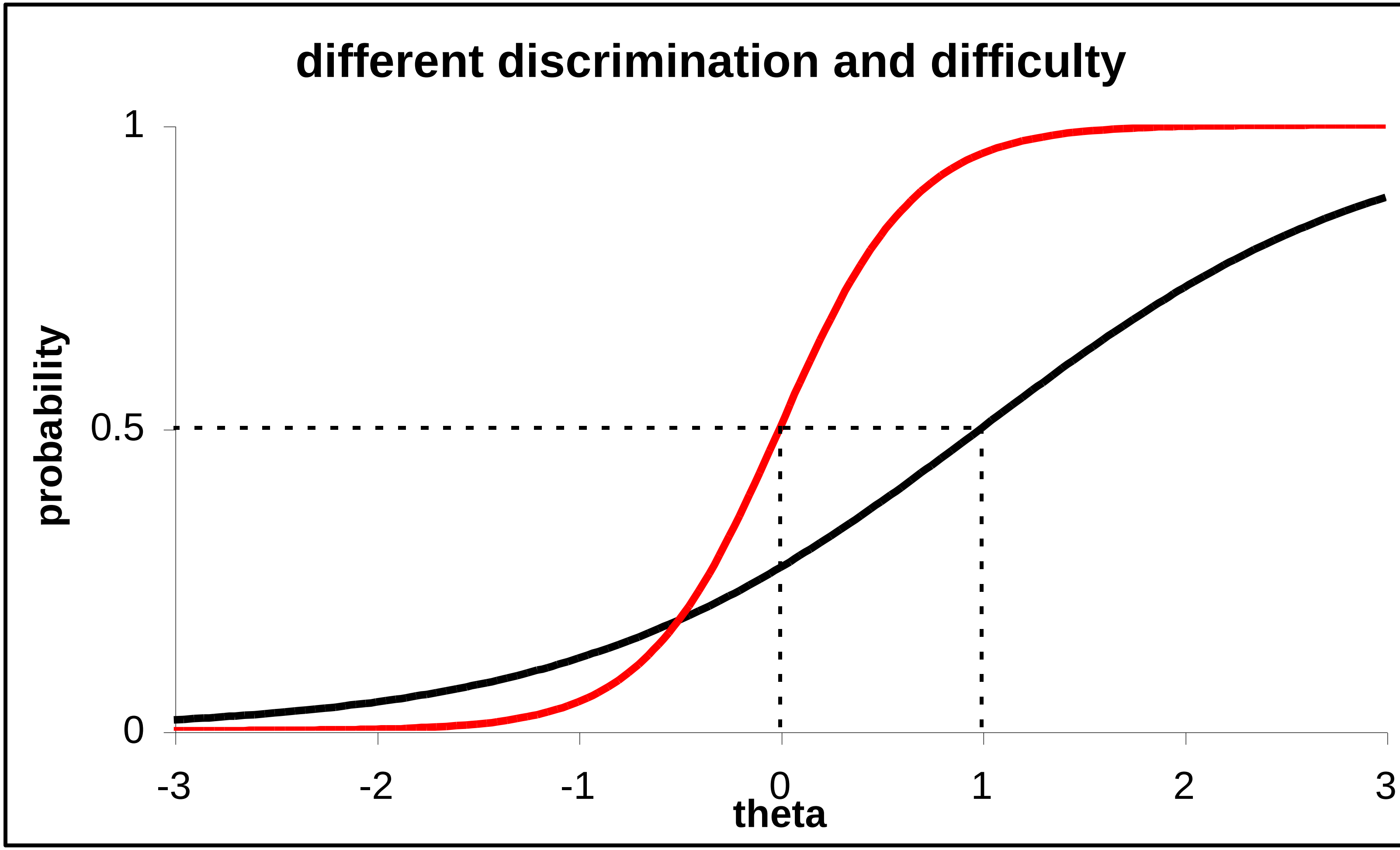
The narrative revisited

- The Rasch model as a narrative assumes
  - That all items discriminate equally
  - That the latent trait θ is unidimensional
  - That there is local stochastic independence
- In most software, the statistical tests have little power if unidimensionality or local independence are not true.

Discrimination (1)



Discrimination (2)



Discrimination and difficulty

$$f_i(\theta) = P(X_i = 1 | \theta) = \frac{\exp[a_i(\theta - \beta_i)]}{1 + \exp[a_i(\theta - \beta_i)]}, \quad (a_i > 0)$$

This is known as the two-parameter logistic model (2PLM)

Rasch model: $a_1 = a_2 = \cdots = a_k$

Local Independence: an example(1)

- Population: children in the age 6 to 14
- Test X: size of the feet
- Test Y: score on a reading comprehension test
- What is the sign of the correlation $\rho(X,Y)$ in this population?
  - Negative?
  - Zero?
  - Positive?
- Why?

Local Independence: an example(2)

- The correlation is positive because
  - The older the children, the bigger their feet
  - The older the children, the better they read
- The variation in age ‘explains’ the correlation
- Proof: in a population of children of the same age (i.e., local) the correlation will vanish

Local independence in IRT

- In an arbitrary group (e.g., a class), the correlation between the answers on item *i* item *j* is (usually) positive, i.e., $\rho(X_i, X_j) > 0$
- In a population of students with the same value of *θ,* this correlation is zero, i.e.,
$$\rho(X, Y \mid \theta) = 0$$
- This is an assumption, and the test to find out if it is tenable is very difficult because we cannot form a group of people with the same ability (ability is latent, i.e., not observable)

Another look at conditional independence

- Given the latent variable, the probability of a correct answer must not depend on the answer to another item.
- Remember multiple matching items: if Tallin is assigned to Portugal, the probability of a correct answer for Lisbon is zero

Iceland	Tallin
Portugal	Bucharest
Roumania	Sofia
Bulgaria	Lisbon
Estonia	

Multidimensionality

- Example 1: a test of ‘communicative competence’ containing Reading items and Listening items
  - Probably better to analyse Reading and Listening separately
- Example 2: a Reading Comprehension test, consisting of (say) six text fragments and 5 questions per fragment.
  - In a fragment about history, students with more interest in or knowledge of history will probably do better than students weaker in this respect, even if their reading ability is equal

The testlet or task problem

- A set of items belonging together for some obvious reason (matching, text fragment) is often called an item bundle or a testlet.
- One can avoid the influence of lack of independence or multidimensionality by considering the testlet as a partial credit item instead of as a collection of binary items.
- And we have a beautiful model for this: the partial credit model

The Partial Credit Model (PCM)

- Possible scores are 0, 1, 2,…,*m*
- In the Rasch model *m* = 1 and 1 item parameter
- In the PCM: *m* item parameters

$$\begin{aligned}
&\text{score 1:} \quad \theta - \beta_{i1} \\
&\text{score 2:} \quad 2\theta - (\beta_{i1} + \beta_{i2}) \\
&\text{score 3:} \quad 3\theta - (\beta_{i1} + \beta_{i2} + \beta_{i3}) \\
&\qquad\quad \vdots
\end{aligned}$$

PCM: exact formula (for m = 2)

$$P(X_i = 0 \mid \theta) = \frac{1}{D}$$

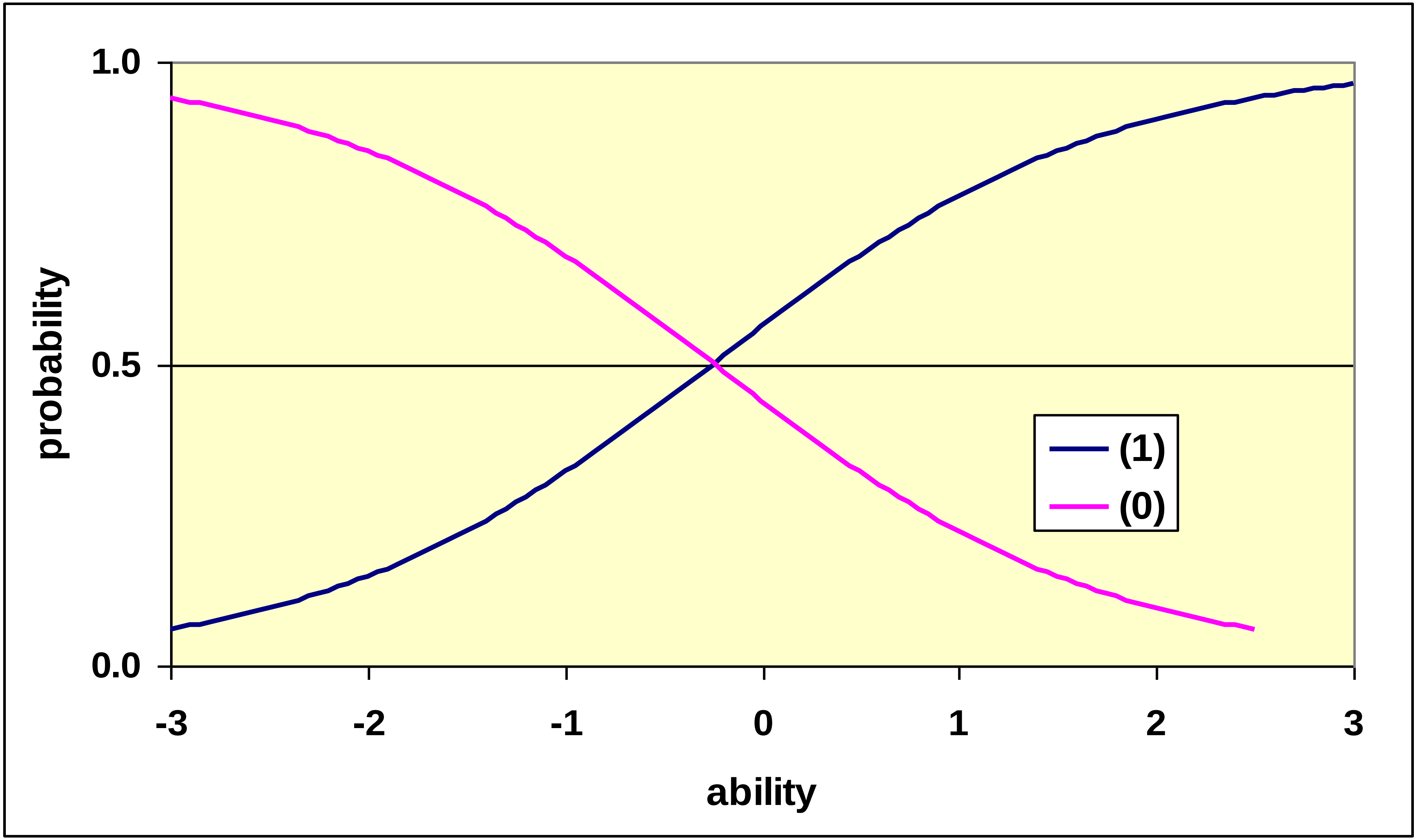
$$P(X_i = 1 \mid \theta) = \frac{\exp(\theta - \beta_{i1})}{D}$$

$$P(X_i = 2 \mid \theta) = \frac{\exp[2\theta - (\beta_{i1} + \beta_{i2})]}{D}$$

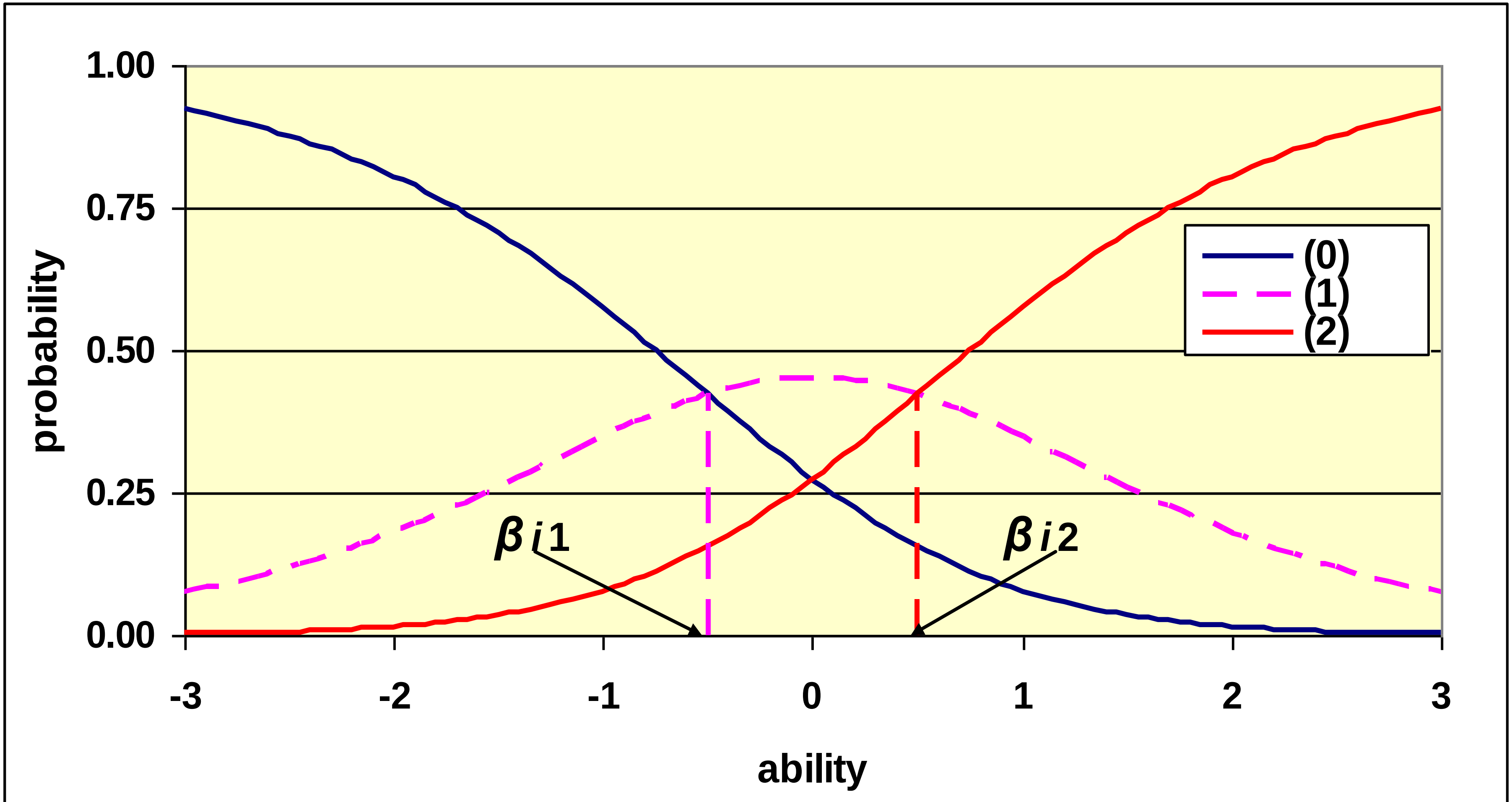
D is the sum of the three numerators.

(This guarantees that the sum of the 3 probabilities equals one.)

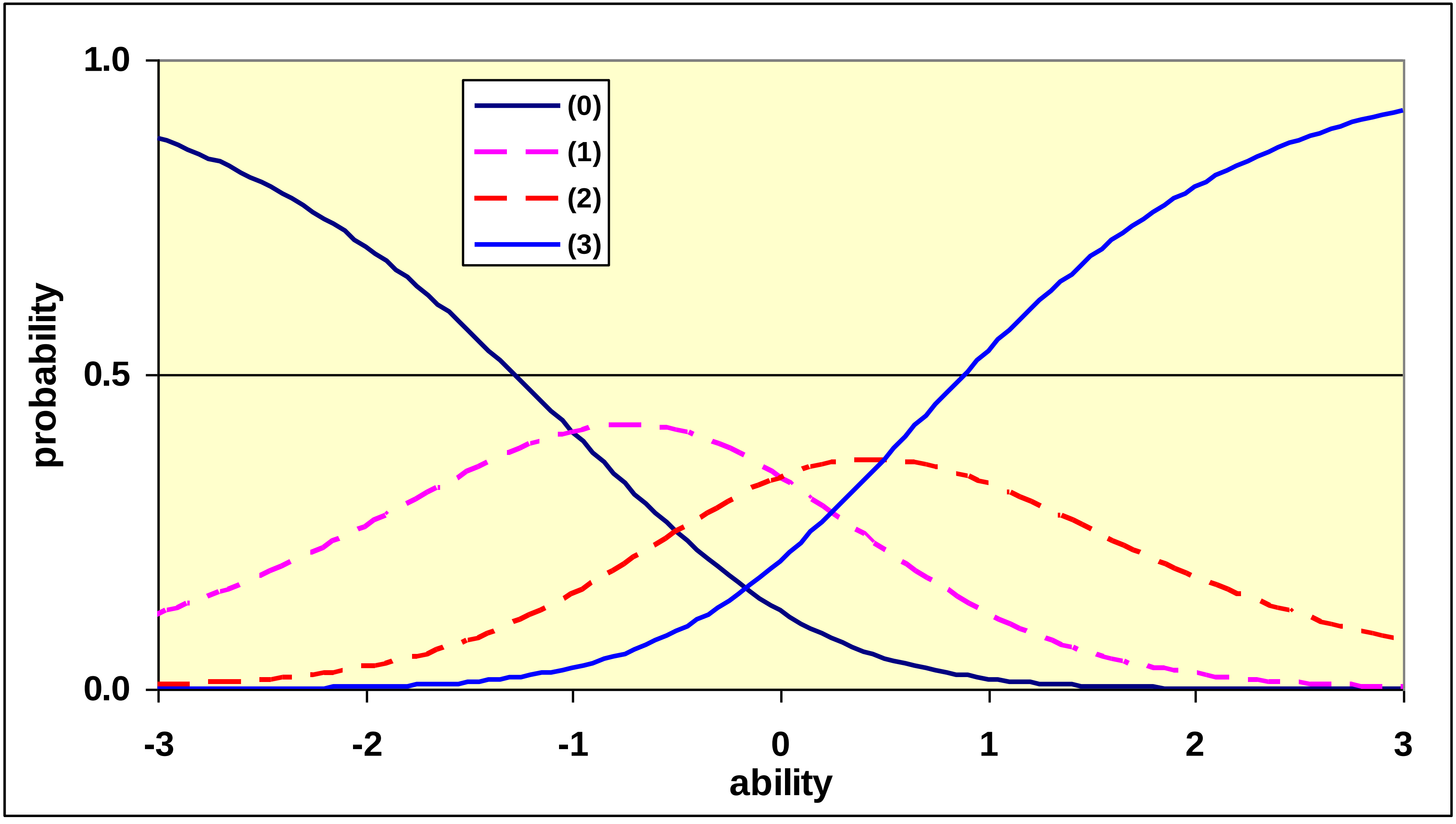
Category reponse functions in the Rasch model



Category reponse functions in the PCM (m = 2)



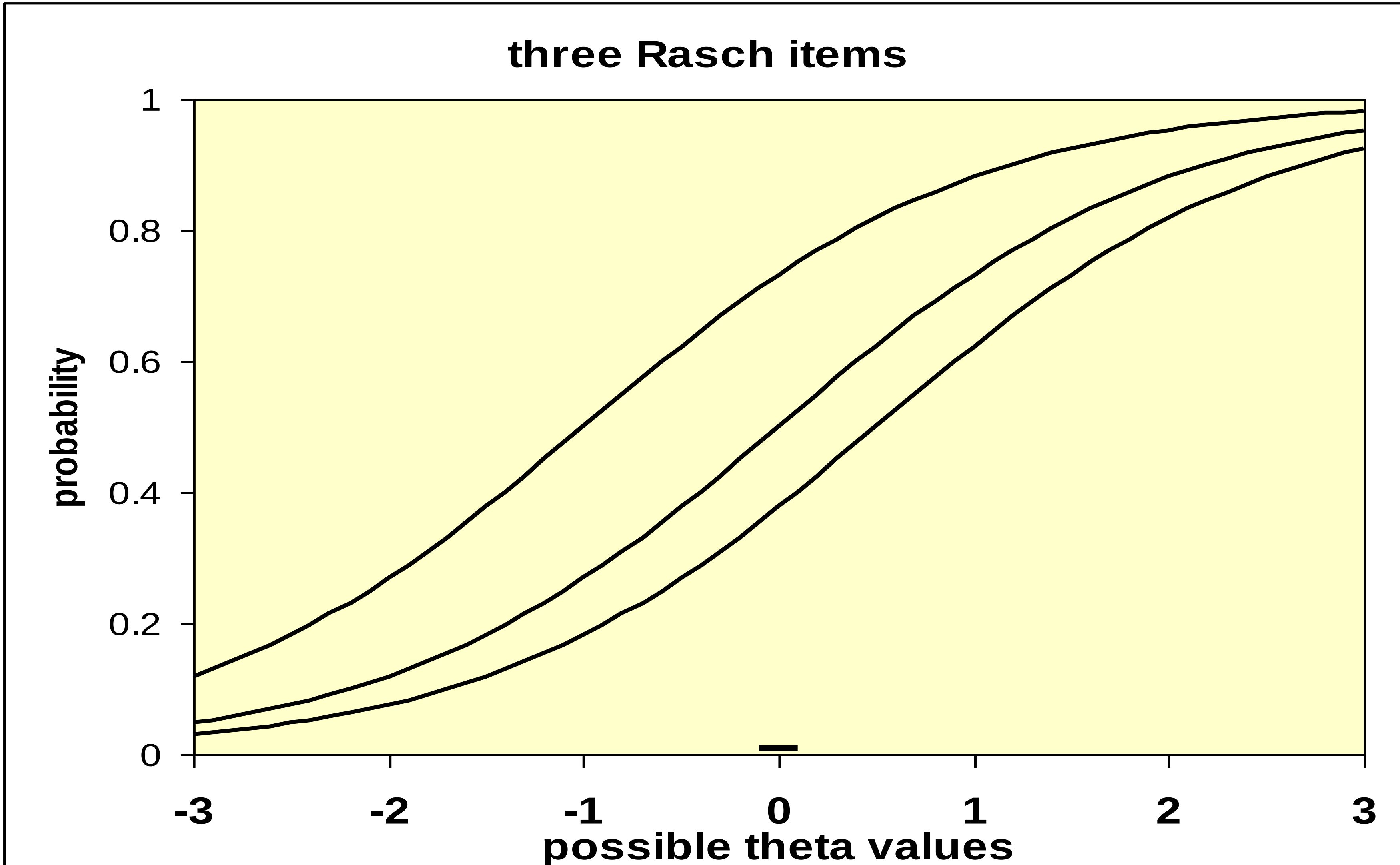
Category reponse functions in the PCM (m = 3)



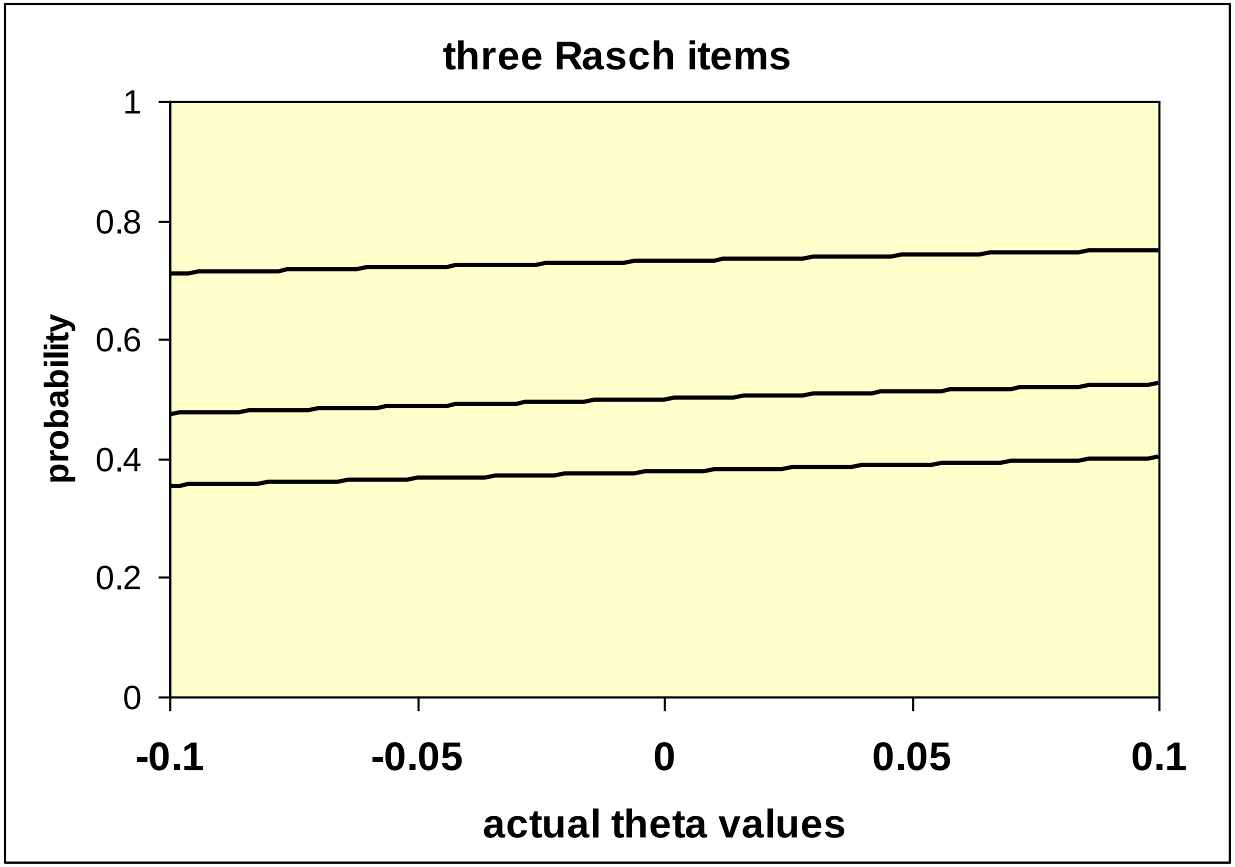
The big advantage of IRT

- Performances on different tests can be meaningfully compared
  - If the tests measure the same construct
  - If they are calibrated together
  - If they have items in common
- Partial credit items and binary items can appear in any mixture
  - Do not exaggerate *m* in the PCM
- But…

IRT is not the ultimate salvation



What happens if the variation of the abilities is small?



- Reliability goes down
- Validity of the Rasch model does not imply one has a reliable test

Thank you

For questions related to this webinar, you can write me via e-mail:

norman.verhelst@gmail.com