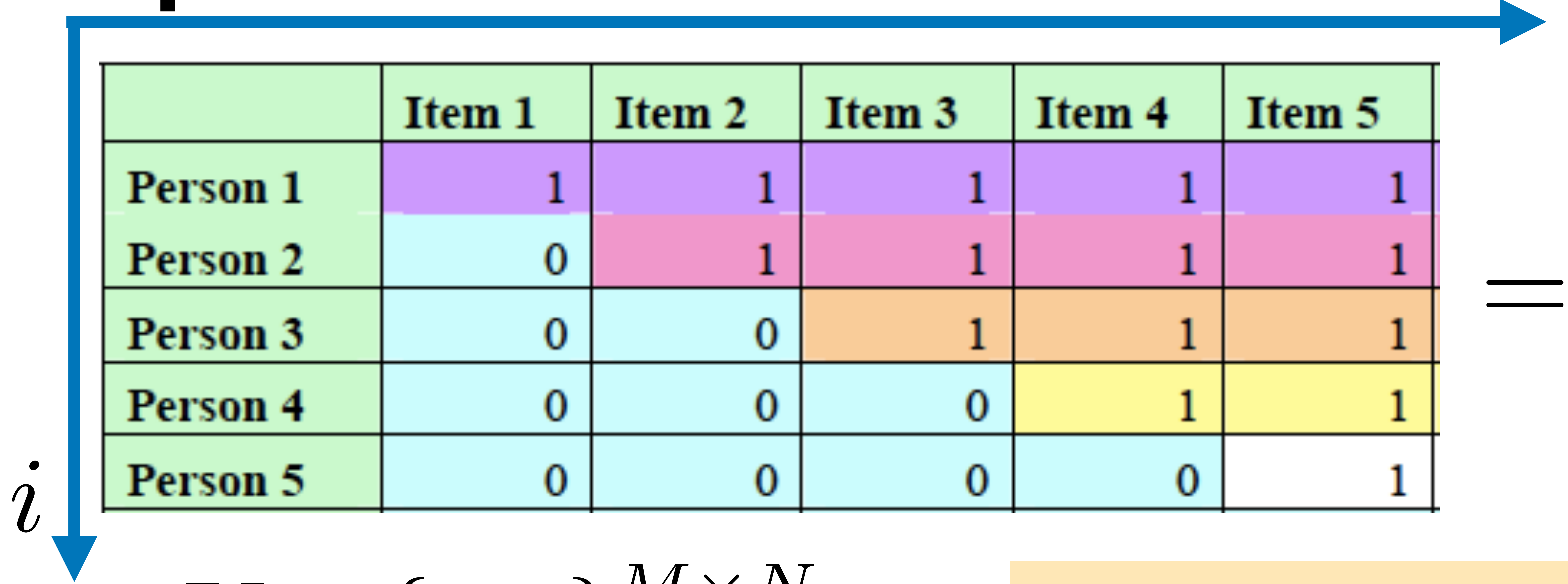
ITEM RESPONSE THEORY (IRT)

Luca Martino

WHAT IS A TEST?

- A SET OF **ITEMS/QUESTIONS/PROBLEMS**
- Here, we just care about a **correct (1) or a wrong answer (0)**
- We can have open answers or multiple answers (etc.). However, we assume that we can perfectly identify a **correct (y=1) or a wrong answer (y=0)**
- We consider ***N items in a Test.***
- We also ***M persons doing the same Test.*** Each person gives us N answers.

Example of data matrix Y



	Item 1	Item 2	Item 3	Item 4	Item 5
Person 1	1	1	1	1	1
Person 2	0	1	1	1	1
Person 3	0	0	1	1	1
Person 4	0	0	0	1	1
Person 5	0	0	0	0	1

$= \mathbf{Y}$

DATA: $\mathbf{Y} \in \{0,1\}^{M \times N}$

$\mathbf{Y}$ is an $M \times N$ matrix

$$[\mathbf{Y}]_{i,j} = y_{i,j} \in \{0,1\}$$

$$i = 1, 2, ..., M$$

$$j = 1, 2, ..., N$$

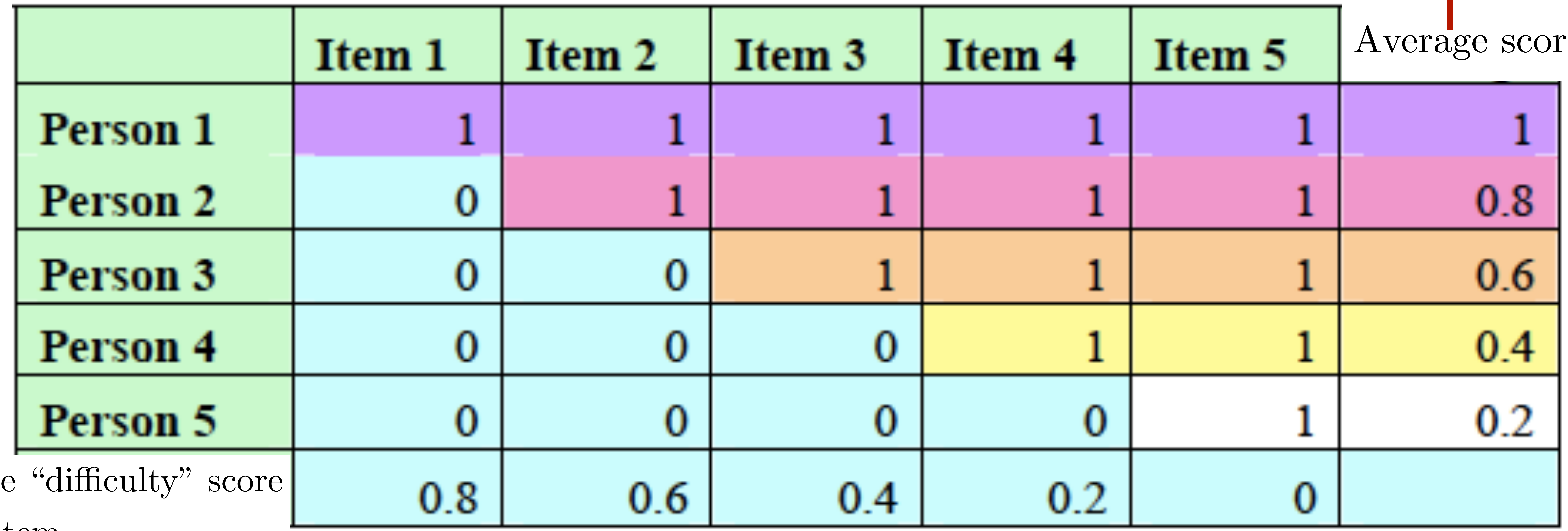
$M = 5$ persons in this example

$N = 5$ items in this example

Some considerations on the previous matrix Y

“tentative student proficiency” (TSD)

Average score



	Item 1	Item 2	Item 3	Item 4	Item 5	
Person 1	1	1	1	1	1	1
Person 2	0	1	1	1	1	0.8
Person 3	0	0	1	1	1	0.6
Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
	0.8	0.6	0.4	0.2	0	

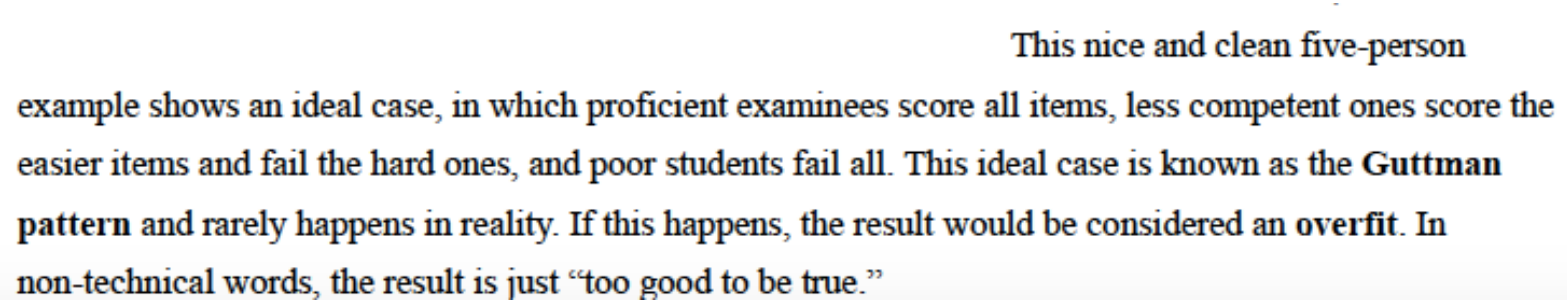
Average “difficulty” score of the item

“tentative item difficulty” (TID)

Some considerations on the previous matrix Y

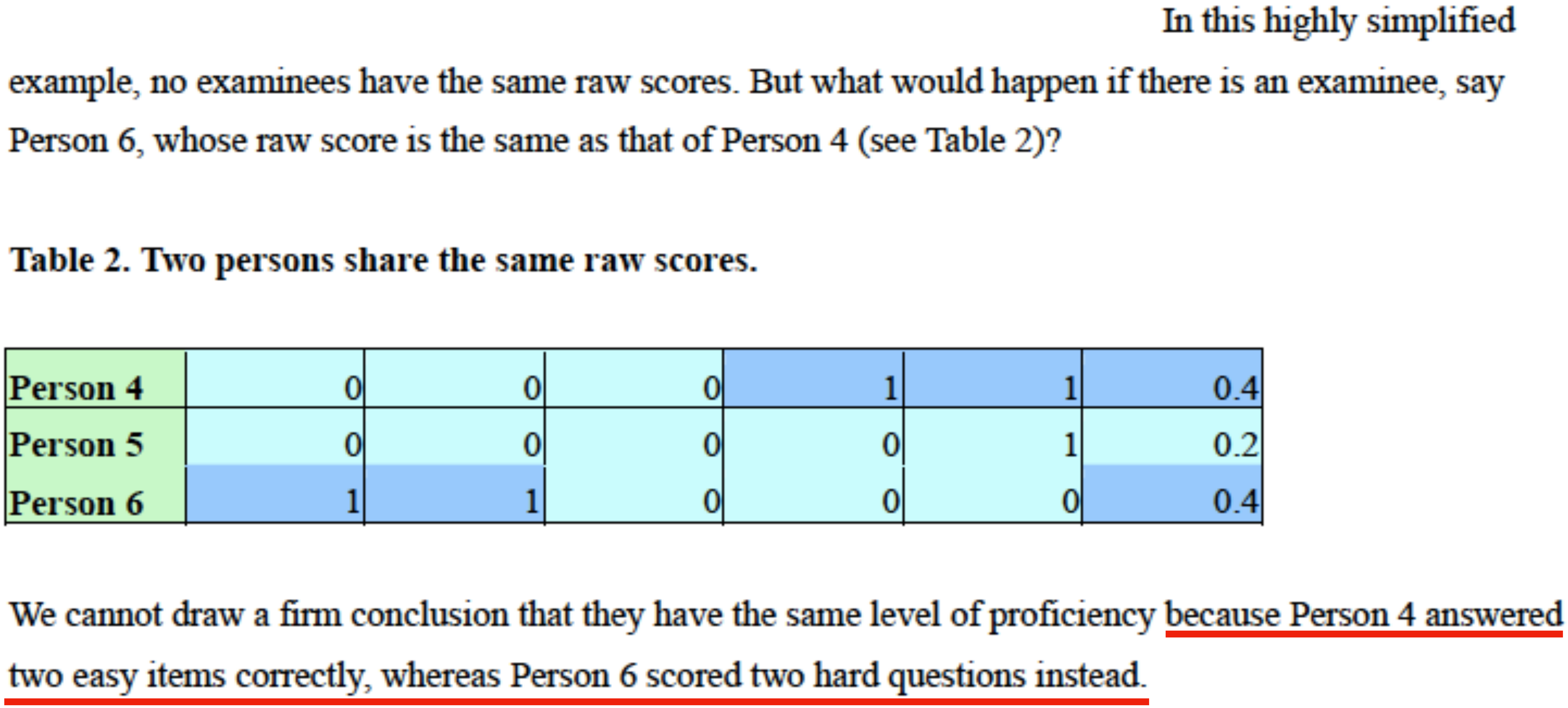
In this example, Person 1, who answered all five items correctly, is **tentatively** considered as possessing 100% proficiency. Person 2 has 80% proficiency, Person 3 has 60%, etc. These scores in terms of percentage are considered tentative because first, in IRT there is another set of terminology and scaling scheme for proficiency, and second, we cannot judge a person's ability just based on the number of correct items he obtained. Rather, the item attribute should also be taken into account.

Some considerations on the previous matrix Y



This nice and clean five-person example shows an ideal case, in which proficient examinees score all items, less competent ones score the easier items and fail the hard ones, and poor students fail all. This ideal case is known as the **Guttman pattern** and rarely happens in reality. If this happens, the result would be considered an **overfit**. In non-technical words, the result is just “too good to be true.”

Some considerations on the previous matrix Y



In this highly simplified example, no examinees have the same raw scores. But what would happen if there is an examinee, say Person 6, whose raw score is the same as that of Person 4 (see Table 2)?

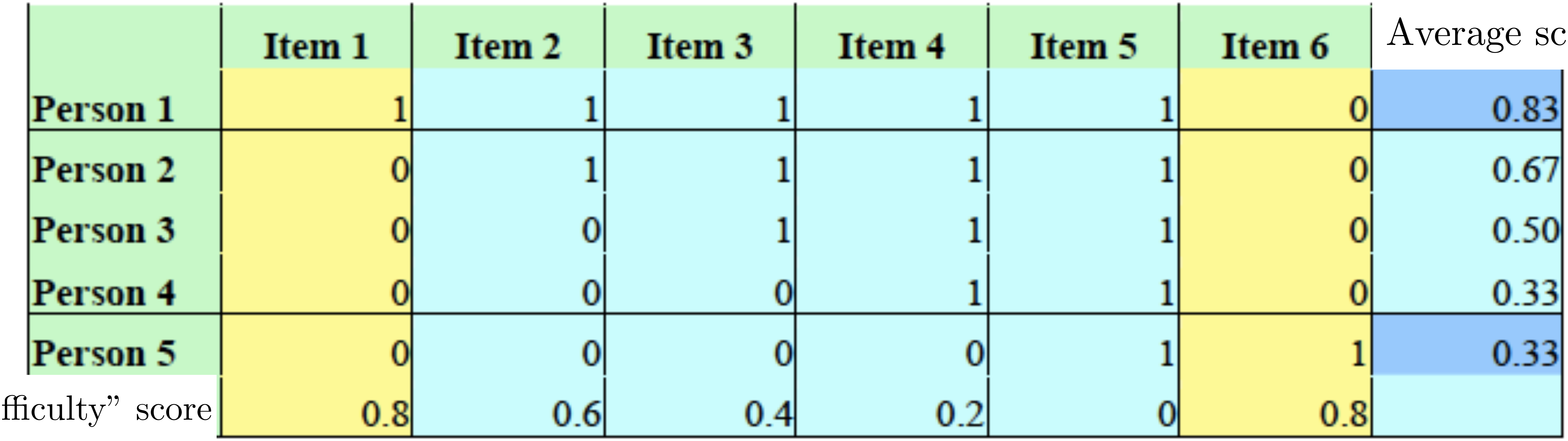
Table 2. Two persons share the same raw scores.

Person 4	0	0	0	1	1	0.4
Person 5	0	0	0	0	1	0.2
Person 6	1	1	0	0	0	0.4

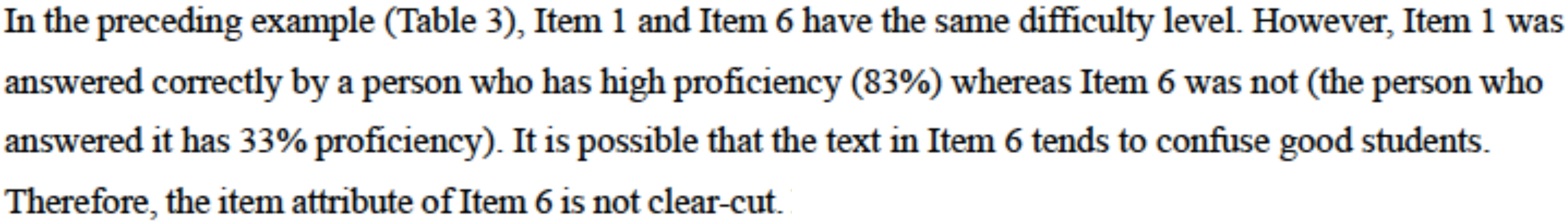
We cannot draw a firm conclusion that they have the same level of proficiency because Person 4 answered two easy items correctly, whereas Person 6 scored two hard questions instead.

Some considerations on the previous matrix Y

Table 3. Two items share the same pass rate.



	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Average score
Person 1	1	1	1	1	1	0	0.83
Person 2	0	1	1	1	1	0	0.67
Person 3	0	0	1	1	1	0	0.50
Person 4	0	0	0	1	1	0	0.33
Person 5	0	0	0	0	1	1	0.33
Average “difficulty” score of the item	0.8	0.6	0.4	0.2	0	0.8	



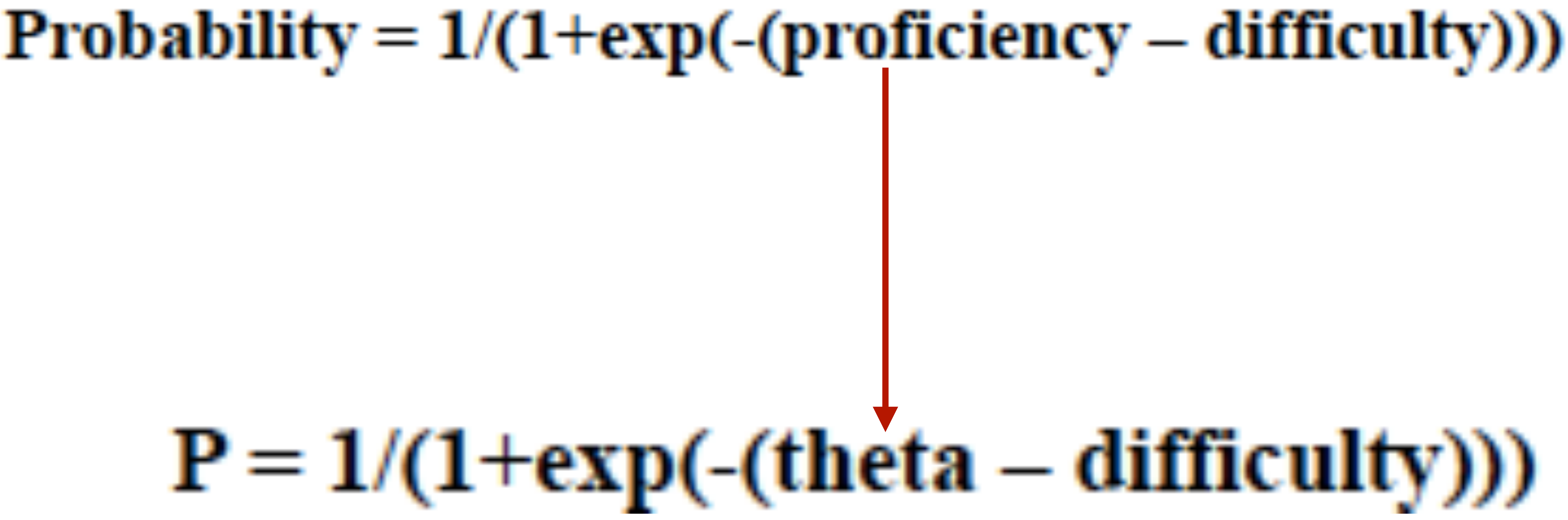
In the preceding example (Table 3), Item 1 and Item 6 have the same difficulty level. However, Item 1 was answered correctly by a person who has high proficiency (83%) whereas Item 6 was not (the person who answered it has 33% proficiency). It is possible that the text in Item 6 tends to confuse good students. Therefore, the item attribute of Item 6 is not clear-cut.

not well-defined, not clear

Main/key point in IRT: the Item Characteristic Curve (ICC)

Also called ITEM RESPONSE FUNCTION (IRF)

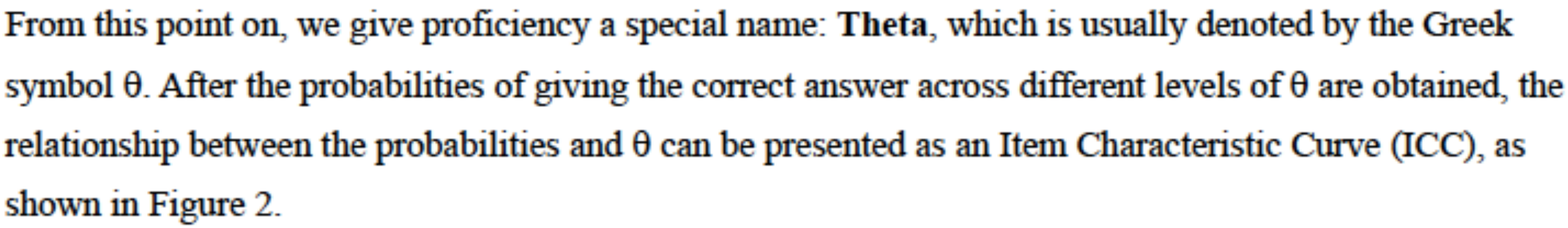
Probability of answering the true response *in a specific item*:



Probability = 1/(1+exp(-(proficiency – difficulty)))

P = 1/(1+exp(-(theta – difficulty)))

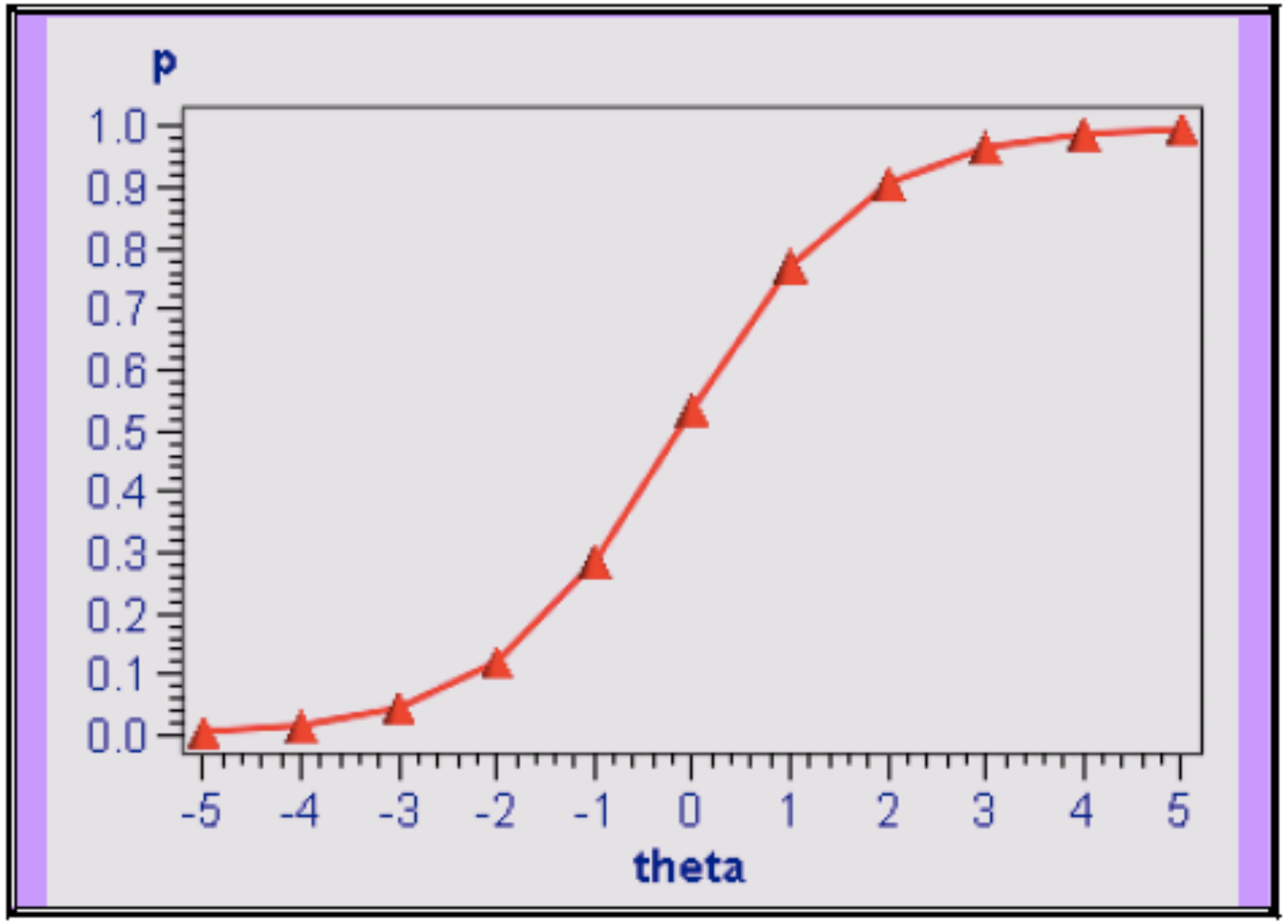
Main/key point in IRT: the Item Characteristic Curve



From this point on, we give proficiency a special name: **Theta**, which is usually denoted by the Greek symbol θ. After the probabilities of giving the correct answer across different levels of θ are obtained, the relationship between the probabilities and θ can be presented as an Item Characteristic Curve (ICC), as shown in Figure 2.

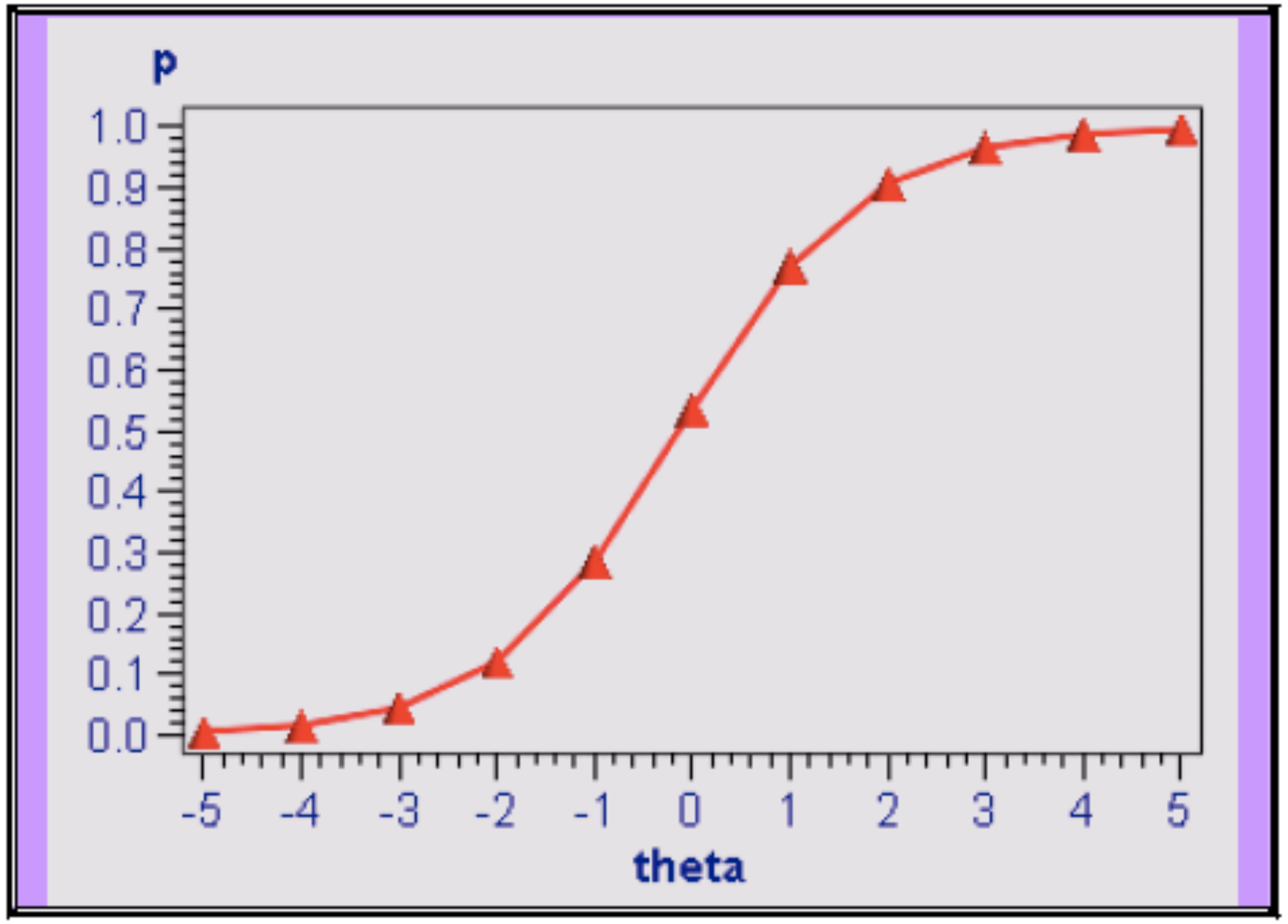
Figure 2. Item Characteristic Curve

$$P = 1/(1+\exp(-(\text{theta} - \text{difficulty})))$$

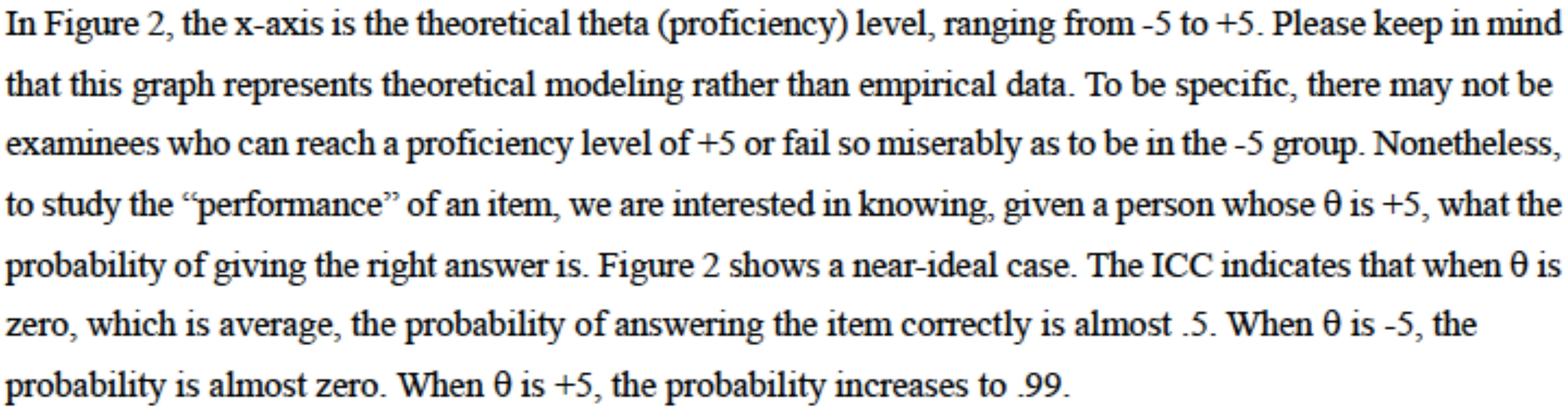


Also called ITEM RESPONSE FUNCTION (IRF)

Main/key point in IRT: the Item Characteristic Curve



θ is the “ability” (proficiency) that we want to estimate


In Figure 2, the x-axis is the theoretical theta (proficiency) level, ranging from -5 to +5. Please keep in mind that this graph represents theoretical modeling rather than empirical data. To be specific, there may not be examinees who can reach a proficiency level of +5 or fail so miserably as to be in the -5 group. Nonetheless, to study the “performance” of an item, we are interested in knowing, given a person whose θ is +5, what the probability of giving the right answer is. Figure 2 shows a near-ideal case. The ICC indicates that when θ is zero, which is average, the probability of answering the item correctly is almost .5. When θ is -5, the probability is almost zero. When θ is +5, the probability increases to .99.

Also called ITEM RESPONSE FUNCTION (IRF)

Item Characteristic Curve (ICC) or ITEM RESPONSE FUNCTION (IRF)

j-th item — i-th person

$\theta_i \Longrightarrow$ Ability of the j-th person

$b_j \Longrightarrow$ Difficulty of the i-th item

One parameter:
Rash Model

$$P_j(\theta_i) = \frac{1}{1+\exp(-(\theta_i - b_j))}$$

$a_j \Longrightarrow$ Discrimination of the i-th item

Two parameters:

$$P_j(\theta_i) = \frac{1}{1+\exp(-a_j(\theta_i - b_j))}$$

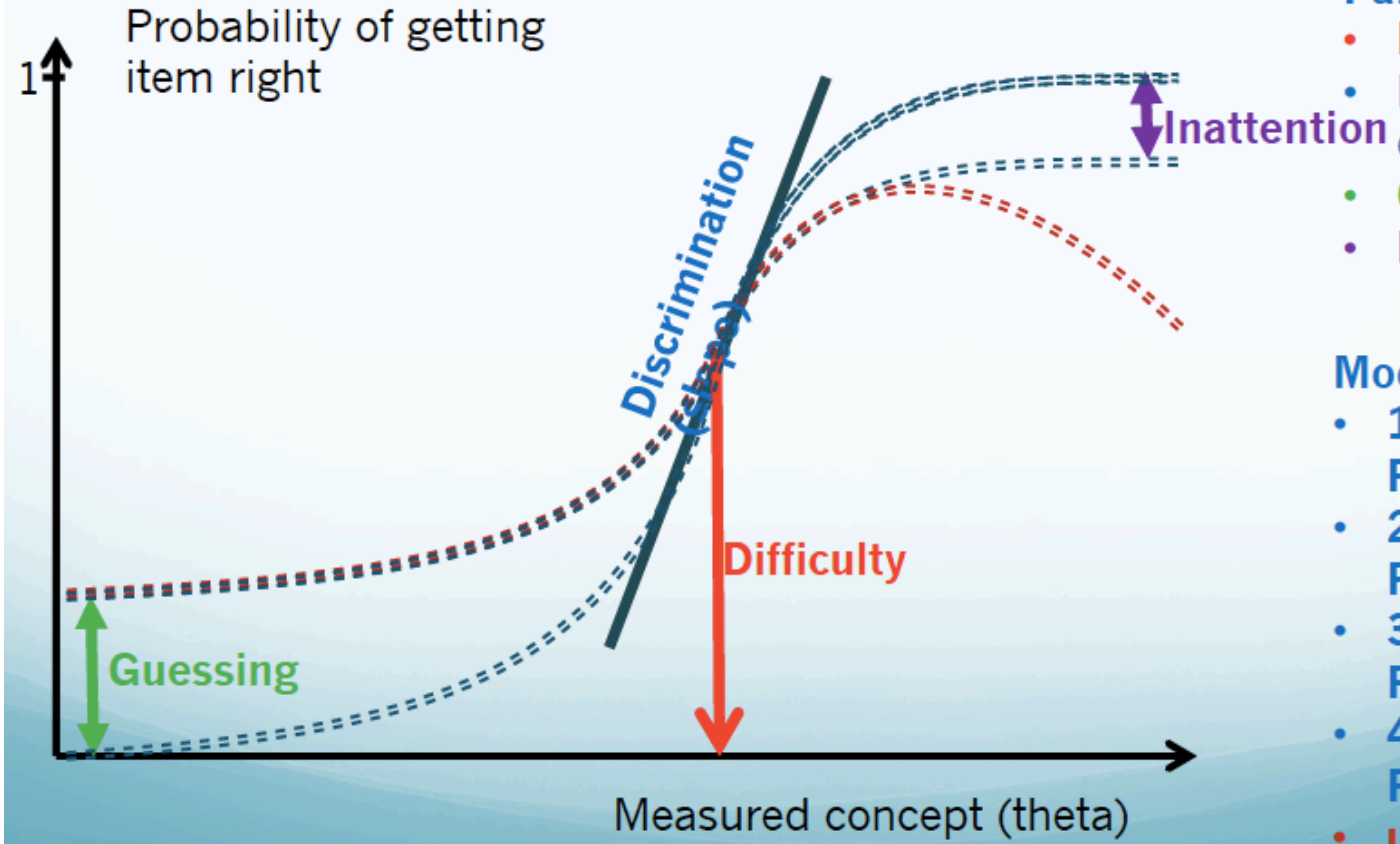
$c_j \Longrightarrow$ Prob. of giving the true answer to the i-th item "just guessing"

Three parameters:

$$P_j(\theta_i) = c_j + (1-c_j)\frac{1}{1+\exp(-a_j(\theta_i - b_j))}$$

Item Response Function

Binary items



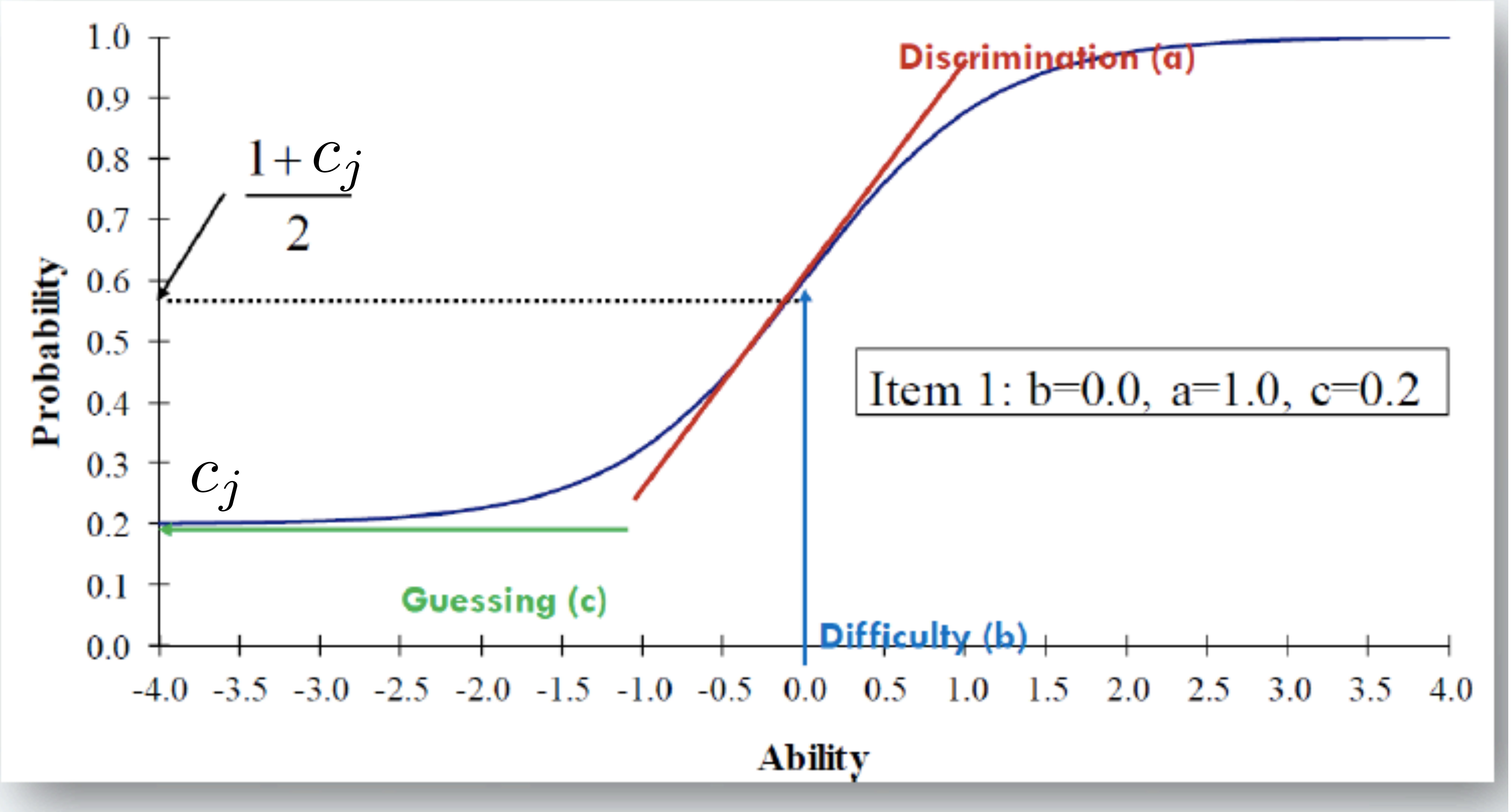
Parameters:

- Difficulty
- Discrimination
- Guessing
- Inattention

Models:

- 1 Parameter
- 2 Parameter
- 3 Parameter
- 4 Parameter
- unfolding

Example of IRF of three-parameter model



One item showing the guessing parameter (c)

For learning the parameters…

Build the likelihood function similarly as in a Logistic Regression… more details ask to the Professor

True - estimated - score V of the test

If each item has value 1:

$$V_i = \sum_{j=1}^{N} P_j(\theta_i)$$

This is the *true (ESTIMATED) score* of the i-th person

It is like a *denoised* version of the score obtained:

$$X_i = \sum_{j=1}^{N} y_{i,j}$$

Characteristic Curve of Test

$$V(\theta) = \sum_{j=1}^{N} P_j(\theta)$$

The Characteristic Curve of Test
gives us a “*true score*” (an effective score) for a theta

Power of the measurement error

$$X_i = \sum_{j=1}^{N} y_{i,j}$$

$$V_i = \sum_{j=1}^{N} P_j(\theta_i)$$

$$e_i = X_i - V_i$$

$$\text{var}[e_i] = \sum_{j=1}^{N} \text{variance of a Bernoulli variable}$$

$$\text{var}[e_i] = \sum_{j=1}^{N} P_j(\theta_i)\,(1 - P_j(\theta_i))$$

Information function of an item

$$I_j(\theta) = \frac{\left[\frac{dP_j(\theta)}{d\theta}\right]^2}{P_j(\theta)\left(1 - P_j(\theta)\right)}$$

Information function of an item

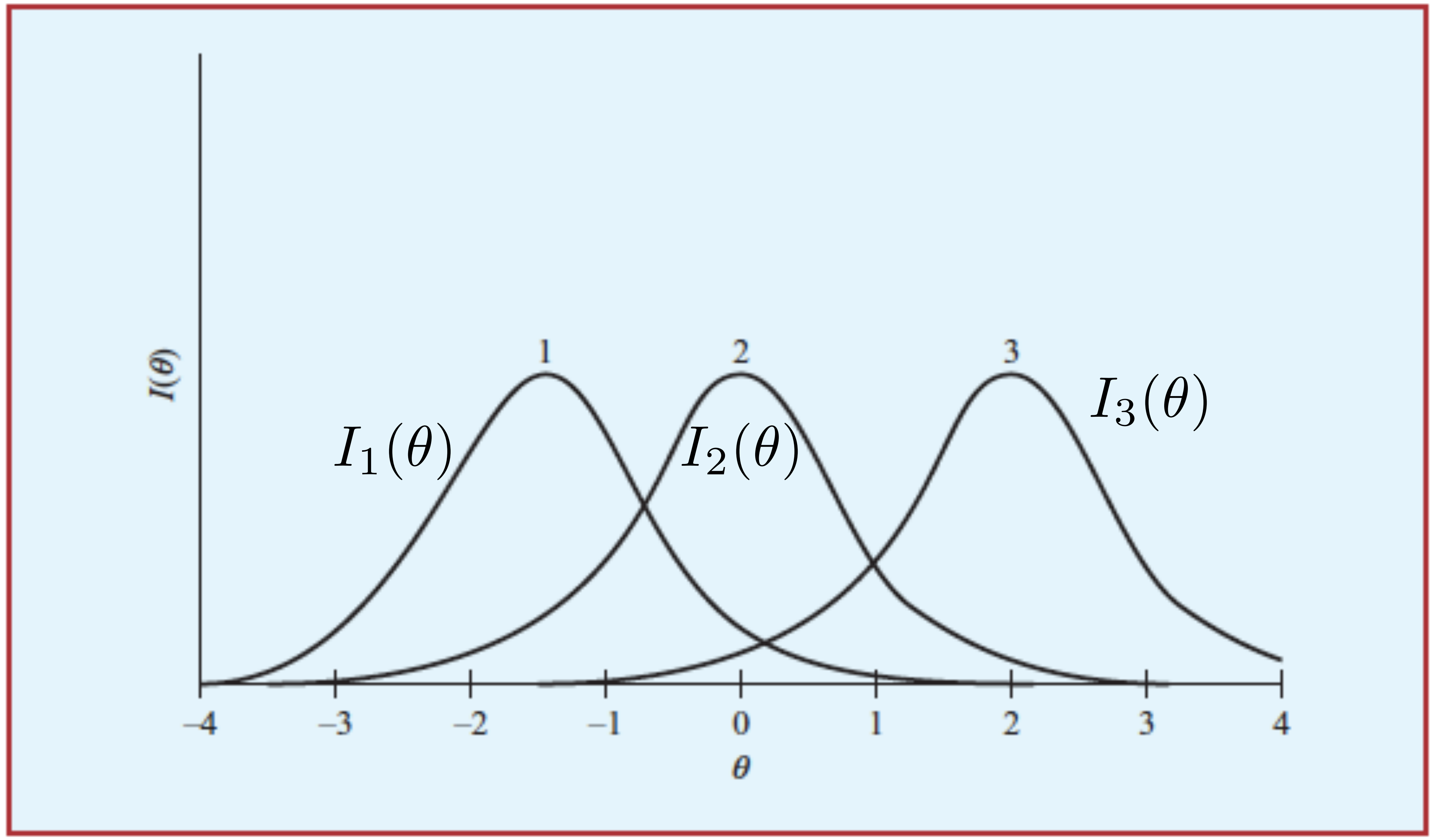
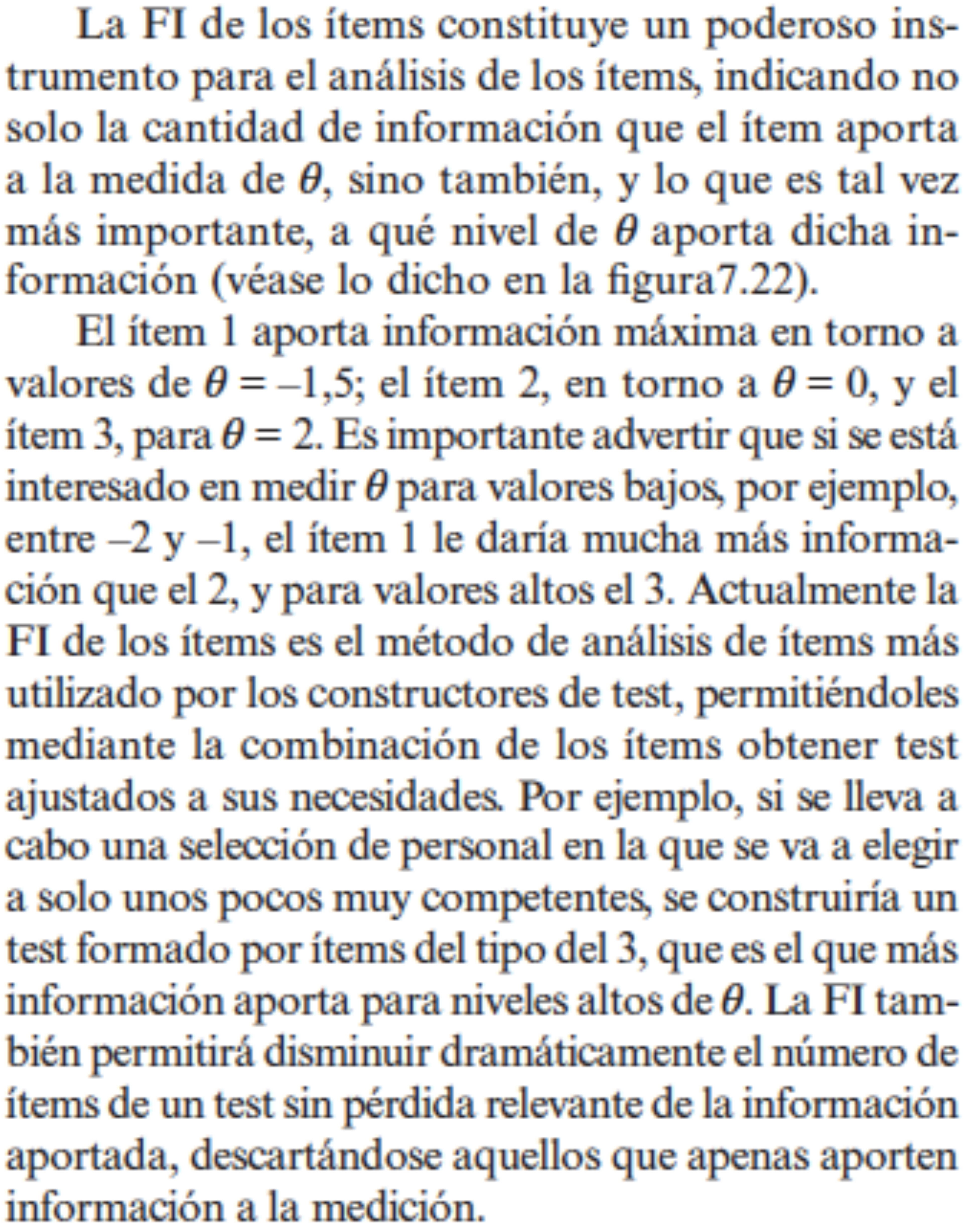


Figura 7.22.—Funciones de información de tres ítems.



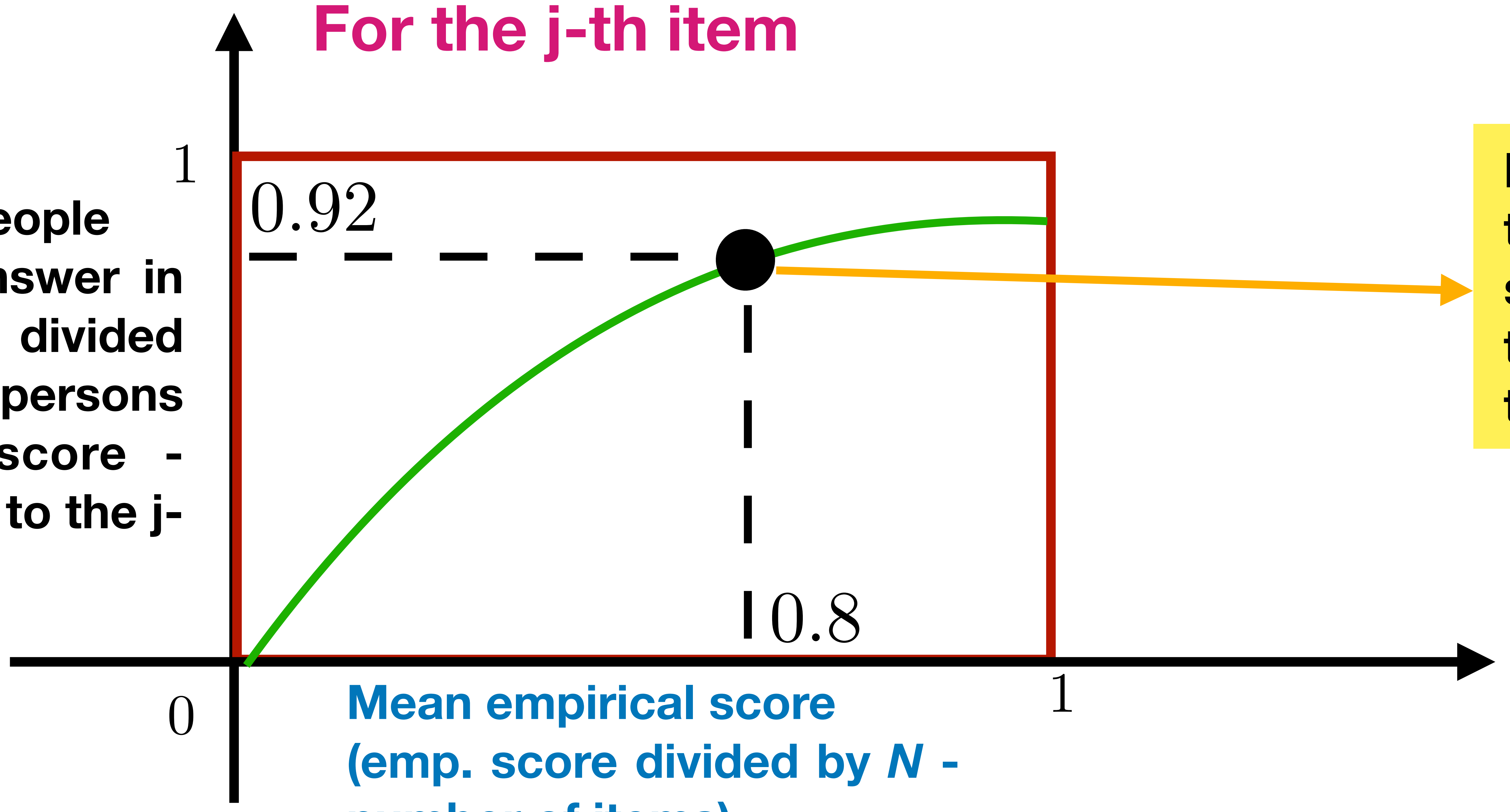
La FI de los ítems constituye un poderoso instrumento para el análisis de los ítems, indicando no solo la cantidad de información que el ítem aporta a la medida de θ, sino también, y lo que es tal vez más importante, a qué nivel de θ aporta dicha información (véase lo dicho en la figura7.22).

El ítem 1 aporta información máxima en torno a valores de $\theta = –1,5$; el ítem 2, en torno a $\theta = 0$, y el ítem 3, para $\theta = 2$. Es importante advertir que si se está interesado en medir θ para valores bajos, por ejemplo, entre –2 y –1, el ítem 1 le daría mucha más información que el 2, y para valores altos el 3. Actualmente la FI de los ítems es el método de análisis de ítems más utilizado por los constructores de test, permitiéndoles mediante la combinación de los ítems obtener test ajustados a sus necesidades. Por ejemplo, si se lleva a cabo una selección de personal en la que se va a elegir a solo unos pocos muy competentes, se construiría un test formado por ítems del tipo del 3, que es el que más información aporta para niveles altos de θ. La FI también permitirá disminuir dramáticamente el número de ítems de un test sin pérdida relevante de la información aportada, descartándose aquellos que apenas aporten información a la medición.

Other curves of interest
(from the classical test theory)

For each item:

in Y-axis ==> (mean) number of people giving a right answer - with a given emp. score
in X-axis ==> (mean) empirical score

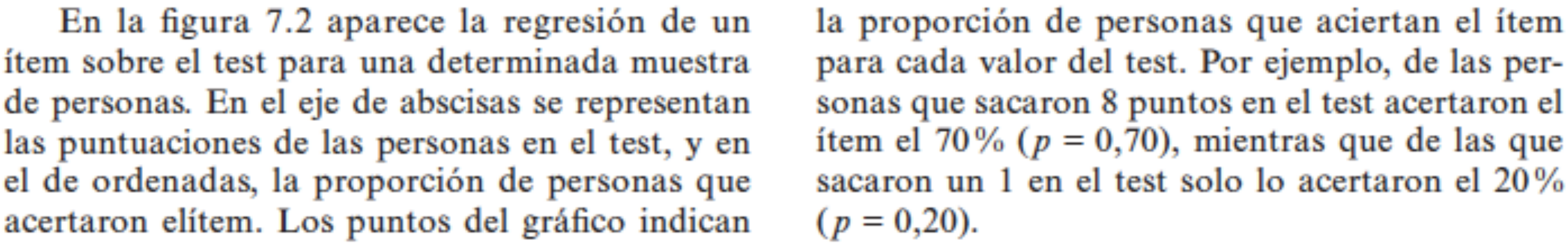


Mean number of people giving the right answer in the j-th item (i.e., divided for the number of persons - with a given score - aswering correctly to the j-th item)

Mean empirical score (emp. score divided by *N* - number of items)

Inside all the persons (let denoted them as *K*) which take a empirical score of 0.8**N*, only 92%=0.92**K* of all them (*K*), have answered correctly to the j-th item.

This curve can recall in some sense the Item Characteristic Curve (ICC) also called ITEM RESPONSE FUNCTION (IRF) in the modern IRT.



En la figura 7.2 aparece la regresión de un ítem sobre el test para una determinada muestra de personas. En el eje de abscisas se representan las puntuaciones de las personas en el test, y en el de ordenadas, la proporción de personas que acertaron elítem. Los puntos del gráfico indican la proporción de personas que aciertan el ítem para cada valor del test. Por ejemplo, de las personas que sacaron 8 puntos en el test acertaron el ítem el 70% ($p = 0{,}70$), mientras que de las que sacaron un 1 en el test solo lo acertaron el 20% ($p = 0{,}20$).

Mean number of people giving the right answer in the j-th item

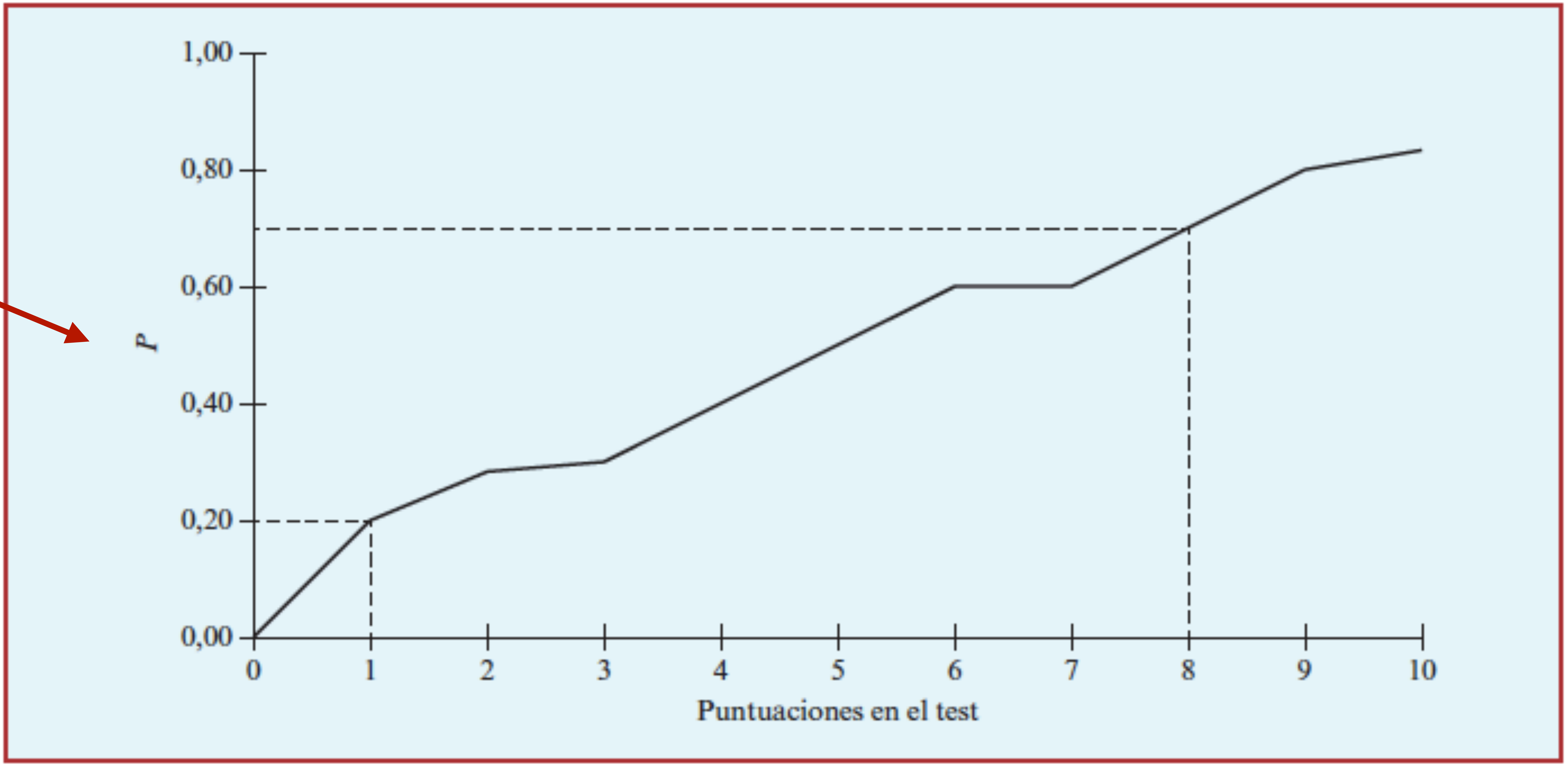


Figura 7.2.—Regresión ítem-test.