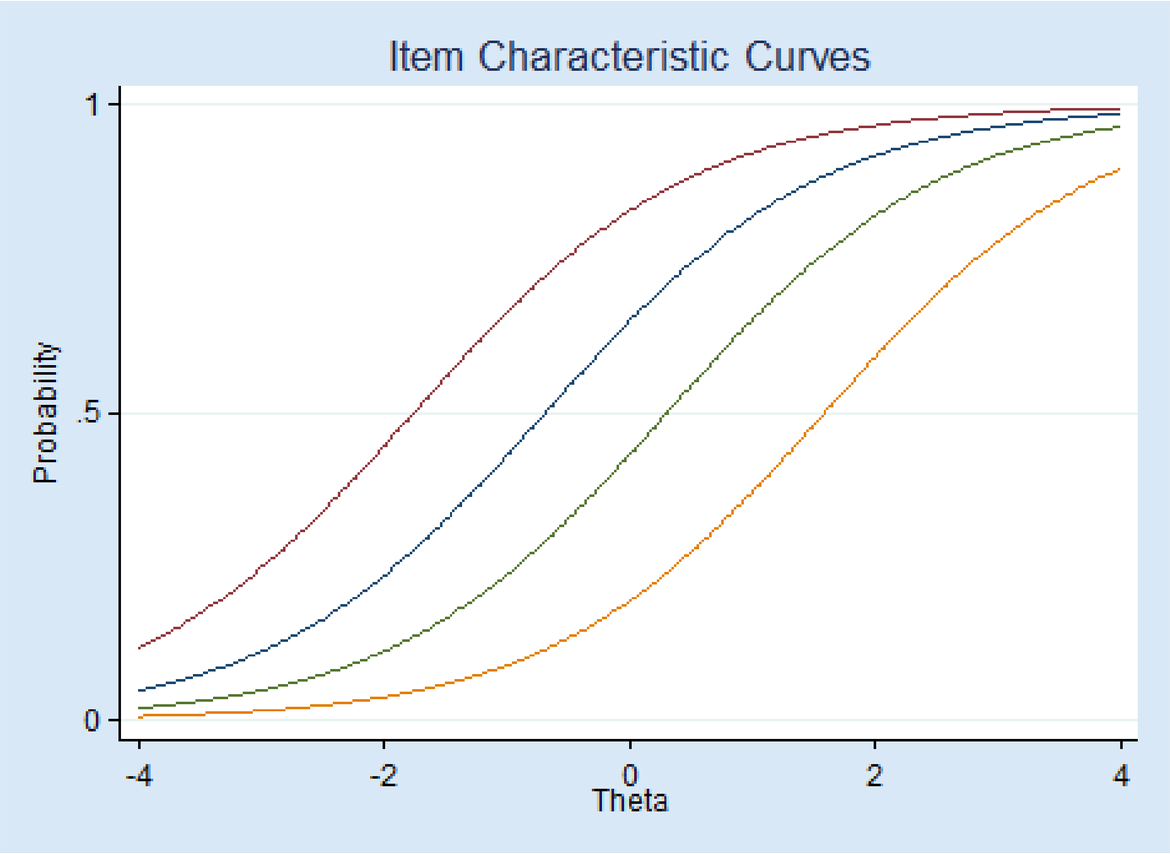
Item Response Theory (IRT) in R



How would you measure height if you couldn’t observe it?

Something that can’t be measured directly might still be measurable indirectly.
Using items, we can measure the unobservable measure.

Check out such measures yourself →
“If You Don't Think This Quiz Can Guess Your Height, Then I DARE YOU To Take It”:
https://www.buzzfeed.com/andrewziegler/quiz-that-can-guess-your-height

What is Item Response Theory?

- a family of statistical measurement models

These models aim to describe the **relationship between** a person's **response** to a given item and the **underlying trait** (for ex. height) the item is used to measure.

- also known as "latent trait theory", "strong true score theory", "fundamental test theory"

- underlying trait = latent variable = unobservable variable = construct = factor

Classical Test Theory: some problems

- the evaluation of the quality of the items and the test as a whole depends on the particular group of persons and the particular group of items used for the evaluation.
    - Item difficulty depends on the group.
    - Item discrimination depends on the group.
- Assumptions of linearity, homogeneity and normality, underlying the use of factor analysis and regression analysis are probably violated.

IRT advantages

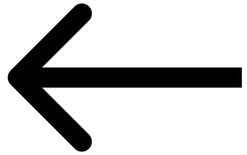
- ability estimation is item independent given a sufficient item pool. In other words, given enough items, trait estimate is independent of items used. It follows that homogeneous item pool is needed.

- measurement error provided conditioned on the latent trait level. It looks on item-level data rather than test-level data.

Most commonly used IRT models

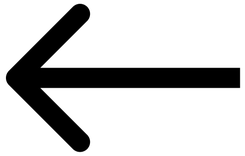
Most common:

- Rasch Models
- 2/3 Parameter Logistic Models



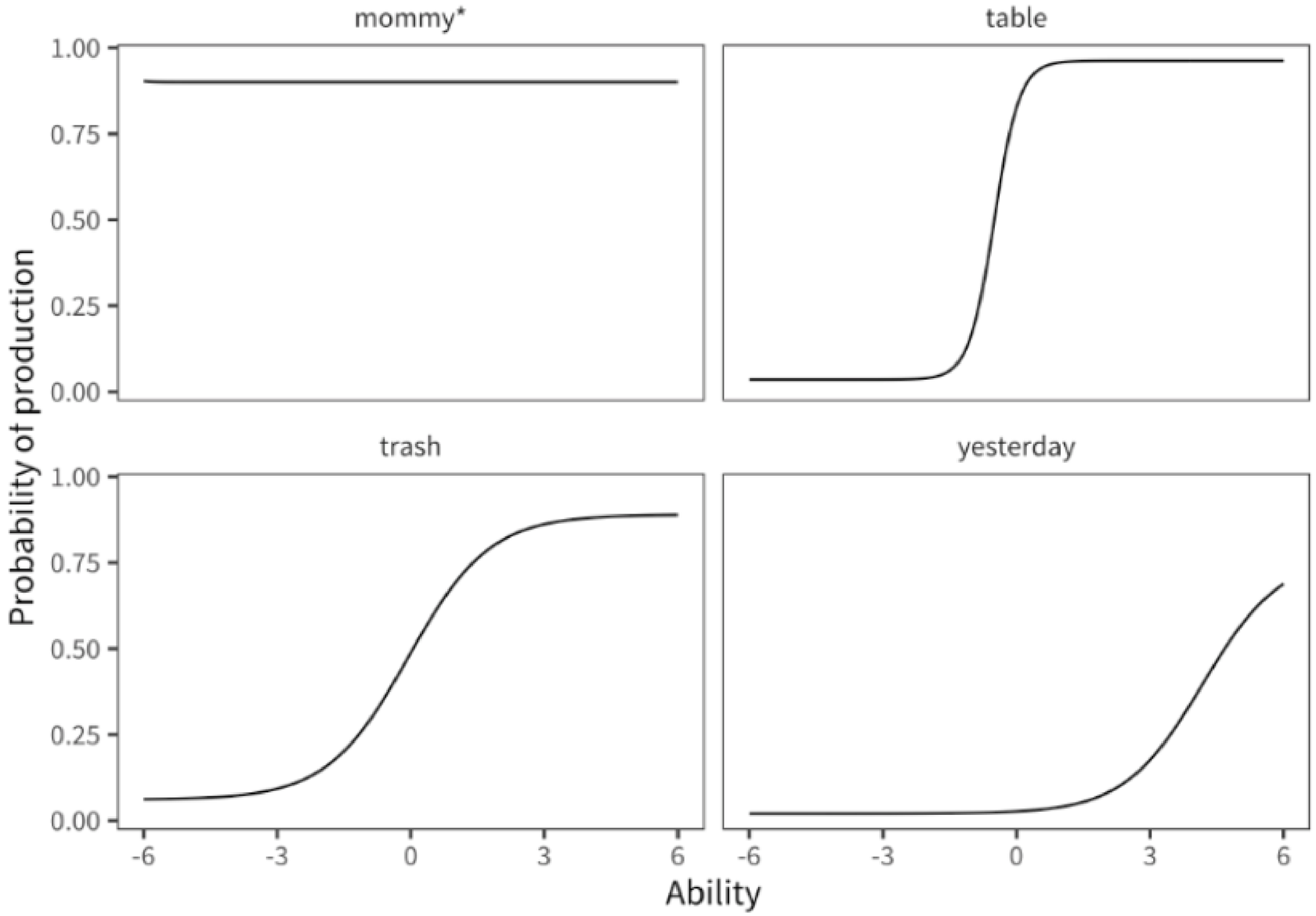
used for dichotomous items (true/false, correct/incorrect)

- (Generalized) Partial Credit Models
  - Graded Response Models
    - Rating Scale Model



used for polytomous items (Likert scale, partial credit items, other ordinal responses)

Example



language ability vs probability of recognizing word

Example

1. Who is the highest ability person? Who is the lowest ability person?
2. Which item is the hardest? Which is the easiest?
3. Which item is the best? Which is the worst?
4. Who has a higher ability between person D and person I?
5. Estimate the probability of person G getting item 2 correct.

person	item 1	item 2	item 3
A	0	0	0
B	1	0	1
C	1	0	0
D	1	0	1
E	1	0	0
F	1	0	1
G	1	NA	0
H	1	0	1
I	1	1	0
J	1	1	1

What is a measurement?

1. You're interested in a latent construct (math ability, extroversion, anxiety etc.)
2. You measure that latent construct by giving people items (which we'll call a test)
3. You do some science with that measurement

How do I get from responses to the latent trait?

Naive approach: The sum score

Assumptions

a. Items are equally difficult.
b. Items are equally related to the latent construct.
c. 1 on all items is positively related to the construct.

Limitations

a. How do I handle missing data?
b. How do I make predictions?
c. How do I make an adaptive test?

child	mommy	yesterday	trash
A	0	0	0
B	1	0	1
C	1	0	0
D	1	0	1
E	1	0	0
F	1	0	1
G	1	NA	0
H	1	0	1
I	1	1	0
J	1	1	1

Item Response Theory (IRT) to the rescue!

Each person p has an ability θ_p

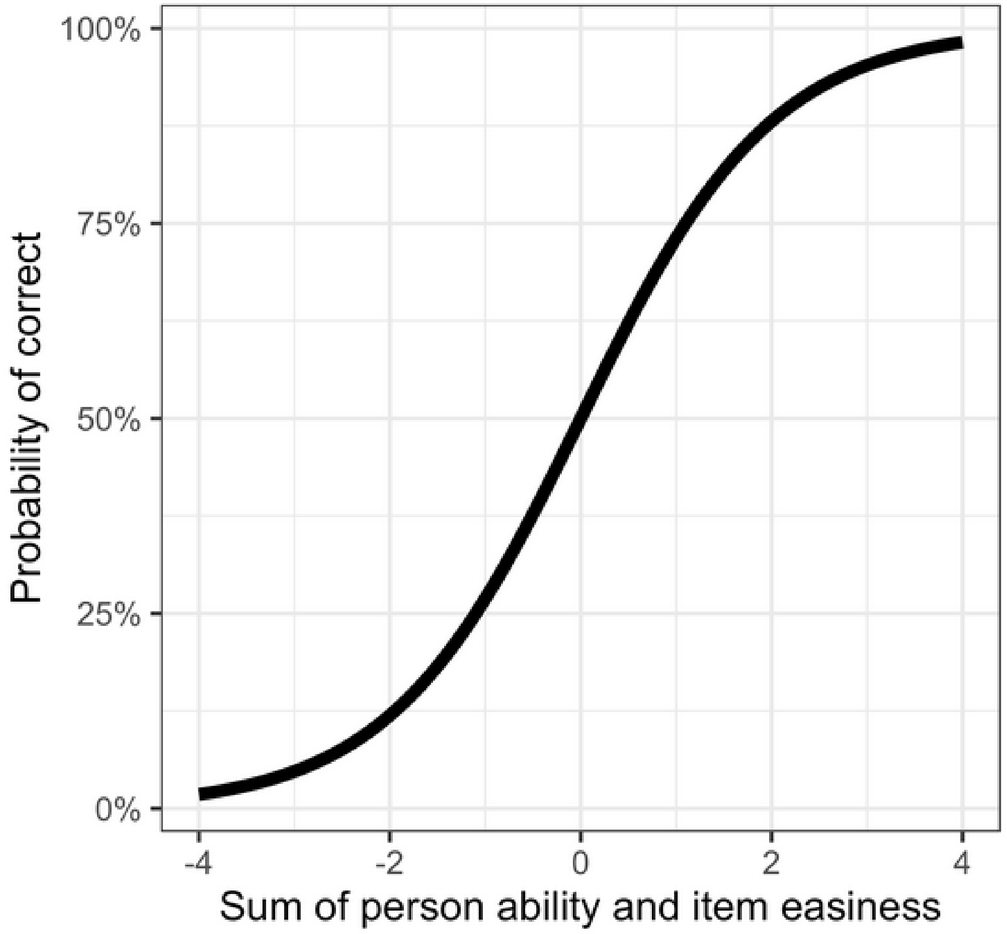
Each item i has an easiness b_i

These combine to give the probability of correct response

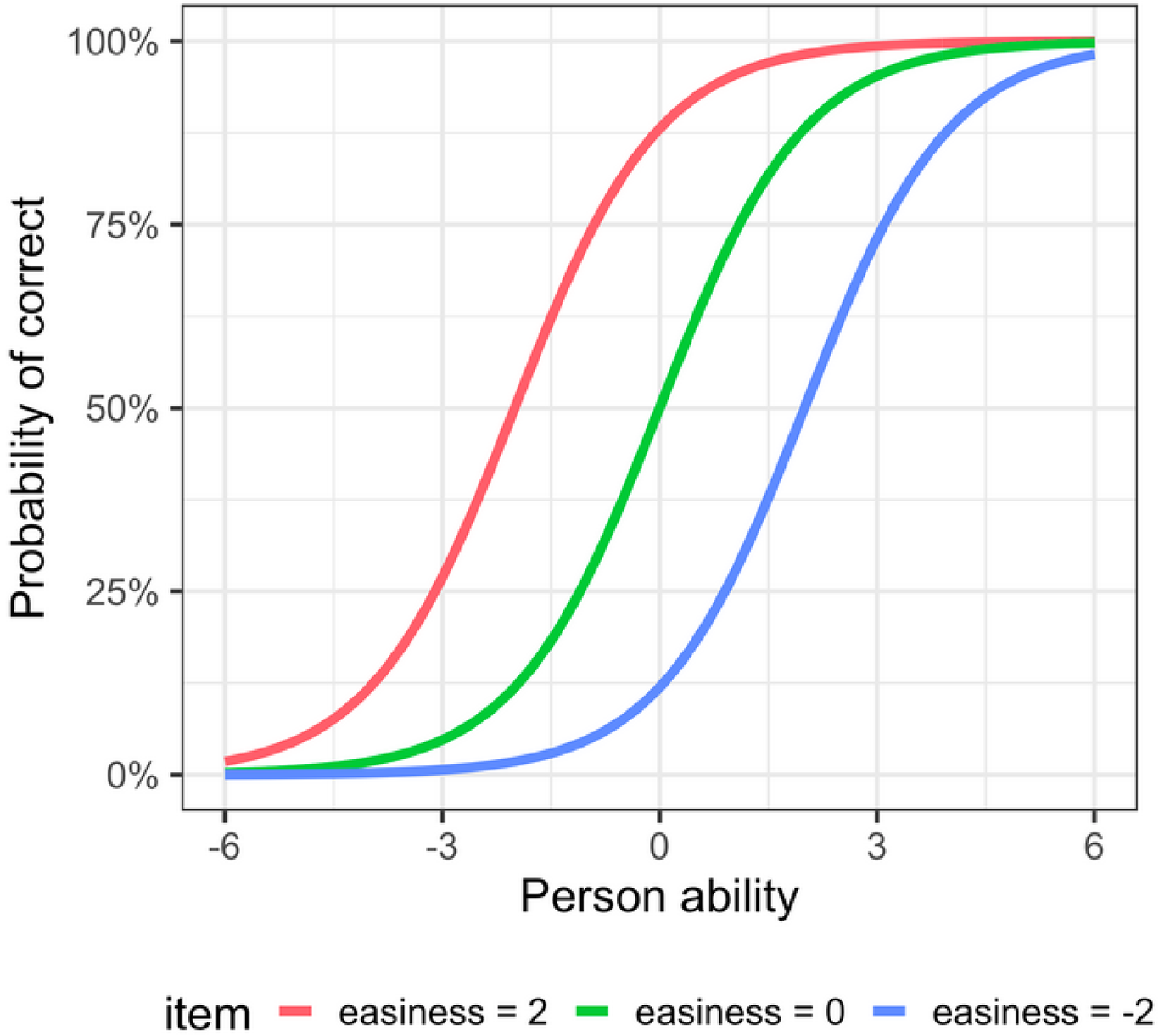
- Consider an encounter between a person of ability θ and an item of difficulty *b*. Since a deterministic response is not acceptable, the response must be expressed in terms of probabilities. This gives us the one-parameter logistic model for the probability of a correct response.
- *If $\theta > b$, $p(1) \rightarrow 1.0$*
- *If $\theta < b$, $p(1) \rightarrow 0.0$*
- *If $\theta = b$, $p(1) \rightarrow 0.5$*

The Sigmoid!

We use the logistic $\sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$ function to map the sum of ability and easiness to probability of correct response



The Item Characteristic Curve



The Rasch Model (1PL, 1PL=1 Parameter Logistic model)

Each person has ability θ_p. Each item has easiness b_i.

$$P(y_{pi} = 1 | \theta_p, b_i) = \sigma(\theta_p + b_i)$$

where

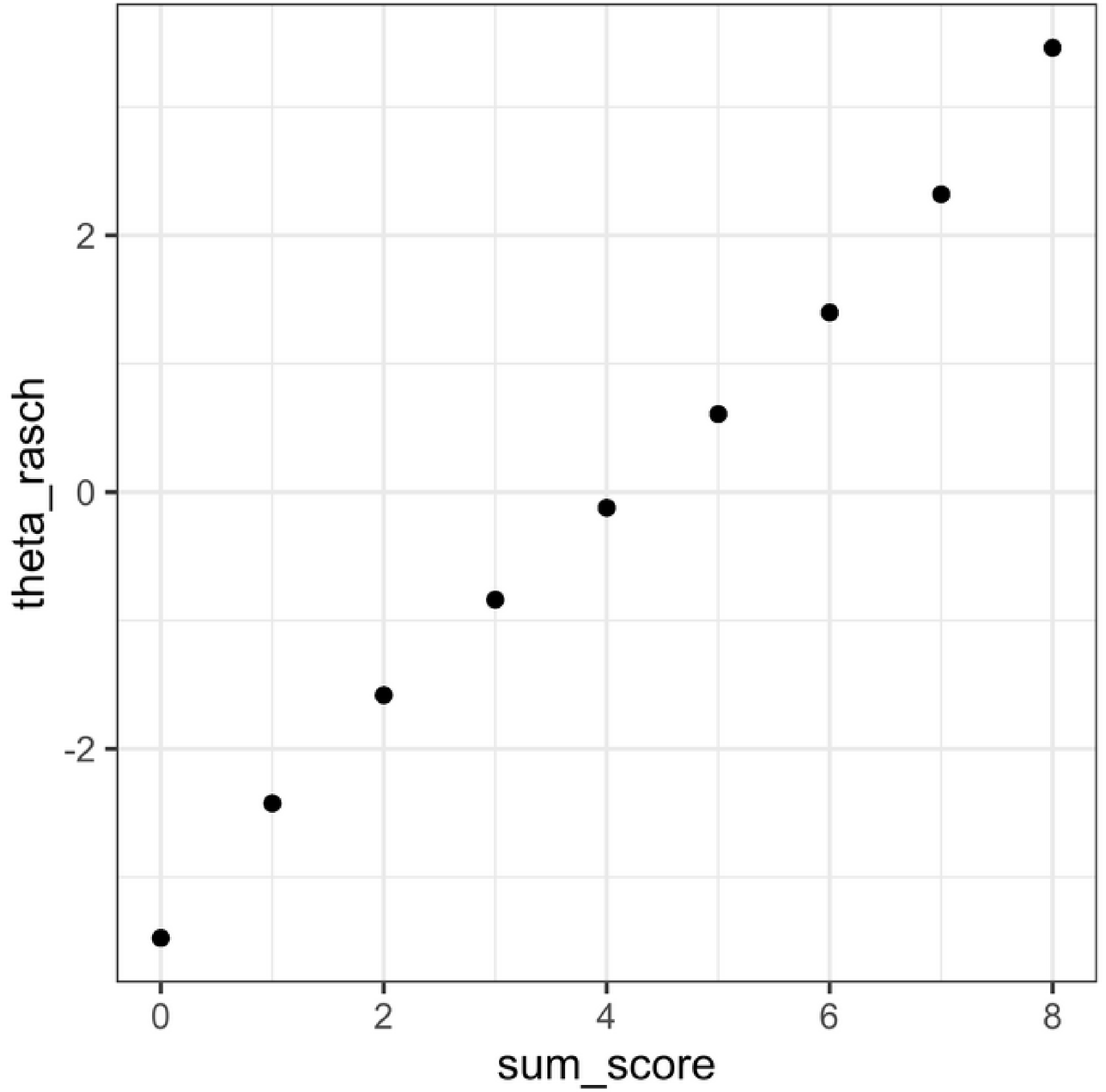
$$\sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

The idea is to fix the easiness and use MLE to estimate the abilities.

IRT assumptions

- **Monotonicity**. The probability of a person endorsing an item increases as the person's latent trait level increases.
- **Unidimensionality**. All items are contributing in the same way to the underlying latent trait.
- **Invariance**. Person trait levels do not depend on which items are administered nor on the particular sample of persons (subject to linear transformation). This enables linking of scales measuring the same construct. We can compare persons even if they responded to different items
- **Local independence**. Item responses are independent given a person's ability. They are uncorrelated after controlling for the latent trait.

In practice, 1PL is not that useful



2-Parameter Logistic Model (2PL)

Each person has ability θ_p. Each item has easiness b_i and discrimination a_i.

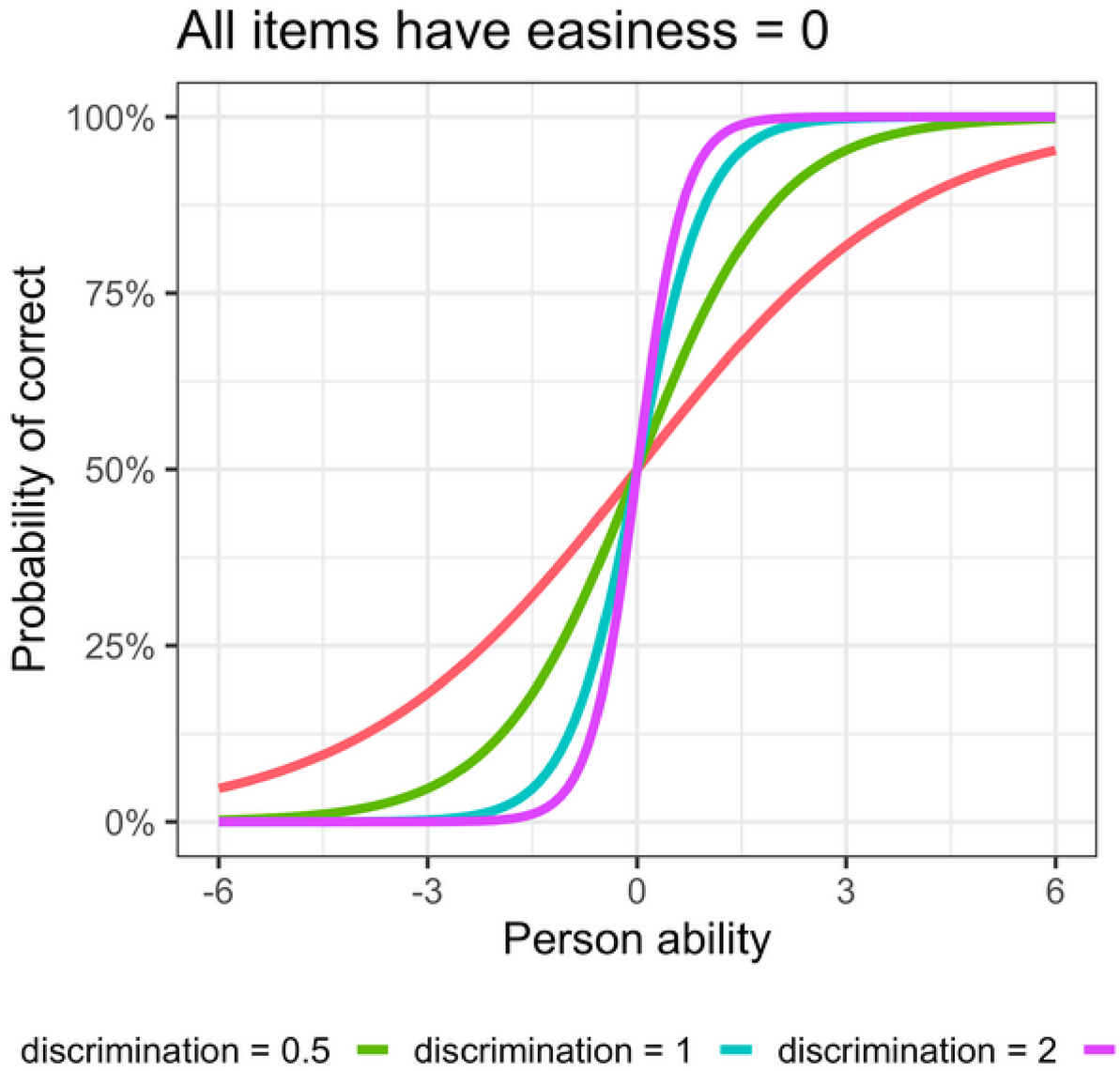
$$P(y_{pi} = 1 | \theta_p, b_i, a_i) = \sigma(a_i \cdot \theta_p + b_i)$$

Note: Intuitively, harder items discriminate better, but in this model they are assumed to be independent.

Note: Any discrimination above 1 is good enough.

Discrimination

The discrimination a_i describes the strength of the relationship between the item and ability

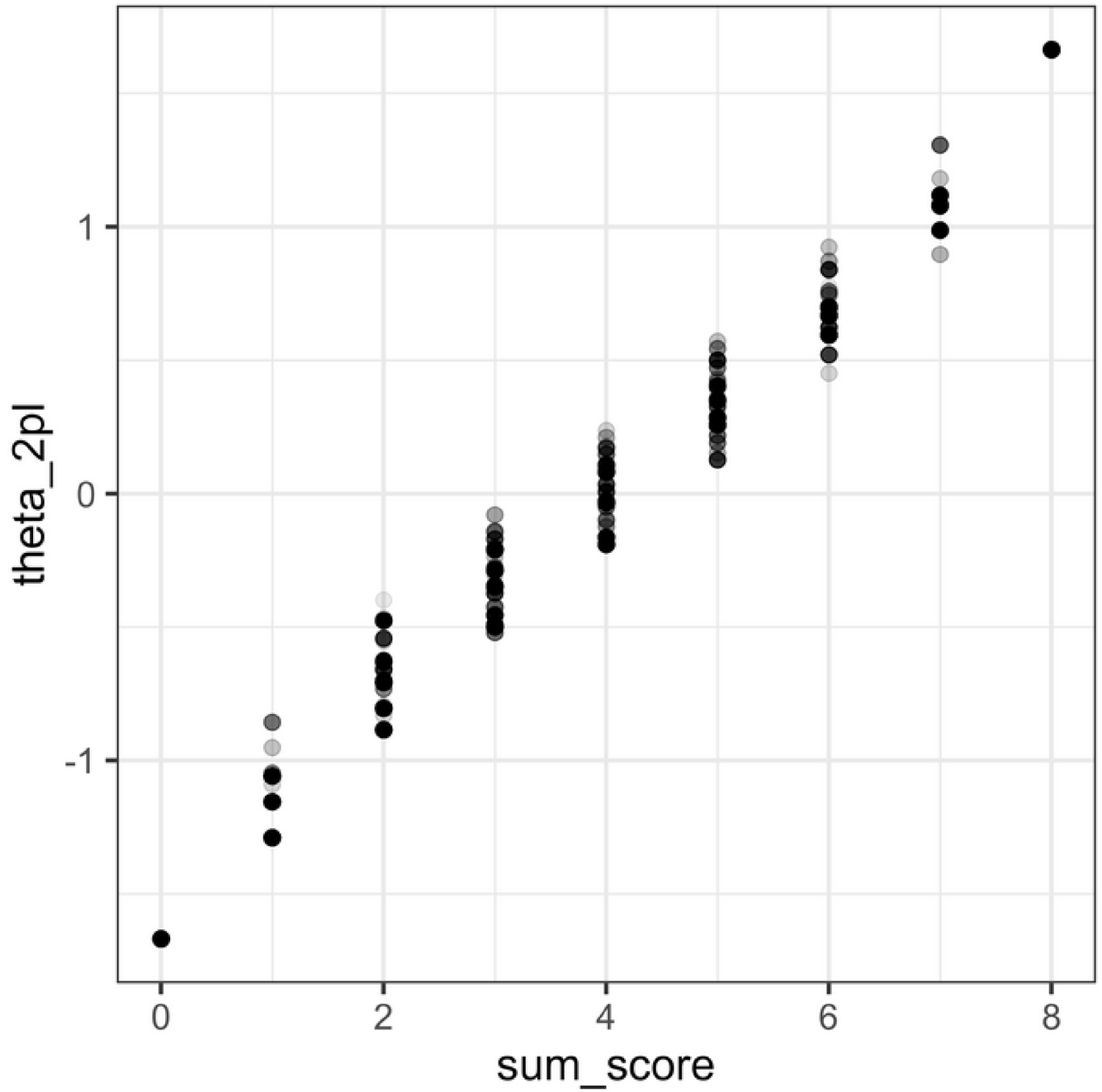


Question: Items with higher discrimination are … ?

A. Worse

B. Better

Fairness



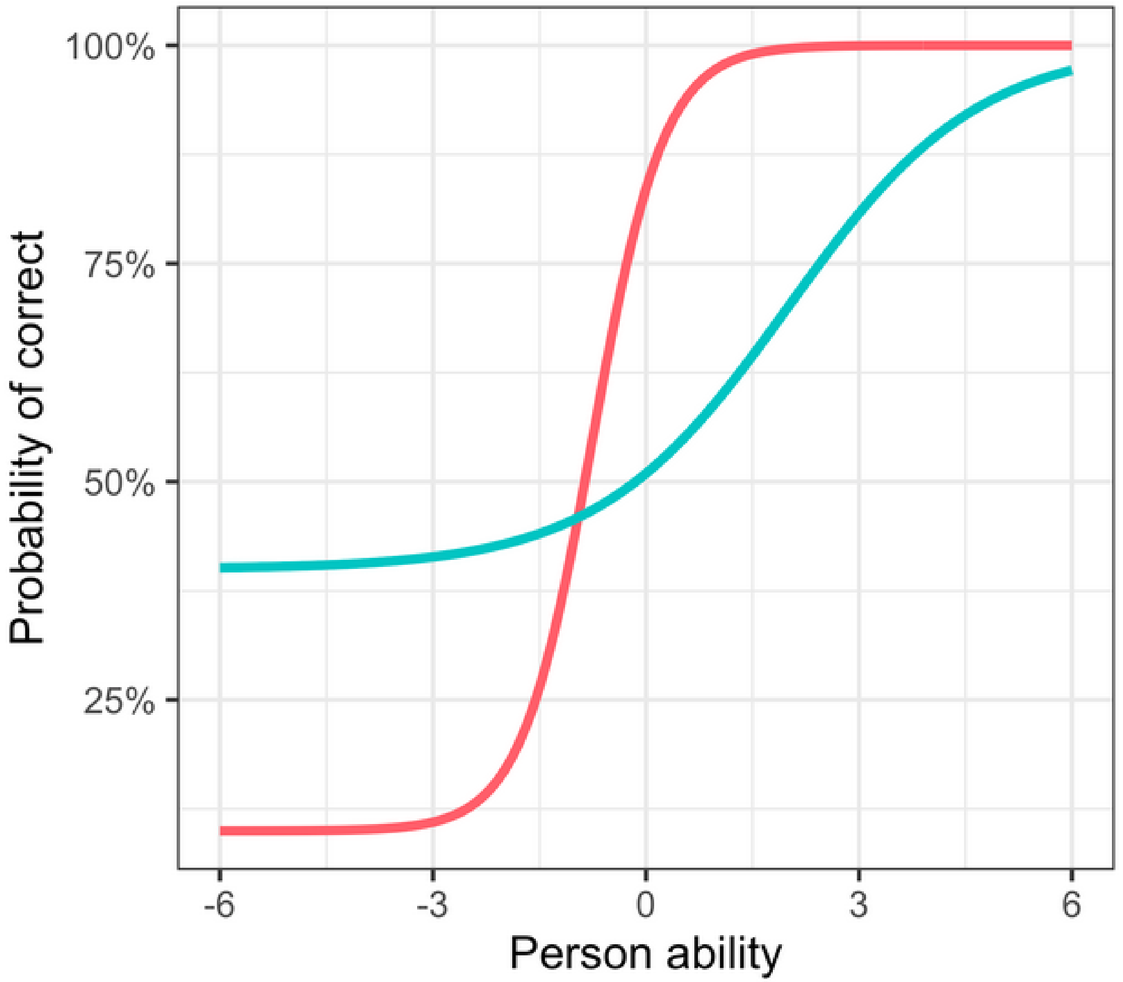
Why stop at two parameters?

Each person has ability θ_p. Each item has easiness b_i, discrimination a_i, and guessability g_i.

$$P(y_{pi} = 1|\theta_p, a_i, b_i, g_i) = g_i + (1 - g_i) \cdot \sigma(a_i \cdot \theta_p + b_i)$$

How to interpret the parameters?

- Easiness is horizontal translation
- Discrimination is slope
- Guessability is starting point at ability negative infinity



Comparing to sum_score