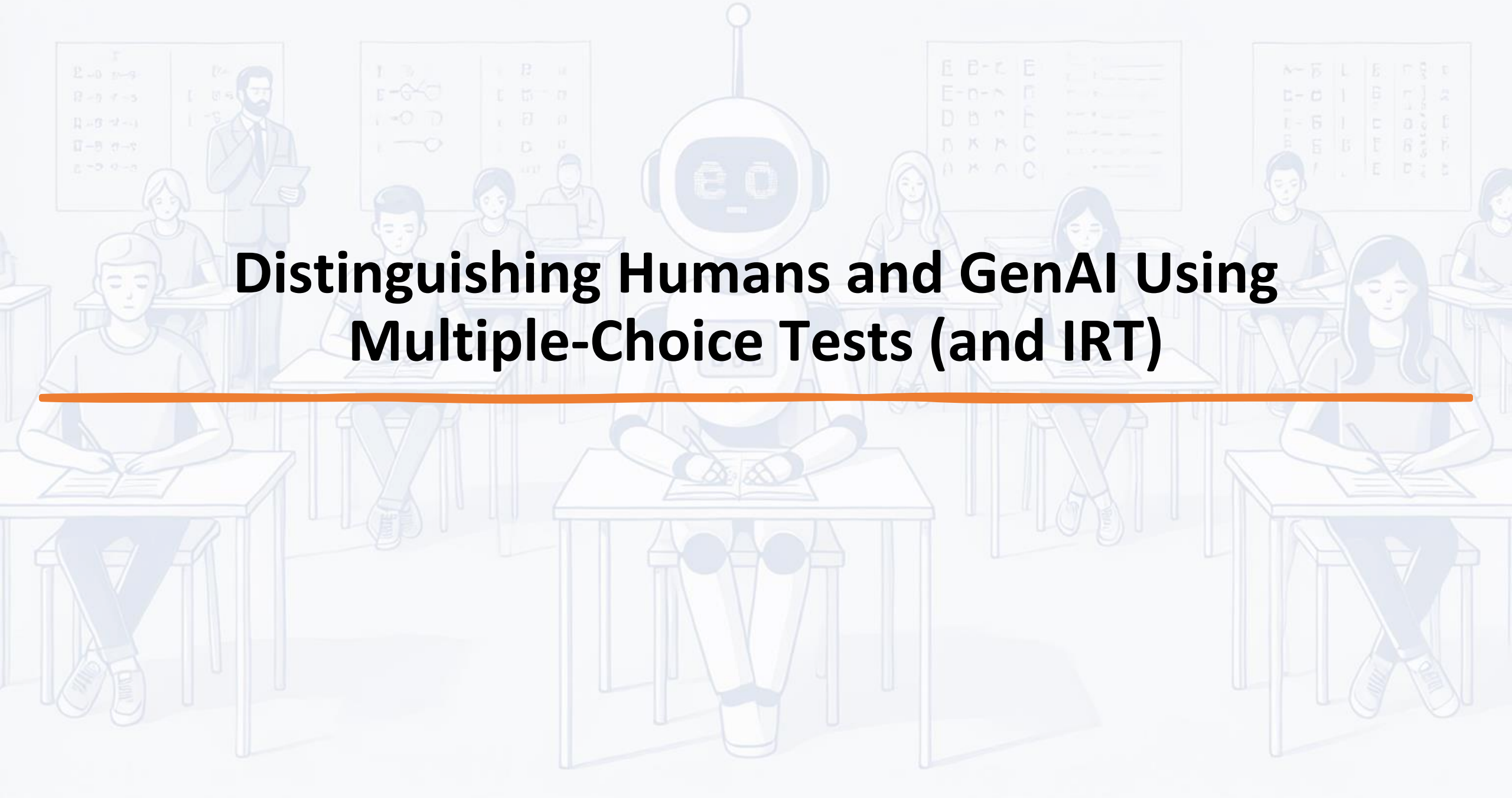


Distinguishing Humans and GenAI Using Multiple-Choice Tests (and IRT)

Remember at March 2023...

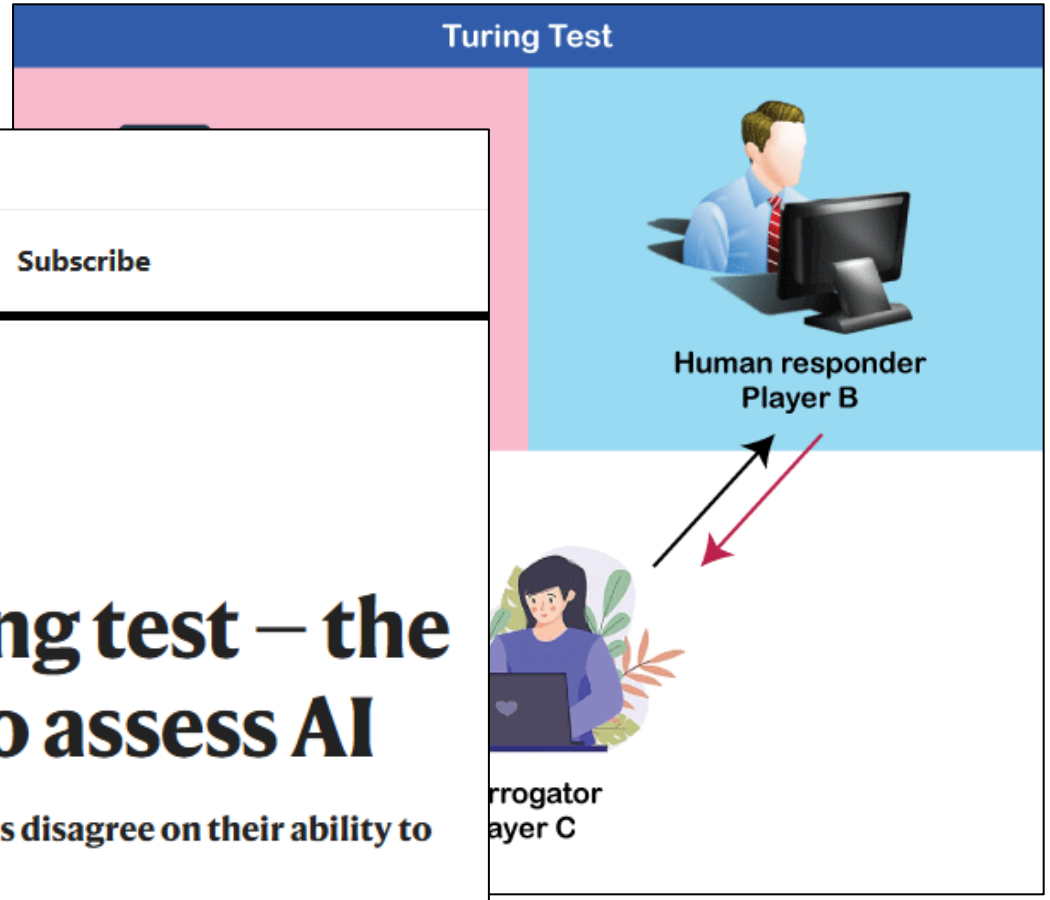
The Washington Post
Democracy Dies in Darkness

How a tiny company with few rules is making fake images go mainstream

Midjourney, the year-old firm behind recent fake visuals of Trump and the pope, illustrates the lack of oversight accompanying spectacular strides in AI

Generative AI Era Challenge

- Multiple modalities, impressive capabilities
- Classica
- Today, more complicated
- Many computational methods to detect GenAI



nature

Explore content | About the journal | Publish with us | Subscribe

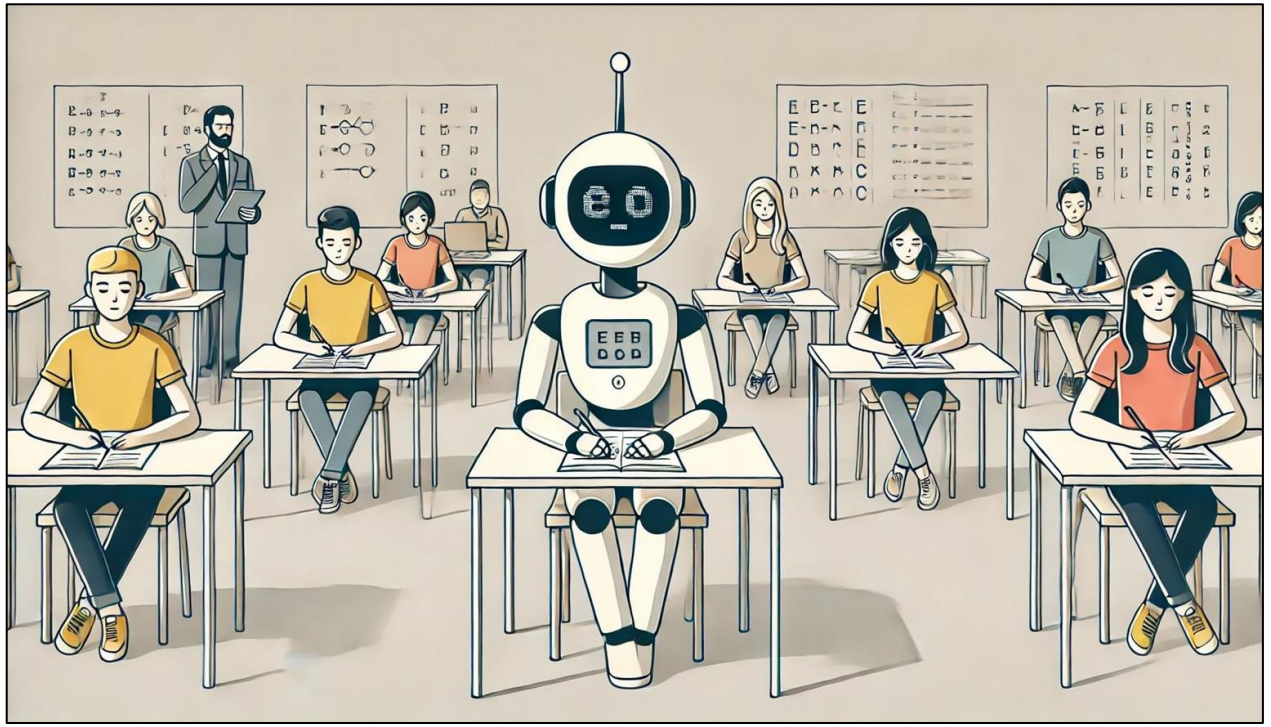
nature > news feature > article

NEWS FEATURE | 25 July 2023

ChatGPT broke the Turing test — the race is on for new ways to assess AI

Large language models mimic human chatter, but scientists disagree on their ability to reason.

In Education

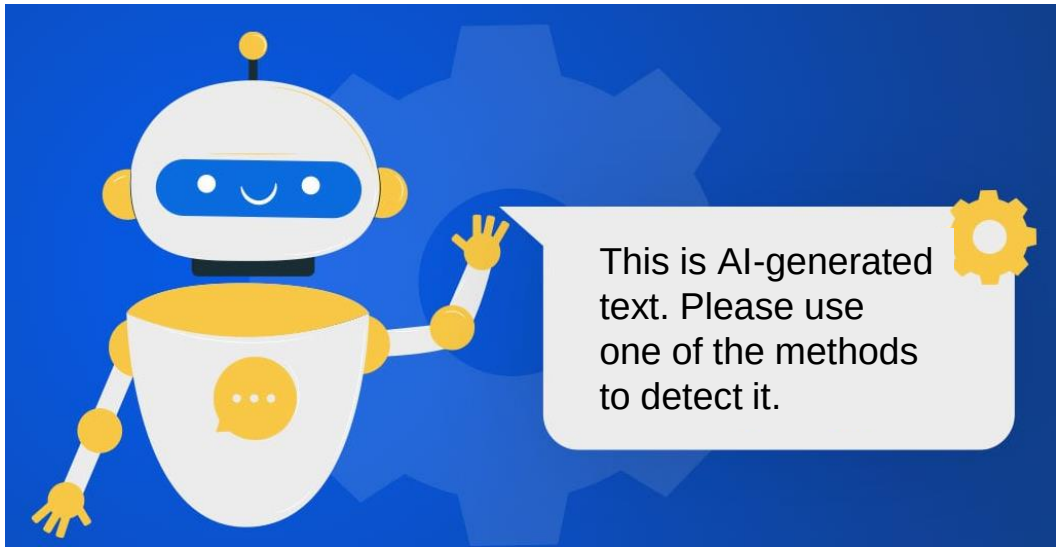


Compromising

- **Learning**
- **Assessment**

In Education

Open questions/essay



Multiple choice questions (MCQ's)

Unexplored

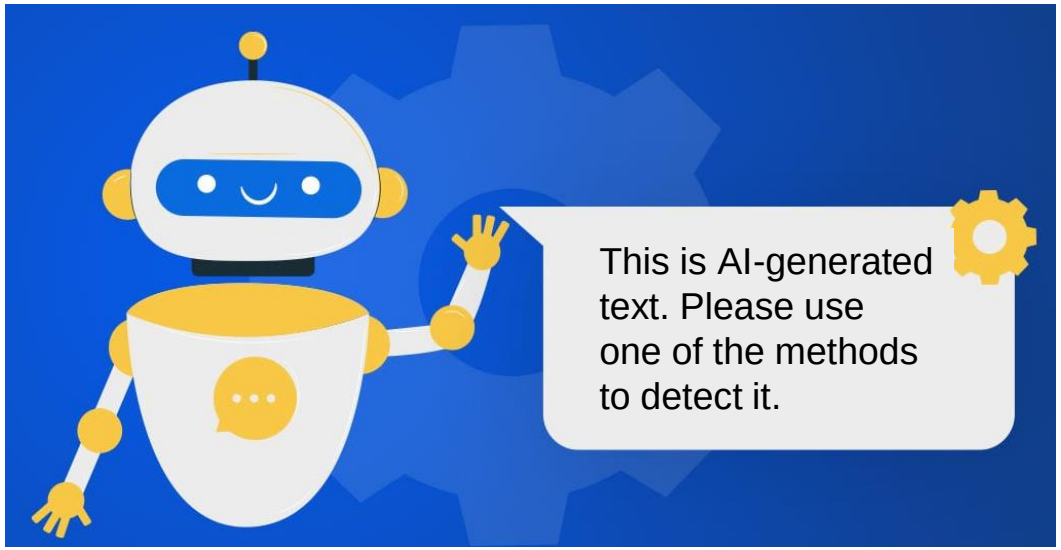
Allegedly "limited bits information"

*Classically the view is that **the whole test** measures something with noise.*

Wu et al. October 2023 , "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions"

In Education

Open questions/essay



Multiple choice questions (MCQ's)

Unexplored

Allegedly "limited bits information"

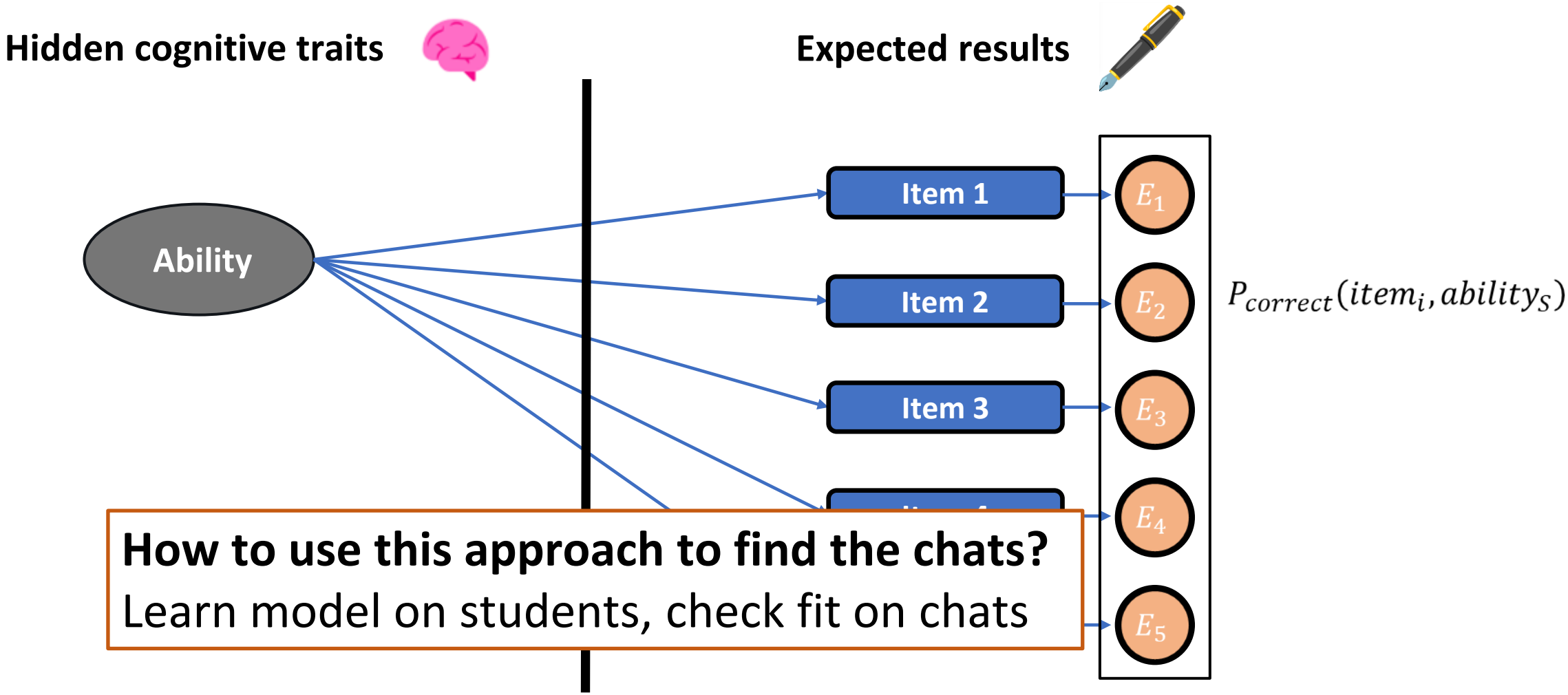
Classically the view is that ***the whole test*** *measures something with noise.*

not useful here

Wu et al. October 2023 , "A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions"

Approach- use Item Response Theory (IRT)

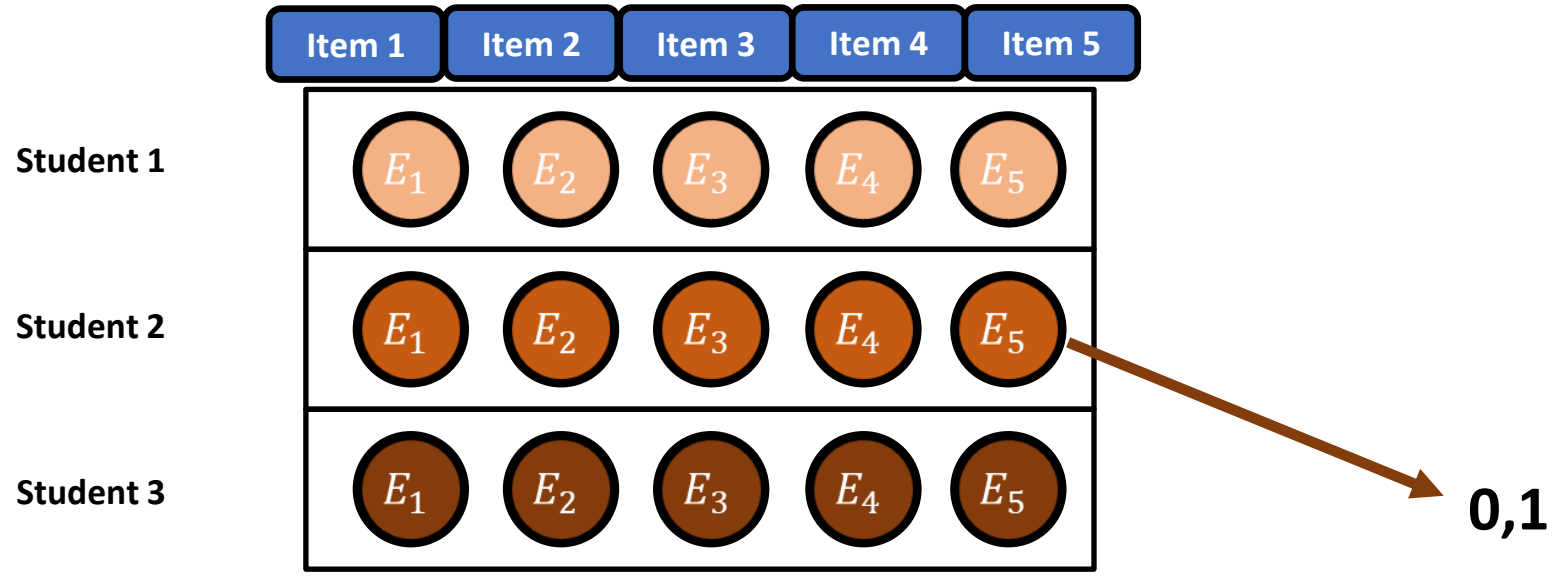
- Test\instrument - a collection of Items that measure the hidden trait(s)



How to use this approach to find the chats?
Learn model on students, check fit on chats

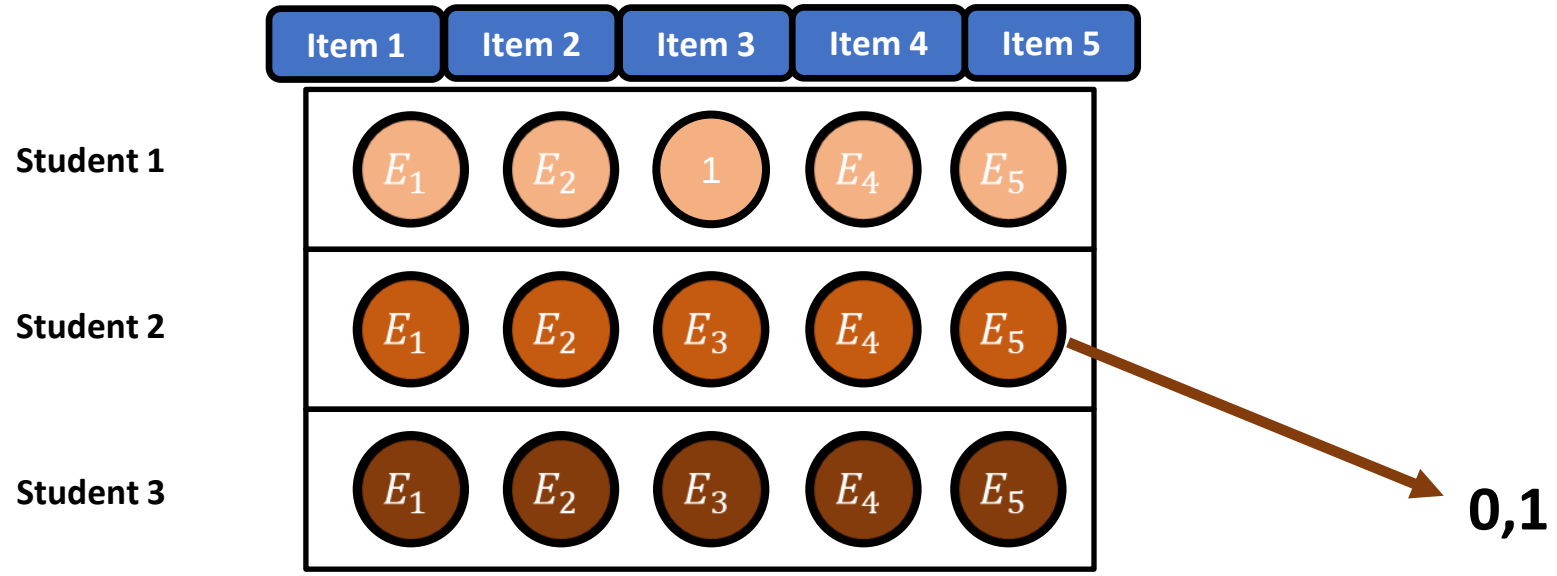
How we will measure ability - example

Response matrix:



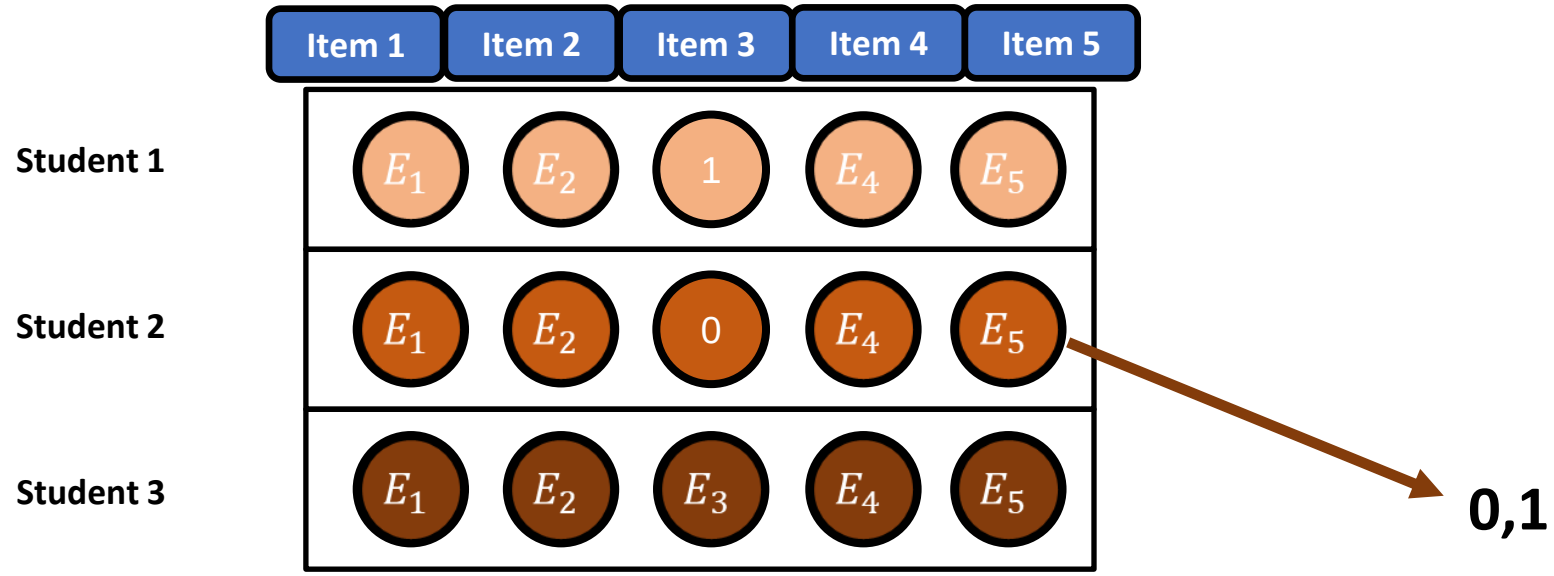
How we will measure ability - example

Response matrix:



How we will measure ability - example

Response matrix:



For more students and items we get this kind of response matrix:

How we will measure ability - example

Response matrix:

	Items												
Students	a	b	c	d	e	f	g	h	i	j	k	l	**score**
A	1	1	1	0	0	0	0	1	1	0	0	1	**6**
B	1	0	1	1	0	0	0	0	1	0	0	0	**4**
C	1	1	1	0	1	0	0	1	1	1	1	1	**9**
D	1	0	1	1	0	0	0	0	1	0	0	1	**5**
E	0	1	1	0	0	1	0	1	1	1	1	1	**8**
F	1	1	1	1	0	0	0	1	1	0	0	1	**7**
G	1	0	1	0	0	1	0	0	1	0	1	1	**6**
H	1	0	1	0	0	0	0	0	0	0	0	1	**3**
I	1	1	1	1	0	0	0	1	1	0	0	1	**7**

How we will measure ability - example

We rearrange the matrix:

	Items												
Students	a	b	c	d	e	f	g	h	i	j	k	l	**score**
A	1	1	1	0	0	0	0	1	1	0	0	1	**6**
B	1	0	1	1	0	0	0	0	1	0	0	0	**4**
C	1	1	1	0	1	0	0	1	1	1	1	1	**9**
D	1	0	1	1	0	0	0	0	1	0	0	1	**5**
E	0	1	1	0	0	1	0	1	1	1	1	1	**8**
F	1	1	1	1	0	0	0	1	1	0	0	1	**7**
G	1	0	1	0	0	1	0	0	1	0	1	1	**6**
H	1	0	1	0	0	0	0	0	0	0	0	1	**3**
I	1	1	1	1	0	0	0	1	1	0	0	1	**7**

Step 1:
score: high to low

How we will measure ability - example

We rearrange the matrix:

	Items												
Students	a	b	c	d	e	f	g	h	i	j	k	l	score
C	1	1	1	0	1	0	0	1	1	1	1	1	**9**
E	0	1	1	0	0	1	0	1	1	1	1	1	**8**
F	1	1	1	1	0	0	0	1	1	0	0	1	**7**
I	1	1	1	1	0	0	0	1	1	0	0	1	**7**
A	1	1	1	0	0	0	0	1	1	0	0	1	**6**
G	1	0	1	0	0	1	0	0	1	0	1	1	**6**
D	1	0	1	1	0	0	0	0	1	0	0	1	**5**
B	1	0	1	1	0	0	0	0	1	0	0	0	**4**
H	1	0	1	0	0	0	0	0	0	0	0	1	**3**

Ability

How we will measure difficulty- example

We rearrange the matrix:

Step 2:
Items: easy to hard

What is an easy item from the data?

Students	a	b	c	d	e	f	g	h	i	j	k	l	score
	Items												
C	1	1	1	0	1	0	0	1	1	1	1	1	**9**
E	0	1	1	0	0	1	0	1	1	1	1	1	**8**
F	1	1	1	1	0	0	0	1	1	0	0	1	**7**
I	1	1	1	1	0	0	0	1	1	0	0	1	**7**
A	1	1	1	0	0	0	0	1	1	0	0	1	**6**
G	1	0	1	0	0	1	0	0	1	0	1	1	**6**
D	1	0	1	1	0	0	0	0	1	0	0	1	**5**
B	1	0	1	1	0	0	0	0	1	0	0	0	**4**
H	1	0	1	0	0	0	0	0	0	0	0	1	**3**

Ability

All students answered correctly

How we will measure difficulty- example

We rearrange the matrix:

Step 2:

Items: easy to hard

	Items												
Students	c	i	a	l	e	f	g	h	b	j	k	d	**Ability**
C	1	1	1	1	1	0	0	1	1	1	1	0	**9**
E	1	1	0	1	0	1	0	1	1	1	1	0	**8**
F	1	1	1	1	0	0	0	1	1	0	0	1	**7**
I	1	1	1	1	0	0	0	1	1	0	0	1	**7**
A	1	1	1	1	0	0	0	1	1	0	0	0	**6**
G	1	1	1	1	0	1	0	0	0	0	1	0	**6**
D	1	1	1	1	0	0	0	0	0	0	0	1	**5**
B	1	1	1	0	0	0	0	0	0	0	0	1	**4**
H	1	0	1	1	0	0	0	0	0	0	0	0	**3**

How we will measure difficulty- example

We rearrange the matrix:

Step 2:

Items: easy to hard

	Items												
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**
F	1	1	1	1	1	1	1	0	0	0	0	0	**7**
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**
D	1	1	1	1	0	0	1	0	0	0	0	0	**5**
B	1	1	1	0	0	0	1	0	0	0	0	0	**4**
H	1	0	1	1	0	0	0	0	0	0	0	0	**3**

How we will measure difficulty- example

We rearrange the matrix:

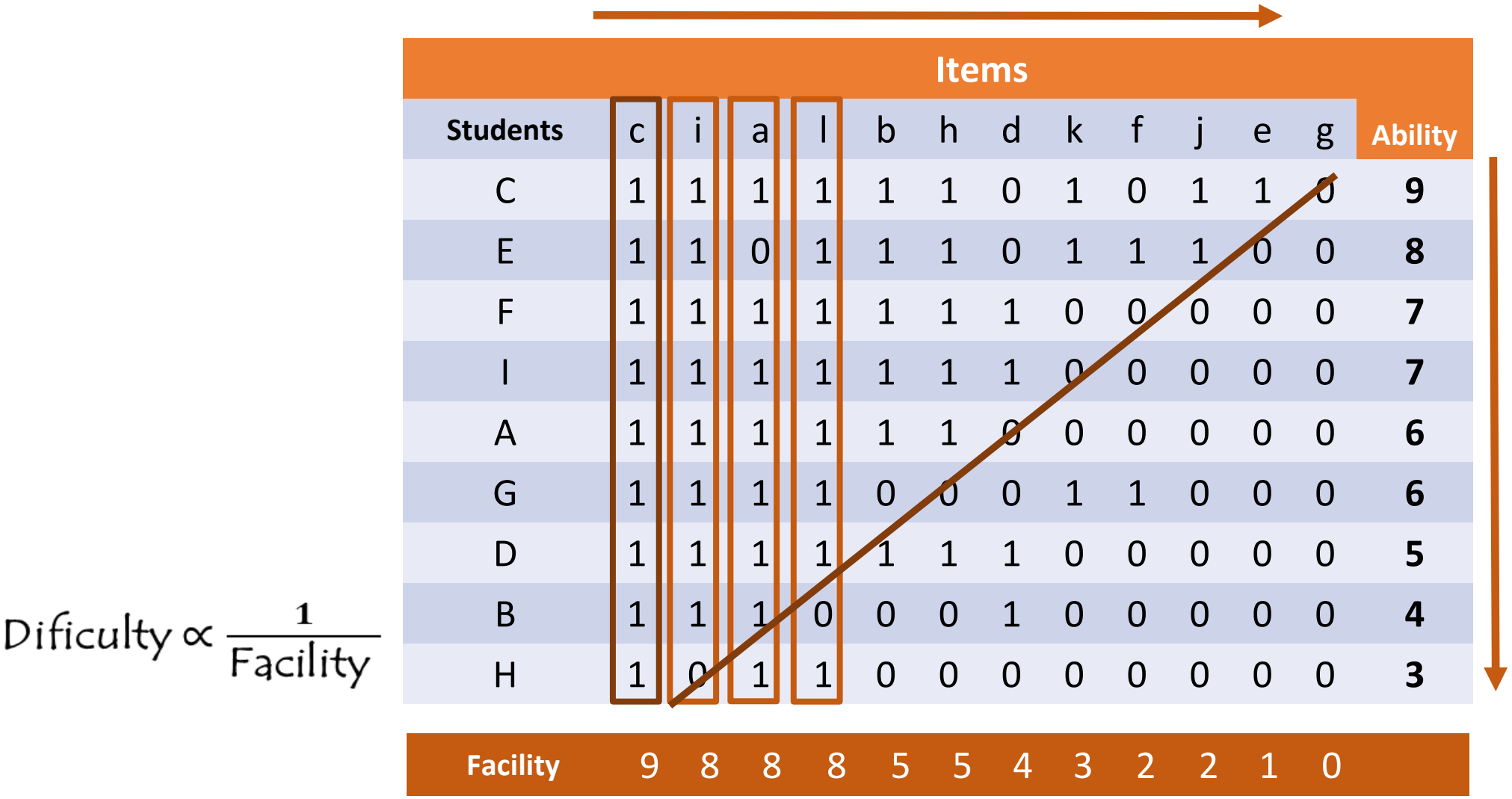
Step 2:

Items: easy to hard

	Items												
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**
F	1	1	1	1	1	1	1	0	0	0	0	0	**7**
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**
D	1	1	1	1	0	0	1	0	0	0	0	0	**5**
B	1	1	1	0	0	0	1	0	0	0	0	0	**4**
H	1	0	1	1	0	0	0	0	0	0	0	0	**3**
Facility	9	8	8	8	5	5	4	3	2	2	1	0	

$$\text{Dificulty} \propto \frac{1}{\text{Facility}}$$

“Expected” pattern



“Expected” pattern

That can have small deviations

Items													
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**
F	1	1	0	1	1	1	1	0	0	0	0	0	**7**
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**
D	1	1	1	1	1	1	1	0	0	0	0	0	**5**
B	1	1	1	1	0	0	1	0	0	0	0	0	**4**
H	1	1	1	1	0	0	0	0	0	0	0	0	**3**
Facility	9	8	8	8	5	5	4	3	2	2	1	0	

$$\text{Dificulty} \propto \frac{1}{\text{Facility}}$$

“Expected” pattern

What happens with large deviations?

	Items												
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**
F	1	1	0	1	1	1	1	0	0	0	0	0	**7**
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**
D	1	1	1	1	1	1	1	0	0	0	0	0	**5**
B	1	1	1	1	0	0	1	0	0	0	0	0	**4**
H	1	1	1	1	0	0	0	0	0	0	0	0	**3**
Facility	9	8	8	8	5	5	4	3	2	2	1	0	

$$\text{Dificulty} \propto \frac{1}{\text{Facility}}$$

“Expected” pattern

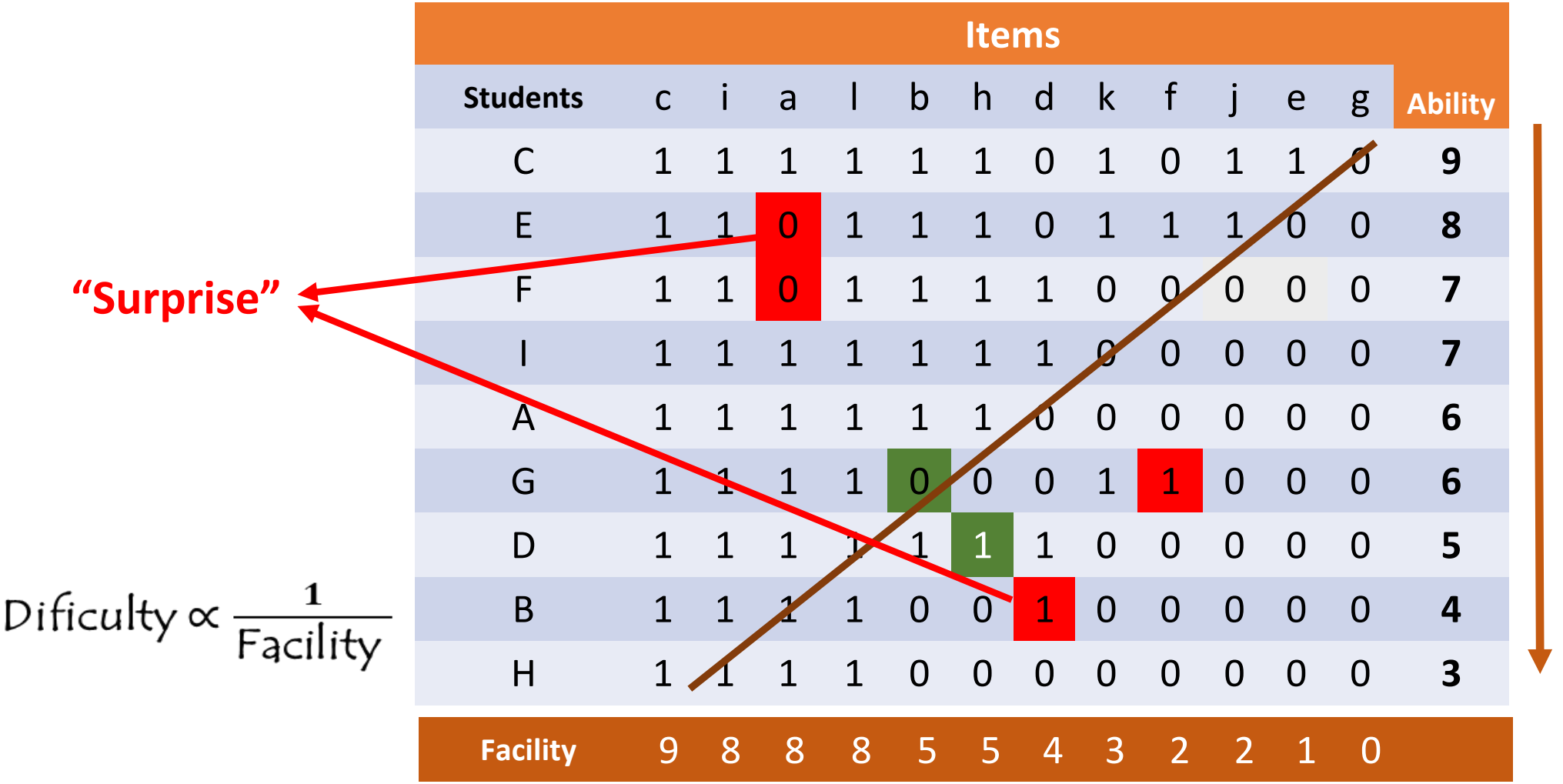
What happens with large deviations?

Items													
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**
F	1	1	0	1	1	1	1	0	0	0	0	0	**7**
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**
D	1	1	1	1	1	1	1	0	0	0	0	0	**5**
B	1	1	1	1	0	0	1	0	0	0	0	0	**4**
H	1	1	1	1	0	0	0	0	0	0	0	0	**3**
Facility	9	8	8	8	5	5	4	3	2	2	1	0	

“Surprise”

$$\text{Dificulty} \propto \frac{1}{\text{Facility}}$$

Calculate an anomaly measure



Calculate an anomaly measure

Students	Items												Ability	Anomaly
	c	i	a	l	b	h	d	k	f	j	e	g		
C	1	1	1	1	1	1	0	1	0	1	1	0	**9**	5
E	1	1	0	1	1	1	0	1	1	1	0	0	**8**	9
F	1	1	0	1	1	1	1	0	0	0	0	0	**7**	4
I	1	1	1	1	1	1	1	0	0	0	0	0	**7**	0
A	1	1	1	1	1	1	0	0	0	0	0	0	**6**	0
G	1	1	1	1	0	0	0	1	1	0	0	0	**6**	6
D	1	1	1	1	1	1	1	0	0	0	0	0	**5**	0
B	1	1	1	1	0	0	1	0	0	0	0	0	**4**	2
H	1	1	1	1	0	0	0	0	0	0	0	0	**3**	0
Facility	9	8	8	8	5	5	4	3	2	2	1	0		

"Surprise"

$$\text{Dificulty} \propto \frac{1}{\text{Facility}}$$

What will happen if we have a new group of student?

	Items													
Students	c	i	a	l	b	h	d	k	f	j	e	g	**Ability**	**Anomaly**
C	1	1	1	1	1	1	0	1	0	1	1	0	9	5
E	1	1	0	1	1	1	0	1	1	1	0	0	8	9
F	1	1	0	1	1	1	1	0	0	0	0	0	7	4
I	1	1	1	1	1	1	1	0	0	0	0	0	7	0
A	1	1	1	1	1	1	0	0	0	0	0	0	6	0
G	1	1	1	1	0	0	0	1	1	0	0	0	6	6
D	1	1	1	1	1	1	1	0	0	0	0	0	5	0
B	1	1	1	1	0	0	1	0	0	0	0	0	4	2
H	1	1	1	1	0	0	0	0	0	0	0	0	3	0
J														
K														
L														
M														

What will happen if we have a new group of student?

Items														
Students	c	i	a	l	b	h	d	k	f	j	e	g	Ability	Anomaly
C	1	1	1	1	1	1	0	1	0	1	1	0	9	5
E	1	1	0	1	1	1	0	1	1	1	0	0	8	9
F	1	1	0	1	1	1	1	0	0	0	0	0	7	4
I	1	1	1	1	1	1	1	0	0	0	0	0	7	0
A	1	1	1	1	1	1	0	0	0	0	0	0	6	0
G	1	1	1	1	0	0	0	1	1	0	0	0	6	6
D	1	1	1	1	1	1	1	0	0	0	0	0	5	0
B	1	1	1	1	0	0	1	0	0	0	0	0	4	2
H	1	1	1	1	0	0	0	0	0	0	0	0	3	0
J	1	1	1	1	1	1	1	0	0	1	1	0	9	
K	1	1	1	1	1	1	0	0	1	0	0	0	7	
L	1	1	1	1	0	0	1	1	1	0	0	0	7	
M	1	1	1	1	0	0	0	0	0	0	0	0	4	

✓ Similar group, similar anomaly measure

What will happen if we have a new group of student?

Students	Items												Ability	Anomaly
	c	i	a	l	b	h	d	k	f	j	e	g		
C	1	1	1	1	1	1	0	1	0	1	1	0	9	5
E	1	1	0	1	1	1	0	1	1	1	0	0	8	9
F	1	1	0	1	1	1	1	0	0	0	0	0	7	4
I	1	1	1	1	1	1	1	0	0	0	0	0	7	0
A	1	1	1	1	1	1	0	0	0	0	0	0	6	0
G	1	1	1	1	0	0	0	1	1	0	0	0	6	6
D	1	1	1	1	1	1	1	0	0	0	0	0	5	0
B	1	1	1	1	0	0	1	0	0	0	0	0	4	2
H	1	1	1	1	0	0	0	0	0	0	0	0	3	0
J														
K														
L														
M														

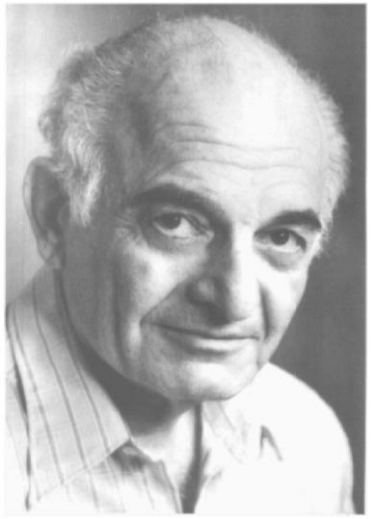
Not Similar

What will happen if we have a new group of student?

Students	Items												Ability	Anomaly
	c	i	a	l	b	h	d	k	f	j	e	g		
C	1	1	1	1	1	1	0	1	0	1	1	0	9	5
E	1	1	0	1	1	1	0	1	1	1	0	0	8	9
F	1	1	0	1	1	1	1	0	0	0	0	0	7	4
I	1	1	1	1	1	1	1	0	0	0	0	0	7	0
A	1	1	1	1	1	1	0	0	0	0	0	0	6	0
G	1	1	1	1	0	0	0	1	1	0	0	0	6	6
D	1	1	1	1	1	1	1	0	0	0	0	0	5	0
B	1	1	1	1	0	0	1	0	0	0	0	0	4	2
H	1	1	1	1	0	0	0	0	0	0	0	0	3	0
J	1	1	0	0	1	0	1	0	1	0	0	1	6	
K	0	0	1	1	0	1	0	0	1	0	1	0	5	
L	0	1	0	0	1	0	1	0	0	1	0	1	5	
M	1	0	0	0	1	1	0	0	1	0	0	1	5	

Not Similar

What will happen if we have a new group of student?



"Guttman error"

	Items													
Students	c	i	a	l	b	h	d	k	f	j	e	g	Ability	Anomaly
C	1	1	1	1	1	1	0	1	0	1	1	0	9	4
E	1	1	0	1	1	1	0	1	1	1	0	0	8	7
F	1	1	0	1	1	1	1	0	0	0	0	0	7	0
I	1	1	1	1	1	1	1	0	0	0	0	0	7	2
A	1	1	1	1	1	1	0	0	0	0	0	0	6	0
G	1	1	1	1	0	0	0	1	1	0	0	0	6	4
D	1	1	1	1	1	1	1	0	0	0	0	0	5	2
B	1	1	1	1	0	0	1	0	0	0	0	0	4	6
H	1	1	1	1	0	0	0	0	0	0	0	0	3	0
J	1	1	0	0	1	0	1	0	1	0	0	1	6	15
K	0	0	1	1	0	1	0	0	1	0	1	0	5	18
L	0	1	0	0	1	0	1	0	0	1	0	1	5	21
M	1	0	0	0	1	1	0	0	1	0	0	1	5	18

G

Our working assumption:

Chatbots "behave" as students with different latent traits

- Two different tests – **with students' responses per item**
  - **Formative -** Chemistry MCQ online *"preparation to matriculation"* test
  - **Summative** - Quantitative Psychometrics part*
- Pollution level -5%
- Three premium version chatbots
- Four anomaly measures *G, G*,U, ZU3- Person Fit Statistics*

*Acknowledgements to National Institute for Testing an Evaluation (NITE) for providing the data

Results

Chatbots “behave” as students with different latent traits

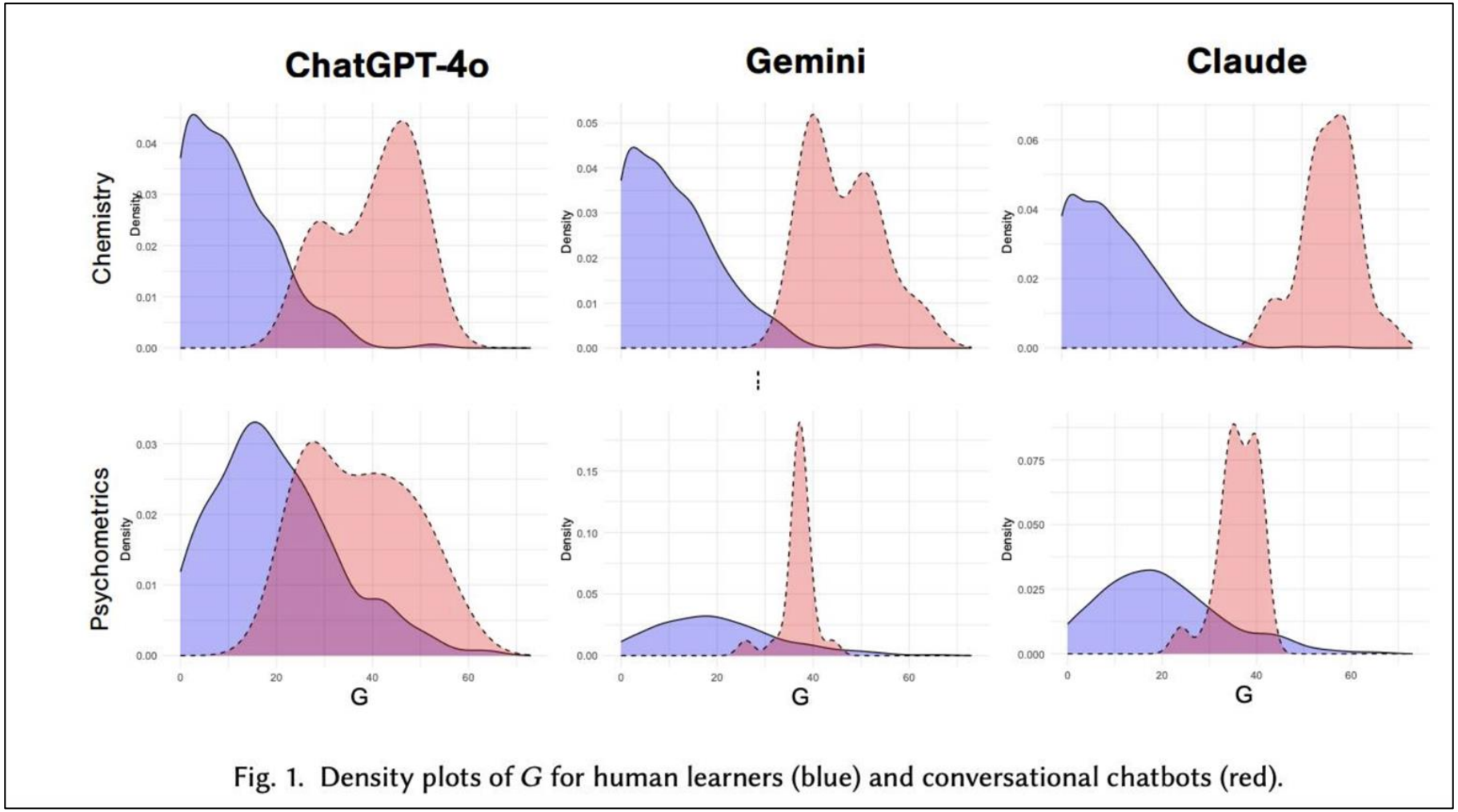
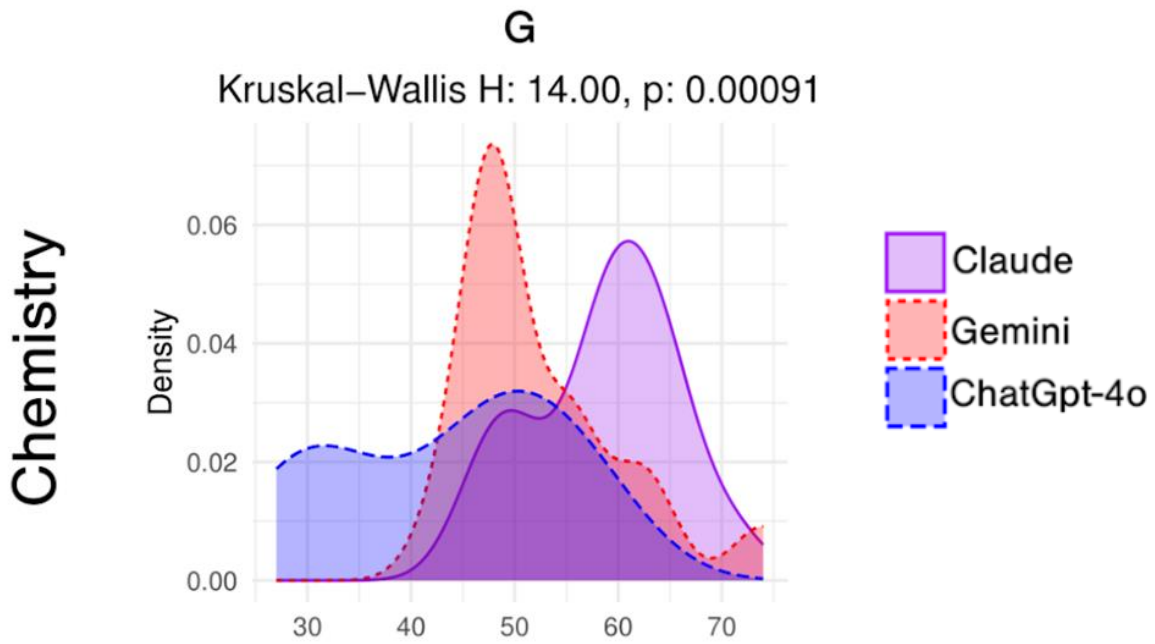


Fig. 1. Density plots of G for human learners (blue) and conversational chatbots (red).

What if the pollution level is increased?

- Results until now: pollution level 5%
- We also tried with 10%, 25%.
- Pollution level ↑ ⇒ Separation humans chats ↓
- Widespread use of chatbots *means* **harder separation**

Do different chatbots behave differently?



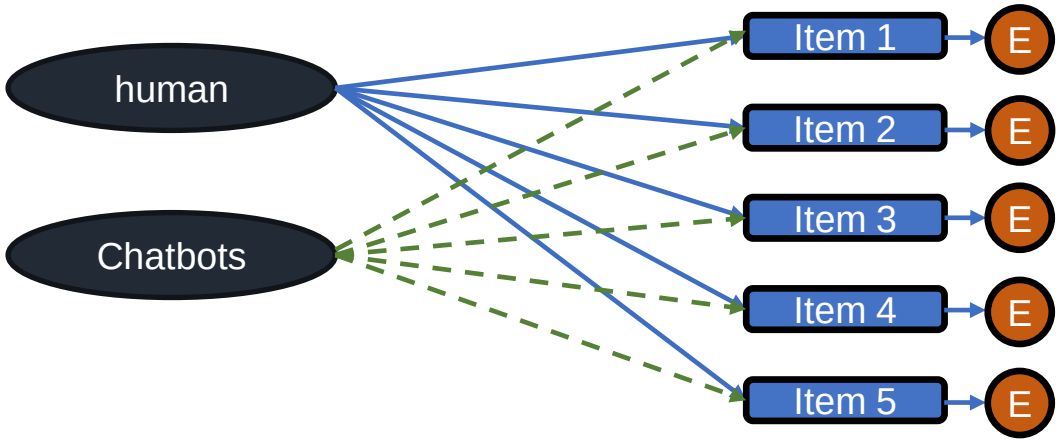
By using the same measures we show:
Chats <u>can be separated</u> using MCQ test

Who from who?
Depends on the test and on the specific measure

To conclude…

- **Humans vs. chatbots** –

  Have different latent traits; can be distinguished using MCQ
- **Chatbots vs. other chatbots**
- Widespread use of chatbots *means* **harder separation**



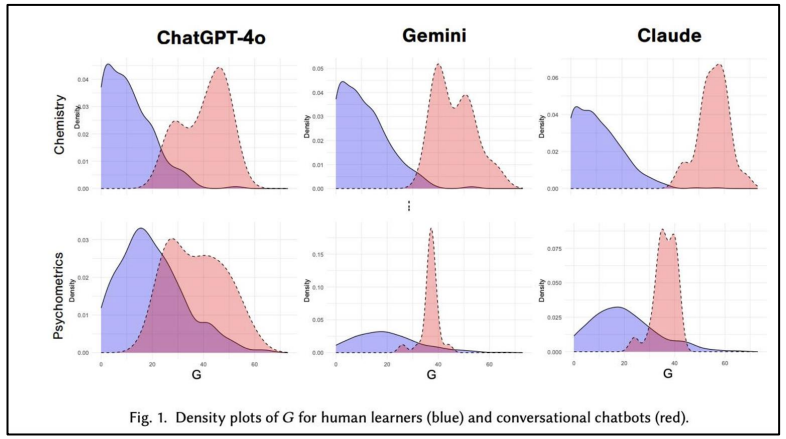
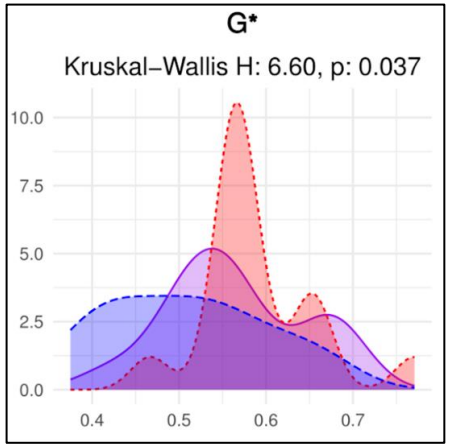


Fig. 1. Density plots of *G* for human learners (blue) and conversational chatbots (red).



THANKS FOR LISTENING

Link to the paper: