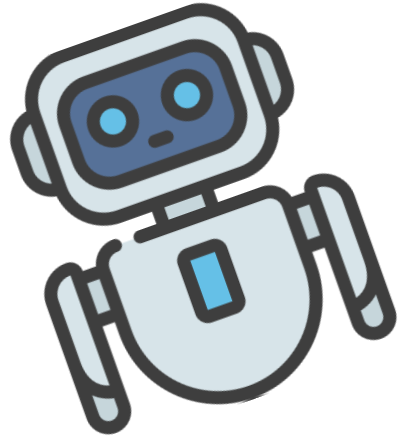


Assessment Design in the AI Era:

Applying Psychometric Theory to Identify Items on Which Humans and Chatbots Diverge

Licol Zeinfeld, Alona Strugatski, Ziva Bar-Dov,

Ron Blonder, Shelley Rap, Giora Alexandron

CASEd Research Group

Generative AI in Education Challenge

- GenAI matches/exceeds student performance on many assessments across domains
- Widespread use
- Learning and assessment may be compromised
- Need to rethink assessment

RESEARCH ARTICLE

ChatGPT performance on multiple choice question examinations in higher education

Philip Newton & Maira Xiromeriti

Assessment & Evaluation in Higher Education • 2024
doi:10.1080/02602938.2023.2299059

"ChatGPT can pass examinations based on multiple choice questions (MCQs), including those used to qualify doctors, lawyers, scientists etc."

Newton & Xiromeriti (2024) • Abstract • Pragmatic scoping review

The Guardian

Int

Universities Students

Higher education

UK universities warned to 'stress-test' assessments as 92% of students use AI

Survey of 1,000 students shows 'explosive increase' in use of generative AI in particular over past 12 months

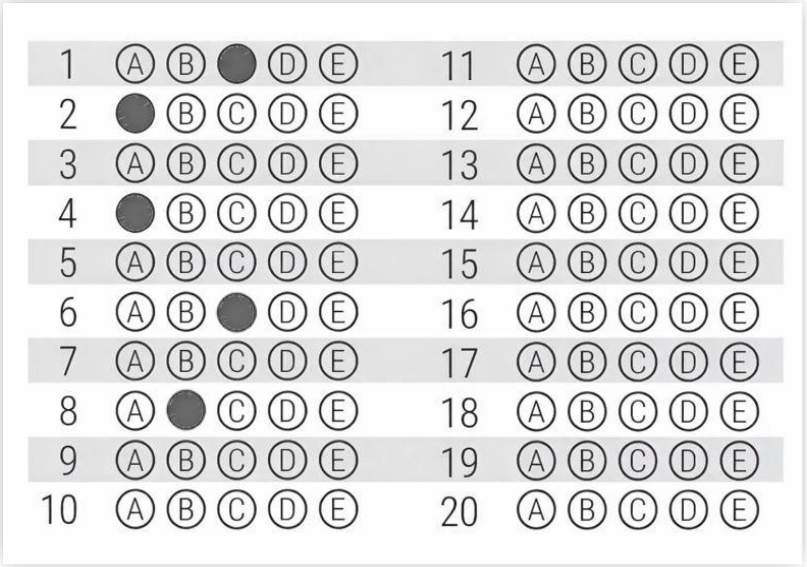
Sally Weale *Education correspondent*

Wed 26 Feb 2025 00.01 GMT

Share

Rethinking Assessment

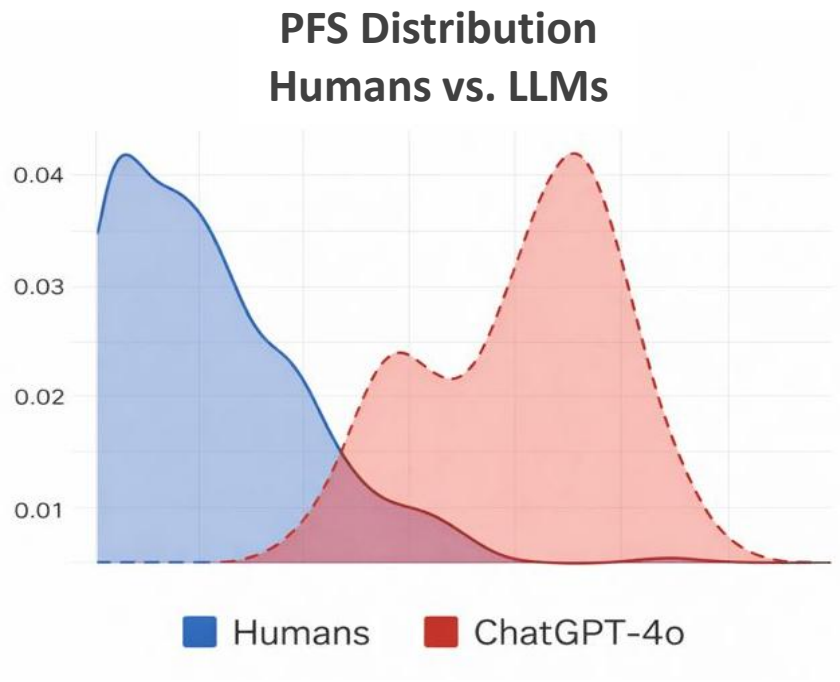
- Adapt assessment to growing GenAI presence
- Understand LLMs strengths and weaknesses
  - Current LLM evals rely on descriptive statistics
  - Psychometrics offer principled ways to study GenAI performance
- A lot of work on LLM responses to text-rich prompts
- Very little work on multiple-choice assessments



Applying psychometrics to evaluate GenAI capabilities*

- Using person-fit statistics (PFS) to distinguish between humans and LLM 'test-takers' on MCQ assessments
- **No information on the item** and the **task dimensions** that distinguish humans/LLMs

=> We **need item-level analysis**



* Strugatski, A., & Alexandron, G. (2025). Applying IRT to Distinguish Between Human and Generative AI Responses. LAK '25

Item Level Analysis Using Differential Item Functioning (DIF)

- DIF are methods to identify items on which two groups differ

Example:

  - A math test where some items rely heavily on reading comprehension
  - Second language learners (SLLs) may be **disadvantaged** compared to native speakers of the same math skill
  - SLLs are termed the 'focal' group, and native speakers the 'reference' group

- **We apply DIF to LLMs**: LLMs are the 'focal' group and Humans are the 'reference'

Goals and Research Questions (RQs)

Goals: Examine whether DIF methods can be useful to **identify** items on which humans and LLMs differ and **characterize** the task dimensions that make items easy/difficult for LLMs

Research Questions:

RQ1: Can DIF techniques identify items that function differently for humans and LLMs?

RQ2: What are the key characteristics that subject-matter experts identify in items on which LLMs exhibit differential performance compared to humans?

Two DIF Methods being used

Mantel–Haenszel (MH-DIF):

- Simple method that is based on odds-ratio for each ability level
- Requires relatively less response data
- Identifies DIF that is uniform across all ability levels

	LLMs	Humans
Correct	a	b
Incorrect	c	d

odds(LLM) = a/c

odds(human) = b/d

$$\text{odds ratio} = \frac{\text{odds(LLM)}}{\text{odds(human)}}$$

Logistic Regression-based (LR-DIF):

- Fits a logistic regression function to the responses
- Can identify more sophisticated DIF patterns (non-uniform, sign-change)
- Requires more data

Instruments & Data

Instruments & human data:

- Chemistry exam (931 high school students)
- Math sections from the Psychometric exam* (~4800 examinees)

GenAI models:

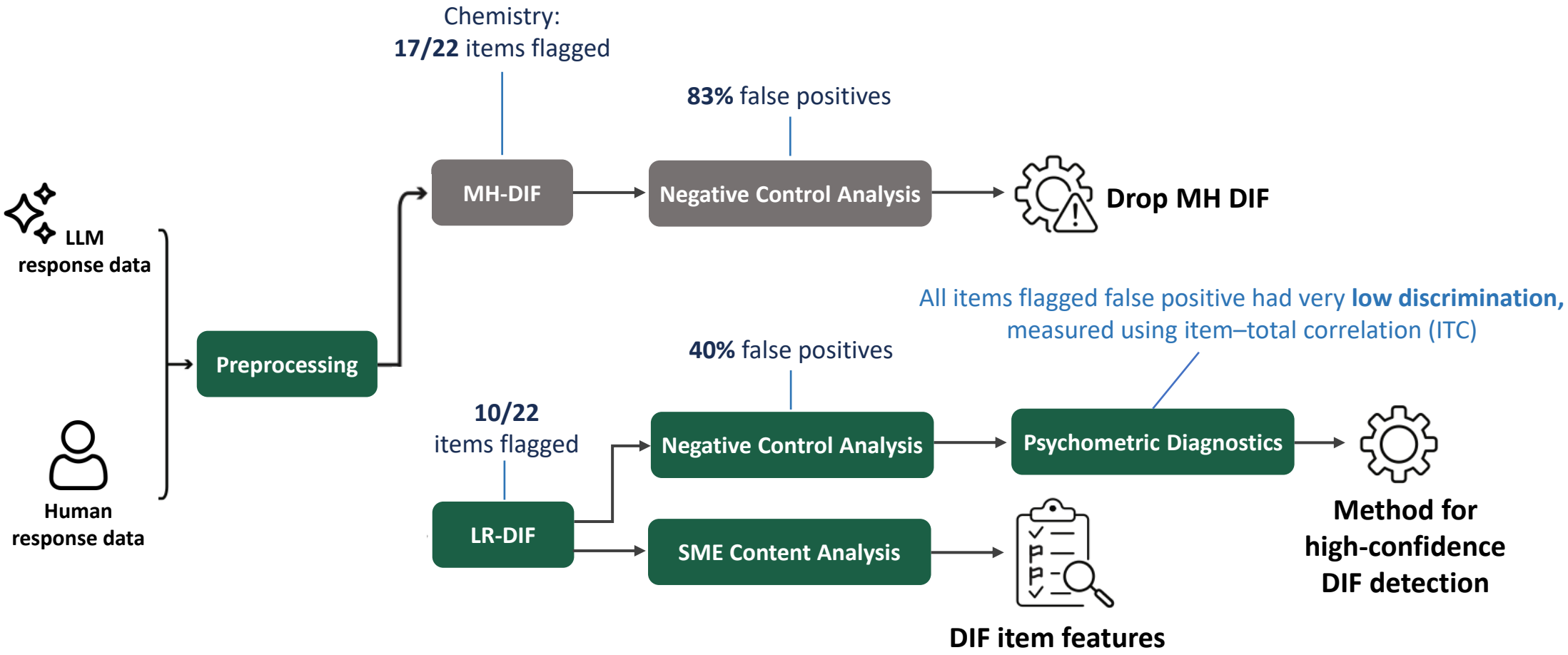
- ChatGPT-4o & 5.2, Gemini 1.5 & 3 Pro, Claude 3.5 & 4.5 Sonnet

Claude

Gemini

*Thanks NITE for providing access to the data used in this research

Procedure


Chemistry:
17/22 items flagged
83% false positives
MH-DIF
Negative Control Analysis
Drop MH DIF
LLM
response data
Preprocessing
Human
response data
All items flagged false positive had very low discrimination,
measured using item–total correlation (ITC)
40% false positives
10/22
items flagged
Negative Control Analysis
Psychometric Diagnostics
LR-DIF
SME Content Analysis
Method for
high-confidence
DIF detection
DIF item features

Results – LR-DIF (partial results for RQ1 and RQ2)

RQ1:

LR-DIF flagged items:

Inst.	POS	NEG
Psych.	$\bar{1}1$, $\tilde{1}3$-$\tilde{1}5$, $\tilde{3}4$, $\bar{4}0$	$\tilde{1}6$, $\bar{2}1$, $\tilde{2}5$, $\tilde{3}1$
Chem.	$\tilde{1}2$, $\bar{1}6$, $\bar{1}7$, $\bar{1}9$	$\bar{5}$-$\bar{7}$, $\tilde{1}4$, $\bar{1}5$, $\tilde{2}2$

uniform **nonuniform**

POS:

- LLM **overperforms** when compared to humans at same ability level

NEG:

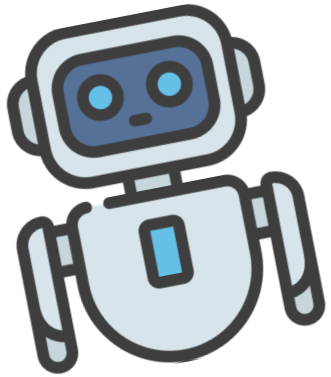
- LLM **underperforms**

RQ2:

- POS (LLM overperformance): “strength in tasks that prioritize rule-governed reasoning, conceptual clarity, canonical chemical principle”
- NEG (LLM underperformance): “These items place greater demands on visual interpretation, linguistic nuance, and the execution of complex, multi-step procedures”

To conclude…

- LR-DIF on items with reasonable (ITC $\geq$ 0.2) discrimination properties is a **robust method** to identify items on which humans and LLMs differ
- Using DIF-identified items can inform item analysis around GenAI and identify **task dimensions** that make items easy/difficult for LLMs
- This can inform assessment redesign that is **GenAI-resilient**



Thank You!