AIME Presentation

Hotaka Maeda

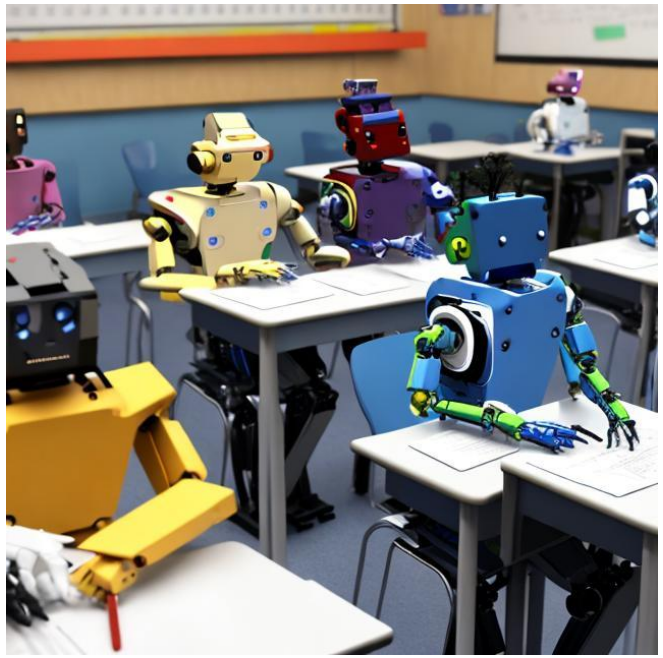
Senior Psychometrician

May 14th, 2025

smarter BALANCED

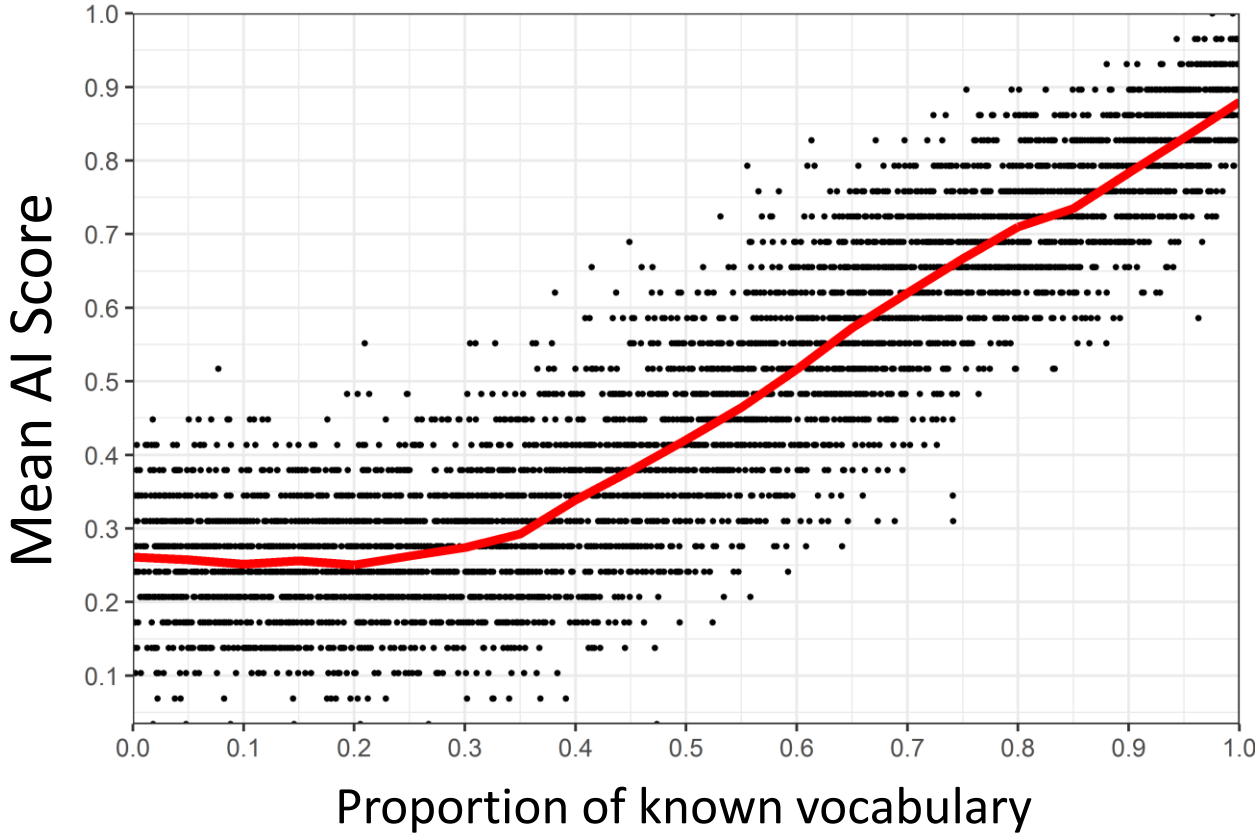
My 2022

- Learned AI through Jeremy Howard's fast.ai
- BERT can take tests
- θ distribution vs artificial 'intelligence'?



NCME 2023:
Field-testing items using AI

- Create 1,000 RoBERTa models with varied θ: Random proportion $U(0,1)$ of the 50,265 token embedding weights set to 0
- Used AI item responses to calibrate new MCQ items



https://arxiv.org/abs/2310.11655

NCME 2024:
Field-testing items using AI (v2)

- Assigned θ to 61 DeBERTa-v3-large models, and fine-tuned it to output 2PL IRT model probabilities
- Generated item response data to do:
  - item calibration with anchors, distractor analysis, dimensionality analysis, scoring, item proportion correct, item discrimination

Maeda, H. (2024). Field-Testing Multiple-Choice Questions With AI Examinees: English Grammar Items. *Educational and Psychological Measurement, 85*(2), 221-244.

Bottleneck for field-testing real items with AI?

Bottleneck for field-testing real items with AI?

DIF

Finding Words Associated with DIF

Predicting DIF using LLMs and Explainable AI

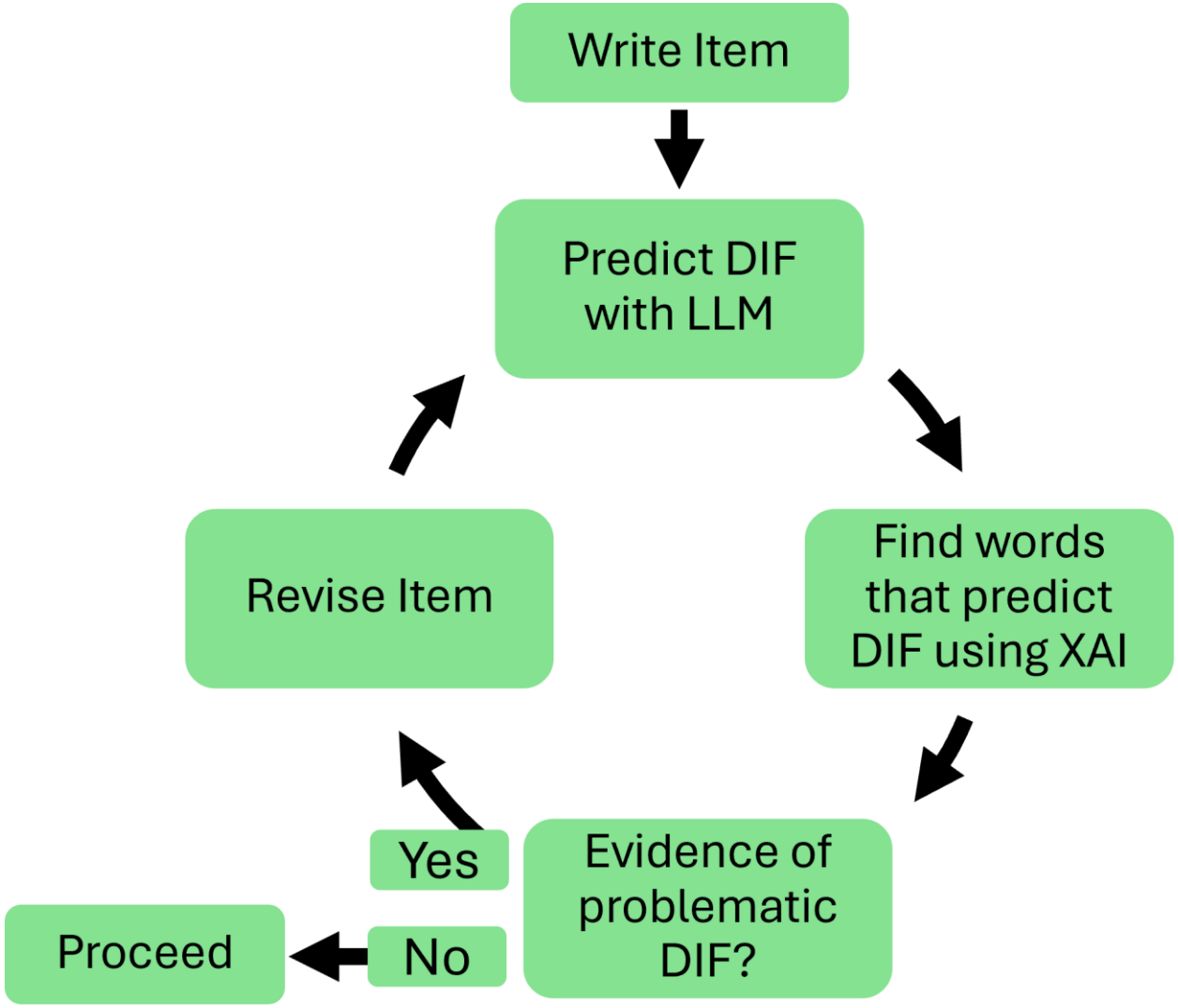
Hotaka Maeda - Smarter Balanced
Yikai (EK) Lu - University of Notre Dame

DIF Analysis

- Differential item functioning (DIF) – attempt to find biased items

1. Psychometricians: identify correct response probability that depend on demographics, given examinee ability
2. Item content developer/SME: identify qualitative cause of bias

My Vision


Write Item
Predict DIF with LLM
Find words that predict DIF using XAI
Evidence of problematic DIF?
Yes
No
Proceed
Revise Item

Purpose

1. Predict DIF from the item text by training (fine-tune) an encoder transformer language model
2. Then, use “explainable AI” (XAI) methods to identify words associated with DIF

- Impact
  - Help review traditional DIF results
  - Avoid sample size issue
  - Provide immediate item writing/revision feedback (3 years in advance for SB)
  - Understand how DIF manifests qualitatively

METHODS

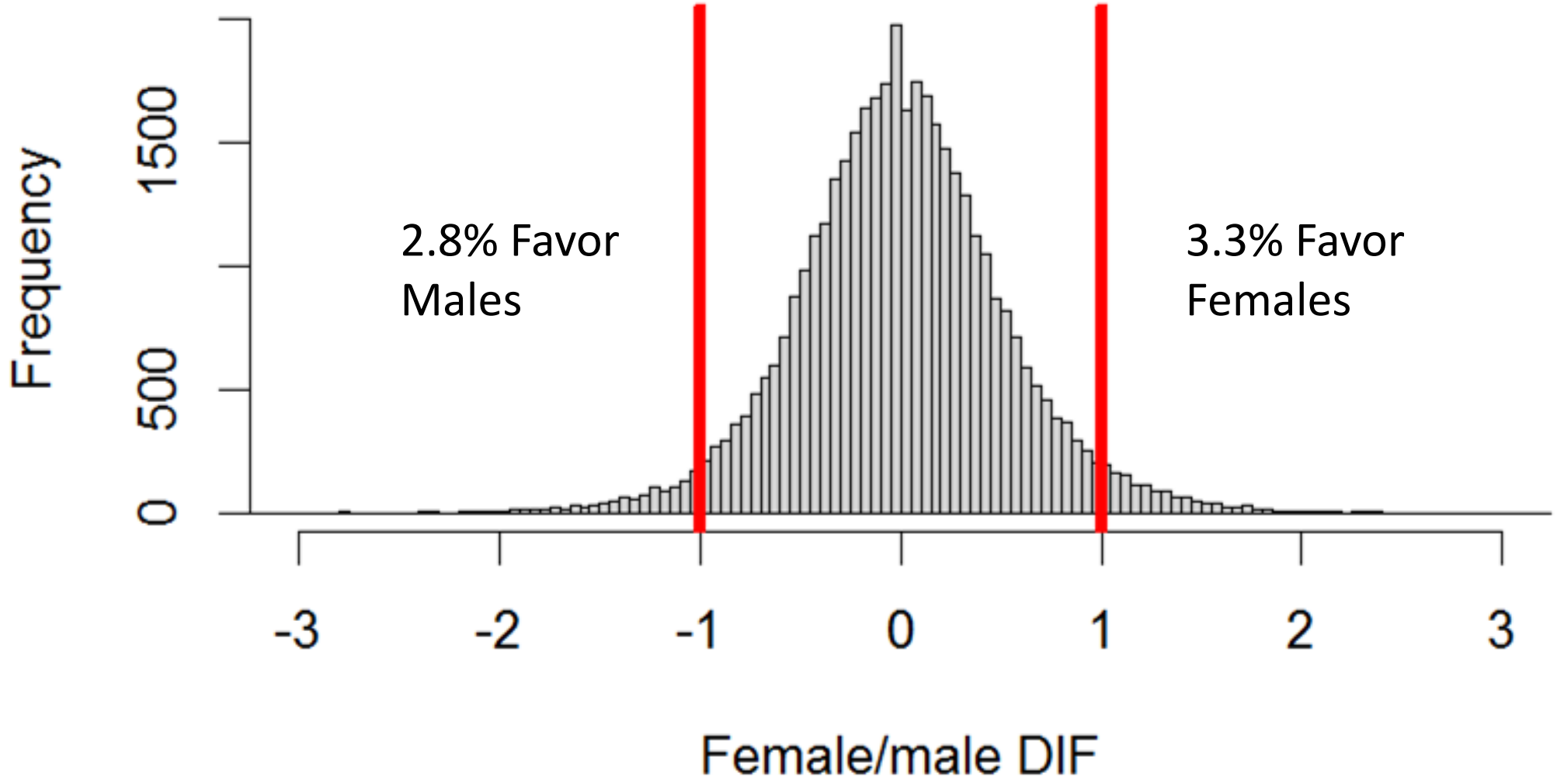
Item Data

- 42,180 English language arts & math summative state assessment items
- Grades 3 to 11
- Variety of item types
- Field tested (calibration, DIF)
- 80% training, 10% evaluation, 10% testing data

DIF Data

- Gender, race/ethnicity, SES, English language learner, disability
- Binary: Mantel-Haenszel delta difference (MH)
  - ≤ -1 favors reference group
  - ≥ 1 favors focal group
- Polytomous: Standardized mean difference effect size (ES)
  - ≤ -0.17 favors reference group
  - ≥ 0.17 favors focal group
- ES/0.17 ≈ MH
- N ≥ 100 examinees per group per item (usually overall N >1500)

Gender DIF Statistic Histogram



Continuous DIF Prediction Method

Item id	Item text	DIF
1	What inference can be made about the narrator's feelings toward.....	-0.6
2	…	…

Predict DIF from Item Text

Mean squared error loss

i = items
N = number of items
Y = DIF

$$\text{MSE} = \frac{\sum(\hat{Y}_i - Y_i)^2}{N}$$

3 Category Prediction Method

3 Probabilities found using DIF and SE

Item id	Item text	DIF	DIF SE	Favor Reference Group	No DIF	Favor Focal Group
1	What inference can be made about the narrator's feelings toward.....	-0.6	0.5	.2	.8	.001
2	…	…	…	…	…	…

3 Category Prediction Method

Item id	Item text	Favor Reference Group	No DIF	Favor Focal Group
1	What inference can be made about the narrator's feelings toward.....	.2	.8	.001
2	…	…	…	…

Predict 3 Probabilities from item text

Cross entropy loss

i = items
g = 3 groups
N = number of items

$$CEL = \frac{-\sum_i \sum_g P_{ig} \log \hat{P}_{ig}}{N}$$

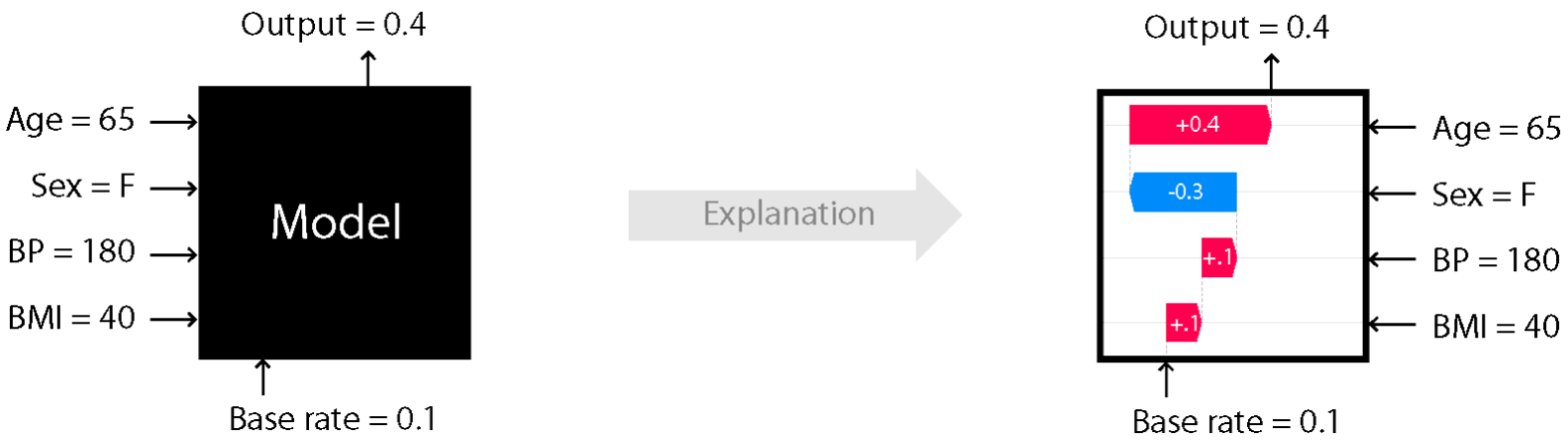
DIF Modeling: Prediction

- Fine-tune DeBERTa V3-large transformer encoder LLM (focal/reference group pairs separately)
  - Continuous model: From item text, predict DIF value, as a continuous variable
  - Category model: From item text, predict 3 DIF probabilities

DIF Modeling: XAI with SHAP

- Association of each word with predicted DIF value ("word attributions")
- Continuous model returns 1 attribution per token
- Categorical model returns 3 attributions per token
  - attribution = ifelse(a_Ref > 0, yes = -1*a_Ref, no = 0) +

    ifelse(a_Foc > 0, yes = a_Foc, no = 0)

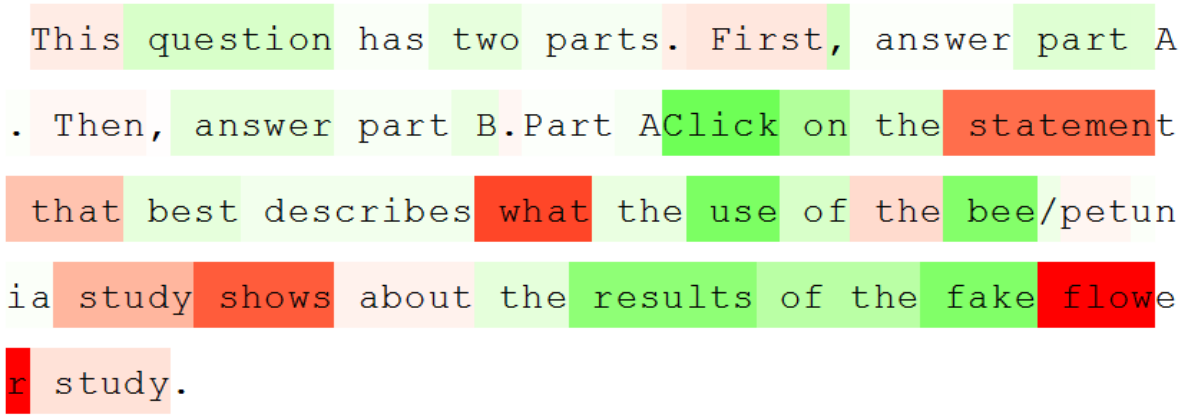


RESULTS

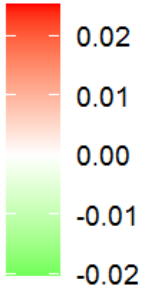
Based on testing data

Example Item without DIF Continuous Model

Continuous Model



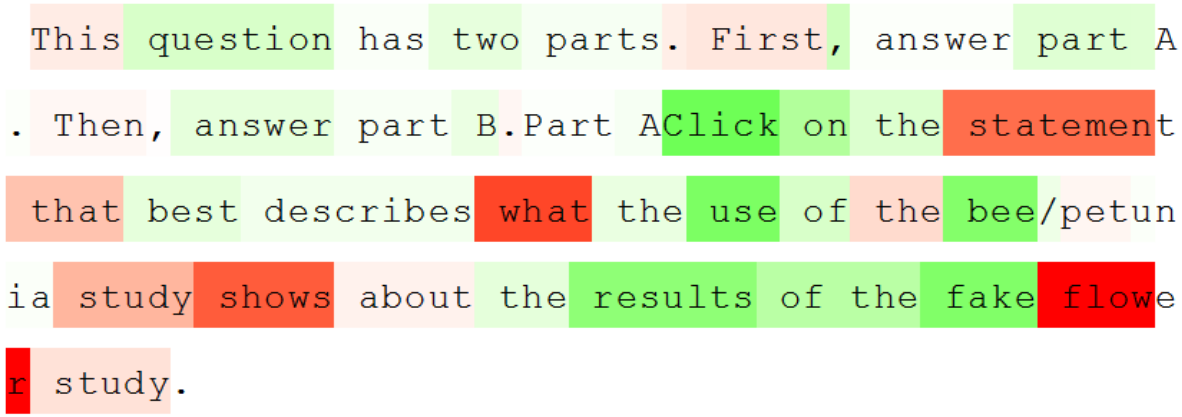
Favors Female Students



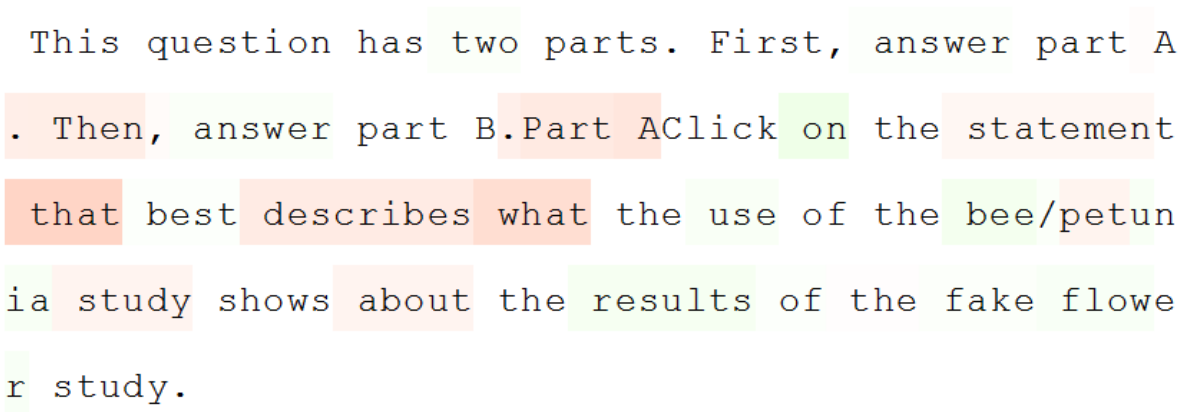
Favors Male Students

Example Item without DIF
Continuous vs Categorical Model

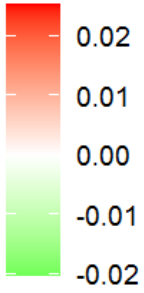
Continuous Model



3 Category Model



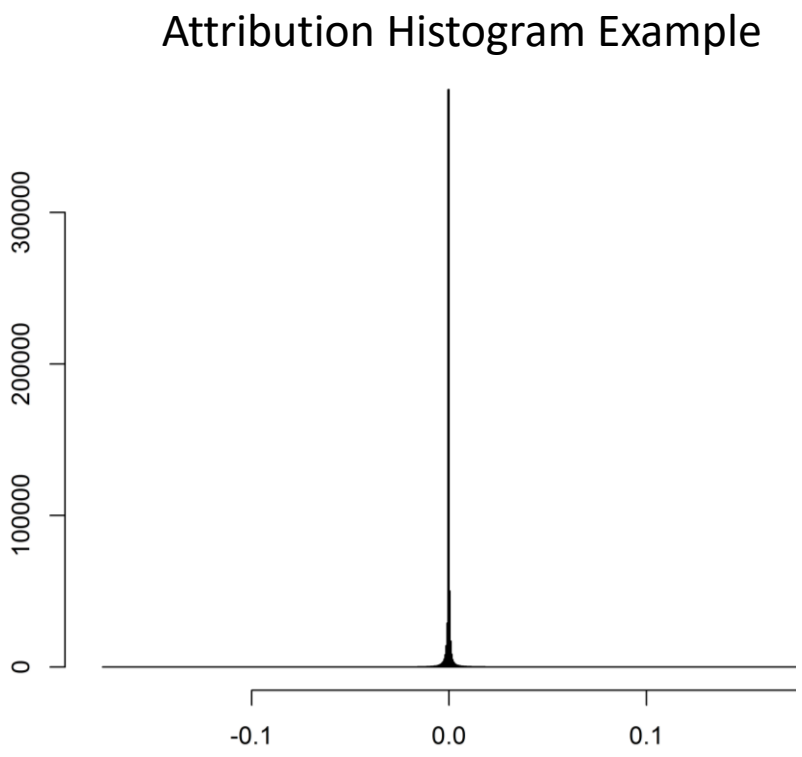
Favors Female Students



Favors Male Students

Female/Male Group Models

Model	R^2	Attribution Kurtosis	Correlation: Attribution & DIF
Categorical	.32	515	.20
Continuous	.33	90	.09



Categorical Model Summary

Fine-tuned each model twice with different seed, averaged the output

Focal/Reference Group	Prediction R^2	Attribution Reliability
Female/Male	.32	.75
Asian/White	.20	.75
Black/White	.11	.68
Hispanic/White	.16	.70
Native/White	.04	.21
Lower SES/Non-LSES	.12	.70
Students w/Disabilities/Non-SWD	.11	.61
English learner/Non-EL	.08	.67

Example Item
- Favors Asian Students

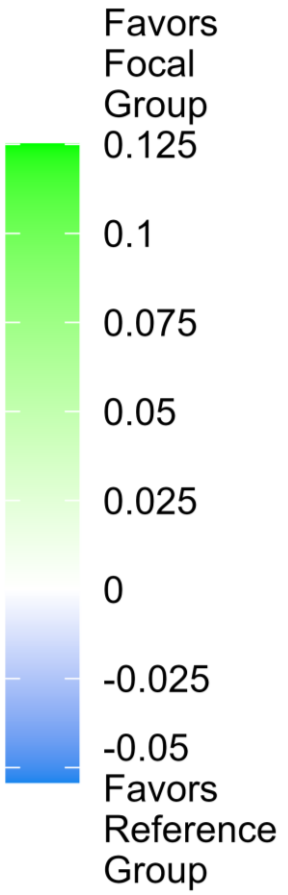
Observed DIF = 1.1
Predicted Favoring:
Asian p = .30
White p = .02

Select the two sentences that are punctuated correctly.[SEP]While I was growing up in the Midwest my favorite question to hear from my parents was "Guess where we're going this time?"[SEP]Although by that point, my parents had the whole vacation planned out; the moment they told me, I started looking up the location to see what activities were available.[SEP]When I was eight my family voted on a vacation to New York City where we stayed in downtown Times Square. Then later when I was ten we flew to Florida again, this time we departed on a cruise to Mexico, Jamaica and the Bahamas for a second time.[SEP]The average life expectancy is seventy years on this planet, this planet has so many different geological features, different climates and different cultures.[SEP]The places I have already visited make my curiosity even greater, and I think that it's important to view the world and ways of life from a different point of view.[SEP]Last year when I was sixteen we went on another cruise where we sailed the Western Caribbean to Puerto Rico, the Bahamas yet again and St Thomas.

Example Item
- Favors Students without Disabilities

Observed DIF = -1.8
Predicted Favoring:
  w/ Disabilities p = .03
  w/o Disabilities p = .39

What is the median number of dog walks for the first 9 weeks?



Top Words Associated with DIF

Foc/Ref Group	Favors Focal	Favors Reference
Female/Male	narrator, message, text, reader, summarize, relationship	growth, decay, equal, option, rounded, number
Asian/White	spelling, spelled, factor, capitalization, multiplication, punctuated	when, read, two, an, mr, click, to, sentence
Black/White	div, multiplying, quotient, multiplication, equation, divide	grams, aaron, equal, parts, an, enter
Hispanic/White	enter, div, equation, quotient, rational, multiplication	rounded, punctuated, shade, phrases, parts, growth
Low SES/Non-LSES	box, irrational, select	equal, rounded, measure, degrees, word, answer
Students with Disabilities/Non-SWD	div, size, unknown, equation, makes, true	directions, argumentative, farm, student, performance, club
English Learners/ Non-EL	from, equations	round, scored, task, mean, read, performance

Discussions

- Most detected DIF seemed to be construct-relevant
- Applications:
  - 1) During item writing
  - 2) During traditional DIF item review
  - 3) When sample size requirements cannot be met for traditional DIF analyses
- Limitations
  - Correlation, not causation (word replacement ≠ DIF elimination)
- AI can be used to fight bias

Preprint Paper

Finding Words Associated with DIF

Hotaka Maeda
Yikai (EK) Lu
hotaka.maeda@smarterbalanced.org

https://arxiv.org/abs/2502.07017