Examinee and Assessment Characteristics Matter: Where Do We Go From Here in the Age of Digital Assessments?

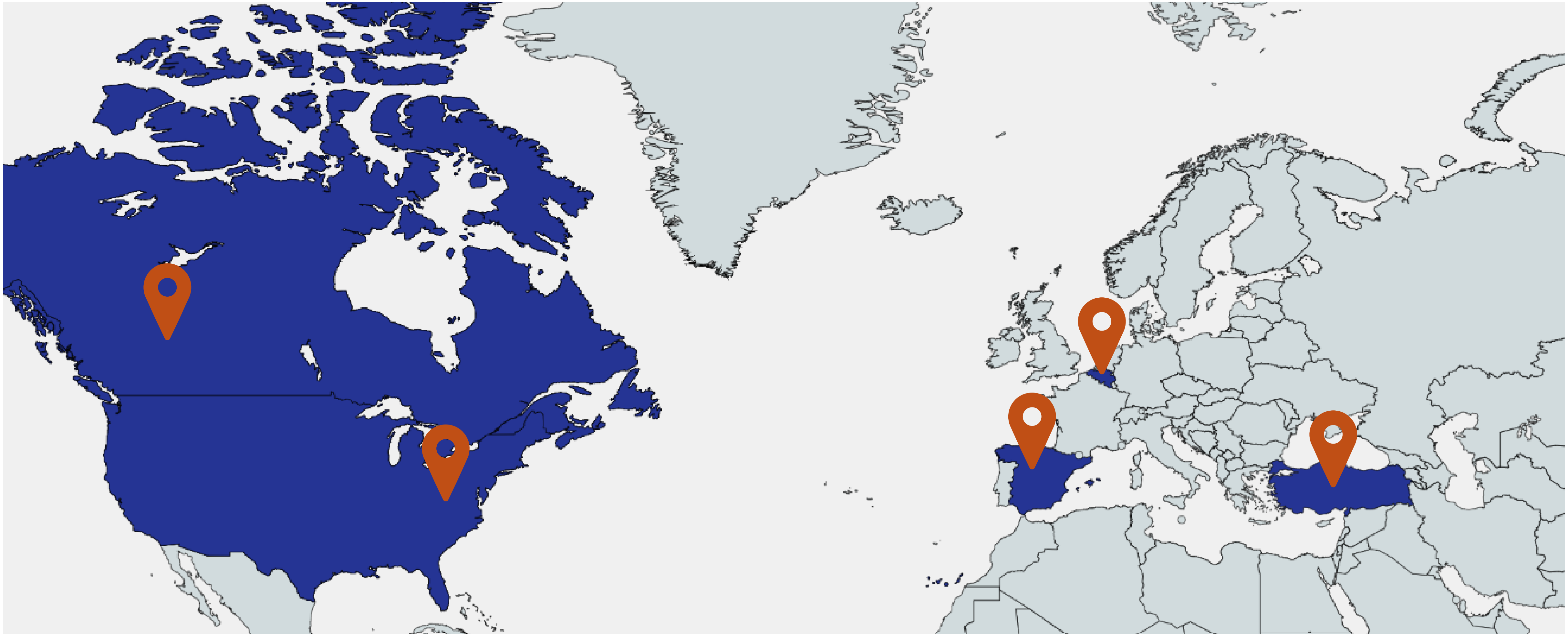
Guher Gorgun
Measurement, Evaluation, and Data Science
Center for Research in Applied Measurement and Evaluation
University of Alberta

Research Interests

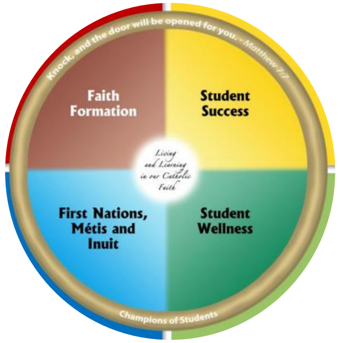
Assessment and examinee characteristics interact

Items as building blocks of assessments

New methods for enhancing assessment practices of diverse populations



Survey development and psychometric analysis



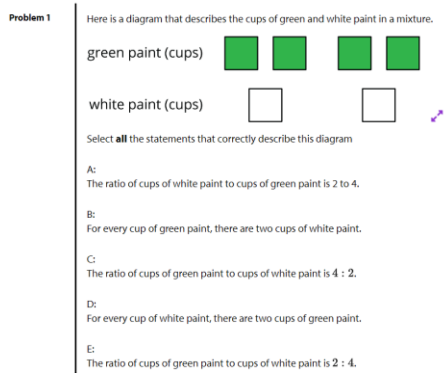
SPOT
Student Perspectives of Teaching
UNIVERSITY OF ALBERTA

C.R.A.M.E

Item difficulty in gamified learning platforms

EdTeKLA
eyeread

Learner modeling in online platforms


Problem 1

Here is a diagram that describes the cups of green and white paint in a mixture.

green paint (cups)

white paint (cups)

Select **all** the statements that correctly describe this diagram

A:
The ratio of cups of white paint to cups of green paint is 2 to 4.

B:
For every cup of green paint, there are two cups of white paint.

C:
The ratio of cups of green paint to cups of white paint is 4 : 2.

D:
For every cup of white paint, there are two cups of green paint.

E:
The ratio of cups of green paint to cups of white paint is 2 : 4.

ASSISTments

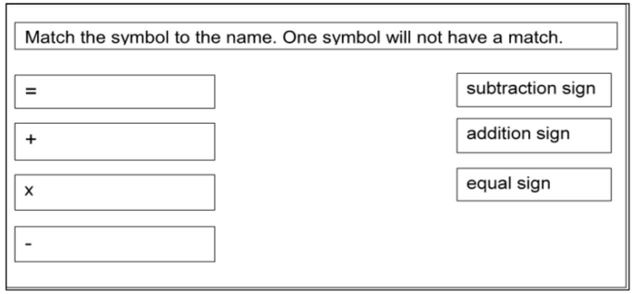
Predicting cognitive engagement in online discussion forums

EdTeKLA

Teacher-centered automated essay scoring

EdTeKLA
ALBERTA INNOVATES

Test-taking engagement in low-stakes and alternate assessments


Match the symbol to the name. One symbol will not have a match.

=	subtraction sign
+	addition sign
x	equal sign
-	

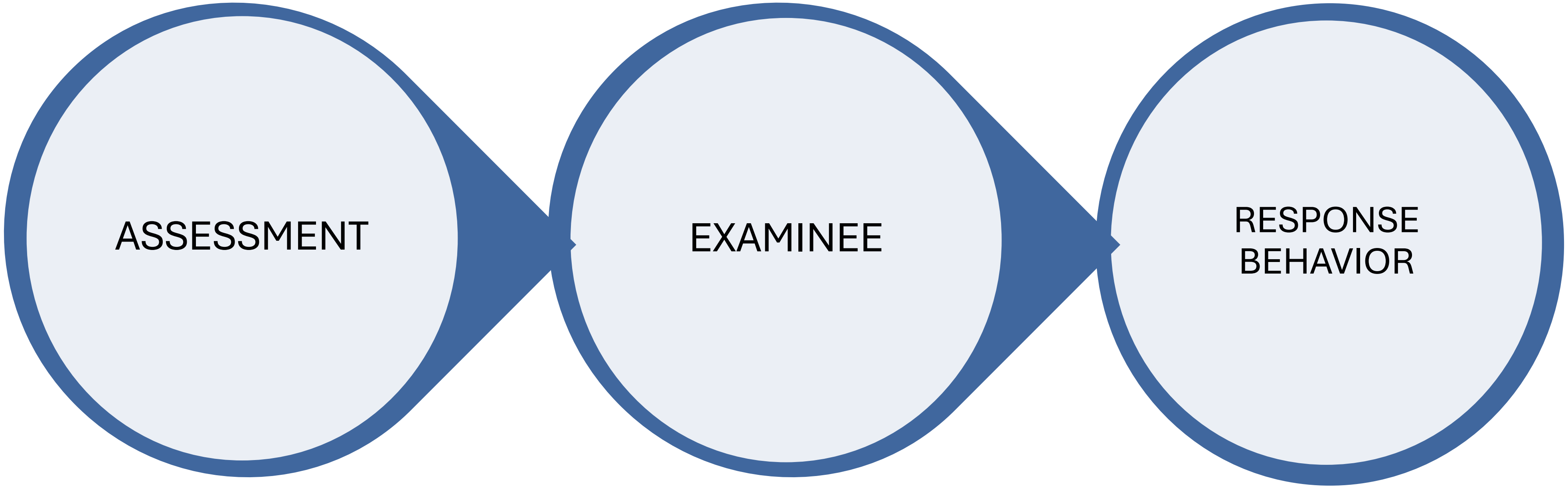
DYNAMIC® LEARNING MAPS
KU ACHIEVEMENT & ASSESSMENT INSTITUTE
Accessible Teaching, Learning, & Assessment Systems

Research Interests

Assessment and examinee characteristics interact

Items as building blocks of assessments

New methods for enhancing assessment practices of diverse populations



HIGH-STAKES

Important consequences for examinee

LOW-STAKES

No direct consequences for examinees

MOTIVATION

FATIGUE

TEST-WISENESS

TEST ANXIETY

EFFORT

ENGAGED RESPONDING

SOLUTION BEHAVIOR

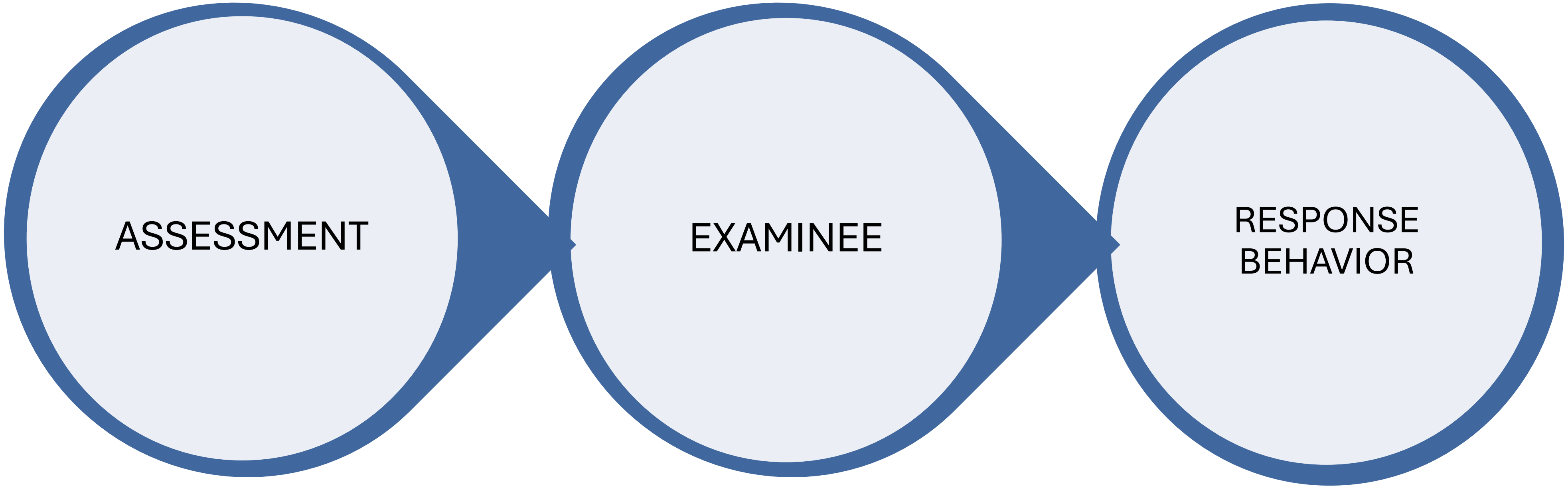
RAPID GUESSING

CHEATING

CARELESS RESPONDING

IDLE RESPONDING

Ulitzsch, E., Yildirim-Erbasli, S. N., **Gorgun, G.,** & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. British Journal of Mathematical and Statistical Psychology, 75(3), 668-698. https://doi.org/10.1111/bmsp.12272



HIGH-STAKES

Important consequences for examinee

LOW-STAKES

No direct consequences for examinees

MOTIVATION

FATIGUE

TEST-WISENESS

TEST ANXIETY

EFFORT

ENGAGED RESPONDING

SOLUTION BEHAVIOR

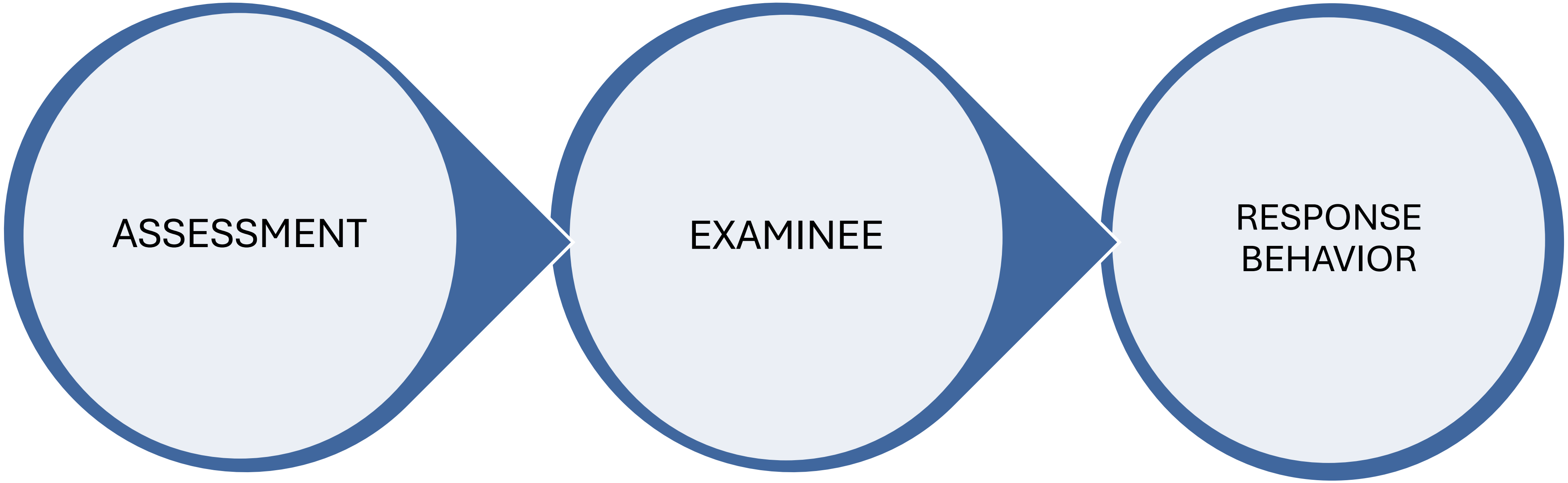
RAPID GUESSING

CHEATING

CARELESS RESPONDING

IDLE RESPONDING

Ulitzsch, E., Yildirim-Erbasli, S. N., **Gorgun, G.,** & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. British Journal of Mathematical and Statistical Psychology, 75(3), 668-698. https://doi.org/10.1111/bmsp.12272



HIGH-STAKES

Important consequences for examinee

LOW-STAKES

No direct consequences for examinees

MOTIVATION

FATIGUE

TEST-WISENESS

TEST ANXIETY

EFFORT

ENGAGED RESPONDING

SOLUTION BEHAVIOR

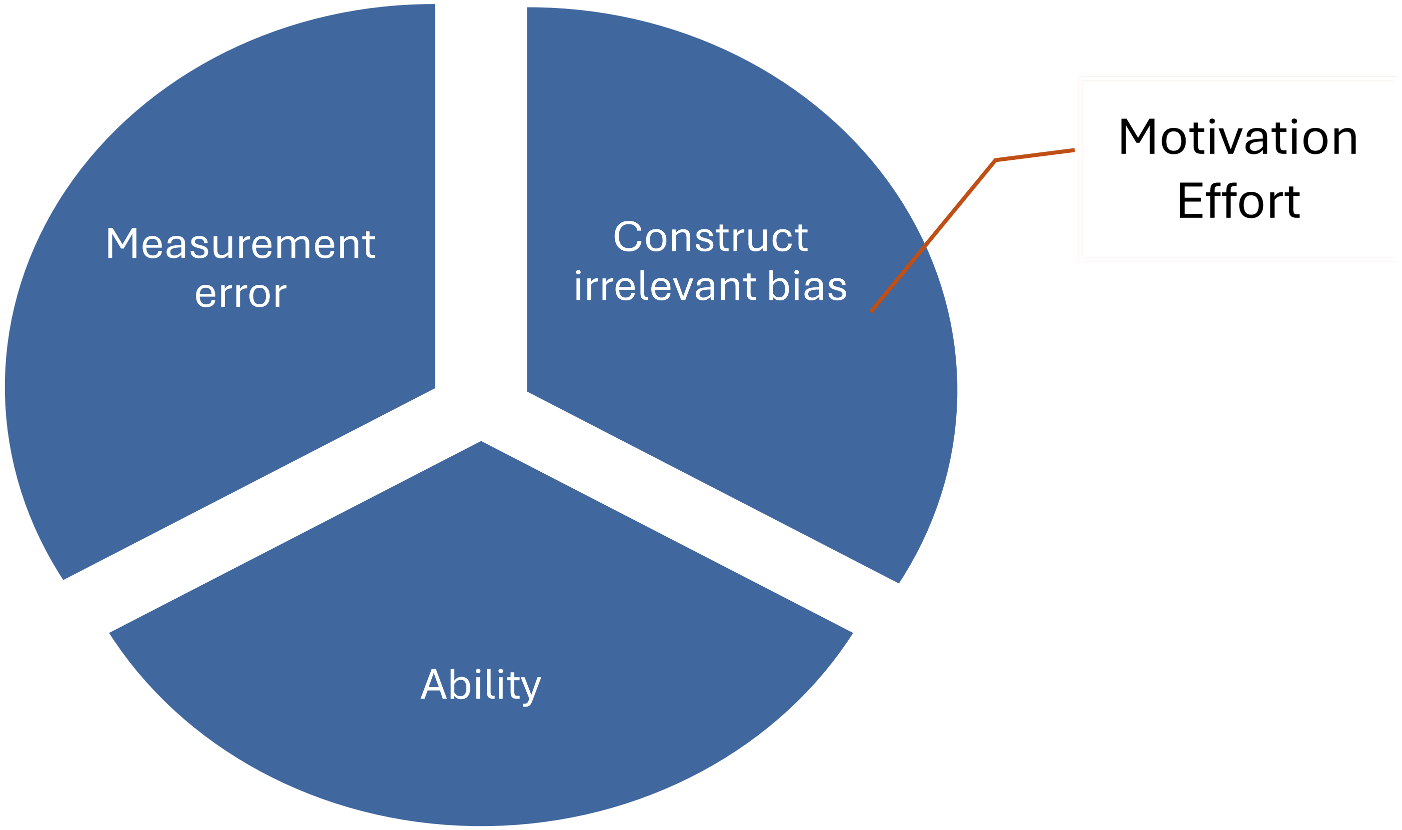
RAPID GUESSING

CHEATING

CARELESS RESPONDING

IDLE RESPONDING

Ulitzsch, E., Yildirim-Erbasli, S. N., **Gorgun, G.,** & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. British Journal of Mathematical and Statistical Psychology, 75(3), 668-698. https://doi.org/10.1111/bmsp.12272



“CONSTRUCT-IRRELEVANT” VARIANCE

Gorgun, G., & Bulut, O. (2022). Identifying aberrant responses in intelligent tutoring systems: An application of anomaly detection methods. *Psychological Testing and Assessment Modeling, 64*(4), 359-384.

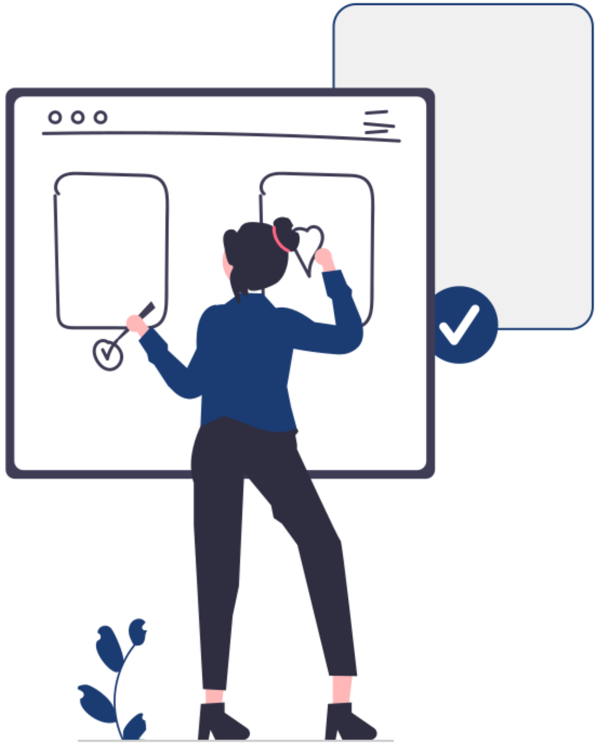
Why should we care about construct-irrelevant variance?

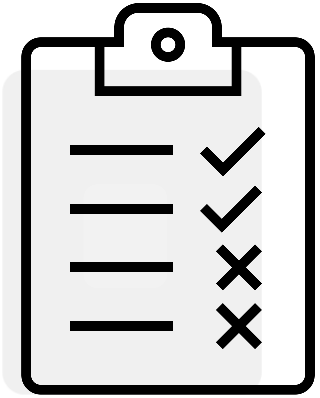
- Poor data quality
- Not-reached or omitted items
- Estimation and classification errors
- Lower prediction accuracy
- Unjustified inferences about examinees

Gorgun, G., & Bulut, O. (2022). Considering disengaged responses in Bayesian and deep knowledge tracing. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.*), Artificial intelligence in education. Posters and late-breaking results, workshops and tutorials, industry and innovation Tracks, practitioners' and doctoral consortium* (pp. 591-594). Lecture Notes in Computer Science, vol 13356. Springer, Cham. https://doi.org/10.1007/978-3-031-11647-6_122

Digital assessments can provide **process** and **product** information about examinees

How an examinee attempted the item





Whether the response is correct

Zumbo, B. D., & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research* (Vol. 26). Springer International Publishing.

Response time

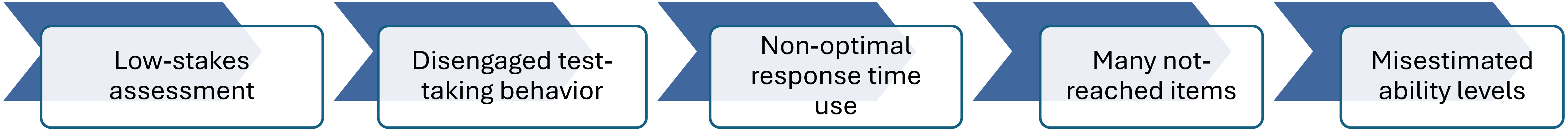
Number of actions

Action sequences

Eye tracking

Sensor data

Study 1: Minimizing the influence of disengagement on ability estimation



Use engagement in the operationalization of ability (Tijmstra & Bolsinova, 2018)

Use response times to identify optimal time users and disengaged responders (Wise & Ma, 2012)

Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement, 81*(5), 847-871. https://doi.org/10.1177/0013164421991211

A novel scoring approach blending **response times** with **response accuracy**

- Recode not-reached as missing (**NA**) or incorrect (**IN**)
- Find response time distributions for each item separating correct and incorrect response time distributions
- Assign partial scores based on response time use

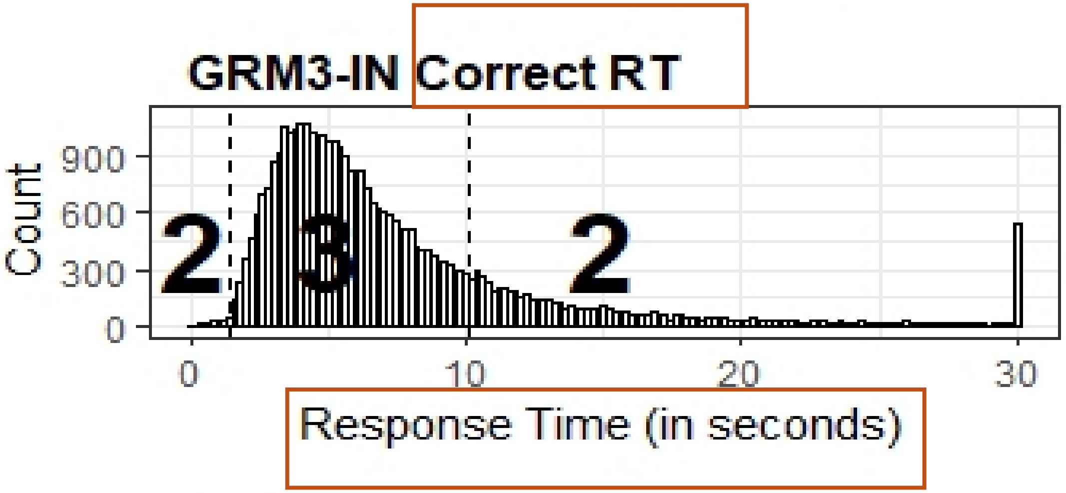
Original scoring method

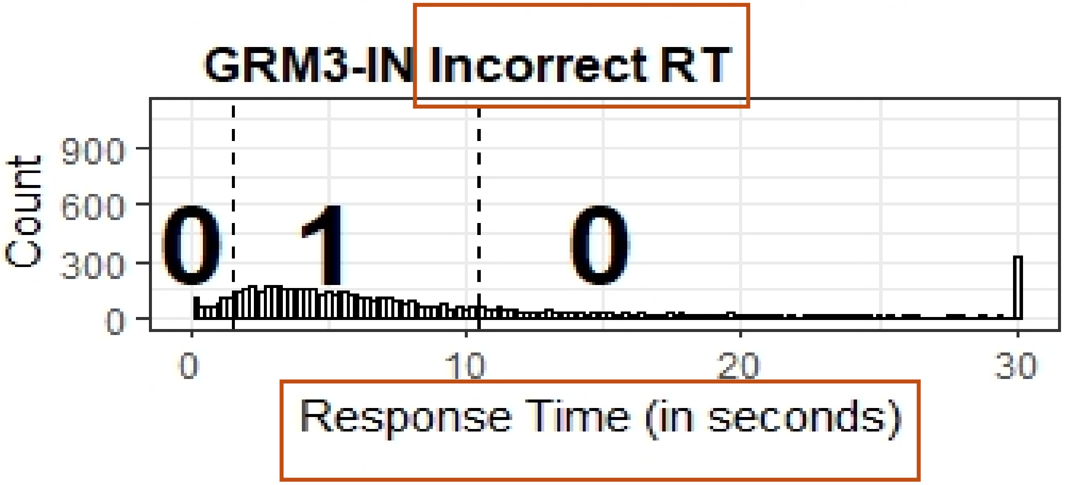
	Correct	Incorrect
Optimal time	1	0
Disengaged	1	0

New scoring method

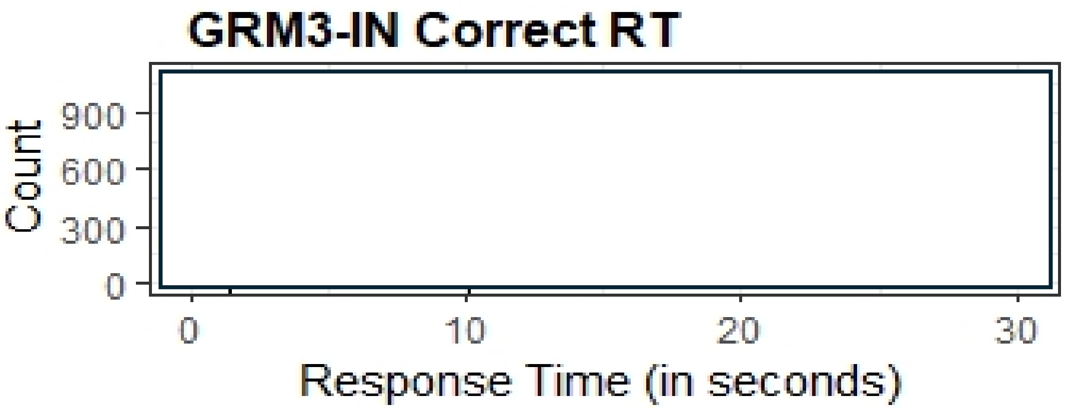
	Correct	Incorrect
Optimal time	3	1
Disengaged	2	0

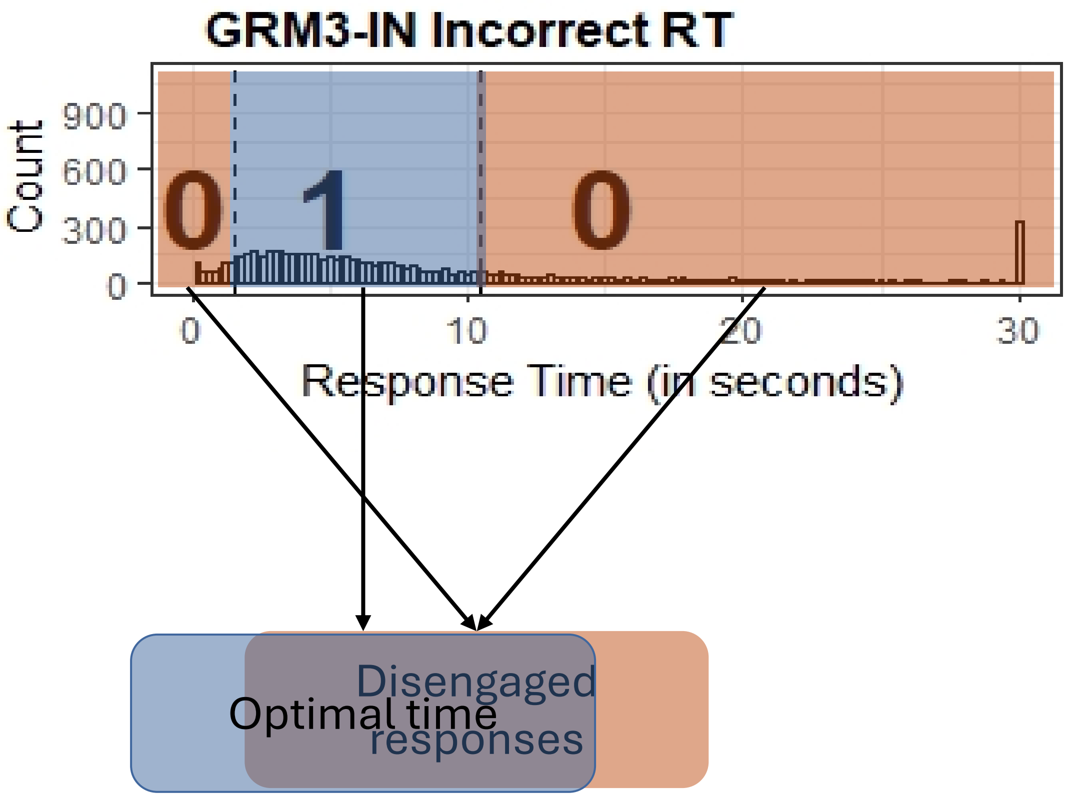
How it works?


GRM3-IN Correct RT
Count
900
600
300
0
2
3
2
0
10
20
30
Response Time (in seconds)

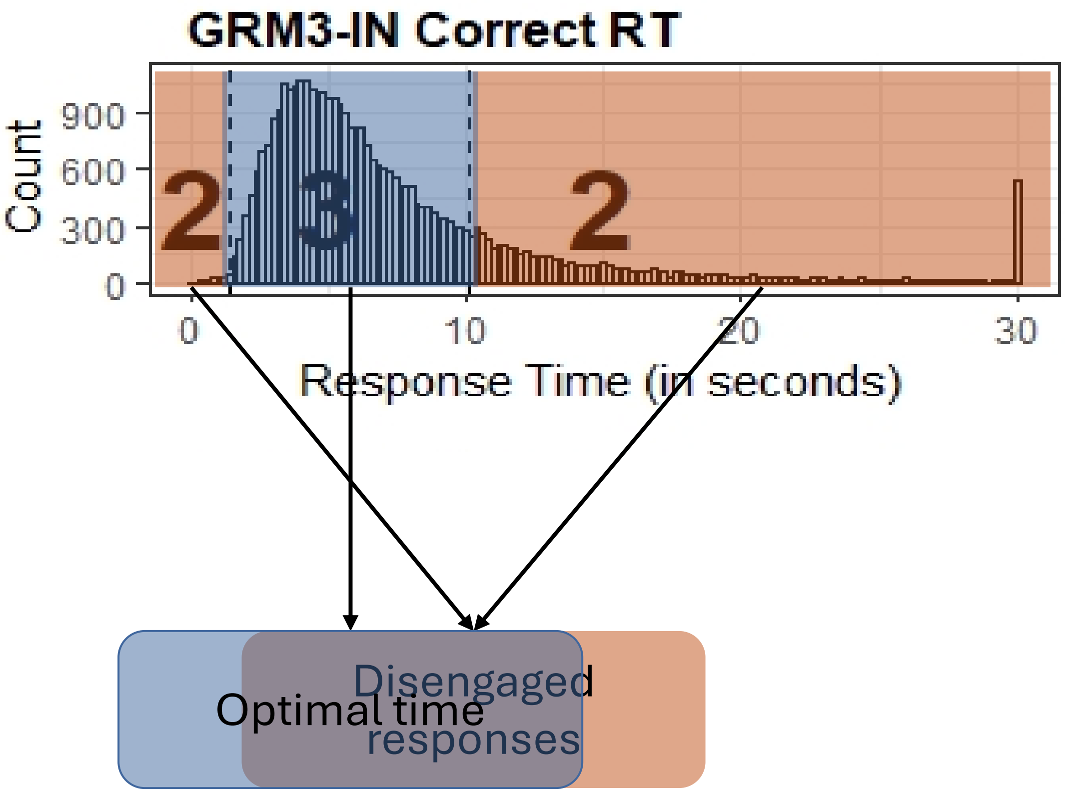

GRM3-IN Incorrect RT
Count
900
600
300
0
0
1
0
0
10
20
30
Response Time (in seconds)

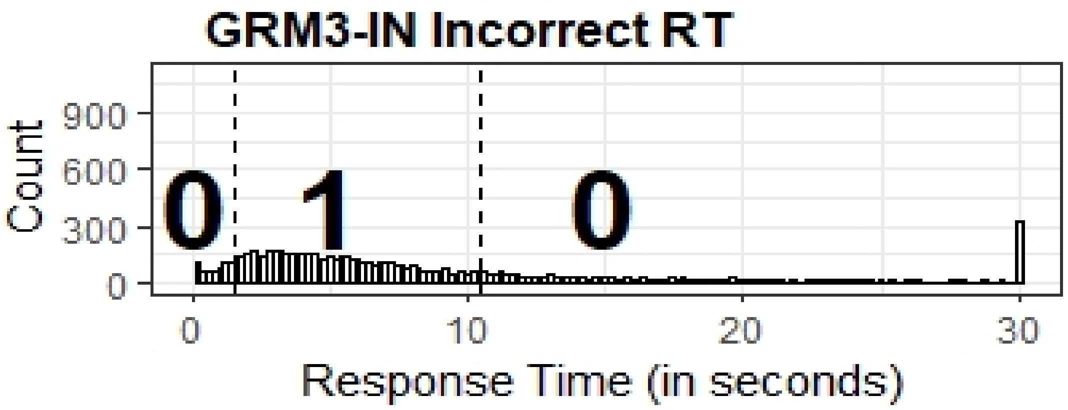
How it works?


GRM3-IN Correct RT
Count
900
600
300
0
0
10
20
30
Response Time (in seconds)


GRM3-IN Incorrect RT
Count
900
600
300
0
0
1
0
0
10
20
30
Response Time (in seconds)
Optimal time
Disengaged responses

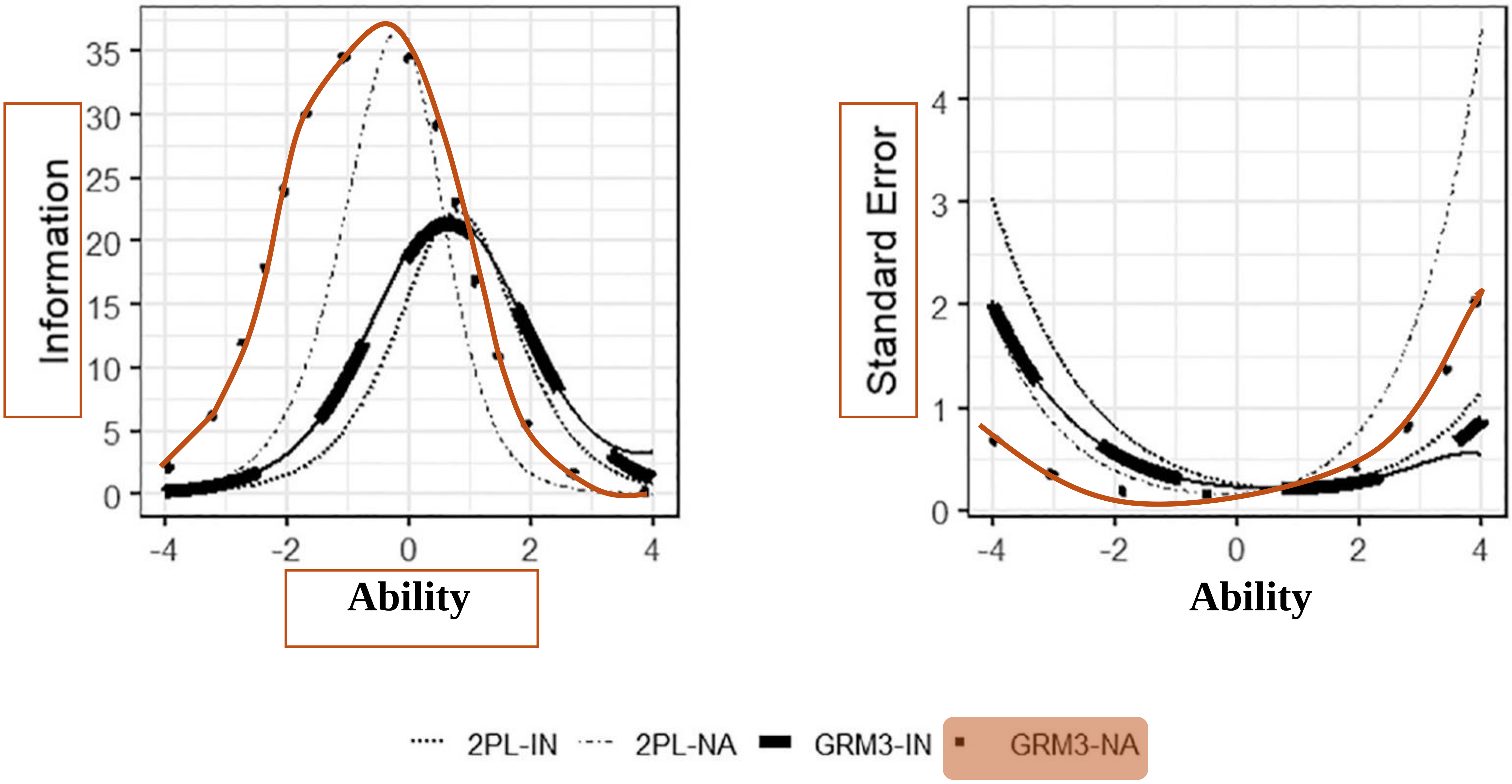
How it works?


GRM3-IN Correct RT
Count
900
600
300
0
2
3
2
0
10
20
30
Response Time (in seconds)
Optimal time
Disengaged responses


GRM3-IN Incorrect RT
Count
900
600
300
0
0
1
0
0
10
20
30
Response Time (in seconds)

Scoring models compared

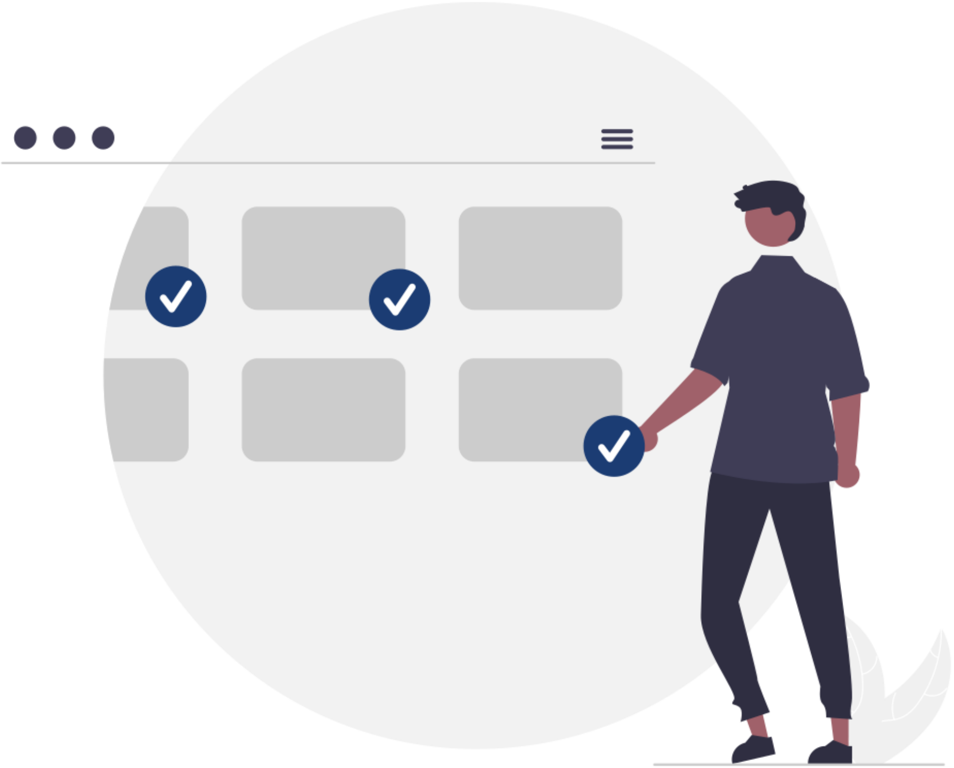
	Not-reached	Scoring approach
2PL-NA Baseline	Missing (NA)	Original
2PL-IN Baseline	Incorrect (IN)	Original
GRM3-NA	Missing (NA)	New
GRM3-IN	Incorrect (IN)	New



Information
35
30
25
20
15
10
5
0
-4
-2
0
2
4
Ability
Standard Error
4
3
2
1
0
-4
-2
0
2
4
Ability
2PL-IN
2PL-NA
GRM3-IN
GRM3-NA

Limitation:

Applied **after** the assessment is over as a **reactive** response



Study 2: Alleviating disengagement by enhancing the adaptivity of the assessment

- In computerized adaptive tests we select the most informative item based on the examinees' interim ability

- We developed a new item selection algorithm, so **ability** and **engagement** can be considered jointly

Gorgun, G., & Bulut, O. (2023). Incorporating test-taking engagement into the item selection algorithm in computerized adaptive tests. *Large-Scale Assessments in Education, 11*, 27. https://doi.org/10.1186/s40536-023-00177-5

In computerized adaptive test, we used response data to calibrate the item parameters

Student ID	Item1	Item2	Item3	Item4	Item5
1	1	0	1	1	1
2	1	1	1	0	1
3	1	0	1	1	0
4	1	1	1	1	0
5	1	0	1	1	1
6	1	1	1	1	0
7	0	0	1	1	0
8	1	0	1	1	0
9	1	1	1	1	0
10	1	1	1	1	1

Item ID	Discrimination	Difficulty
Item1	2.267	-1.010
Item2	2.320	-0.810
Item3	2.064	-1.584
Item4	2.920	-0.988
Item5	1.7042	-0.123

The most informative item is selected based on interim ability maximizing item information function

$$I_{j,\theta}(\theta_i) = a_{j,\theta}^2 \, P_{j,\theta}\,(\theta_i)\left(1 - \, P_{j,\theta}(\theta_i)\right)$$

Item ID	Information Function
Item1	0.065
Item2	0.100
Item3	0.034
Item4	0.157
Item5	0.236

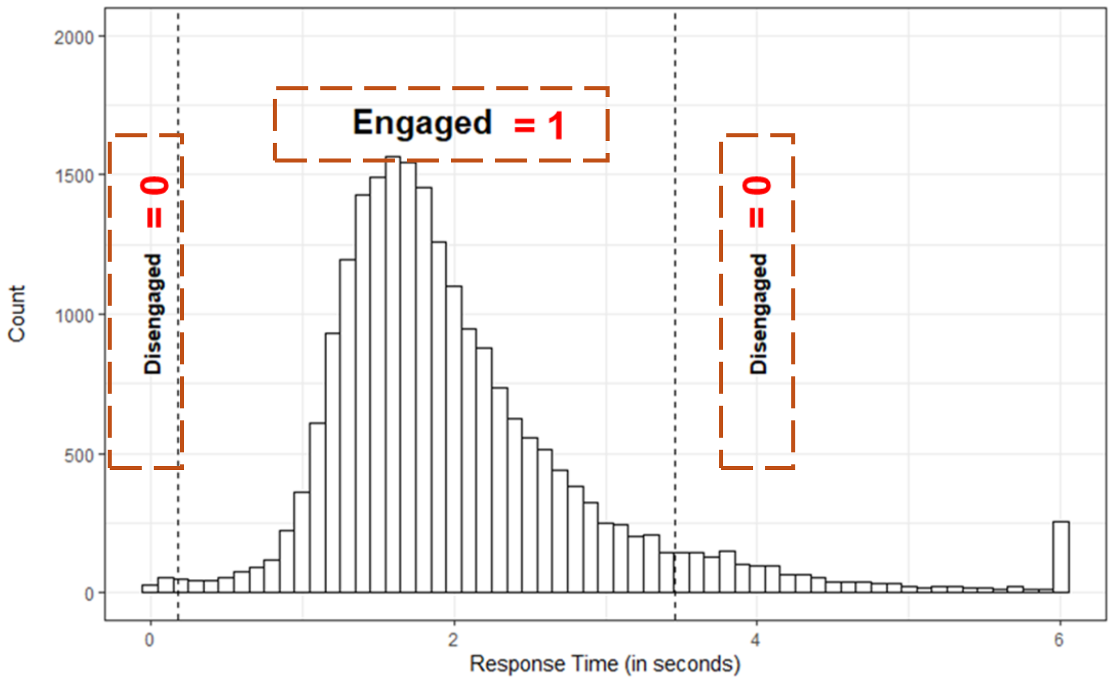
θ_i: the ability of examinee *i*

$a_{j,\theta}$: discrimination index of item *j*

$P_{j,\theta}$: the probability of answer item *j*

We create binary engagement data using response times

Student ID	Item1	Item2	Item3	Item4	Item5
1	0.474	1.101	0.645	1.123	1.062
2	1.612	1.384	0.769	1.608	1.258
3	0.629	1.062	1.421	-0.070	1.791
4	0.951	1.381	1.437	0.470	1.367
5	0.825	1.169	0.470	0.571	0.793
6	0.657	0.934	0.733	1.035	1.329
7	0.239	1.329	0.265	0.700	1.493
8	-0.08	-1.007	0.621	-1.632	-1.37
9	0.671	0.637	0.062	0.581	1.278
10	0.466	0.597	0.261	0.451	1.244



We use the binary engagement data to calibrate engagement parameters

Student ID	Item1	Item2	Item3	Item4	Item5
1	1	0	0	1	1
2	1	1	1	1	1
3	0	0	0	0	0
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	0	0	0	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Item ID	Discrimination	Difficulty
Item1	1.518	-0.221
Item2	1.710	-0.465
Item3	0.598	-1.238
Item4	1.565	-0.267
Item5	1.435	-0.448

The new algorithm selects the next item that maximizes both ability and engagement **item information functions**

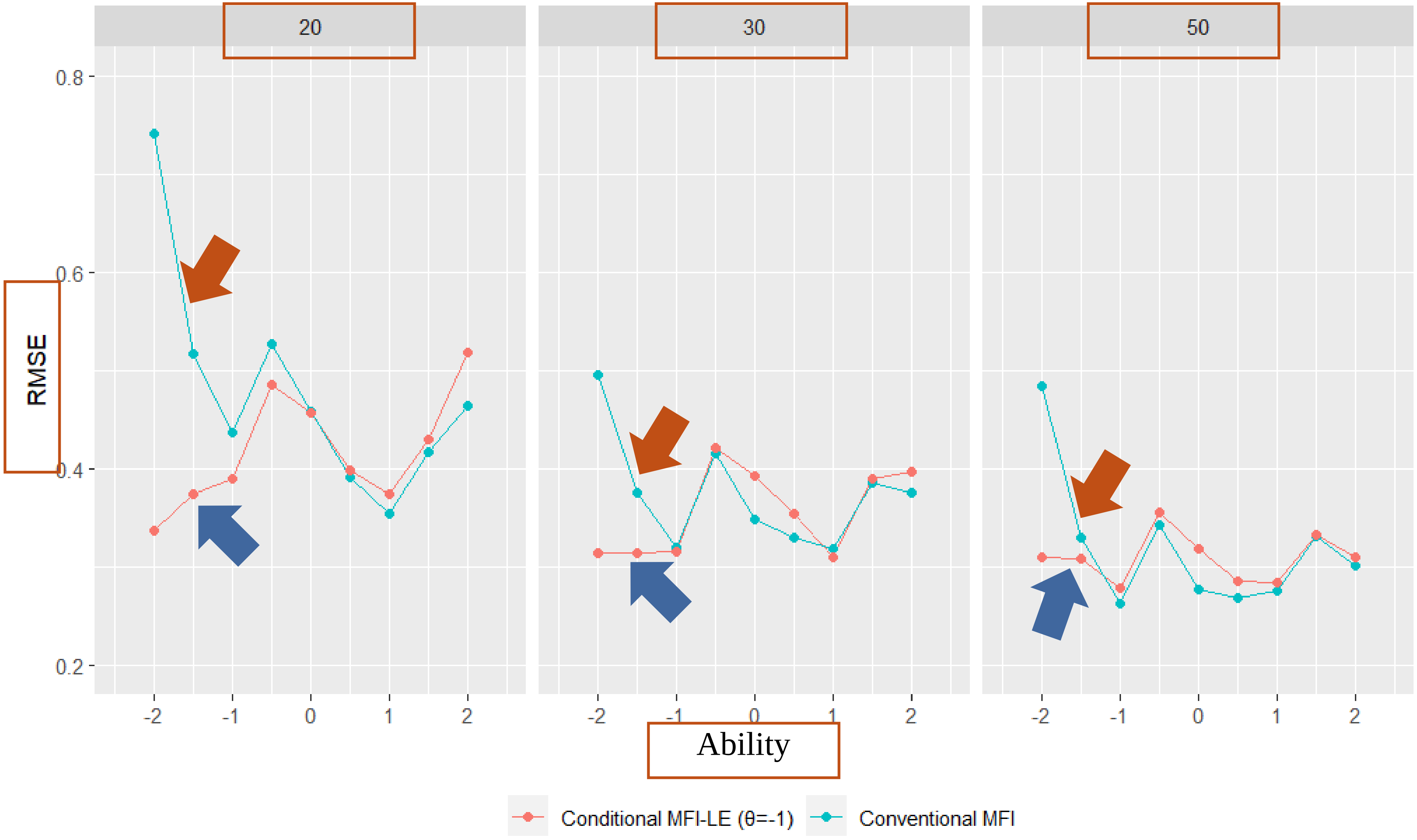
Ability Information Function

$$\boxed{I_{j,\theta}(\theta_i) = a_{j,\theta}^2\, P_{j,\theta}\,(\theta_i)\left(1 - \;P_{j,\theta}(\theta_i)\right)}$$

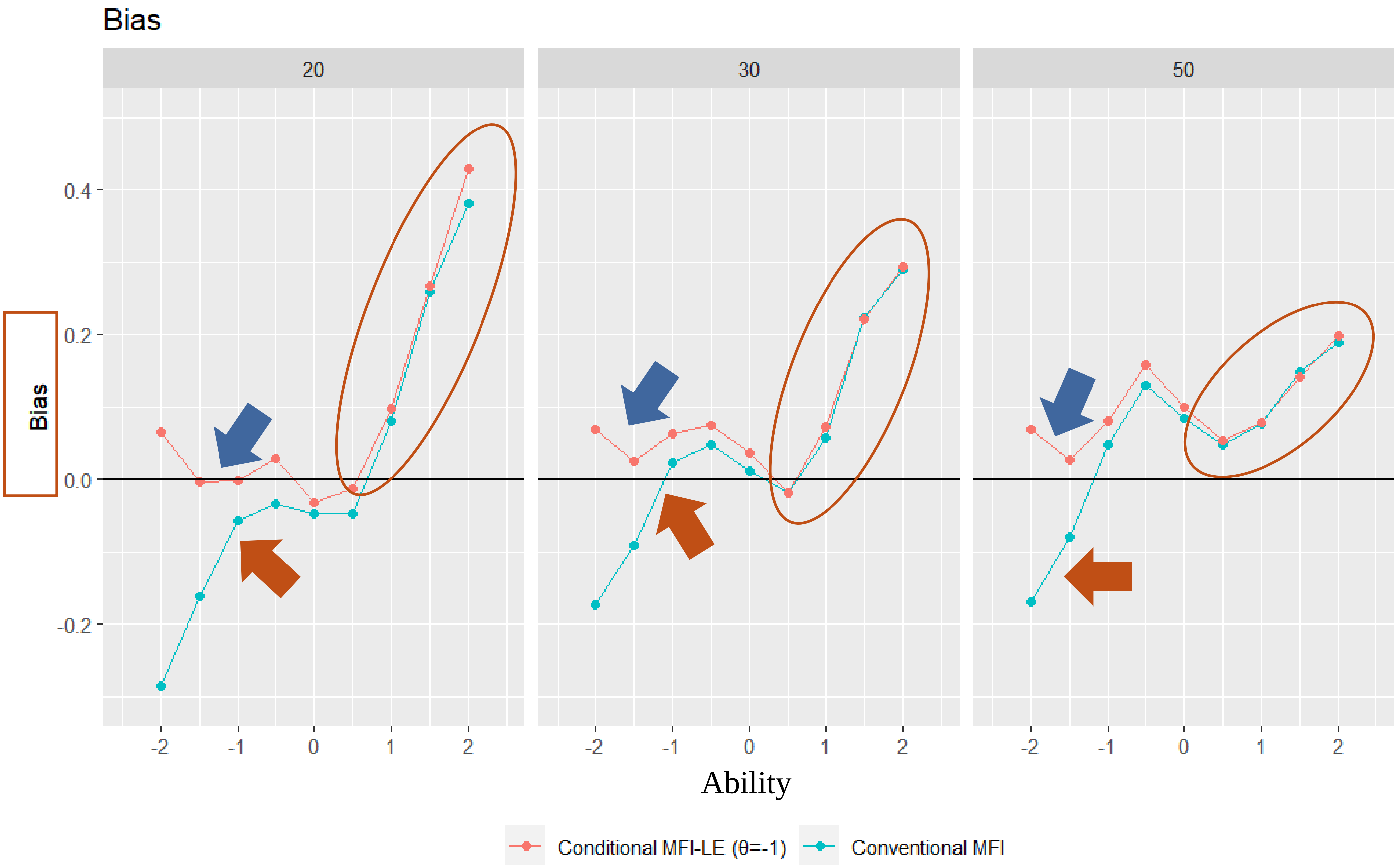
Engagement Information Function

$$\boxed{I_{j,e}(\theta_{e.i}) = a_{j,e}^2\, P_{j,e}\,(\theta_{e.i})\left(1 - \;P_{j,e}(\theta_{e.i})\right)}$$

RMSE



20
30
50
RMSE
0.8
0.6
0.4
0.2
-2
-1
0
1
2
Ability
Conditional MFI-LE (θ=-1)
Conventional MFI



Bias
20
30
50
0.4
0.2
0.0
-0.2
Bias
-2
-1
0
1
2
Ability
Conditional MFI-LE (θ=-1)
Conventional MFI

Research Interests

Assessment and examinee characteristics interact

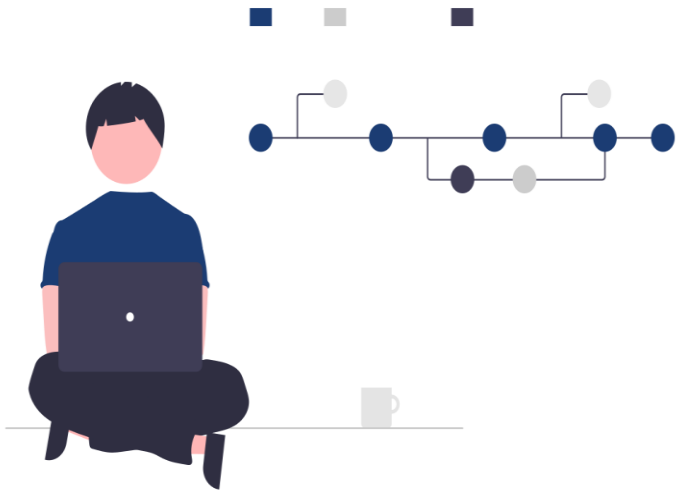
Items as building blocks of assessments

New methods for enhancing assessment practices of diverse populations

Adaptivity in an assessment requires **many** and **diverse** items

Created items need to go through an extensive evaluation process

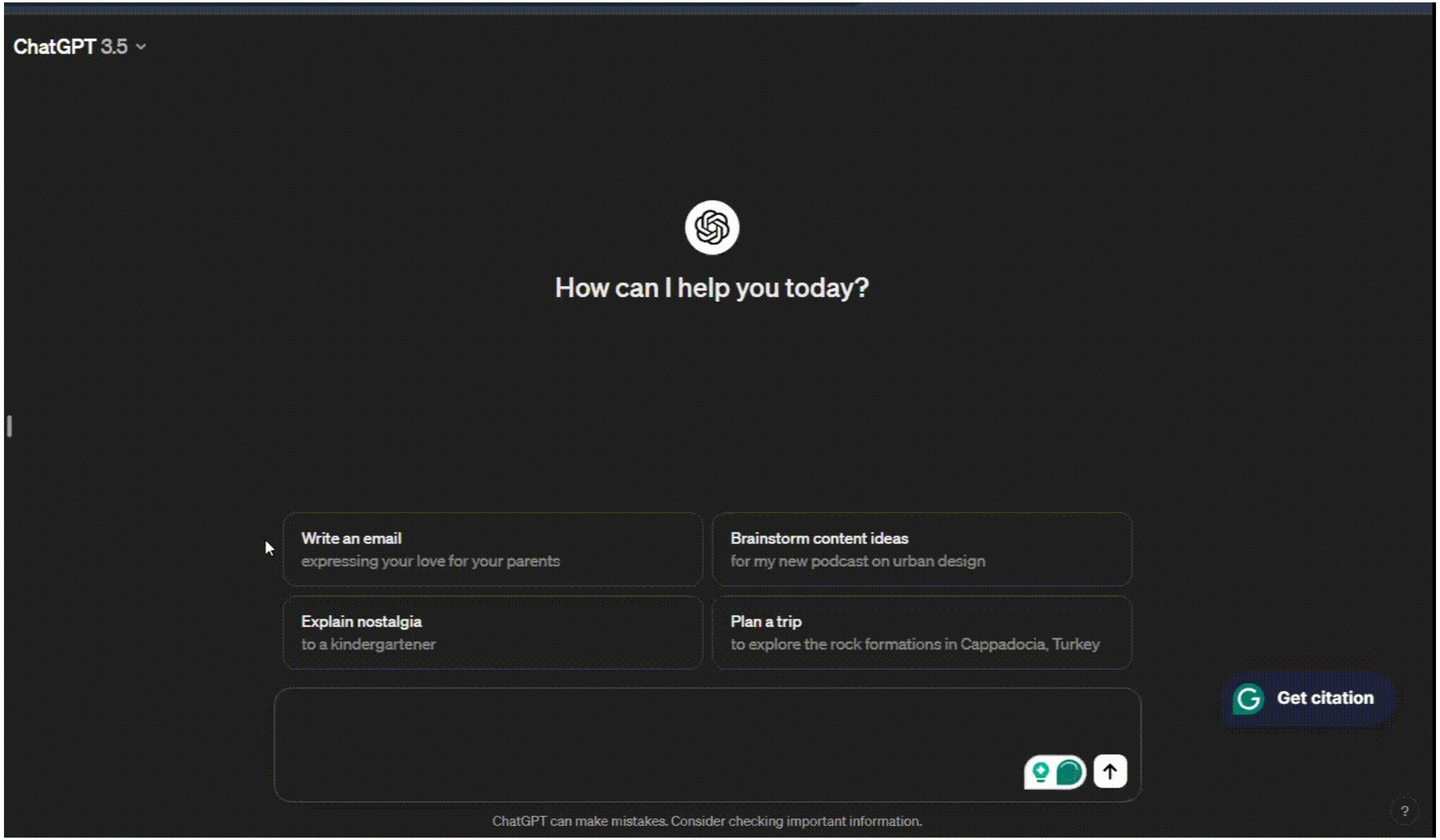
Unknown item quality



Resource-intensive process

Gorgun, G. & Bulut, O. (revise & resubmit). Exploring quality criteria and evaluation methods in automated question generation: A comprehensive survey.

Leveraging large-language models to facilitate item analysis



How do large language models work?

It is finally spring in Edmonton because it stopped …

raining	.025
snowing	.672
.	
.	
.	
blooming	.004
singing	.001

Study 3: Item analysis with large-language models as a filtering process

- *N* = 1825 cloze items

- Items were rated as either *Good* or *Bad*
  - Bad *n* = 1102
  - Good *n* = *723*

- Instruction-tuned Llama-2
  - 20% used as test set (*n* = 363)

Enhancing the precision of a large-language model by providing instructions, examples, and output for a specific task

Gorgun, G. & Bulut, O. (under review). Using instruction-tuned large-language models to evaluate automatically generated questions: A feasibility study.

Treat Llama-2 as a content expert

You are a **content expert** helping us understand the cloze question quality. A good question tests a **key concept** from the sentence, is **reasonable to answer**, **unambiguous**, and **specific**. A bad question is unreasonable to answer, too broad, ambiguous, or lacks specificity and depth.

Is this question good or bad?

Instruction

- You are a content expert helping us understand the cloze question quality. ... Is this question good or bad?

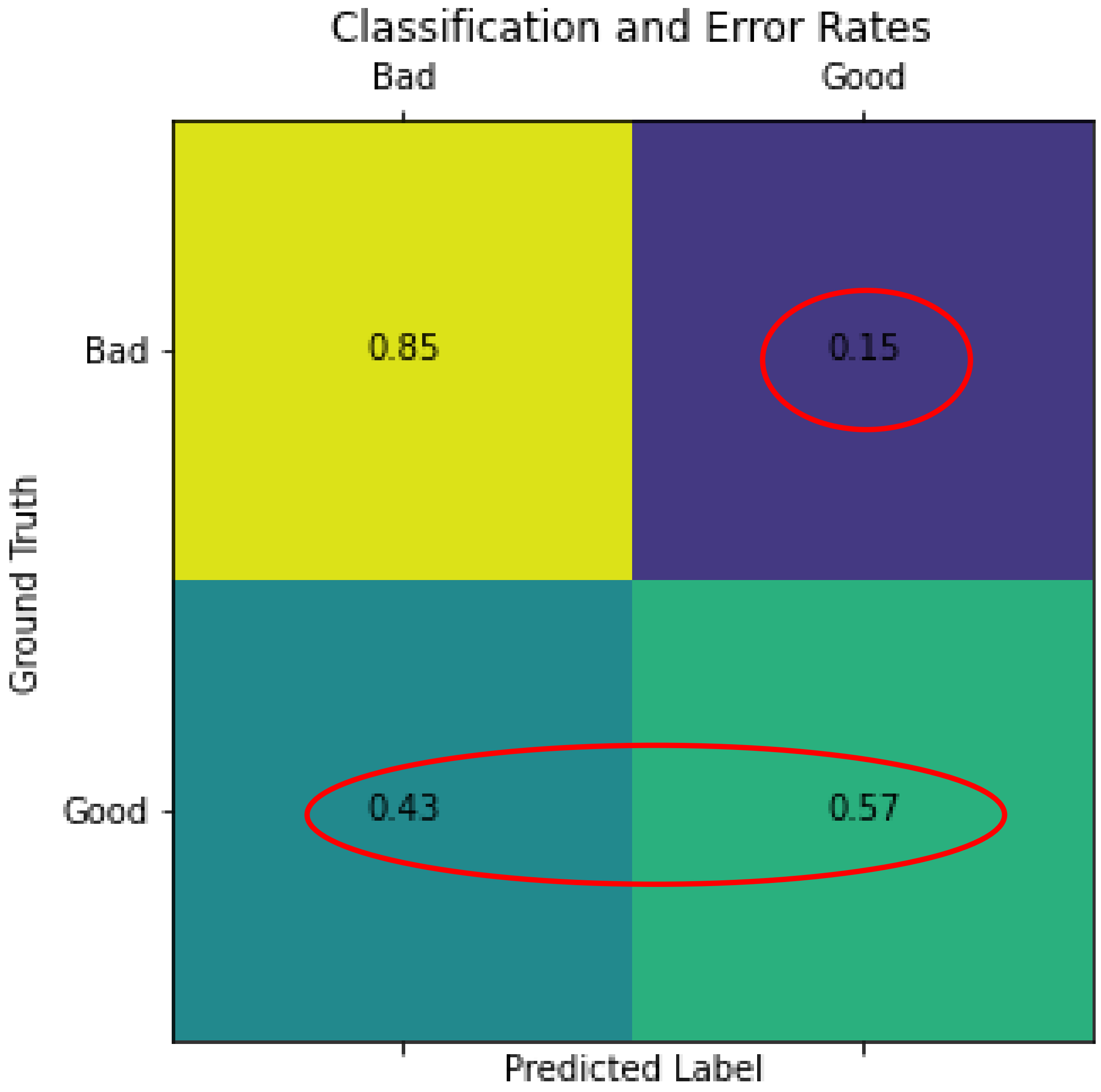
Input

- The Iron Age north of the Alps is divided into the Pre-Roman Iron Age and _______.

Output

- Bad

Using >1400 examples in training set, we instruction-tuned Llama-2 and evaluated how well it can predict item quality on the test set



Llama-2 performed well for filtering out bad questions from the item bank.

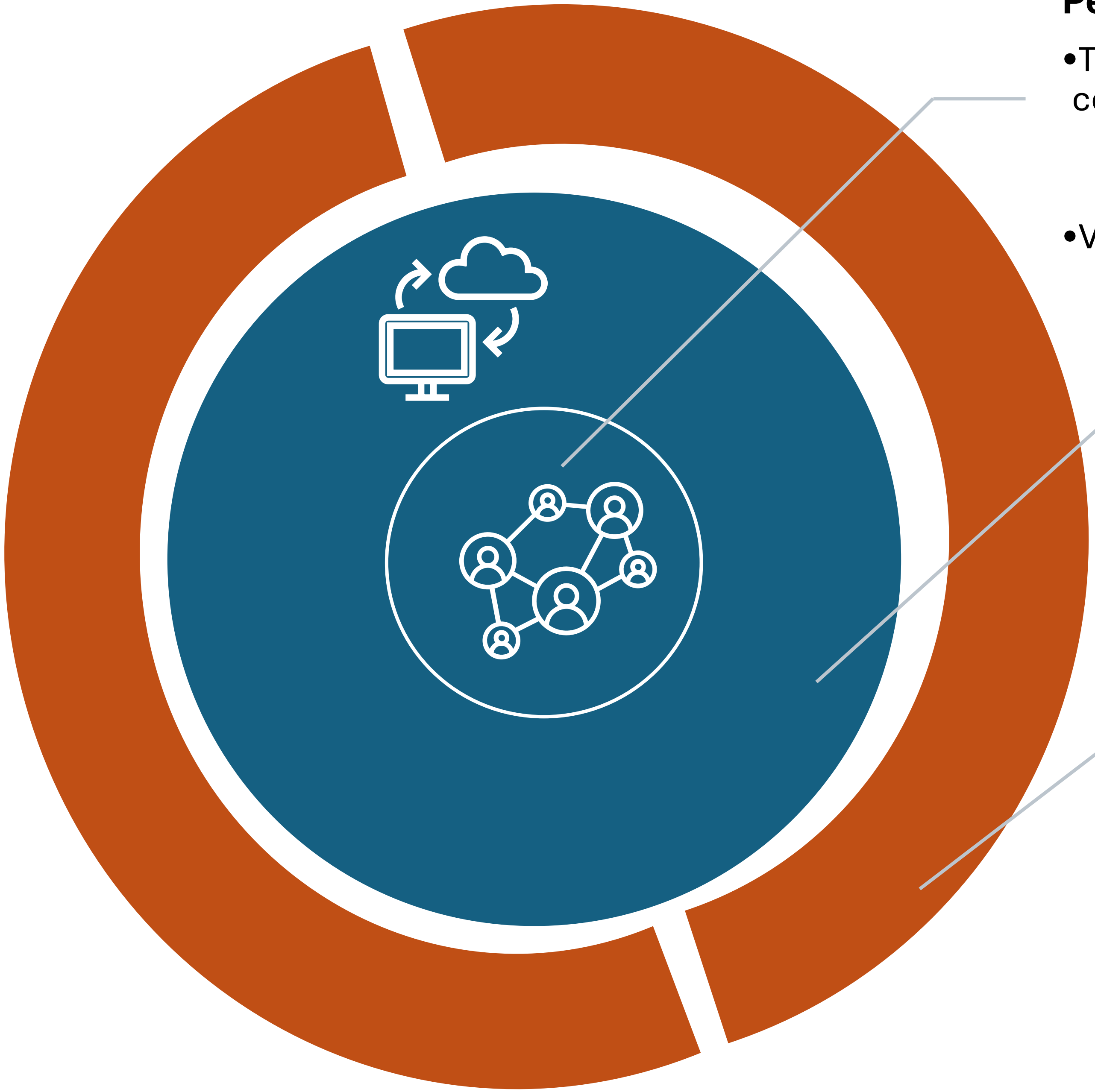
However, almost half of good items were classified as bad.

Research Interests

Assessment and examinee characteristics interact

Items as building blocks of assessments

New methods for enhancing assessment practices of diverse populations



Personalized assessments

- The role of large-language models and computational models?
  - Diverse and inclusive item development
  - Adapting assessments to diverse populations
- Validity framework?

Fusion of computational methods with psychometrics

- Fairness in models trained with past human data
- Validity of new methods

Promote transdisciplinary communication among industry partners and academic researchers

- Reiterate psychometric rigor in learner modeling
- Establish collaboration among practitioners and researchers

UBC EdTeKLA C.R.A.M.E eyeread ETS

Key contributions

Modelled learner behavior (e.g., engagement) across various educational contexts

Enhanced adaptivity of assessments to promote better assessment experiences for examinees

Leveraged theories and concepts in computer science, psychology, and learning sciences to introduce novel psychometric methods

gorgun@ualberta.ca

guhergorgun.com