The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid | July 11-15, 2022
VIRTUAL CONFERENCE

A Robust Computerized Adaptive Testing Approach in Educational Question Retrieval

Yan Zhuang[1,3], Qi Liu [1,2,3*], Zhenya Huang[1,3], Zhi Li[1,3], Binbin Jin[4], Haoyang Bi[1,3], Enhong Chen[1,3], Shijin Wang[3,5]

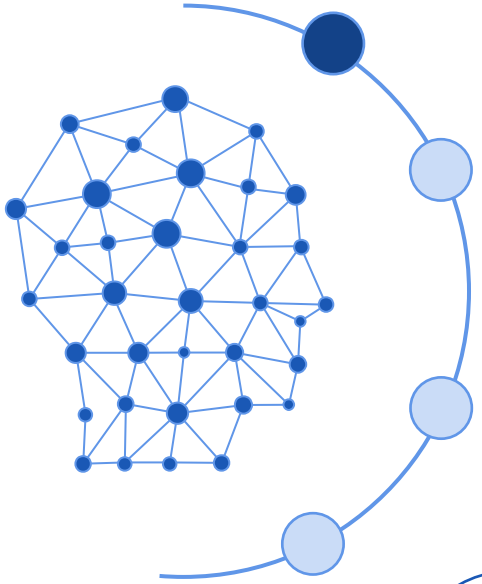
[1]Anhui Province Key Laboratory of Big Data Analysis and Application, School of Data Science & School of Computer Science and Technology, University of Science and Technology of China (USTC); [2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center; [3] State Key Laboratory of Cognitive Intelligence;
[4] Huawei Cloud Computing Technologies Co., Ltd; [5] iFLYTEK AI Research (Central China), iFLYTEK, Co., Ltd

Reporter: Yan Zhuang

SIGIR-2022

中国科学技术大学
University of Science and Technology of China

Outline



1 Background

2 Methodology

3 Experiment

4 Conclusion

Background

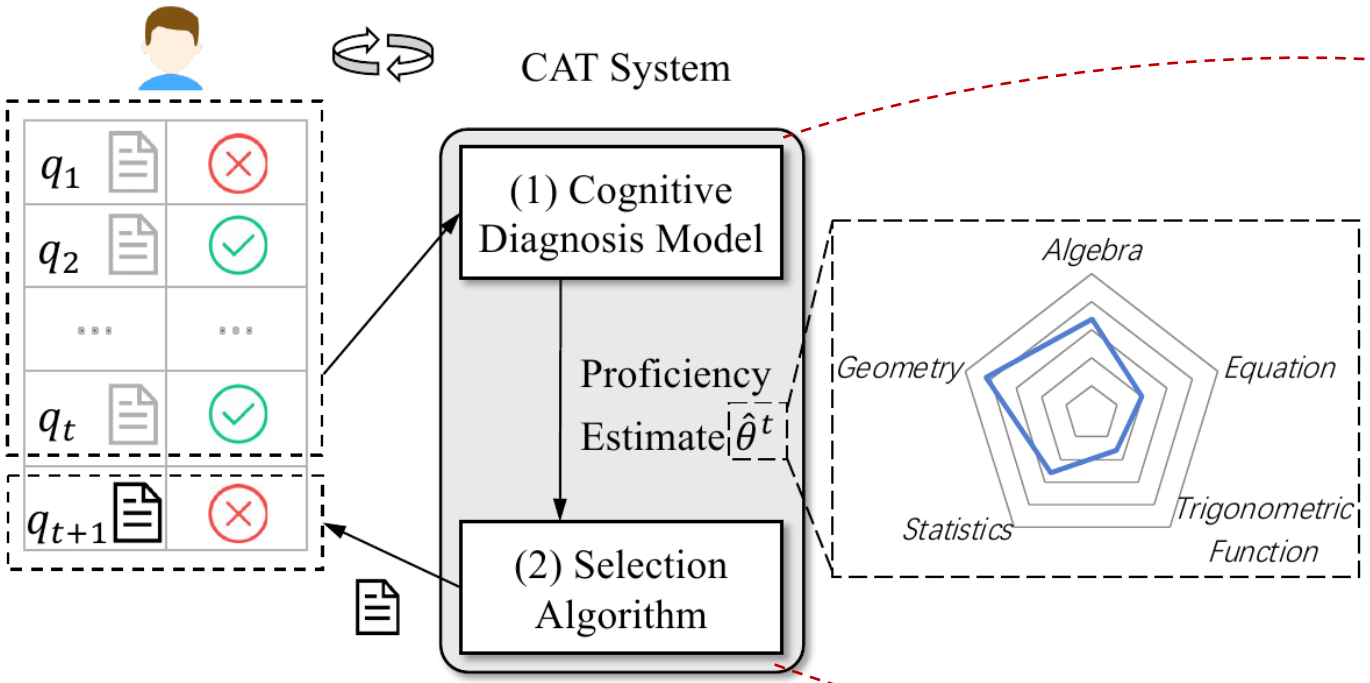
How to accurately and efficiently measure student’s proficiency?

- Paper-and-pencil Examination
  - Too many questions - inefficient and boring
  - Fixed time/place - inflexible

- Computerized Adaptive Testing (CAT)
  - Personalization and reduce test length
  - Flexible time/place

Background

Typical CAT procedure.



(1) Cognitive Diagnosis Model (CDM), which first estimates the student's current proficiency $\hat{\theta}^t$ based on previous t responses.

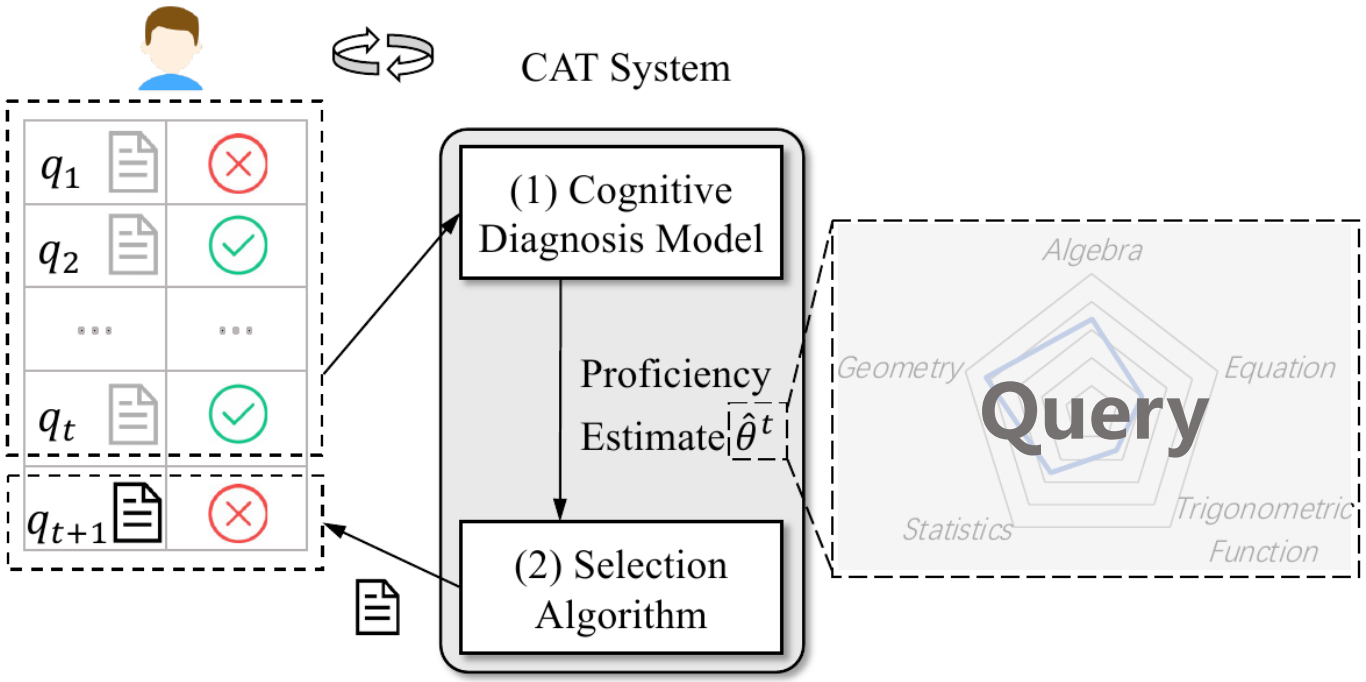
(2) The Selection Algorithm then retrieves the next question, guided by his/her current proficiency estimate $\hat{\theta}^t$ above.

Goal:

- Measuring student' s proficiency accurately
- Reducing test length

Background

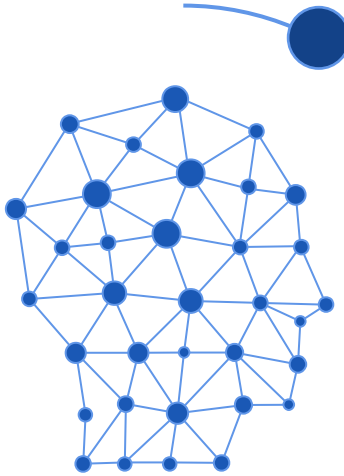
Typical CAT procedure.



◆ **Drawback：** The selection algorithm is inefficient if the query ($\hat{\theta}^t$) is not close to student's true proficiency θ_0

Poor Robustness

Outline



1. Background
2. Methodology
3. Experiment
4. Conclusion

Methodology

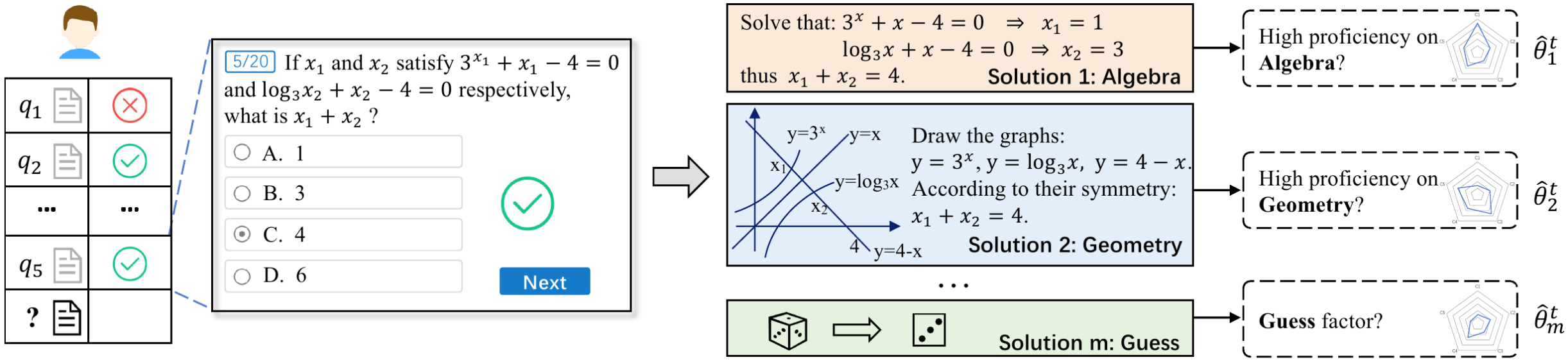
Findings

- **Student is "multi-facet " : Student's previous responses often correspond to multiple estimates instead of the singleton:**
  - **Example 1: One student correctly answers a simpler question (difficulty = 3) but wrong answers a harder one (difficulty = 8) -> his/her proficiency [3, 8]**
  - **Example 2: There are multiple solutions to a question. Each one corresponds to a potential proficiency.**

Methodology

Findings

◆ **Student is "multi-facet " : Student's previous responses often correspond to multiple estimates instead of the singleton:**



Methodology

Proposed Approach

- **Multiple estimates $\{\hat{\theta}_1^t, \hat{\theta}_2^t \ldots, \hat{\theta}_k^t\}$ will be generated at each step as student's multi-facet perspective.**
  - Using their average $\theta^* = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^t$ as a new query (replacing $\hat{\theta}^t$) for question selection, ensuring that $\|\theta^* - \theta_0\| \to 0$

$$\left\|\theta^* - \theta_0\right\|^2 = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left\|\hat{\theta}_i - \theta_0\right\|^2}_{\text{① Accuracy}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left\|\hat{\theta}_i - \theta^*\right\|^2}_{\text{② Diversity}}$$

Methodology

Proposed Approach

- **Multiple estimates $\{\hat{\theta}_1^t, \hat{\theta}_2^t \ldots, \hat{\theta}_k^t\}$ will be generated at each step as student's multi-facet perspective.**
  - Using their average $\theta^* = \frac{1}{m}\sum_{i=1}^{m} \hat{\theta}_i^t$ as a new query (replacing $\hat{\theta}^t$) for question selection, ensuring that $\|\theta^* - \theta_0\| \to 0$

$$\left\|\theta^* - \theta_0\right\|^2 = \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left\|\hat{\theta}_i - \theta_0\right\|^2}_{\text{① Accuracy}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\left\|\hat{\theta}_i - \theta^*\right\|^2}_{\text{② Diversity}}$$

smaller **larger**

Methodology

Proposed Approach

- At step t, we adjust the optimization function of $\{\hat{\theta}_1^t, \hat{\theta}_2^t \ldots, \hat{\theta}_k^t\}$, by adding diversity-regularization term $\phi(\theta_i)$ to the commonly used MLE target:

$$\hat{\theta}_i^t = \arg\min_{\theta_i} \mathcal{L}_{MLE}(\theta_i) - \lambda\psi(\theta_i) \quad \text{for } i = 1, \ldots, m.$$

$$\psi(\theta_i) = \frac{1}{2}\left\|\theta_i - \theta_i^*\right\|^2, \quad \theta_i^* = \frac{1}{i-1}\sum_{k=1}^{i-1}\hat{\theta}_k^t,$$

Methodology

Proposed Approach

- At step t, we adjust the optimization function of $\{\hat{\theta}_1^t, \hat{\theta}_2^t \ldots, \hat{\theta}_k^t\}$, by adding diversity-regularization term $\phi(\theta_i)$ to the commonly used MLE target:

$$\hat{\theta}_i^t = \arg\min_{\theta_i} \mathcal{L}_{MLE}(\theta_i) - \lambda\psi(\theta_i) \quad \text{for } i = 1, \ldots, m.$$

Accuracy

Diversity

$$\psi(\theta_i) = \frac{1}{2}\left\|\theta_i - \theta_i^*\right\|^2, \quad \theta_i^* = \frac{1}{i-1}\sum_{k=1}^{i-1}\hat{\theta}_k^t,$$

Methodology

Theoretical Analysis

- Such estimator's desirable statistical properties: asymptotic unbiasedness, efficiency, and consistency.

THEOREM 1 (ASYMPTOTIC UNBIASEDNESS AND EFFICIENCY). *The CDM's Fisher information on θ_0 is denoted as $I(\theta_0)$. When $\lambda < I(\theta_0)$ and $m \to \infty$, the estimator θ^* is asymptotically unbiased, that is,*

$$\mathbb{E}\left[\theta^*\right] = \theta_0. \tag{9}$$

→ **Unbiasedness**

Further, it is asymptotically efficient, with an asymptotic variance: $\mathbb{V}\mathrm{ar}\left[\theta^*\right] = \frac{1}{tI(\theta_0)}$, *which is equal to Cramér–Rao lower bound [6].*

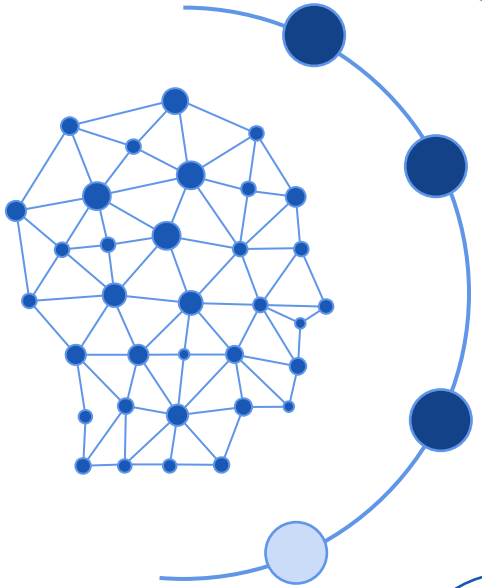
→ **Efficiency**

THEOREM 2 (CONSISTENCY). *Given any arbitrary small positive quantity ϵ, when $\lambda < I(\theta_0)$ and $m \to \infty$, the estimator θ^* is consistent, that is,*

$$\lim_{t\to\infty} \Pr\left\{\left|\theta^* - \theta_0\right| \geq \epsilon\right\} = 0. \tag{10}$$

→ **Consistency**

Outline



1. Background
2. Methodology
3. Experiment
4. Conclusion

Experiment

Setups

- **Dataset**
  - Real-world datasets from three **online tutoring system**
  - Involving two classic CDM**: IRT and NCDM**
- **Comparison Methods**
  - **Traditional information/uncertainty-based:**
    - FSI, KLI, MAAT (active learning)
  - **Deep Learning :**
    - BOBCAT
- **Evaluation Metrics:** AUC, ACC, MSE

Experiment

Results

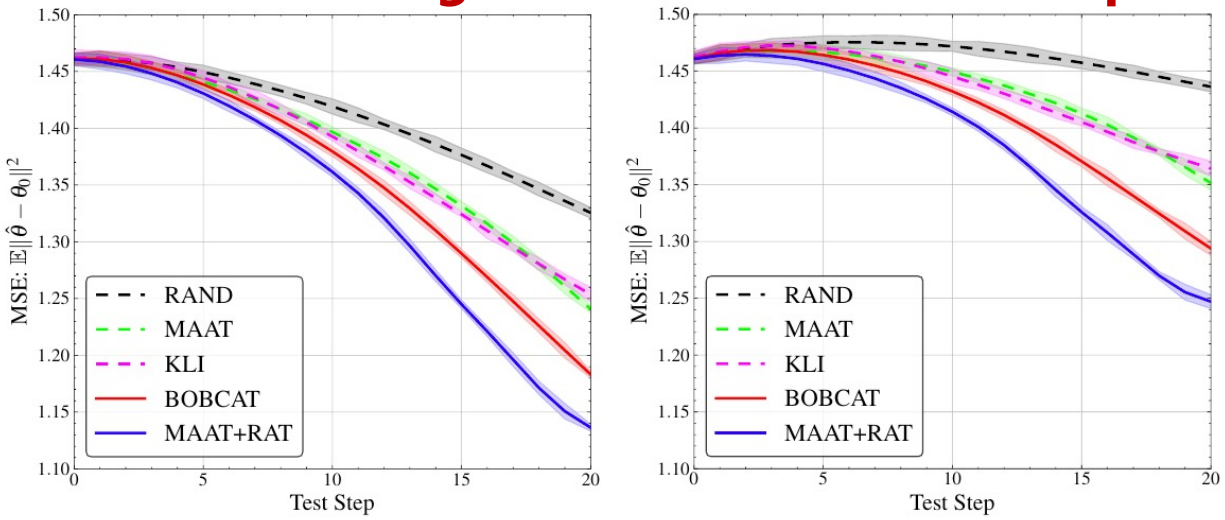
Our proposed RAT is general and achieves the best performance on all datasets and all types of CDMs.

Dataset	Junyi						Eedi						Math					
CDM	IRT			NCDM			IRT			NCDM			IRT			NCDM		
Metric	ACC (%)						ACC (%)						ACC (%)					
Step	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
Random	70.30	71.73	72.11	70.28	71.96	73.12	62.83	65.88	68.62	62.16	66.30	69.22	72.57	73.88	80.31	72.11	76.12	81.40
FSI	71.25	72.93	74.02	–	–	–	64.63	67.72	70.54	–	–	–	74.07	78.63	83.63	–	–	–
KLI	71.37	72.98	74.92	–	–	–	64.57	67.14	70.08	–	–	–	73.42	77.40	83.14	–	–	–
MAAT	72.31	73.31	75.22	72.44	73.17	75.47	64.86	67.38	71.42	64.22	68.13	71.70	75.84	77.37	82.53	76.36	79.87	82.81
BOBCAT	73.25	73.81	75.89	**73.54**	74.13	76.32	65.58	68.14	72.20	66.30	69.56	72.31	77.19	79.90	82.66	78.36	81.00	85.04
FSI+RAT	73.46	74.82	76.10	–	–	–	66.10	70.39	73.17	–	–	–	77.36	80.75	**84.92**	–	–	–
KLI+RAT	73.76	**75.88**	**77.19**	–	–	–	66.01	70.27	**73.55**	–	–	–	**78.09**	**81.19**	84.57	–	–	–
MAAT+RAT	**73.78**	75.35	76.92	73.10	**75.30**	**77.13**	**66.14**	**70.42**	73.25	**67.35**	**71.65**	**73.37**	77.14	79.71	83.87	**78.38**	**81.14**	**85.05**
Metric	AUC (%)						AUC (%)						AUC (%)					
Step	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20	5	10	20
Random	72.83	73.18	75.32	72.55	74.46	76.87	65.48	68.63	72.20	66.00	69.82	72.55	67.82	67.61	76.90	67.98	70.50	76.97
FSI	73.70	74.28	76.16	–	–	–	67.27	70.72	74.50	–	–	–	69.56	73.13	78.15	–	–	–
KLI	73.91	74.41	76.07	–	–	–	67.10	70.33	73.89	–	–	–	69.82	73.28	78.28	–	–	–
MAAT	74.16	75.32	77.35	75.27	75.91	78.32	67.19	70.32	74.74	67.13	71.36	74.73	69.10	73.90	78.89	69.67	75.15	78.90
BOBCAT	75.99	76.25	78.49	**75.81**	76.33	79.64	68.43	71.03	75.76	69.11	72.01	76.13	70.62	74.32	79.19	**71.17**	74.54	79.58
FSI+RAT	**76.56**	76.64	78.86	–	–	–	**68.93**	**73.12**	75.99	–	–	–	70.89	76.17	79.38	–	–	–
KLI+RAT	76.33	**77.94**	**79.67**	–	–	–	68.90	72.99	76.03	–	–	–	**71.03**	76.01	**80.66**	–	–	–
MAAT+RAT	75.67	77.74	79.41	75.33	**77.06**	**79.83**	**68.93**	73.05	**76.09**	**70.39**	**73.88**	**76.63**	70.44	**77.41**	79.14	70.44	**76.40**	**80.63**

Experiment

Robustness Evaluation

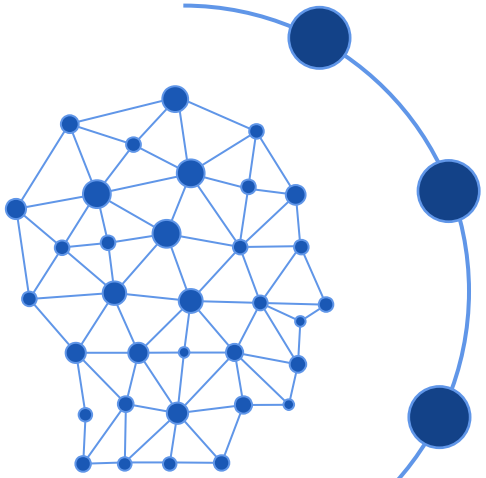
- We artificially generate θ_0 and simulate student-question interaction process and expose this simulated CAT to various perturbations:
  - **Guess factors: The label is changed from 0 to 1 with 25% probability.**
  - **Slip factors: The label is changed from 1 to 0 with 5% probability.**



(a) Guess: 25%, Slip: 5%.

(b) Guess: 50%, Slip: 10%.

Outline



1. Background
2. Formulation
3. Experiment
4. Conclusion

Conclusion

Conclusion

- **Present a new optimization criterion called RAT for educational measurement.**
  - Generic and Robust
- **Such new estimator in RAT possesses highly desirable statistical properties**
  - asymptotic unbiasedness, efficiency, and consistency
- **Conduct extensive experiments with real-world educational datasets**
  - Efficiency, Robustness

Thanks for your listening!

For more details, please refer to our paper!

Reporter：Yan Zhuang
zykb@mail.ustc.edu.cn

中国科学技术大学
University of Science and Technology of China