Item reduction for digital literacy assessment: Perspectives from content-expert, psychometrics, and machine-learning

He Ren[1], Qianqian Pan[2], Yuxiao Zhang[3], Weicong Lyu[4]

[1] College of Education, University of Washington, United States
[2] National Institute of Education, Nanyang Technological University, Singapore
[3] College of Education, Purdue University, United States
[4] Faculty of Education, University of Macau, Macau, China

NCME 2025, 4/26, Denver

UNIVERSITY *of* WASHINGTON | COLLEGE OF EDUCATION

Outline

- Background
- Methods
- Study Design
- Results
- Discussions

Background: Digital literacy (DL)

- Digital literacy (DL) is an essential skill for success in education, work, and social interactions

- A digital literacy assessment based on DigComp 2.1 was developed (Law et al., 2023)
  - Assess a broad age range
  - Offer a comprehensive view of individuals' DL skills

Background: The Digital literacy Assessment (DLA)

Properties of the DLA

- Confirmed a unidimensional construct of digital literacy, with strong item discrimination, a wide difficulty range, and high reliability
- A relatively long test

Form 1	Form 2	Form 3
Grade 3 to 5 students	Grade 6 to 9 students	Grade 10 to 12 students
45 items	50 items	51 items

Background: Item reduction

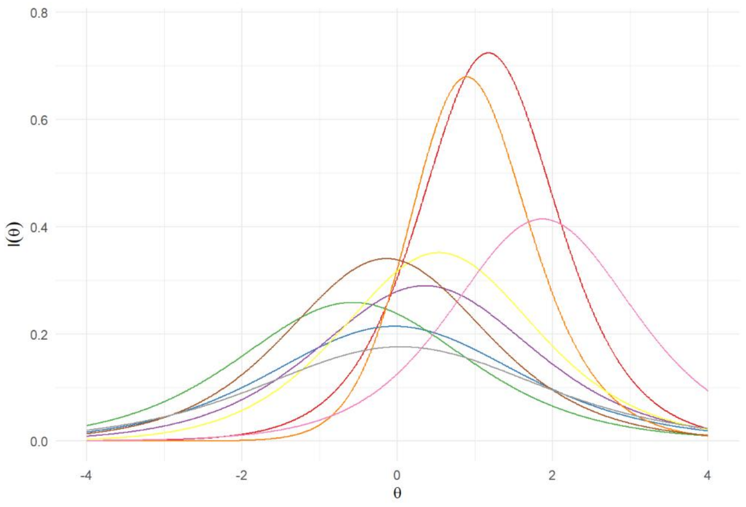
- A real-world problem
  - Longer tests might lead to low response rate, more careless responses
  - Longer tests might lead to higher administration costs
  - Adding new domains to an existing test often requires reducing the number of items in the original test

Methods: Psychometric methods

- Item response theory (IRT)
  - A family of psychometric models that predict' responses by linking students' latent ability and item characteristics
  - Item information is the quantity that measures how precisely an item can assess individual with specific latent traits

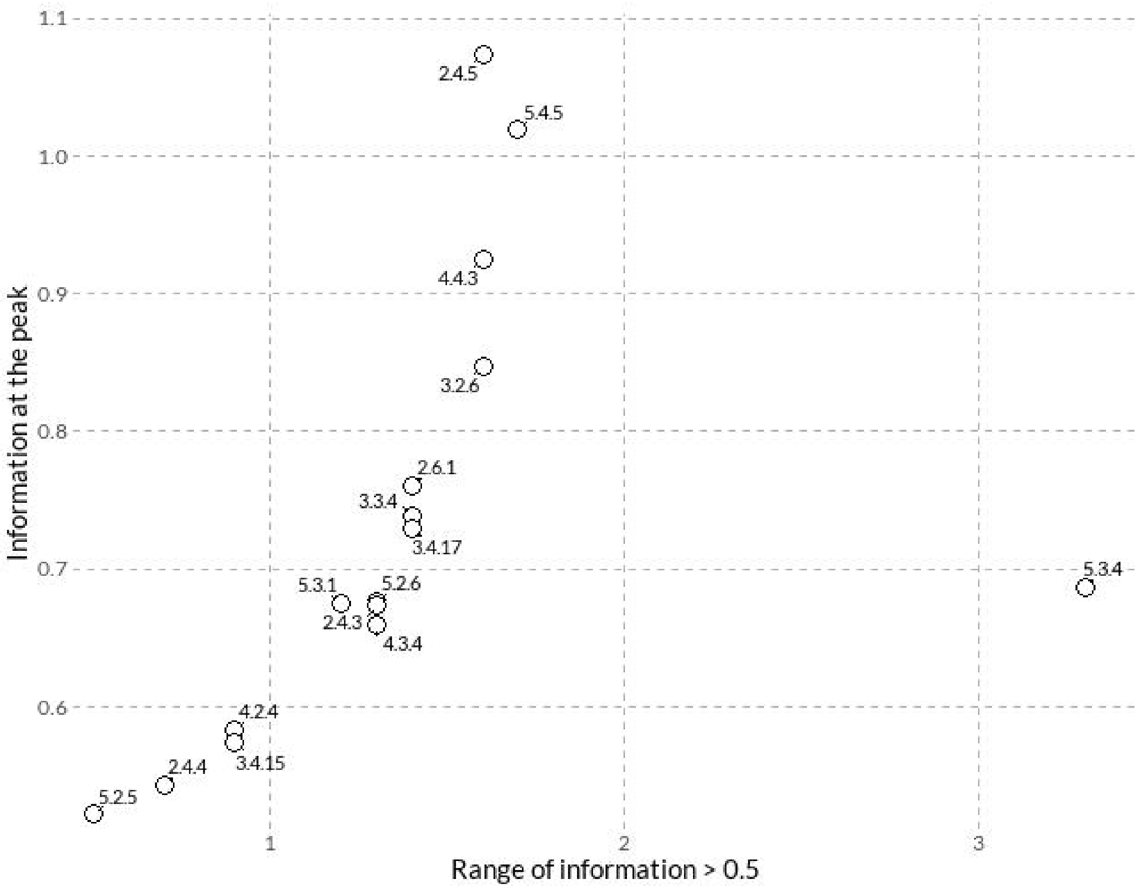
$$I_j(\theta) = \frac{[P_j'(\theta)]^2}{\boxed{P_j(\theta)}[1 - P_j(\theta)]}.$$

Probability of correct response on item *j*



Methods: Psychometric methods

- High information in a wide range
  - Plot of peak information vs. range with information > δ
  - The upper right items are selected



Methods: Machine learning methods

- A feature selection problem
  - Each item is one feature
  - The outcome is student's performance/ability
- Several methods
  - LASSO regression
  - Random forest (RF)
  - Genetic algorithm (GA)

Methods: LASSO regression

- For a linear regression

$$y = \boldsymbol{X\beta}$$

  - The optimization can be expressed as

$$\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2\}$$

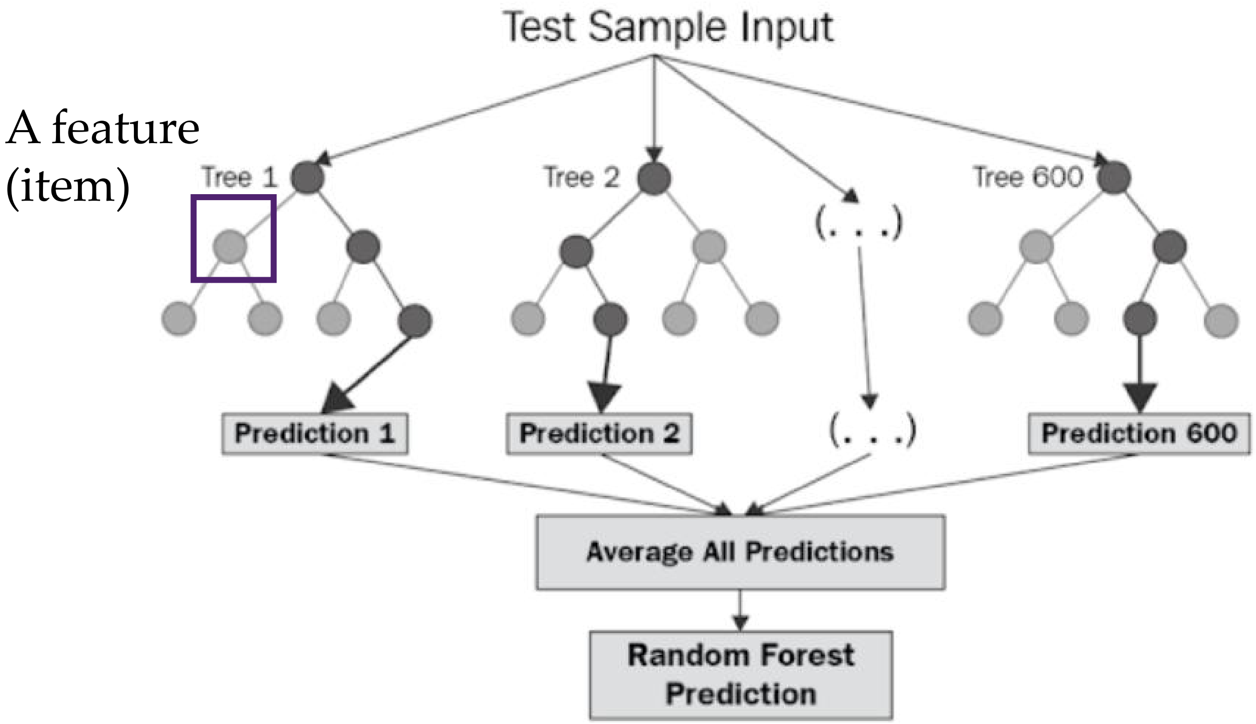
  - LASSO regression optimize

$$\min_{\beta}\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \boxed{\lambda\|\boldsymbol{\beta}\|_1}\}$$

l_1 norm so that unimportant βs will be shrunk to 0

Methods: RF

- An ensemble method with decision tree as the basic unit
- Permutation Feature importance
  - By breaking the relationship between each feature and outcome, we determine the importance of each feature (measured by the decrease in prediction accuracy).



Methods: GA

- Item selection can be represented by an indicator vector

$$(1,0,0,0,1,0,0)$$

The 1st and 5th items are selected out of the 7 items

- Aims to minimize the cost function

$$Cost = l * \underbrace{\boxed{s}}_{\text{Number of selected items}} + (1 - \underbrace{\boxed{R^2_{adj}}}_{\text{Fit for estimating the total score}})$$

- 2^D possible combinations with D items
- A computational search technique

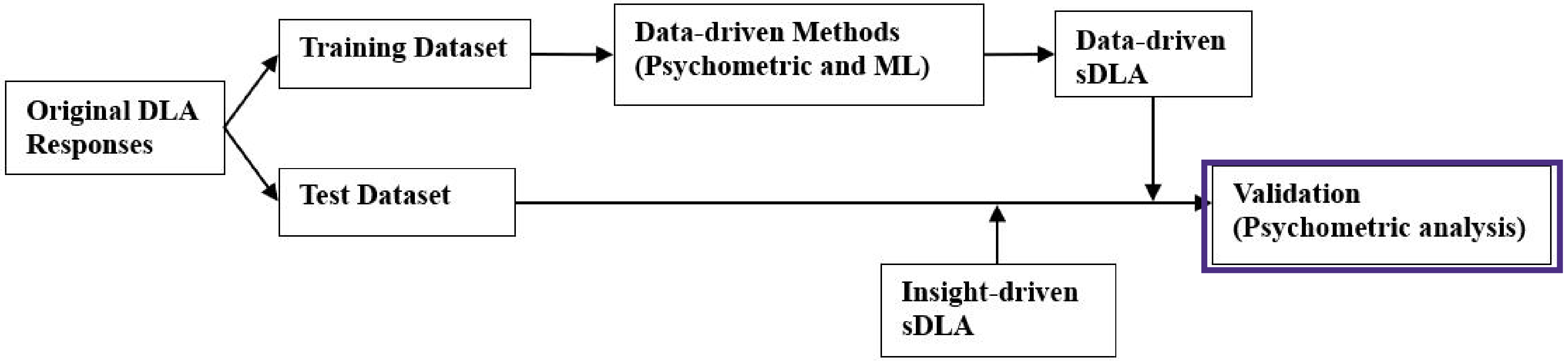
Methods: GA

- Searching procedure
  - Crossover
  - Mutation

Parents		Offsprings		Offsprings
(1,0,0,0,1,0,0)	→	(1,0,0,0,1,0,1)	→	(1,0,0,0,1,0,1)
(0,0,1,0,1,0,1)		(0,0,1,0,1,0,0)		(0,0,1,**1**,1,0,0)
	Crossover		Mutation	

Study Design

- Experts have developed a short DLA, with 10 items per form (Pan et al., 2024)
- Short DLAs with the same length were developed with data-driven methods
- Procedure



IRT analysis was conducted on the test sample with selected and all items, respectively

Results

- Evaluation on ability estimation

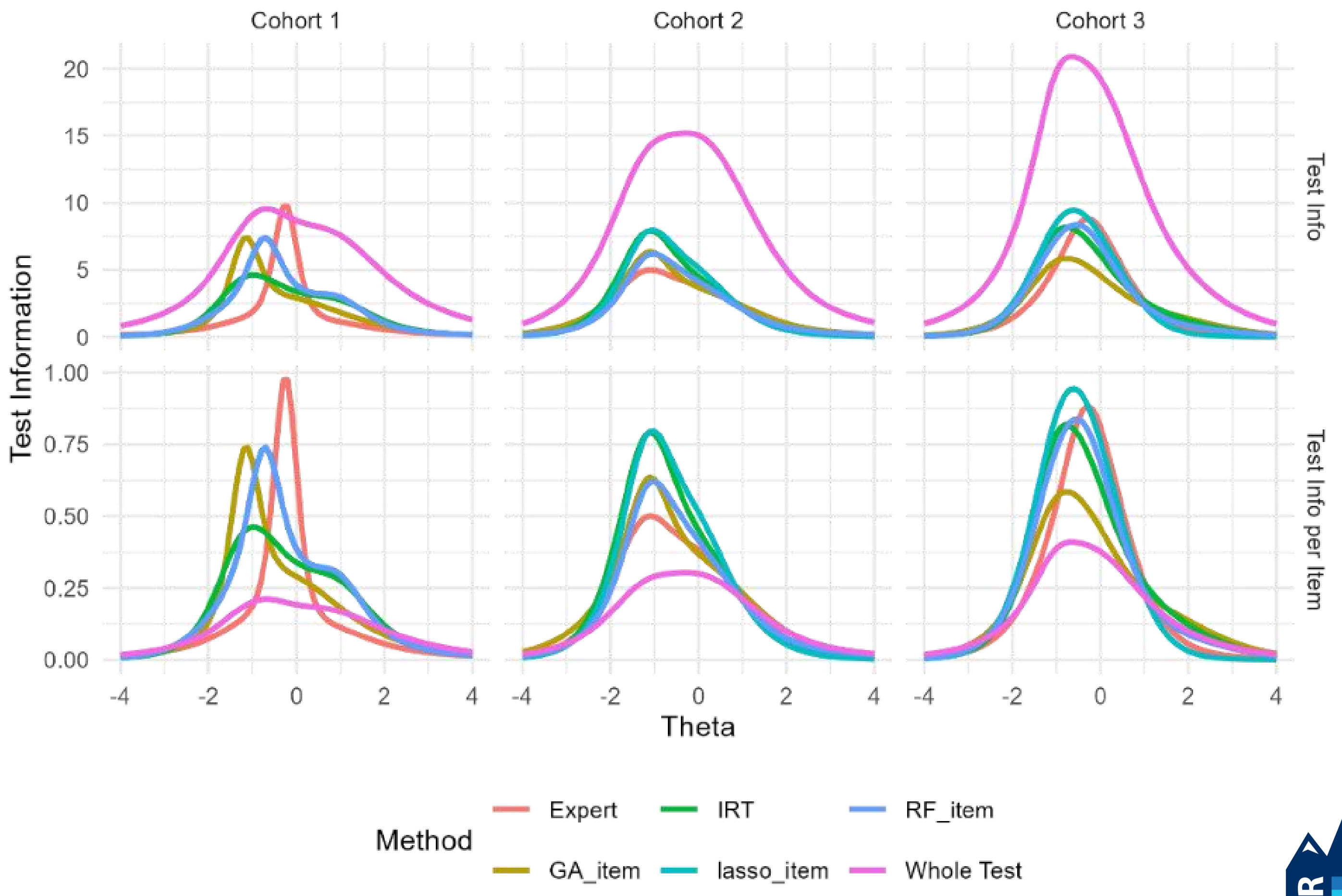
Table 1. Correlations between IRT estimated ability with selected items and full test

	Method	Cohort 1	Cohort 2	Cohort 3
	Expert	0.799	0.910	0.898
Data-driven methods	Item Information-IRT	0.893	0.922	0.924
	LASSO	0.878	0.930	0.924
	GA	0.860	0.920	0.894
	RF	0.878	0.923	0.918

Note. GA: genetic algorithm; RF=random forest

Results

- Evaluation on test information



Results

- Evaluation on content coverage

Table 2. Number of selected items in each domain.

Data-driven methods

Cohort	Method	Domain 1	Domain 2	Domain 3	Domain 4	Domain 5
1	Expert	2	2	2	2	2
	Item Information-IRT	1	2	1	4	2
	LASSO	1	2	1	3	3
	GA	1	2	1	3	3
	RF	1	2	1	3	3
2	Expert	2	2	2	2	2
	Item Information-IRT	0	2	2	4	2
	LASSO	0	1	2	5	2
	GA	2	1	1	3	3
	RF	1	1	2	5	1
3	Expert	2	2	2	2	2
	Item Information-IRT	0	3	3	1	3
	LASSO	0	2	2	3	3
	GA	2	2	1	3	2
	RF	0	1	3	3	3

Note. GA: genetic algorithm; RF=random forest

Discussion

- This study investigated a variety of item reduction methods for a performance-based digital literacy assessment
- All data-driven methods (psychometric–based and ML–based) produced short-form scores that correlated more strongly with the full test score than the expert-driven short form
- A significant limitation of data-driven methodologies is the reduced content coverage
- No single method consistently outperformed the others across all cohorts
- Underscored the necessity of contextual calibration and iterative validation

Thanks!
heren@uw.edu

Personal Homepage

He Ren[1], Qianqian Pan[2], Yuxiao Zhang[3], Weicong Lyu[4]

[1] College of Education, University of Washington, United States
[2] National Institute of Education, Nanyang Technological University, Singapore
[3] College of Education, Purdue University, United States
[4] Faculty of Education, University of Macau, Macau, China