Cambridge Assessment

Is adaptive testing as useful as it claims to be?

http://www.cambridgeassessment.org.uk/our-research/

Tom Benton | March 2019

Aim of research

- Adaptive testing is supposed to be provide more reliable assessment than fixed forms
    - Use IRT to calculate which items are most appropriate for each candidate (item information)
    - Candidates avoid items that are too easy or too hard
    - Supposedly can halve the length of tests without any loss of reliability
- Question 1: Is this is true for the typical assessments we run?
- Question 2: Does it matter which IRT model we use?

Previous research – Optimal test construction

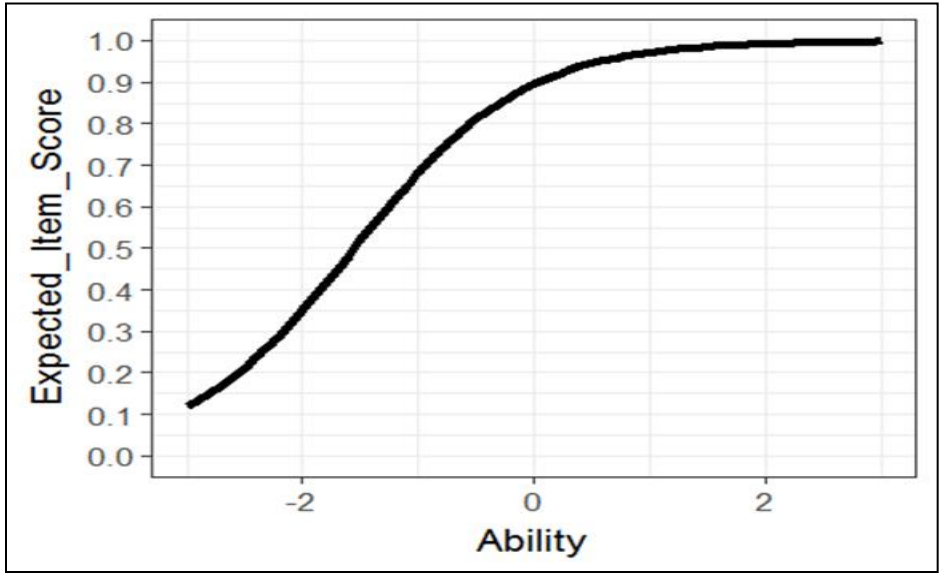
- Previous findings:
    - Optimal construction of **fixed tests** based on more complex IRT (than Rasch) *appeared* to substantially increase reliability
    - However, these gains in reliability often did not translate into the expected gains in predictive validity

Full paper available from

http://www.cambridgeassessment.org.uk/our-research/all-published-resources/conference-papers/

How do we select items using IRT?

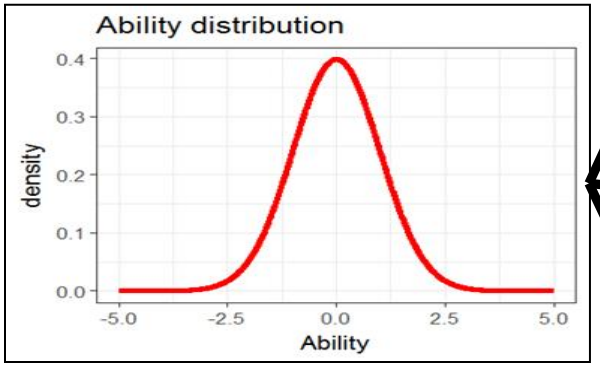
- Allows for fact that item quality is population dependent
- Key statistic is the item information function

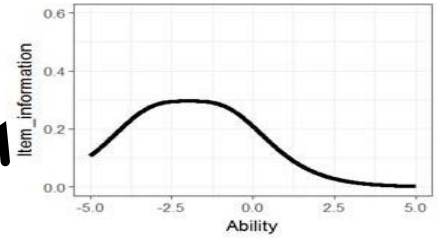


Cambridge Assessment

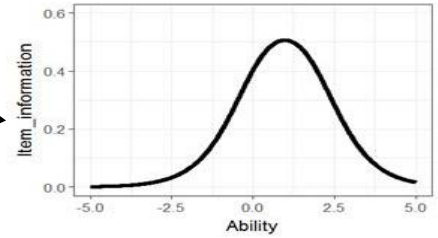
How can we select items for fixed tests using IRT?

For any population we can calculate the expected item information (weighted average) → Select items with best values

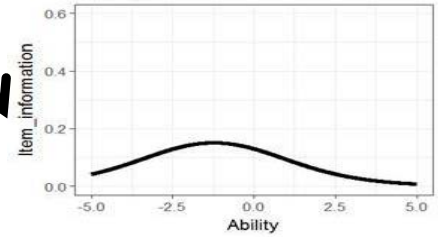




0.20

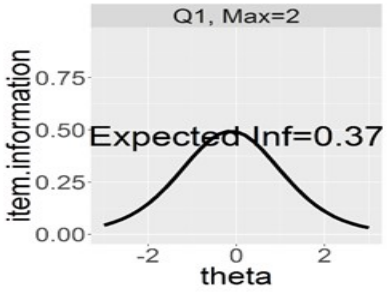


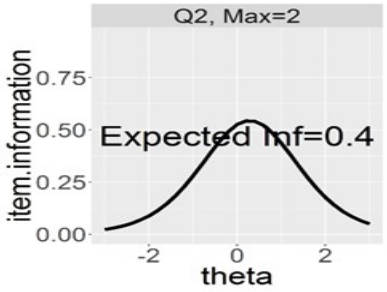
0.35

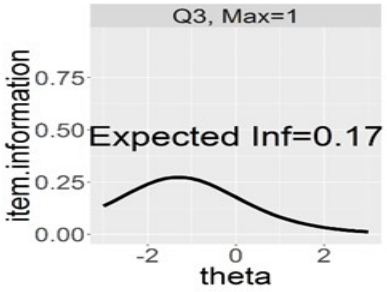


0.12

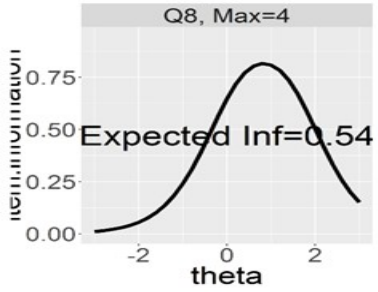
Cambridge Assessment

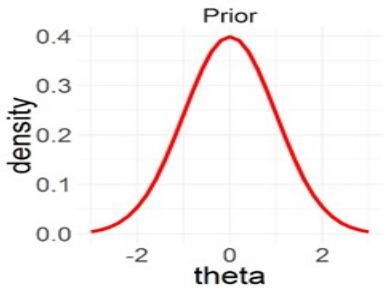


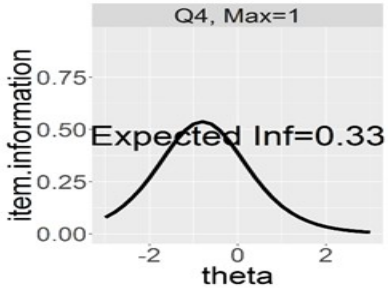


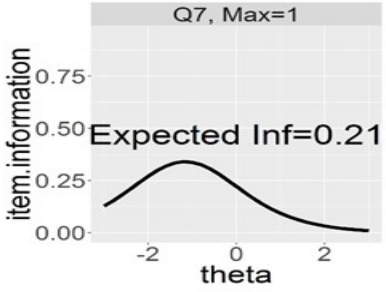


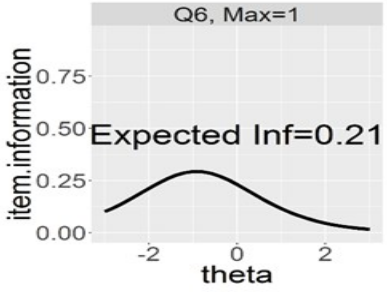
Adaptive item selection (same idea but 1 item at a time)

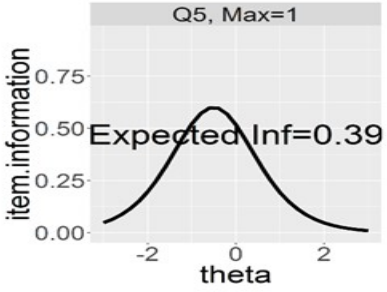


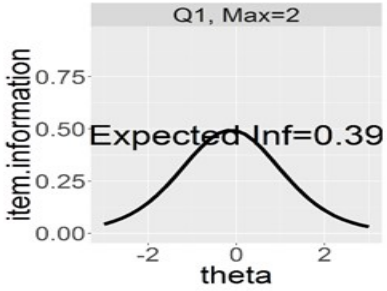


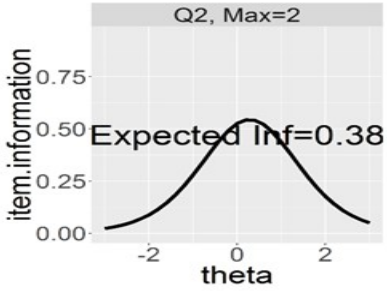


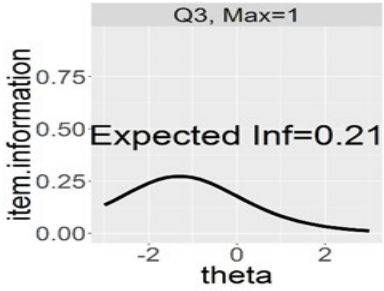


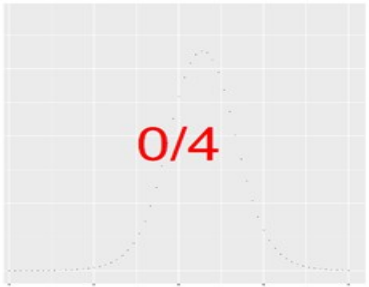


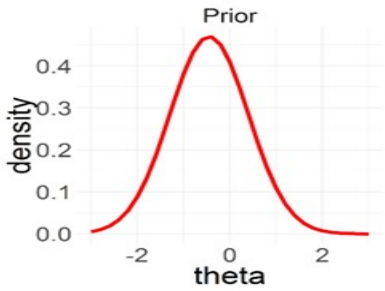


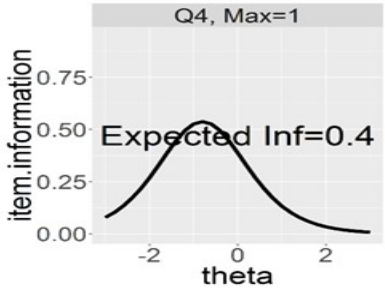

Q1, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.39
-2
0
2
theta

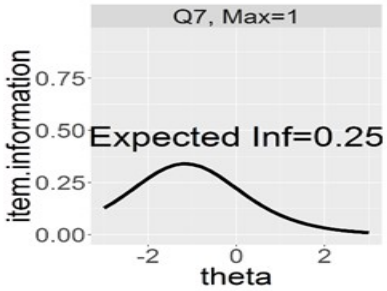

Q2, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.38
-2
0
2
theta

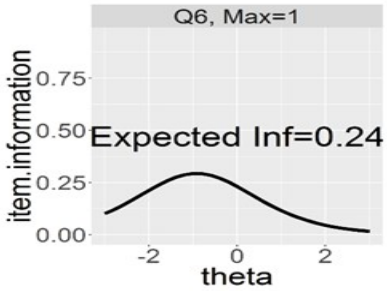

Q3, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.21
-2
0
2
theta

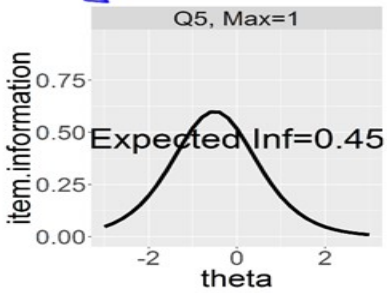

0/4

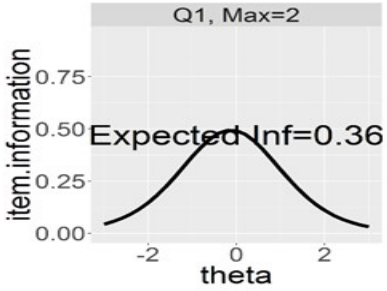

Prior
density
0.4
0.3
0.2
0.1
0.0
-2
0
2
theta

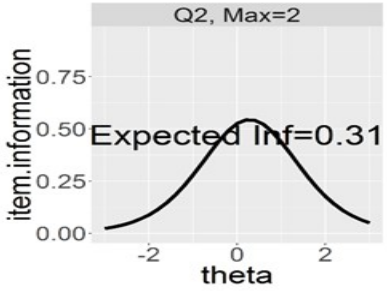

Q4, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.4
-2
0
2
theta

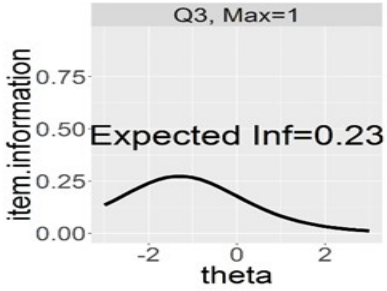

Q7, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.25
-2
0
2
theta

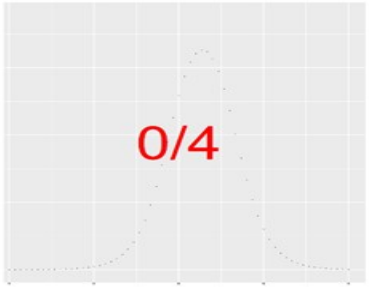

Q6, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.24
-2
0
2
theta

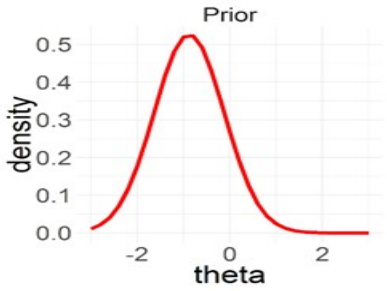

Q5, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.45
-2
0
2
theta

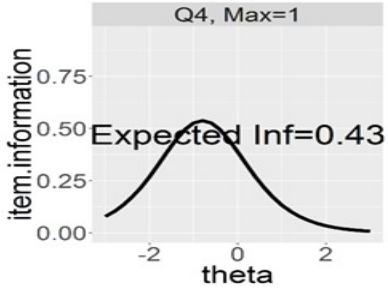

Q1, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.36
-2
0
2
theta

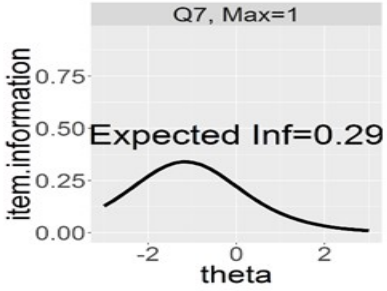

Q2, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.31
-2
0
2
theta

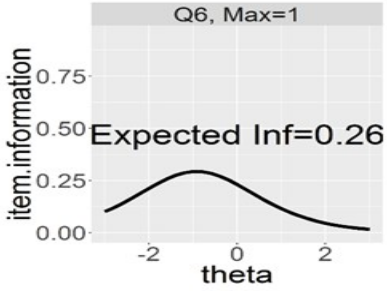

Q3, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.23
-2
0
2
theta

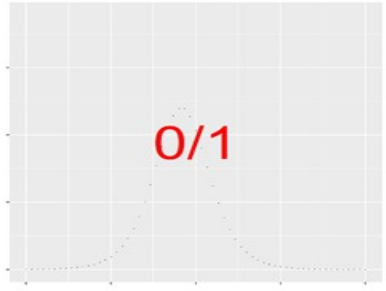

0/4

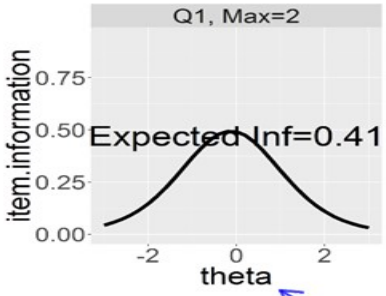

Prior
density
0.5
0.4
0.3
0.2
0.1
0.0
-2
0
2
theta

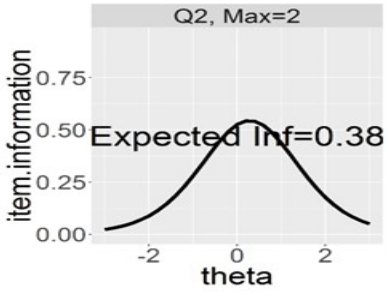

Q4, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.43
-2
0
2
theta

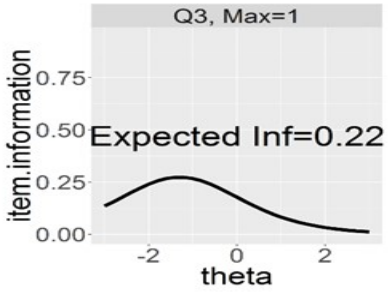

Q7, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.29
-2
0
2
theta

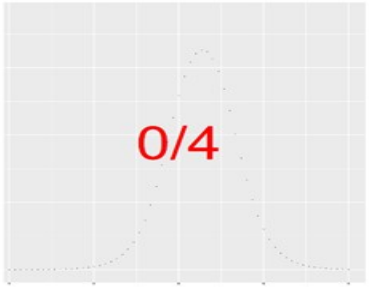

Q6, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.26
-2
0
2
theta

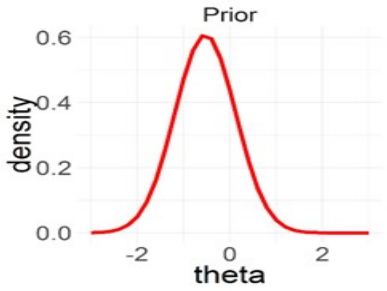

0/1

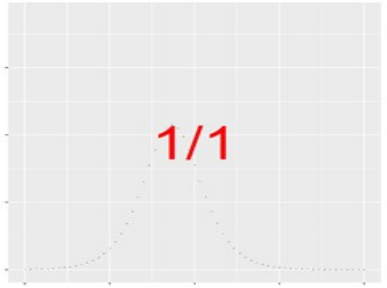

Q1, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.41
-2
0
2
theta

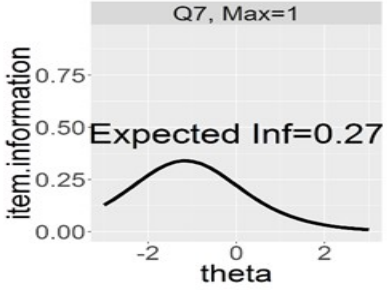

Q2, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.38
-2
0
2
theta

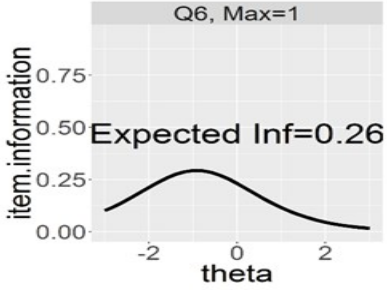

Q3, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.22
-2
0
2
theta

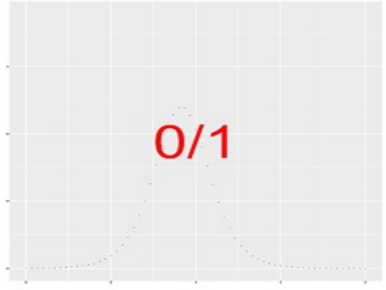

0/4

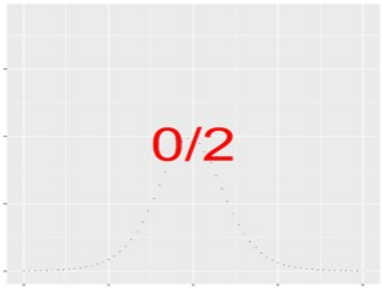

Prior
density
0.6
0.4
0.2
0.0
-2
0
2
theta

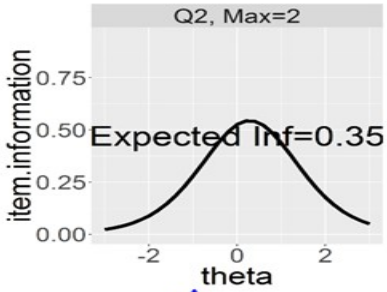

1/1

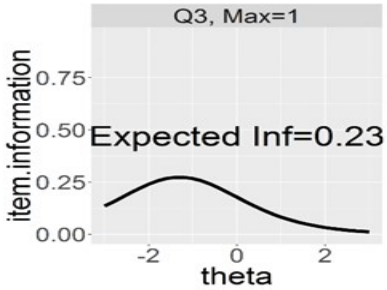

Q7, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.27
-2
0
2
theta

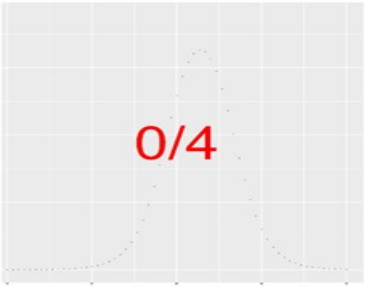

Q6, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.26
-2
0
2
theta

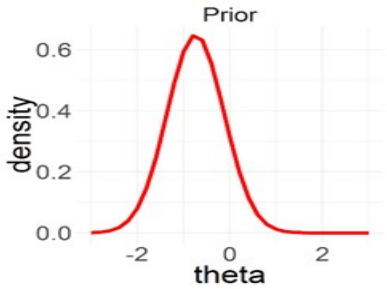

0/1

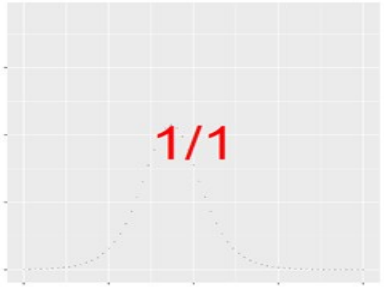

0/2

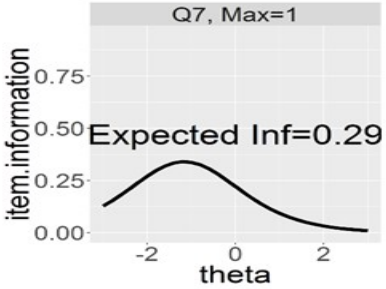

Q2, Max=2
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.35
-2
0
2
theta

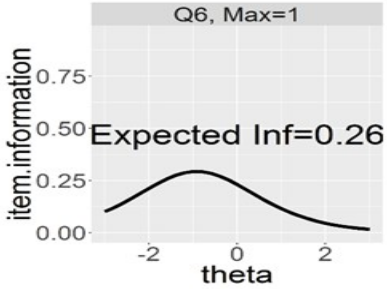

Q3, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.23
-2
0
2
theta

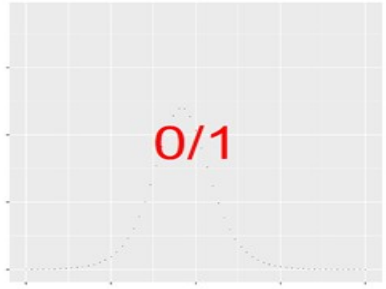

0/4


Prior
density
0.6
0.4
0.2
0.0
-2
0
2
theta


1/1


Q7, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.29
-2
0
2
theta


Q6, Max=1
item.information
0.75
0.50
0.25
0.00
Expected Inf=0.26
-2
0
2
theta


0/1

Choice of IRT models

Various options, including:

- Graded response model (GRM)
    - Different items may have different discriminations
    - Items with higher discrimination parameters (usually) yield more information
    - Scores give more weight to highly discriminating items
- Rasch model
    - All items equally discriminating
    - Focus is purely on which items are of the most appropriate difficulty

Outline of research study

For a bunch of real assessments:

- Simulate adaptive testing with real data (half the length of the original).
    - i.e. only count scores from items chosen by adaptive procedure and ignore the rest
- What happens to estimates of reliability?
- What happens to predictive value? (Predictive validity)
- How does this compare to non-adaptive half length tests?

Experimental method: Data

Analysis based on item level data from **122 real assessments**

- At least 5,000 candidates (median 8,000).
- No optional items
- At least 20 items (median 32)
- No items with negative discriminations
- At least one 1-mark and one multi-mark item (up to 5 marks)
- Maximum available scores ranged from 30 to 80 (median 60)
- Unidimensional (Velicer's method)

Experimental method (creation of test forms)

For each test form analysed…

1) Randomly create a test for each with half the number of marks of the full-length test (e.g. a 30 mark test from a 60 mark test)

- Start with **Random** selection

For example for a full length 60 mark test:

→e.g. Original had: Sixteen 1-mark, Ten 2-mark and Eight 3-mark items

→e.g. Randomly select : Eight 1-mark, Five 2-mark and Four 3-mark items

Experimental method (creation of test forms, ctd.)

2) Create “optimal” **fixed form** tests out of the full-length test with the same distribution of item maxima as the random selection

→ e.g. select the eight *best* 1-mark items, the five *best* 2-mark items and the four *best* 3-mark items

Try various methods of selecting the “best” items:

- IRT expected item information (based on GRM or Rasch)
- Classical statistics (R_rest, Item-total covariance, item variance)
- Stepwise method to maximise Cronbach’s alpha

Experimental method (creation of test forms, ctd)

3) Simulate an **adaptive procedure** for each candidate so that they all answer the same distribution of item maxima as the random selection

Essentially mimics a computer adaptive procedure but works within items of each maxima in turn from highest to lowest.

For example the adaptive procedure might:

→ *Sequentially* assign each candidate four 3-mark items

→ Then *sequentially* assign each candidate five 2-mark items

→ Then sequentially assign each candidate eight 1-mark items

Apply above method for both GRM and Rasch models

Example – 40 mark Biology Test

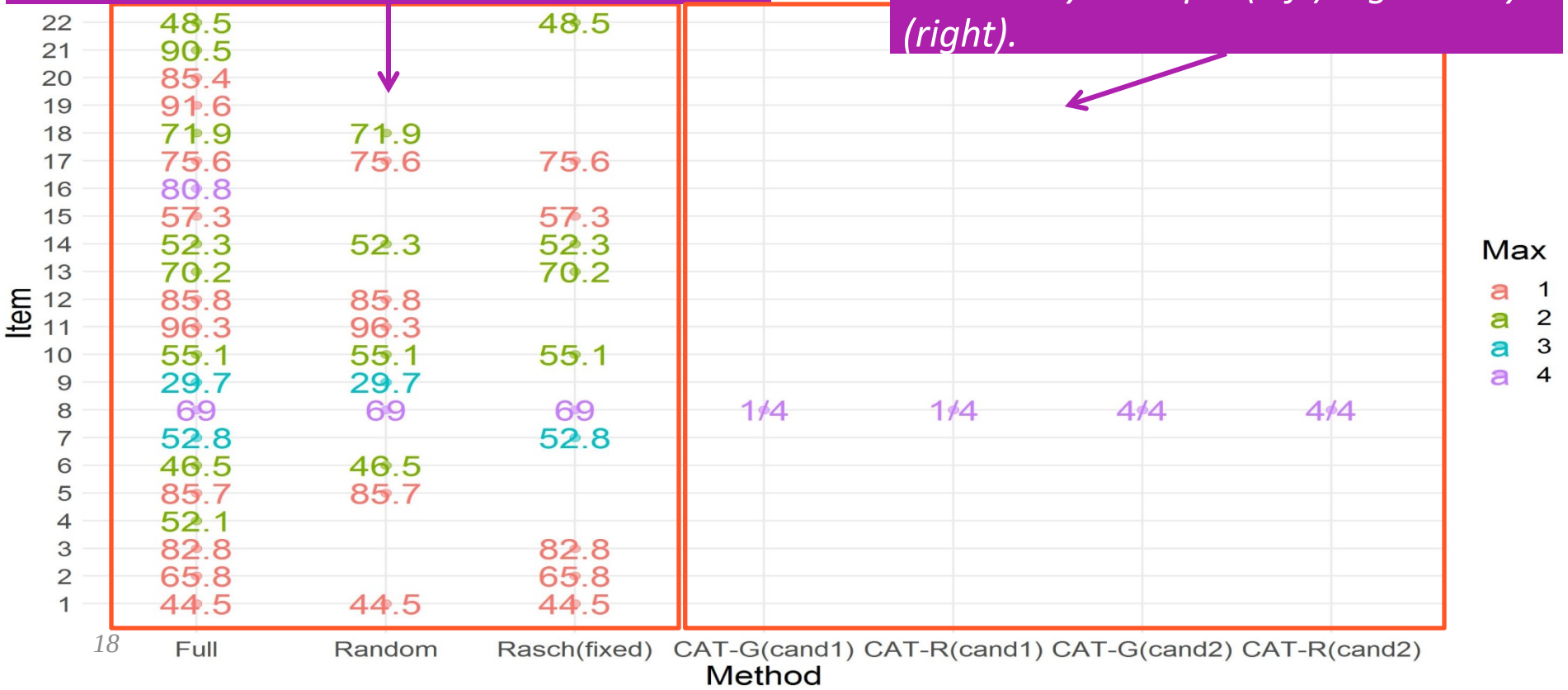
Each methods selects a 20 mark test consisting of

- One 4-mark item (out of 2)
- One 3-mark item (out of 2)
- Four 2-mark items (out of 8)
- Five 1-mark items (out of 10)

Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

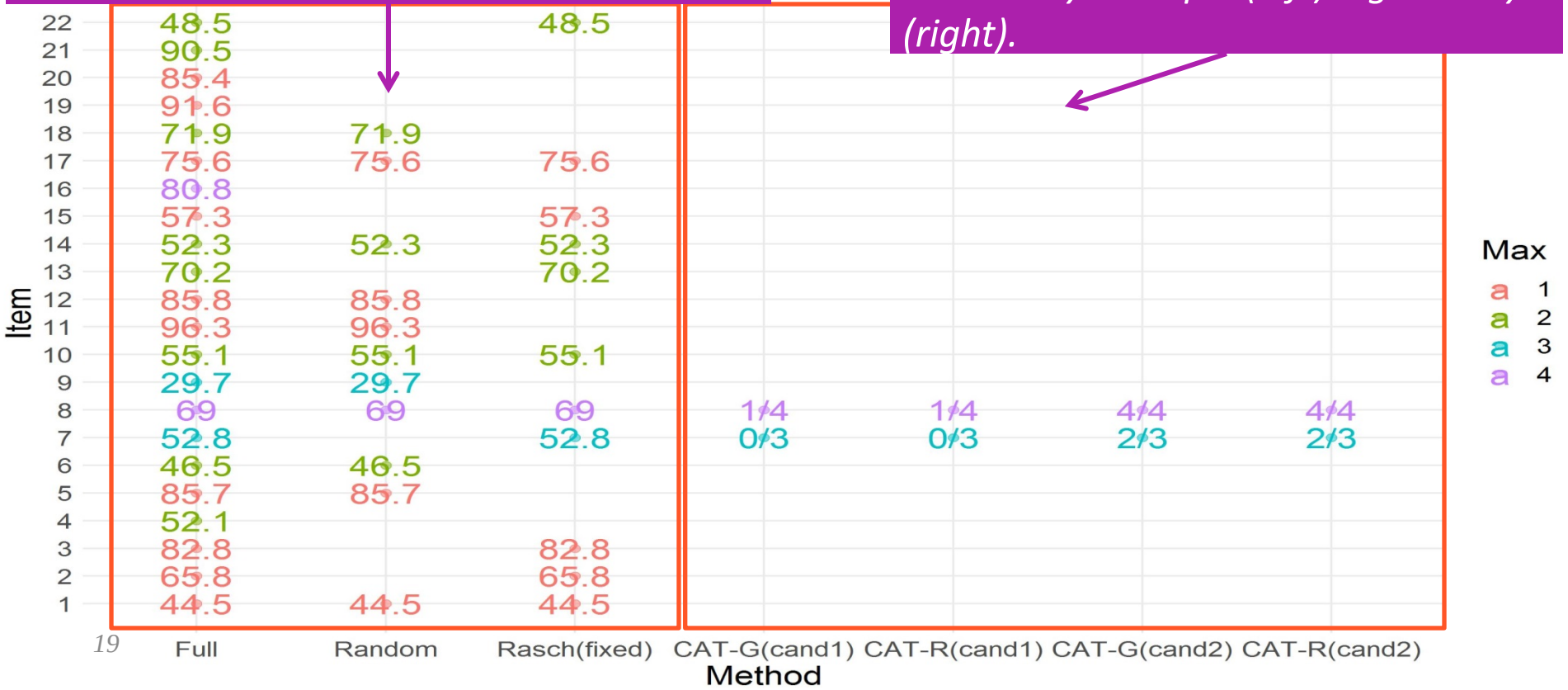
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

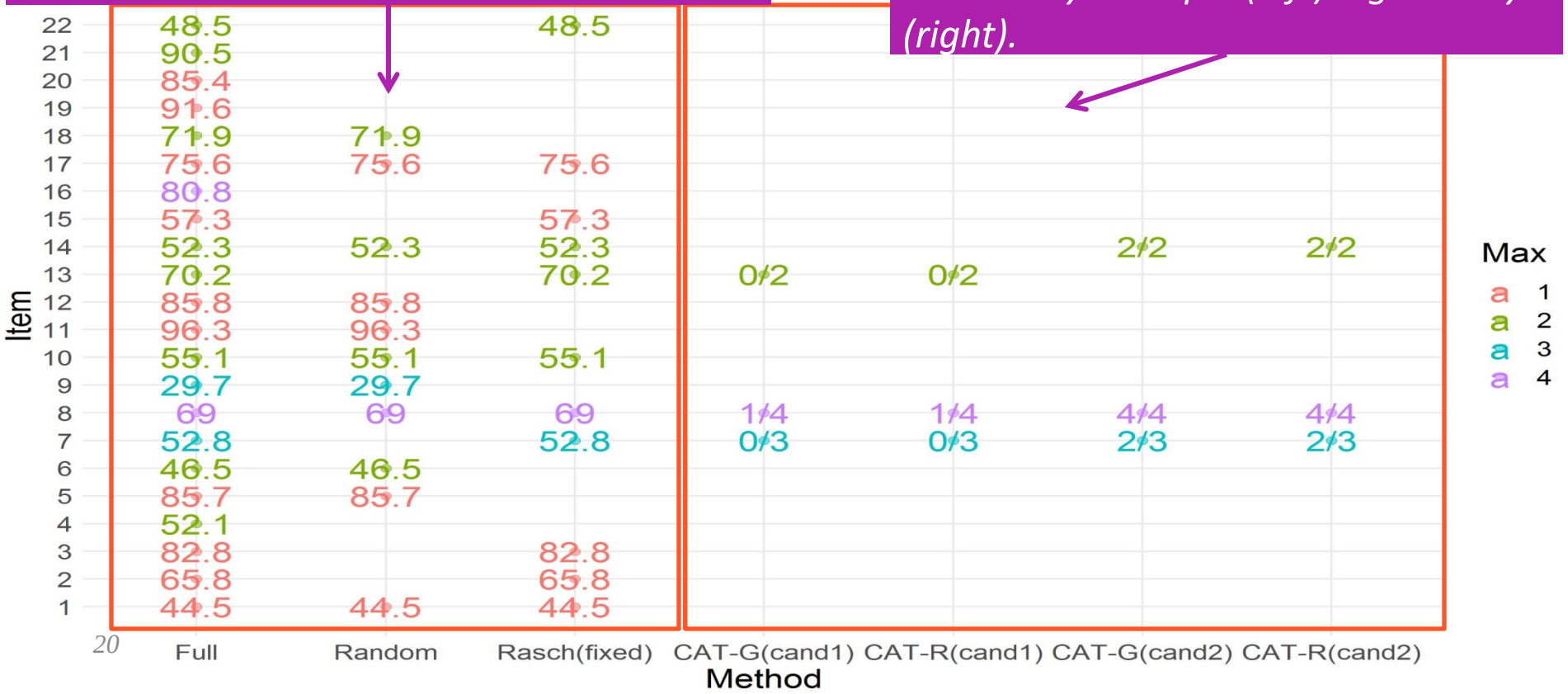
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

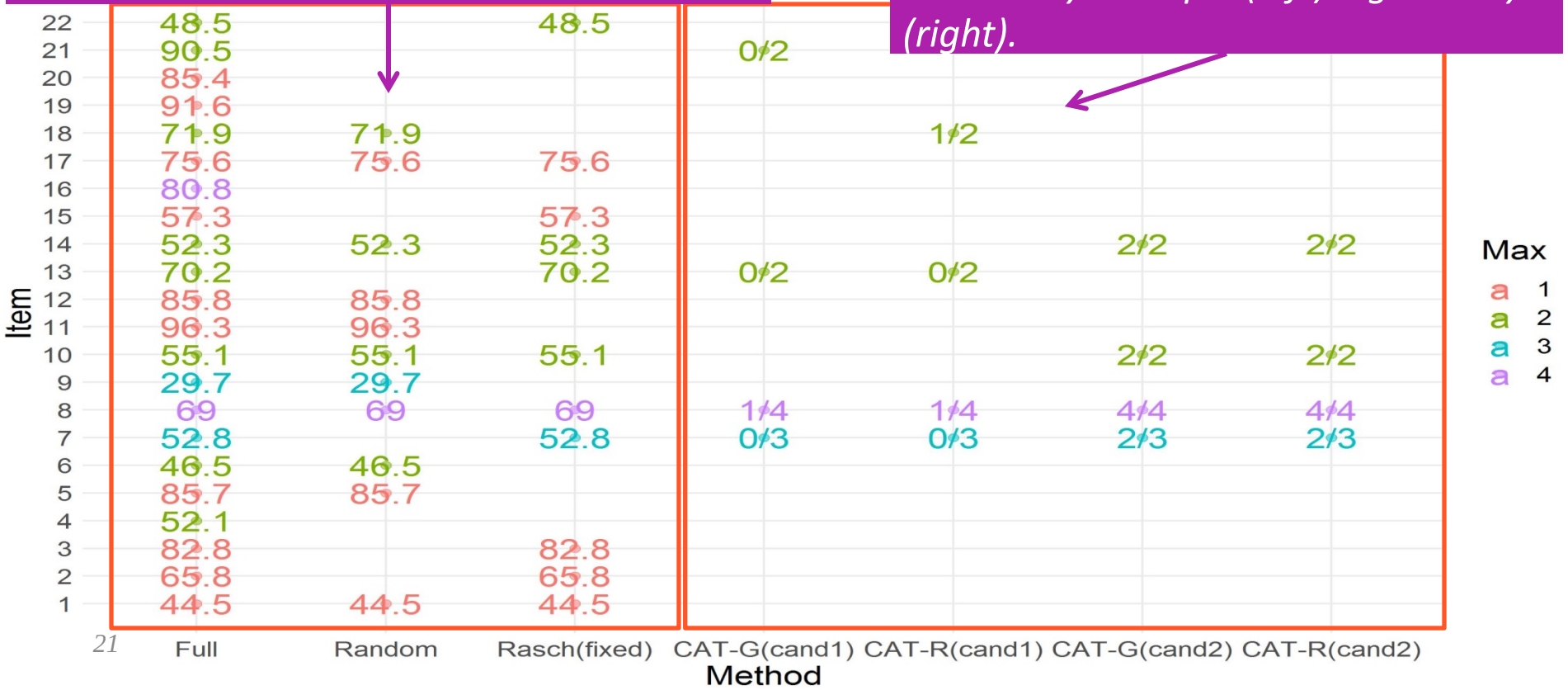
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

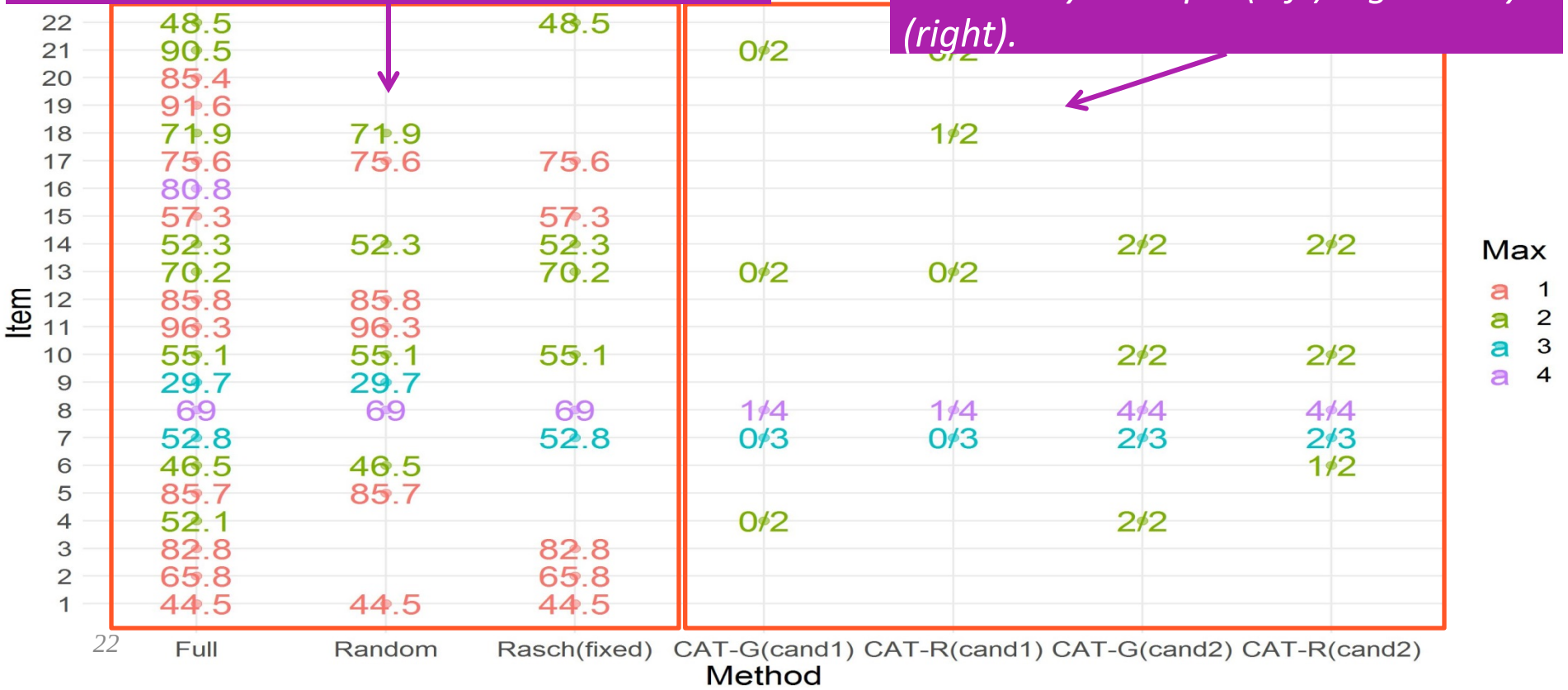
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

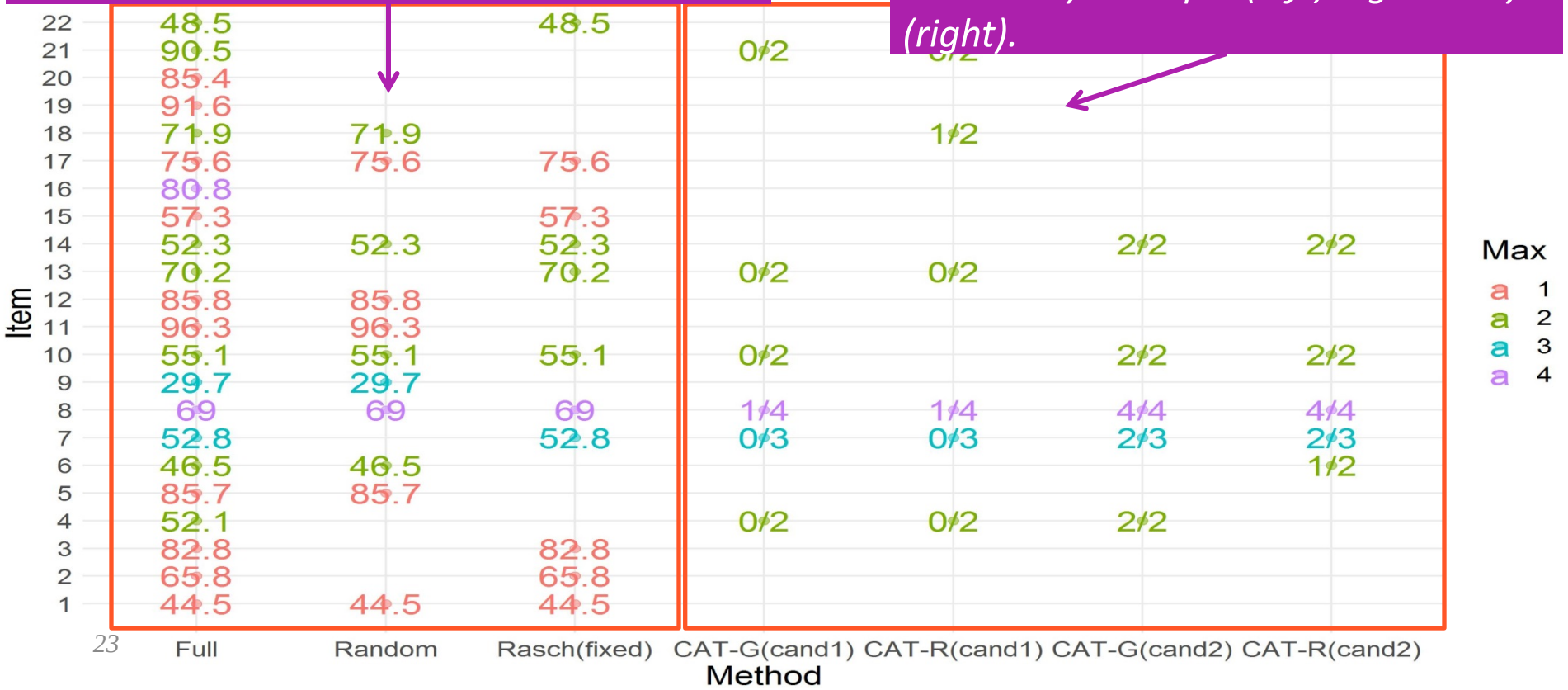
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

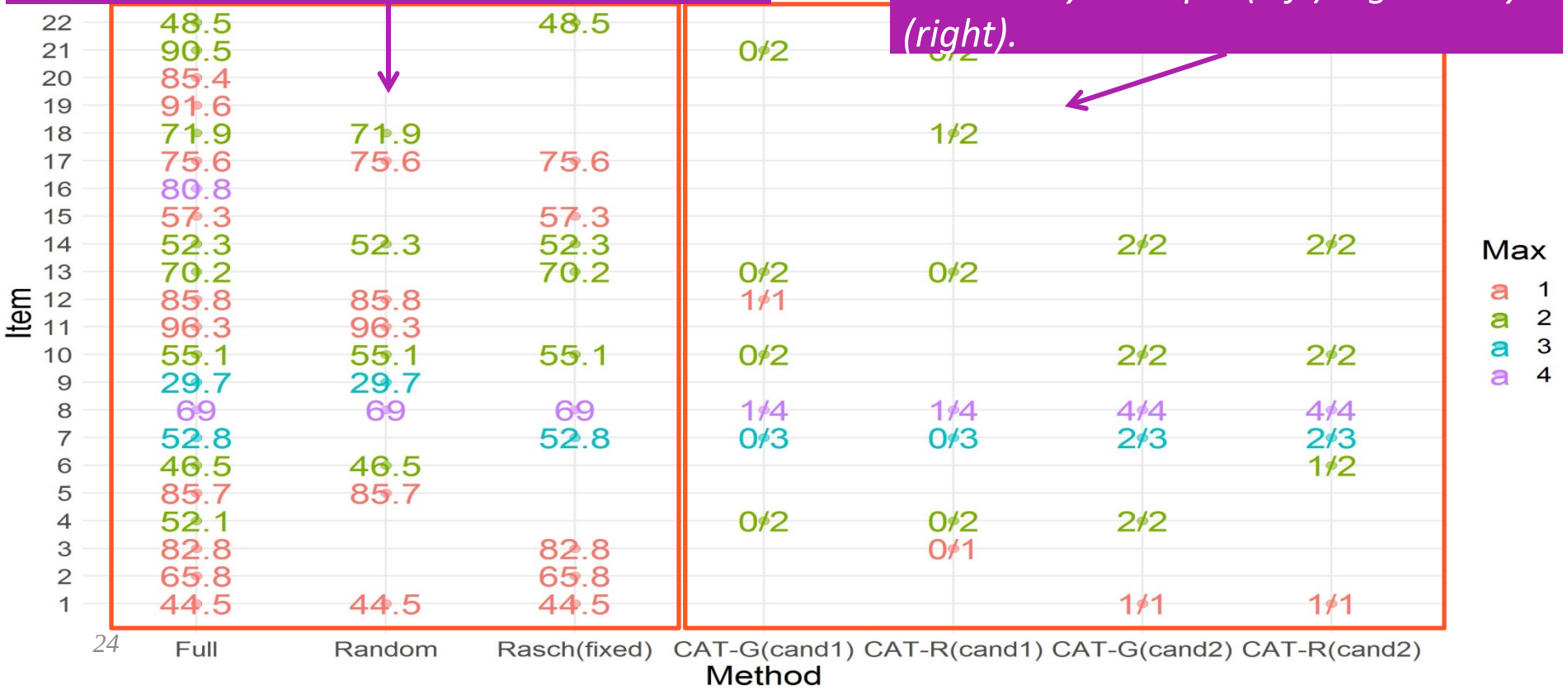
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

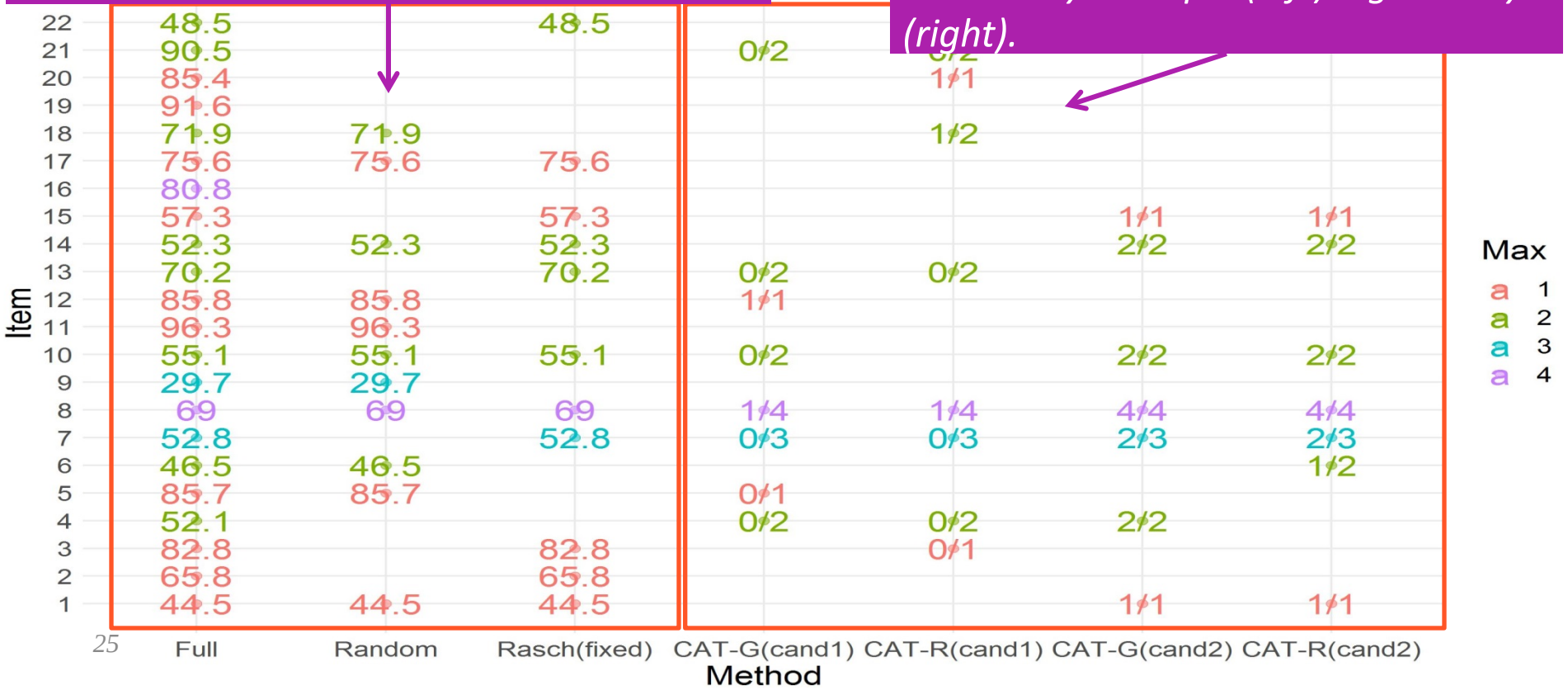
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

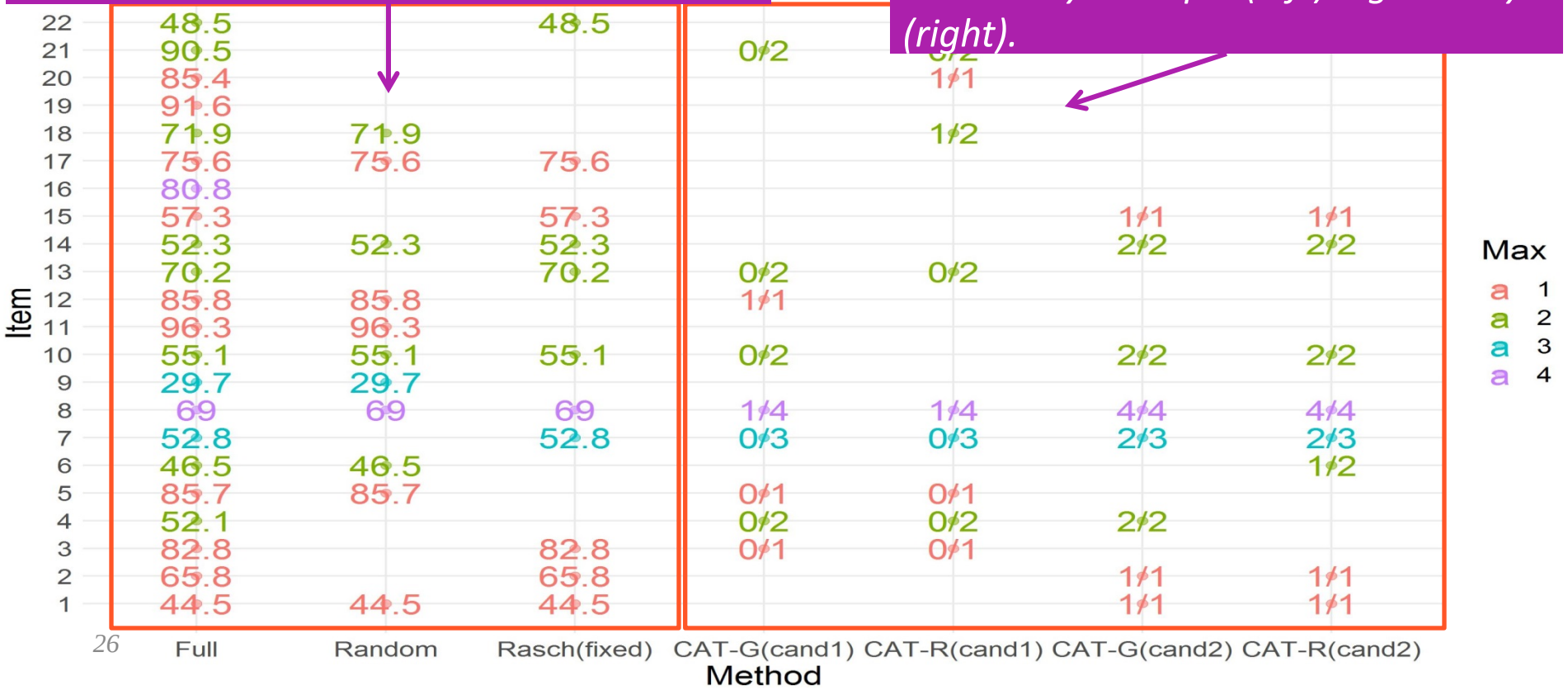
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

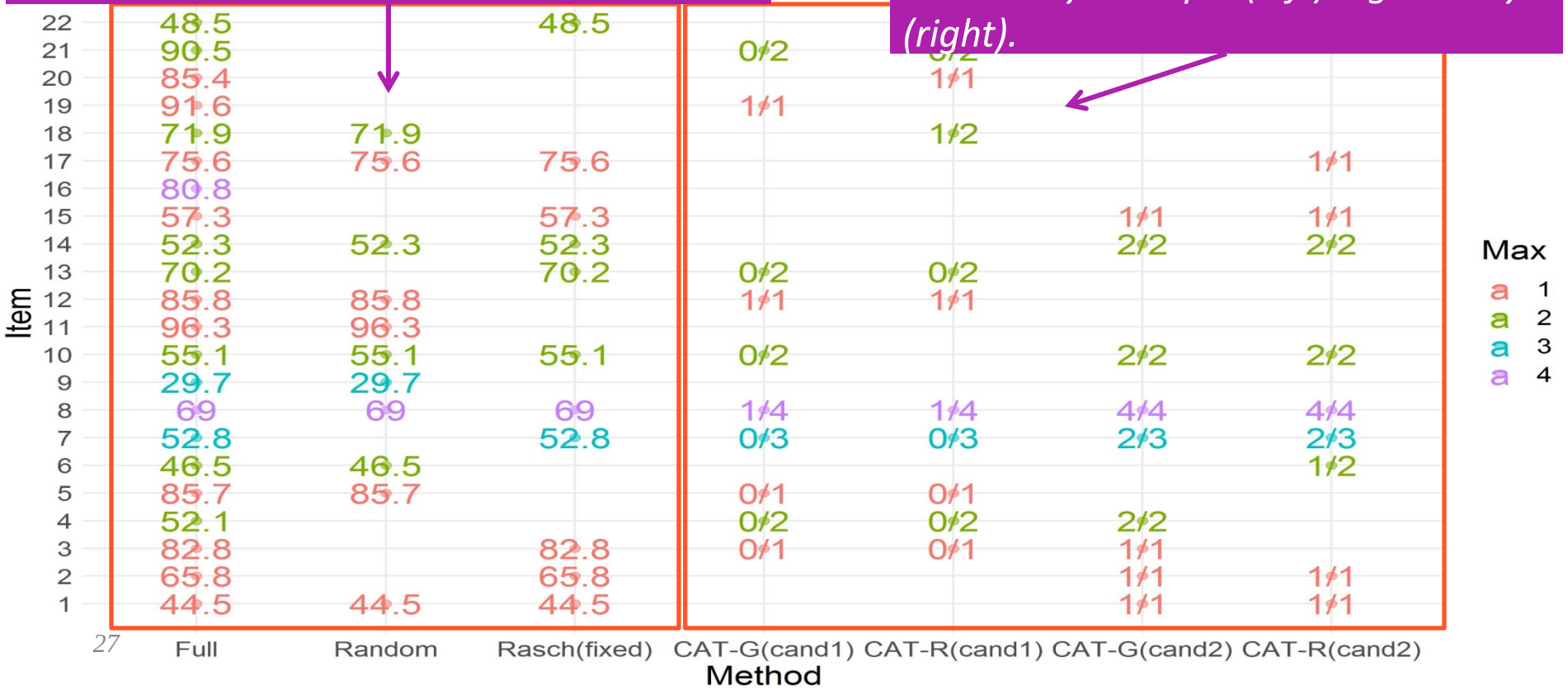
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

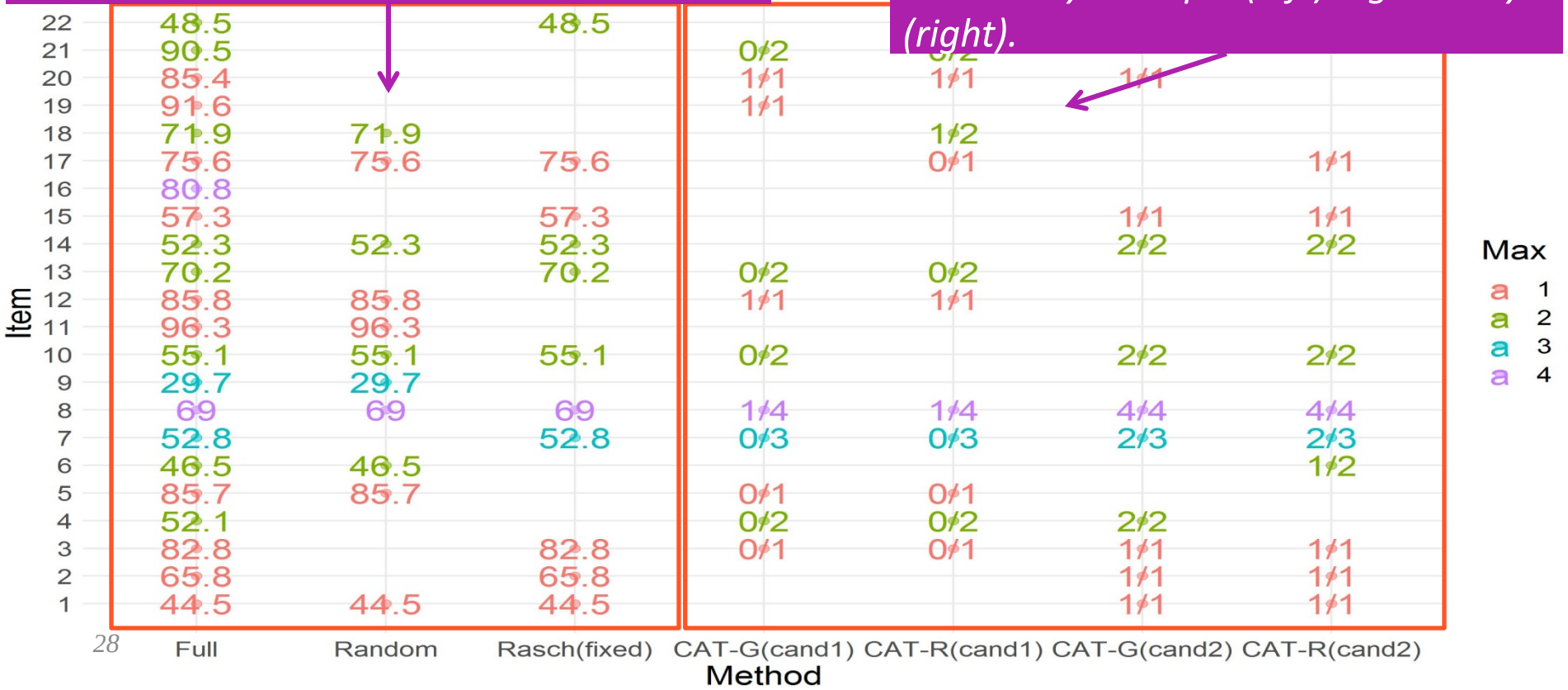
For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Example – 40 mark Biology Test

For fixed format tests the facilities of the chosen items are shown

For CAT selected items vary across pupils with later items chosen by earlier performance (marks shown). *Low ability example (left) high ability (right).*



Experimental method (evaluation of test forms)

1) Calculate scores for each pupil for each test form

- Simple sum-scores for fixed format tests
- IRT-scores (EAP ability estimates) for pseudo adaptive test

2) Calculate reliabilities

- Sum-score reliabilities estimated using IRT(GRM)
- IRT-score reliabilities estimated based on model used to generate them

3) Calculate predictive value

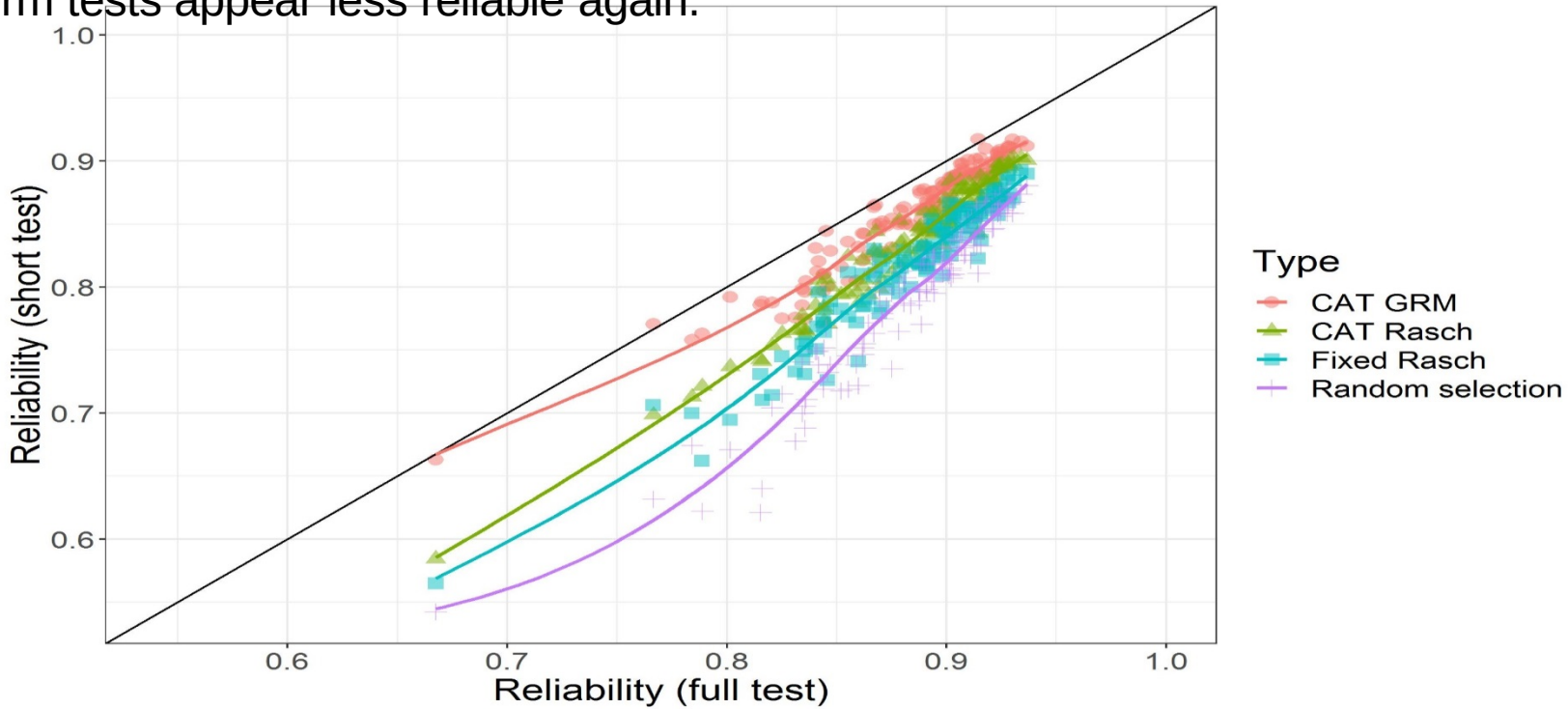
- Correlation with achievement on other assessments
    - ISAWG (roughly ≡ average achievement on other tests taken that month)

Results: Reliability

Pseudo-CAT using GRM to choose items appears *far superior*.

Pseudo-CAT using Rasch model appears less effective.

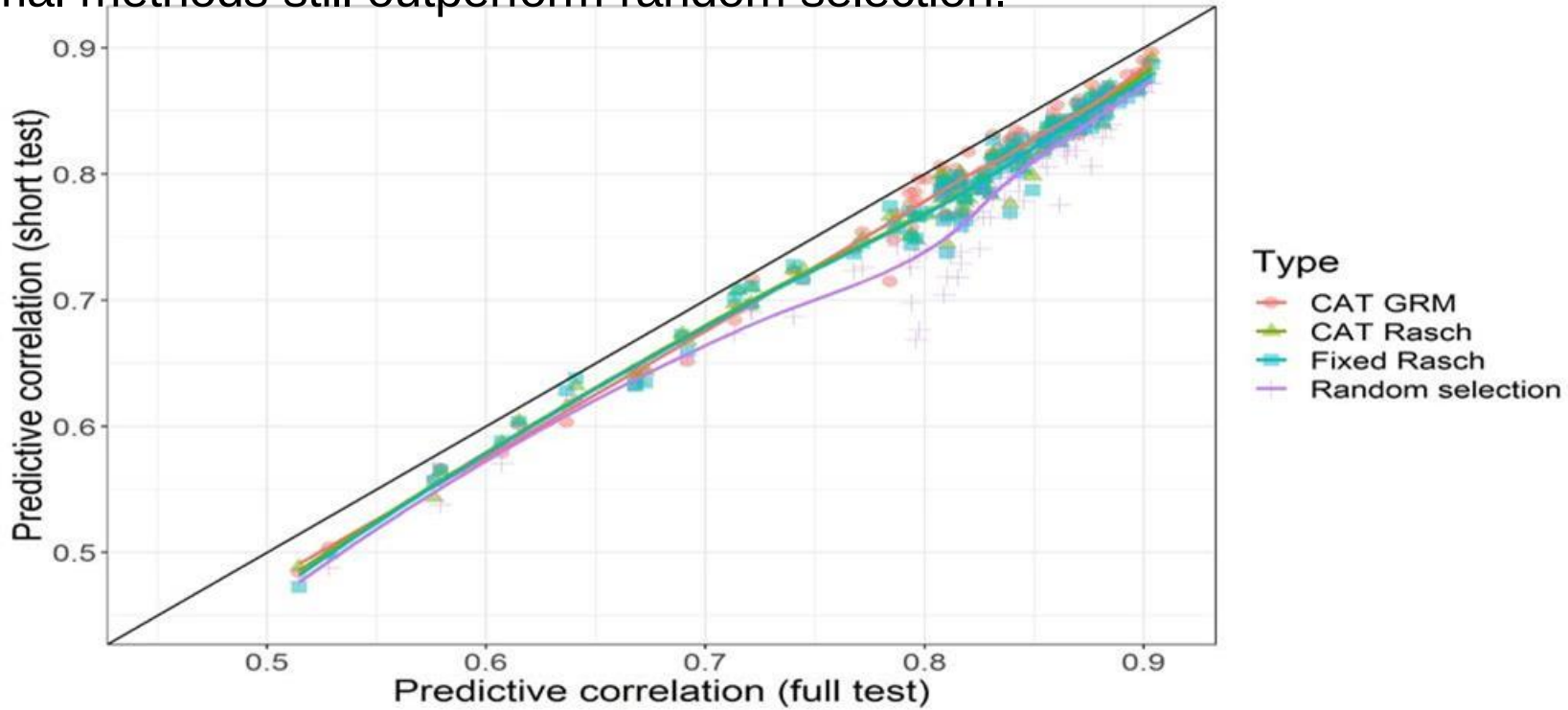
Fixed form tests appear less reliable again.



Results: Predictive value

Advantage of pseudo-CAT approaches over fixed format selection based on Rasch almost entirely disappears.

All optimal methods still outperform random selection.

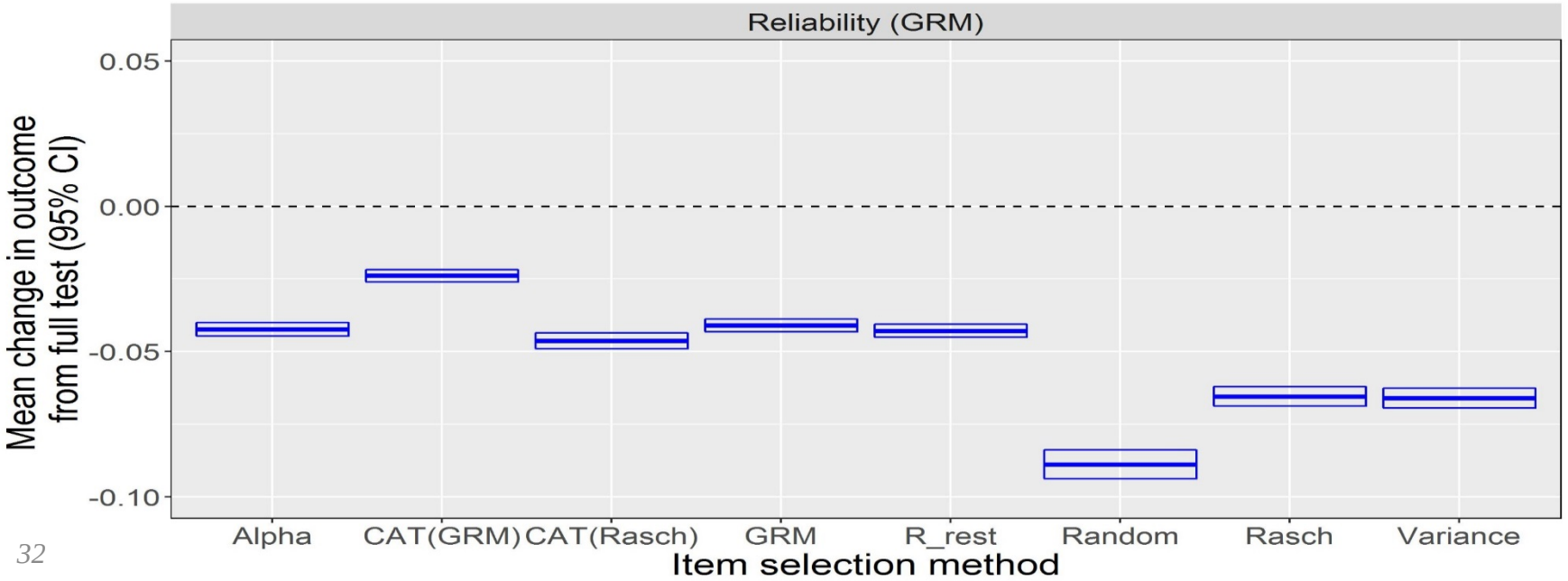


Results: Reliability

Zoom in to look at mean change in reliability from full-length test.

Pseudo-CAT using GRM to choose model appears *far superior*.

Pseudo-CAT using Rasch model appears less effective

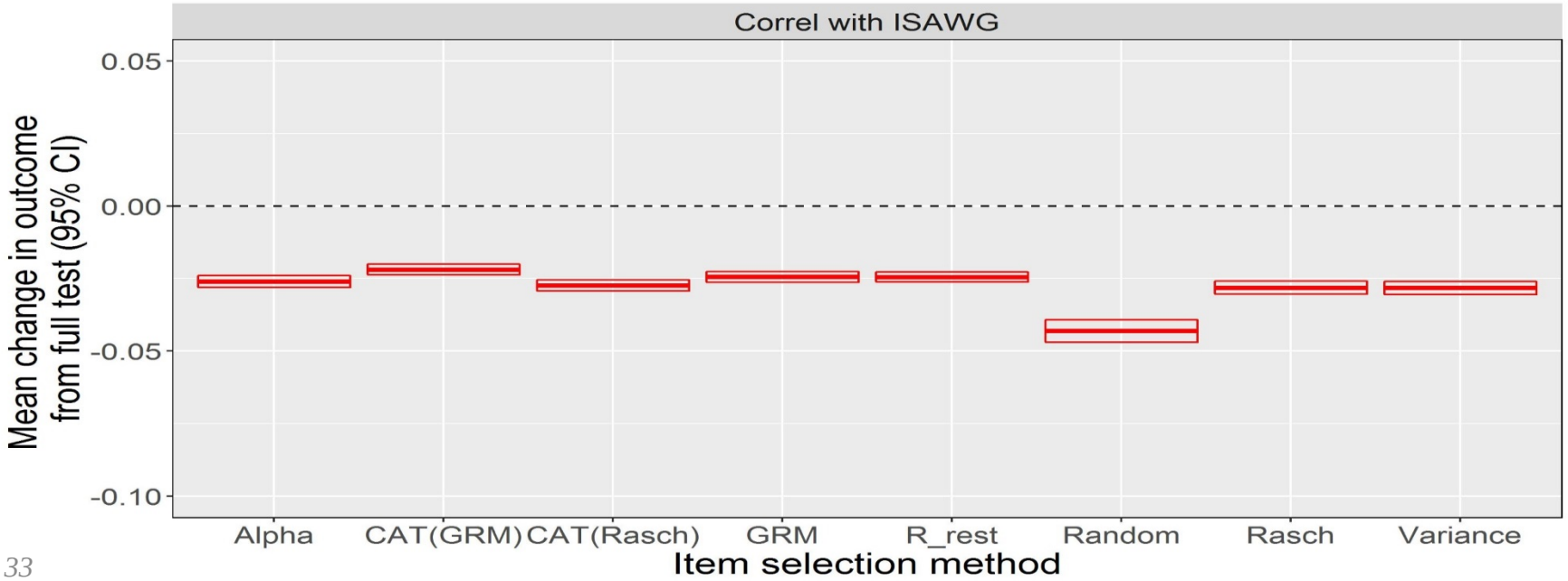


Results: Predictive value

Advantage of GRM (over classical) now almost invisible.

Rasch delivers most of the gain achieved by GRM.

Advantage of CAT itself over fixed tests seems fairly marginal.



Summary

- (Pseudo) adaptive testing does lead to better reliability than fixed format tests
- But differences in predictive value are very slight
    - Not in line with expectations given reliability differences
- Issue (of over claiming about performance improvements) is worst if non-Rasch models used
    - Are items that measure distinct skills (from rest of test) confused with low quality items?
    - Could (non-Rasch) IRT approaches make it less likely that tests include range of skills
    - Does the fact that, in adaptive testing, different students do different items matter more than we like to think?

Discussion

- Improving reliability usually thought of as reducing the influence of random error
    - Almost always conceptualised this way in simulation studies
    - Ought to also improve predictive value → **Not true in practice**
    - Not achieved by switch from sum-score to IRT scoring
- Be careful about making decisions based only on simulations

    → May lead to tests being too short