Optimal pretesting of questions for Swedish national tests in school

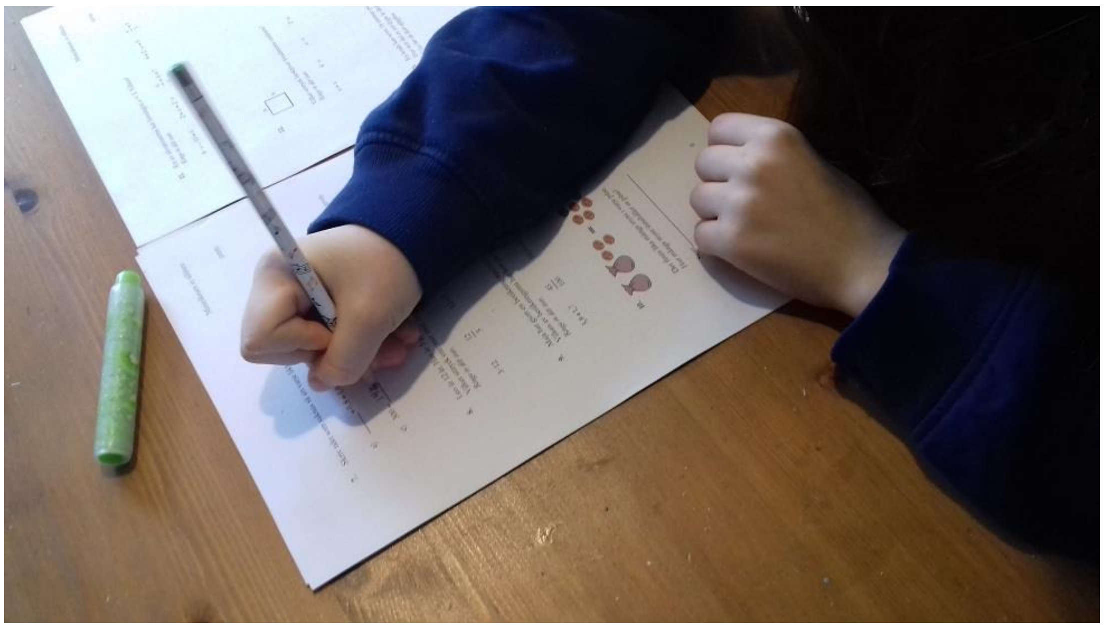
Frank Miller, Stockholm and Linköping University

Joint work with Ellinor Fackle-Fornius

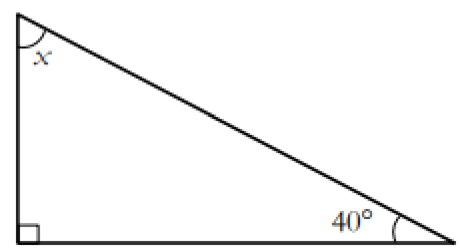
COMPSTAT, Bologna, 2022

National tests in Swedish schools

- National tests are conducted in Grade 3, 6, 9, and ~12
- Here: Mathematics test in Grade 6



d) $\frac{56}{8} = \_\_\_\_\_\_ - 10$



- Voluntary pupils pretest question (“items”) before use

Tests

Pretesting of many new items in Spring 2021

Selection of 85 items

Final pretesting May/June 2022

Pupils

National tests Spring 2022

Now: Optimize design for this pretesting

Pretesting of items

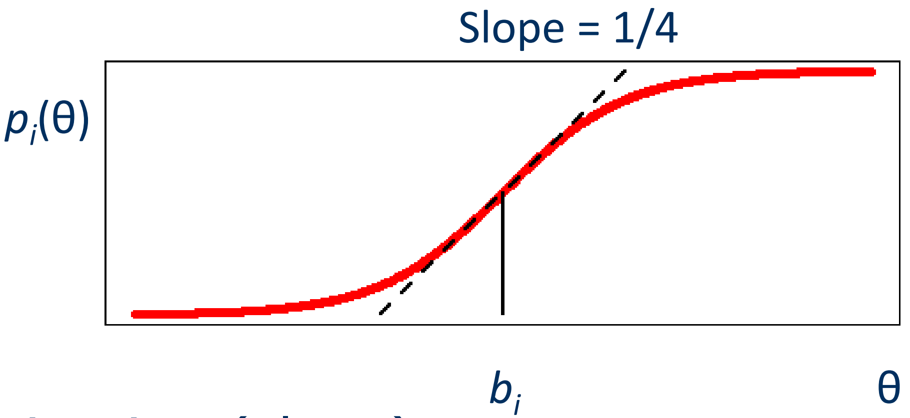
- One important reason to pretest is to estimate item characteristics like difficulty
- Usually, new items are randomly allocated to pupils for pretesting
- Can we improve precision of estimates when we **allocate based on the ability of pupils**?

Model: 1-parameter logistic (1PL, Rasch model)

- Probability to answer item *i* correctly (*i=1,…,n*):

$$p_i(\theta) = P(Y = 1|\theta, b_i) = \frac{1}{1 + \exp\{-(\theta - b_i)\}}$$

- $\theta \in \mathbb{R}$ pupil's ability
- b_i item difficulty

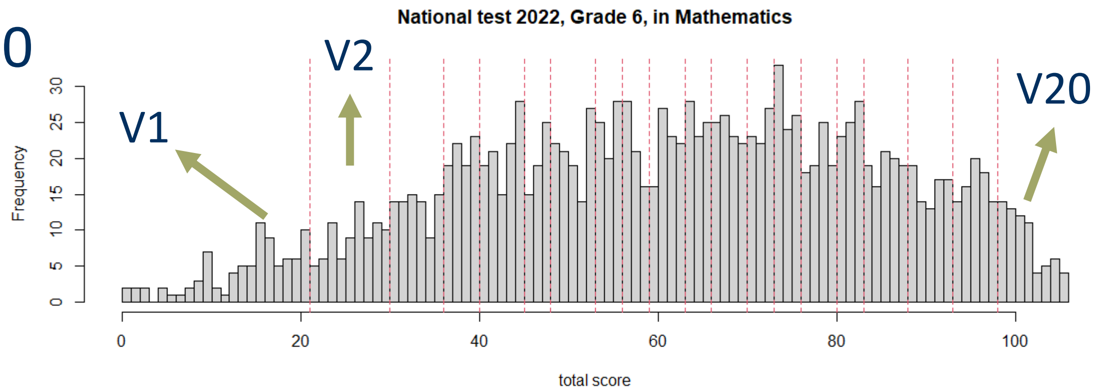


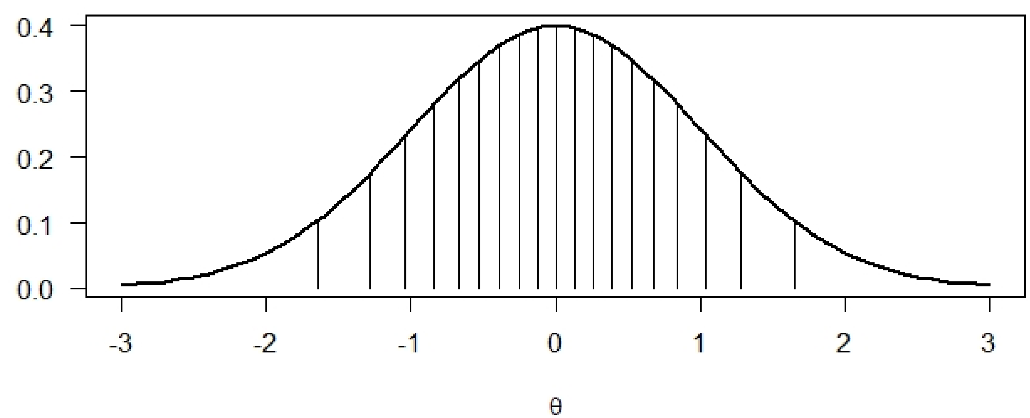
- 2PL model: a_i item discrimination (slope)
- 3PL model: c_i guessing parameter (lower asymptote)

Pupils' results in national test and versions

Based on their results in national test, ~1600 participating pupils were allocated to 20 versions:

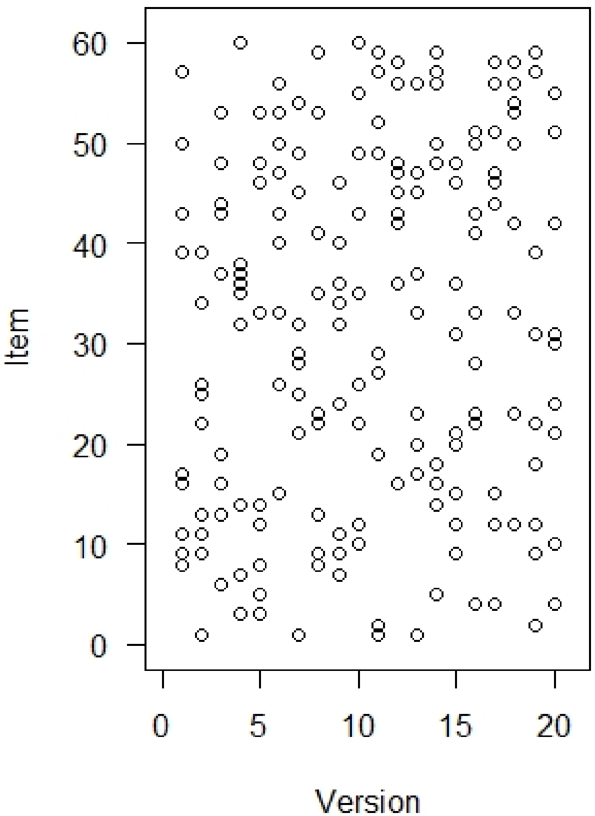
- 5% pupils with lowest result to V1,
- next 5% to V2,
- and so on,
- 5% with highest results to V20





What is a design here?

	V 1	V 2	V 3	…	…	V 20
I 60	0	0	0			0
I 59	0	0	0			0
I 58	0	0	0			0
I 57	1	0	0			0
…	…	…	…			…
I 6	0	0	1			0
I 5	0	0	0			0
I 4	0	0	0			1
I 3	0	0	0			0
I 2	0	0	0			0
I 1	0	1	0			0



Uncertainty of estimates (example: 1PL model)

OPTIC

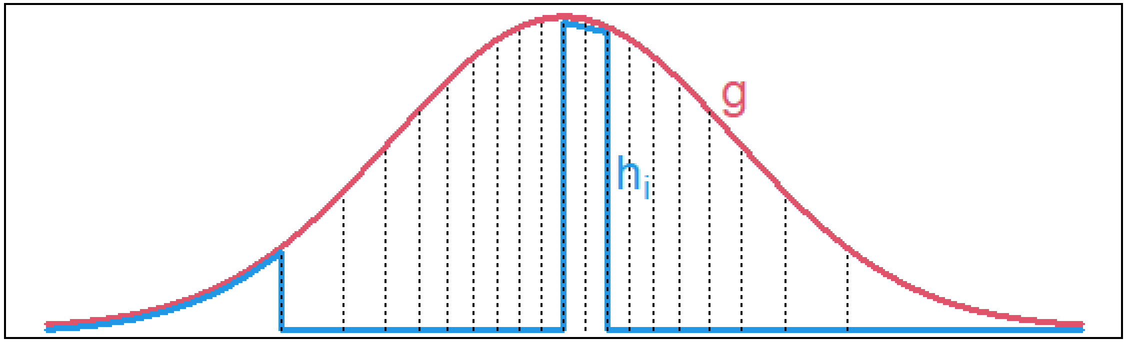
Optimal Calibration of Questions in Computerized Achievement Tests

Stockholms universitet

- Variance of the estimate for item difficulty b_i is in the 1PL model inversely proportional to information:

$$M_i = \int p_i(\theta)\big(1 - p_i(\theta)\big)h_i(\theta)d\theta$$

where $h_i(\theta)$ is sub-population allocated to item *i*



g=population of pupils

- Approach described by Ul Hassan and Miller (2019); based on finite population sampling (Wynn, 1982)

Uncertainty of estimates (example: 1PL model)

$$M_i = \int p_i(\theta)\big(1 - p_i(\theta)\big)h_i(\theta)d\theta$$

- M_i depends on difficulty b_i
- Need some guess for b_i which we get from pretesting done in Spring 2021
- We have many items to be pretested
- D-optimal design: maximize $\prod M_i$

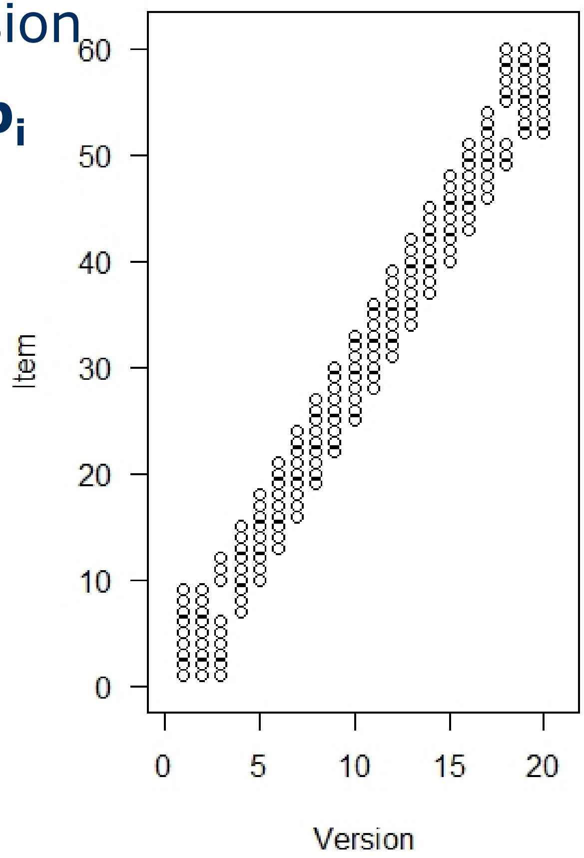
- For other models (2PL, 3PL, ...), variance of parameter estimates is characterized by a matrix M_i; D-optimal design: maximize $\prod \det(M_i)$

Optimal design for illustrating 1PL example

- 60 items, 20 versions, 9 items per version
- 1PL model for all 60 items; **difficulty b_i equidistant between -2 and 2**

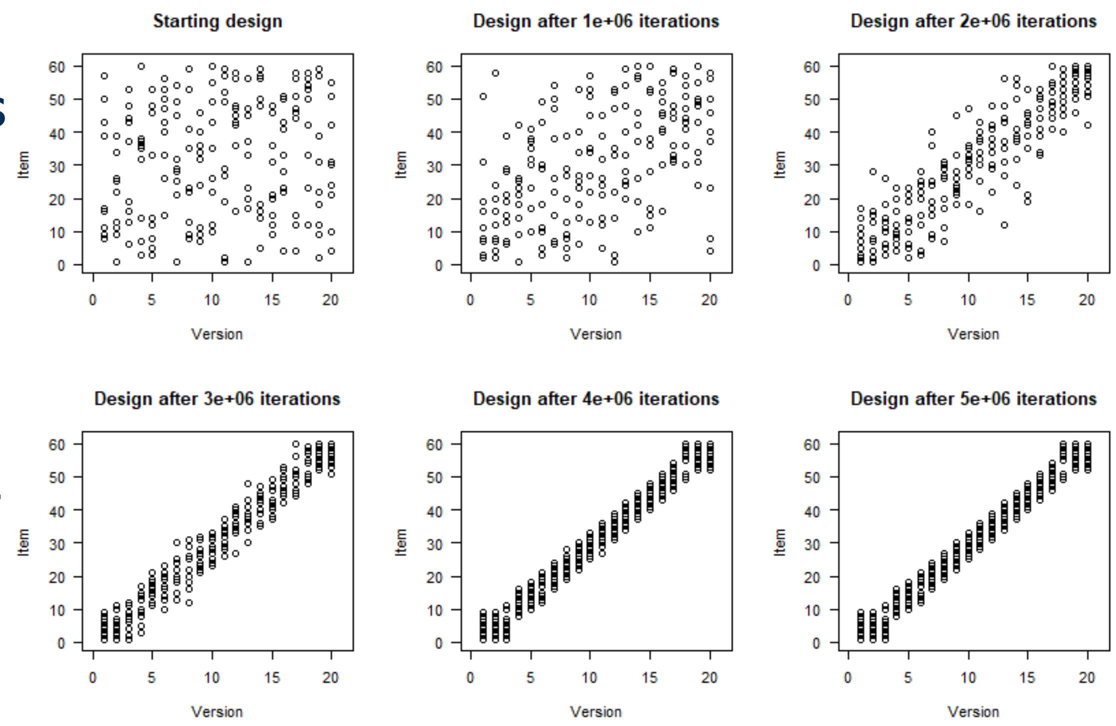
- D-optimal design:
    - Version 1/2: Item 1-9
    - Version 3: Item 1-6, 10-12
    - Version 4: Item 7-15

    …
    - Version 19/20: Item 52-60



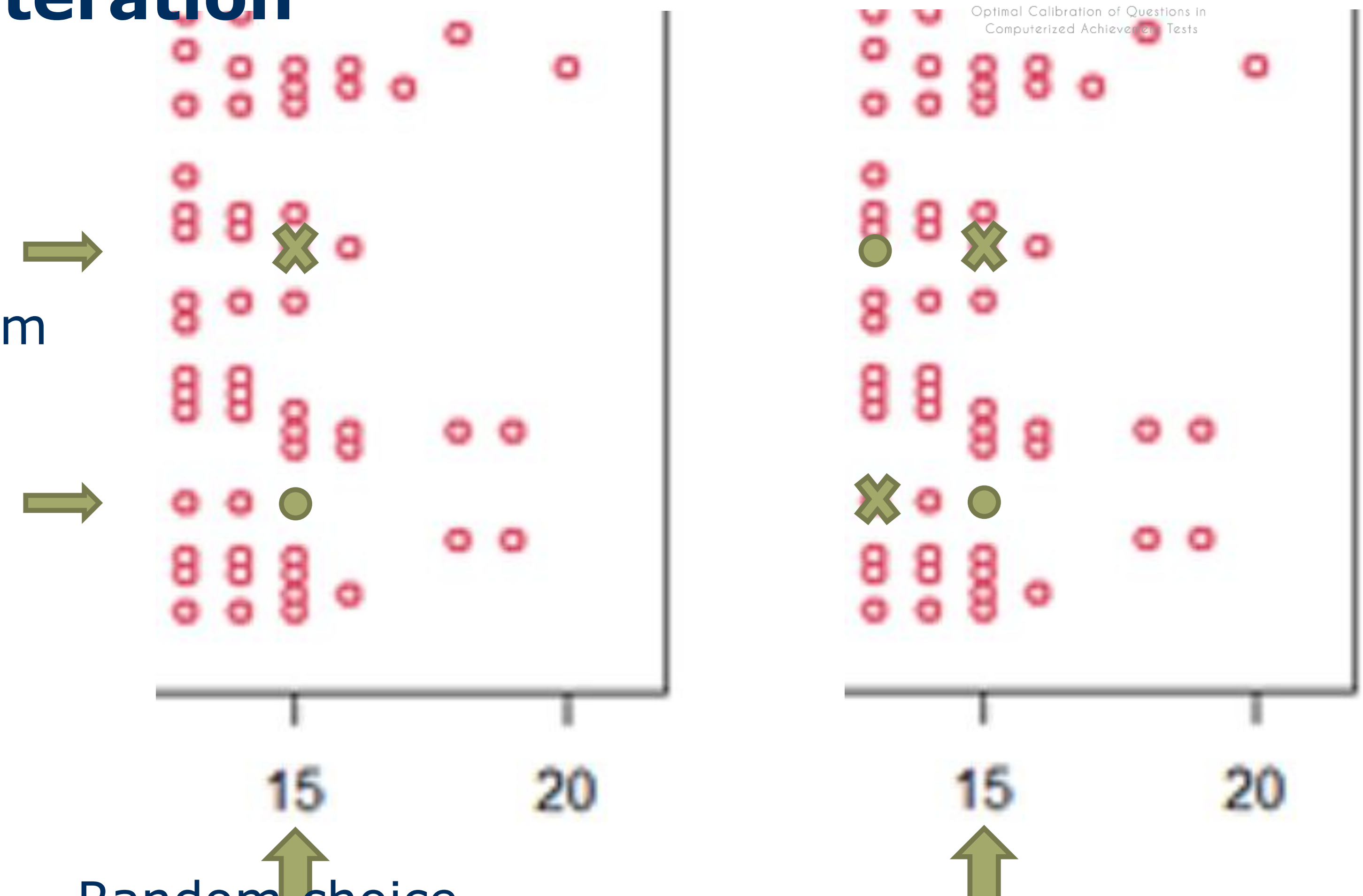
Run of the optimisation algorithm

- Simulated annealing algorithm used
- Random design-changes done in each of millions iterations
- A design change is accepted if it improves variance or – in early iterations with some probability – worsens variance a little



Design-changes in one iteration

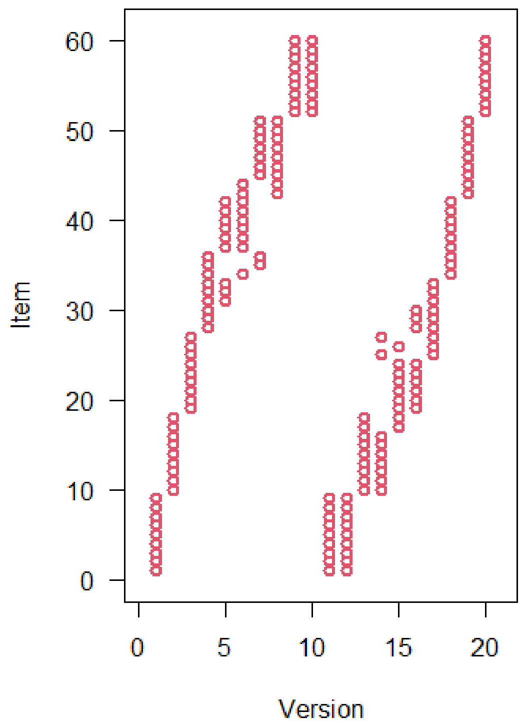
Random choice



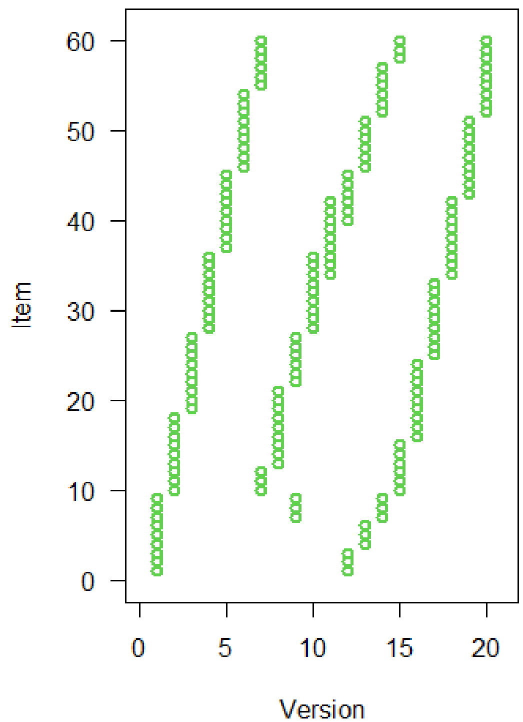
Random choice

Optimal design in examples for other models

2PL model

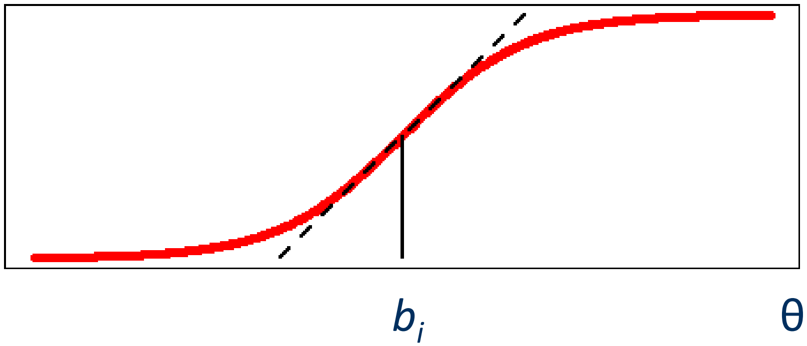


3PL model



Theoretical optimal designs for 1PL, 2PL, 3PL model

- 1PL: D-optimal to observe pupils with $\theta = b_i$



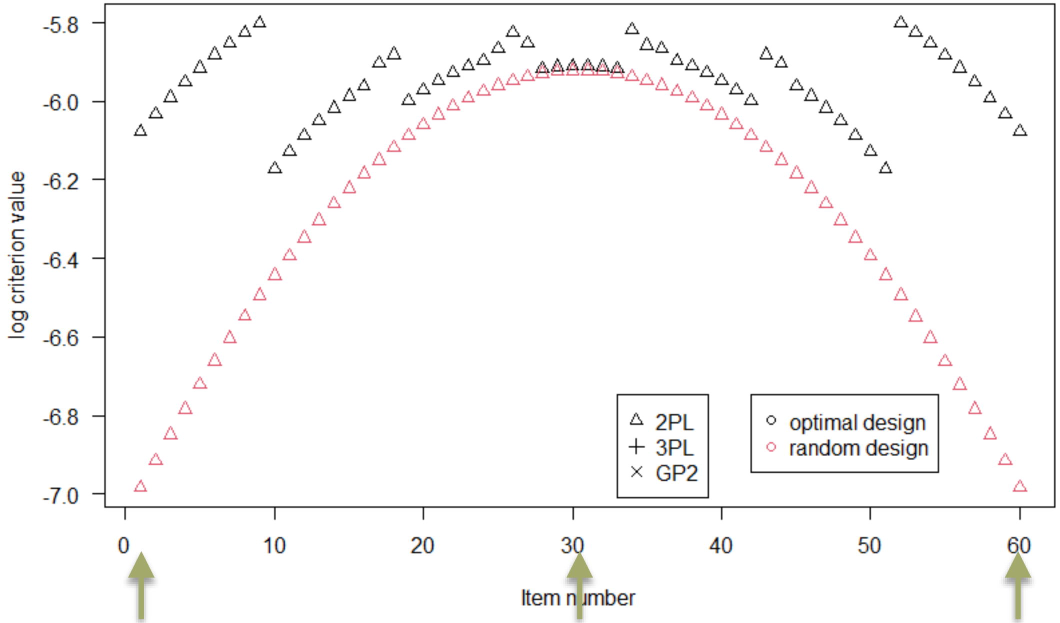
- 2PL: D-optimal to observe pupils with $\theta = b_i \pm const/a_i$ (1/2 in each design point)

  (see Abdelbasit and Plackett, 1983)

- 3PL: D-optimal to observe pupils with $\theta = -\infty$, $\theta = b_i \pm const/a_i$ (1/3 in each design point)

Optimal design for illustrating 2PL example

- Precision of each item after random design (red) and after D-optimal design (black)

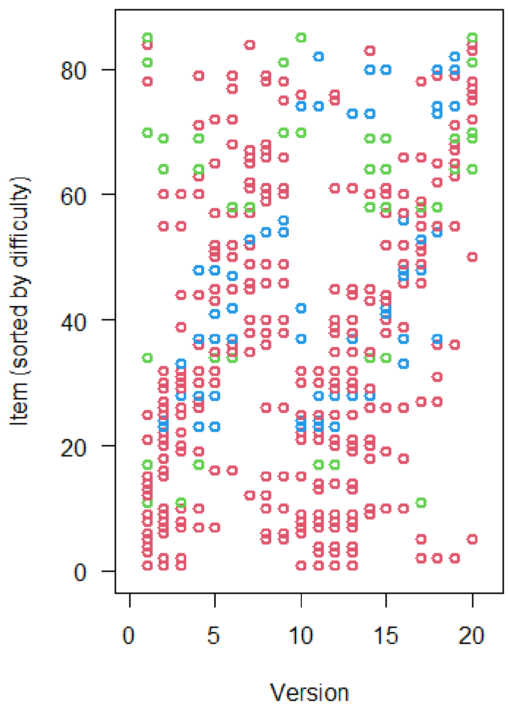


Easiest item, b=-2 | Middle difficulty | Most difficult item, b=2

Final pretesting in May/June 2022

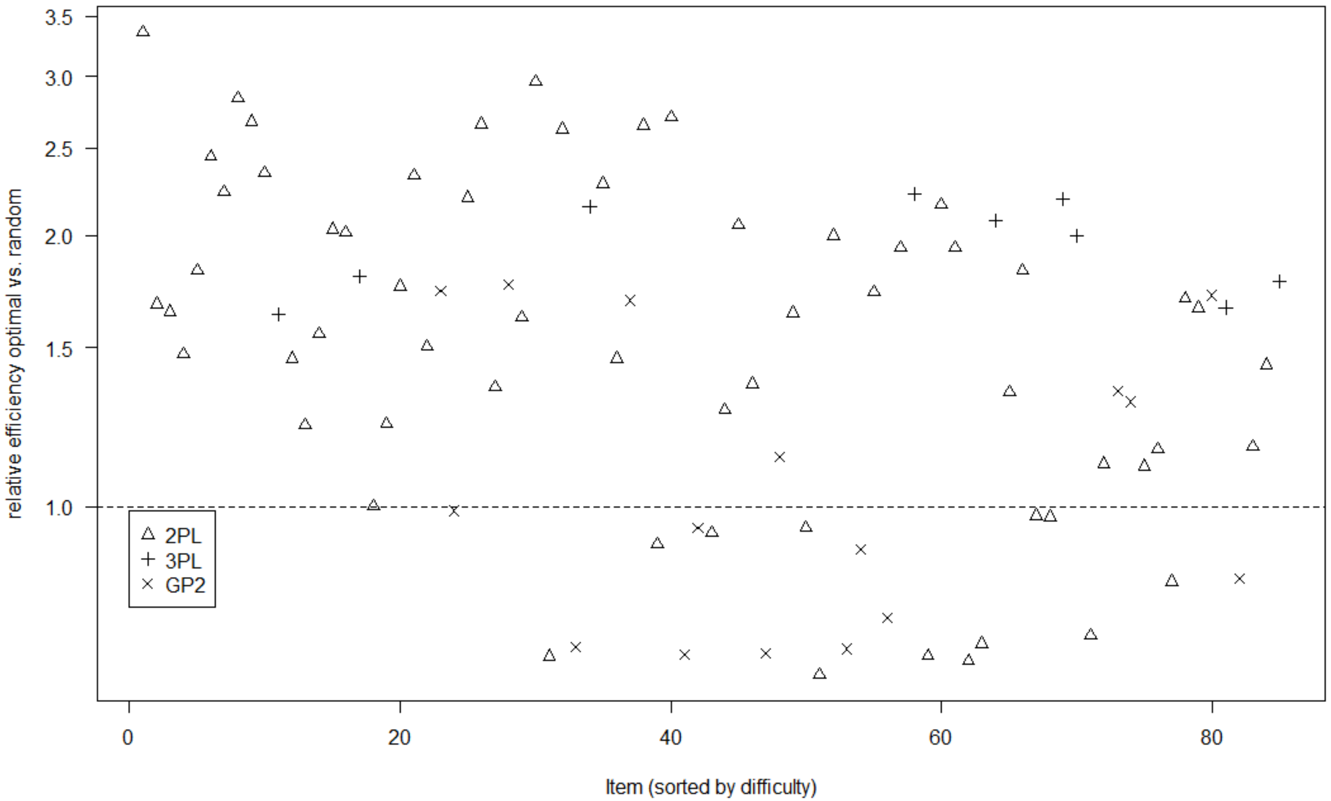
- Examples before were for illustration
- Reality is more complicated:
  - Pretesting items are of **mixed format;** several models are used (2PL, 3PL, Generalized Partial Credit Model GPCM)
  - Some items belong together (e.g. Problem 7a, 7b, 7c; called here “**item groups**”)
  - **Time needed** is different for items; time for each item was pre-estimated by experts; target time of 40 minutes for the test

Optimal design for final pretesting May/June 2022



(red=2PL, green=3PL, blue=GPCM item)

Precision of items after optimal vs. random design



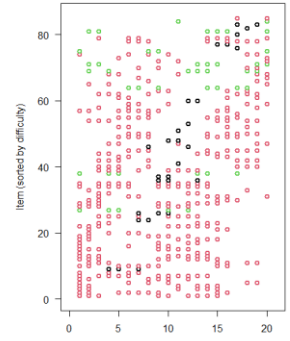
- Relative efficiency optimal vs. random design: 1.44

Tests

OPTIC

Optimal Calibration of Questions in Computerized Achievement Tests

Stockholms universitet



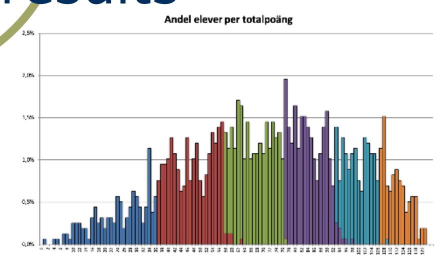
Pretesting of many new items in Spring 2021

Selection of 85 items and design choice

Final pretesting May/June 2022

Pupils' results

National tests Spring 2022



Currently, analysis is ongoing

Thank you!

Joint work

Ellinor Fackle-Fornius (Design elaboration),

Maria Nordlund, Anette Nydahl, Samuel Sollerman (Planning for implementation and conduct of the test)

Support

This work is supported by the Swedish Research Council

References

- Abdelbasit KM, Plackett RL (1983). Experimental design for binary data. *Journal of the American Statistical Association*, **78,** 90-98.
- Ul Hassan M, Miller F (2019). Optimal item calibration for computerized achievement tests. *Psychometrika,* **84**, 1101-1128.
- Wynn H (1982). Optimum submeasures with applications in finite population sampling. In *Statistical decision theory and related topics III*, pages 485-495. Academic Press, New York.