

Item Selection Methods Based on Multiple Objective Approaches for Classifying Respondents Into Multiple Levels

**Maaïke M. van Groen¹, Theo J. H. M. Eggen^{1,2},
and Bernard P. Veldkamp²**

IF: 1.155

Reporter: Yingshi Huang

Applied Psychological Measurement

XX(X) 1–14

© The Author(s) 2013

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621613509723

apm.sagepub.com



- Computerized adaptive test (CAT)
- Computerized classification testing (CCT)

How to assemble a test?

- **Objective:** select those items that provide the most accurate
- **Current methods:** based on one point on the scale and are



multiple cut scores

the peak of the information function to be located **at each of**
(information is gathered throughout a larger part of the ability scale)



need to optimize at a combination of cut scores

item exposure (Simpson & Hetter, 1985)

content control (Van der Linden, 2005)

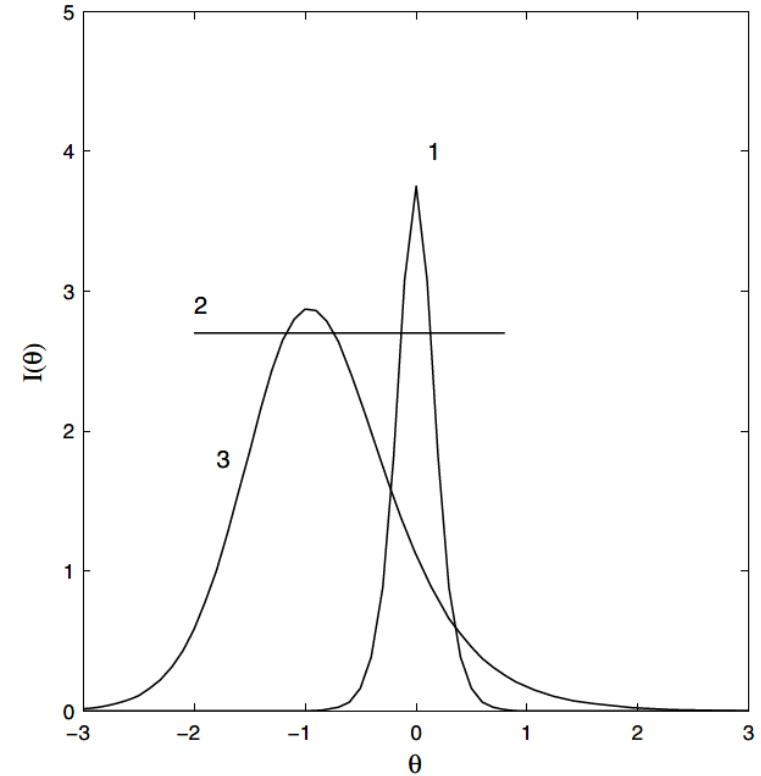
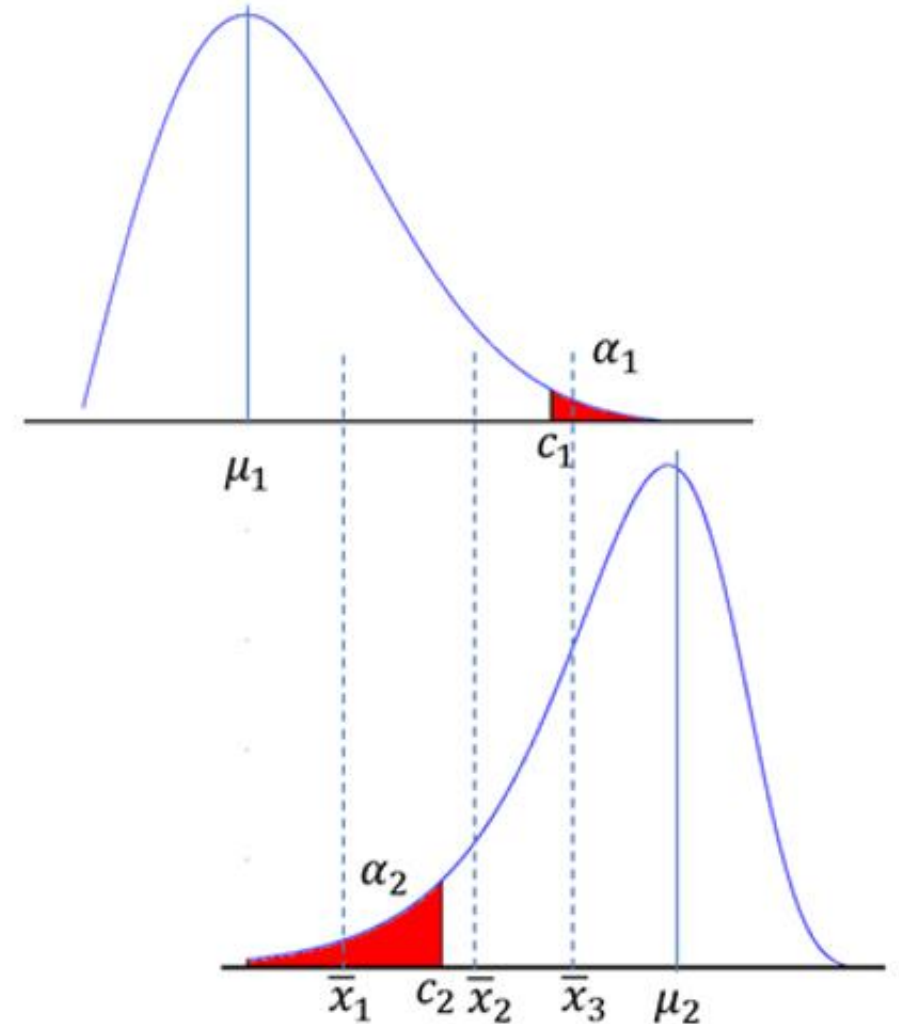
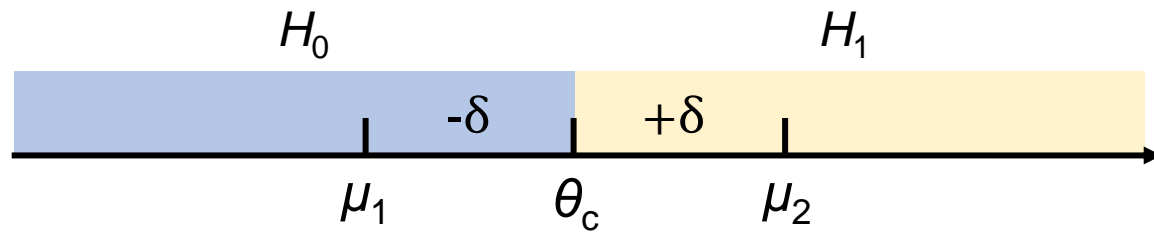


FIGURE 1.6. Examples of three possible targets for a test information function: (1) a test used for admission decisions with cutoff score $\theta = 0$; (2) a diagnostic test over lower range of abilities; and (3) a test with an information function that follows a given population distribution.

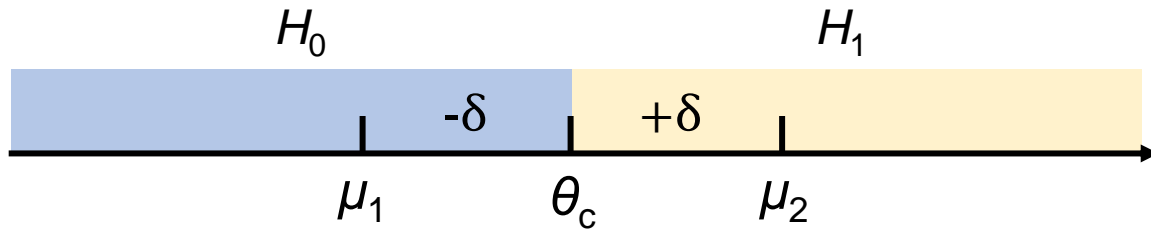
How to stop?

- The sequential probability ratio test (SPRT)



How to stop?

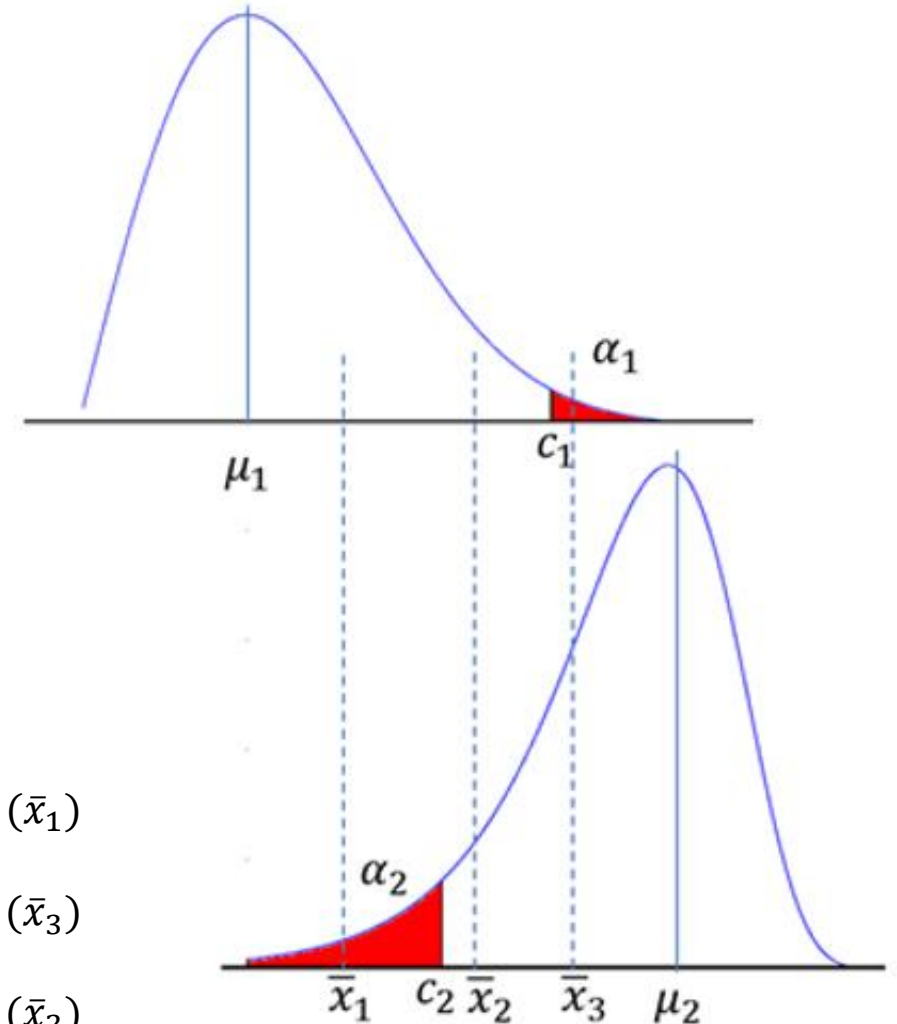
- The sequential probability ratio test (SPRT)



$$L(\theta; \mathbf{x}) = \prod_{i=1}^k p_i(\theta)^{x_i} [1 - p_i(\theta)]^{1-x_i}$$

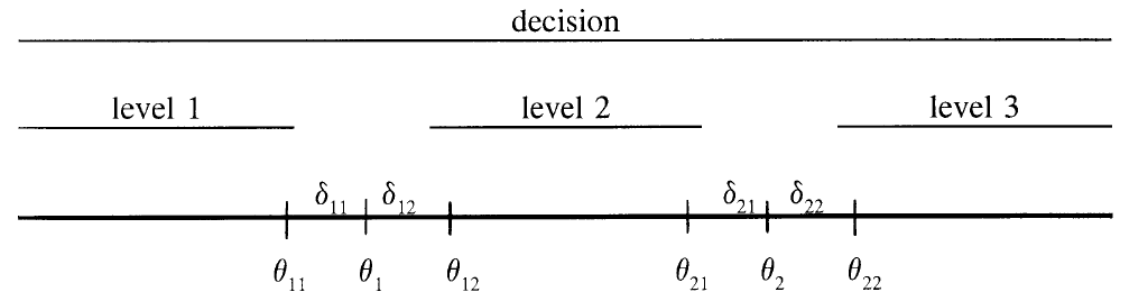
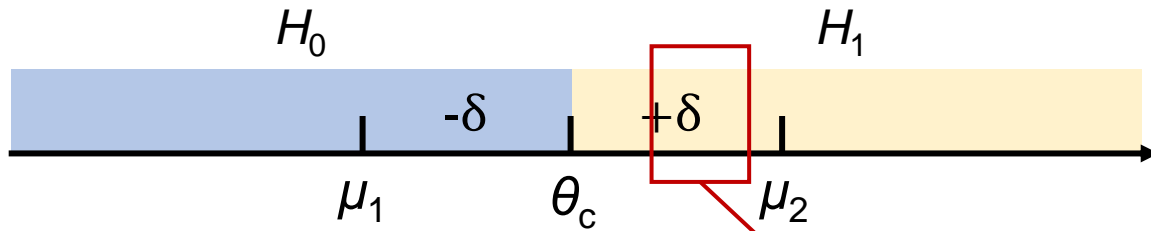
$$LR(\theta_c + \delta; \theta_c - \delta) = \frac{L(\theta_c + \delta; \mathbf{x})}{L(\theta_c - \delta; \mathbf{x})}$$

- ability below the cutting point if $LR(\theta_c + \delta; \theta_c - \delta) \leq \beta(1 - \alpha)$ (\bar{x}_1)
- ability above the cutting point if $LR(\theta_c + \delta; \theta_c - \delta) \geq (1 - \beta)\alpha$ (\bar{x}_3)
- administer another item if $\beta(1 - \alpha) < LR(\theta_c + \delta; \theta_c - \delta) < (1 - \beta)\alpha$ (\bar{x}_2)



How to stop?

- The sequential probability ratio test (SPRT)



$H0_1: \theta \leq \theta_{11}$ (Level 1) $H1_1: \theta \geq \theta_{12}$ (higher than 1);
 $H0_2: \theta \leq \theta_{21}$ (lower than 3) $H1_2: \theta \geq \theta_{22}$ (Level 3).

$$L(\theta; \mathbf{x}) = \prod_{i=1}^k p_i(\theta)^{x_i} [1 - p_i(\theta)]^{1-x_i}$$

$$LR(\theta_c + \delta; \theta_c - \delta) = \frac{L(\theta_c + \delta; \mathbf{x})}{L(\theta_c - \delta; \mathbf{x})}$$

shorter tests
less accurate decisions
overlapping

- ability below the cutting point if $LR(\theta_c + \delta; \theta_c - \delta) \leq \beta(1 - \alpha)$ (\bar{x}_1)
- ability above the cutting point if $LR(\theta_c + \delta; \theta_c - \delta) \geq (1 - \beta)\alpha$ (\bar{x}_3)
- administer another item if $\beta(1 - \alpha) < LR(\theta_c + \delta; \theta_c - \delta) < (1 - \beta)\alpha$ (\bar{x}_2)

- Choosing an item selection method in conjunction with the SPRT

Spray and Reckase (1994)

- maximizing information **at the cutting score**
- in shorter tests than does selecting items at the current ability estimate.

VS

Thompson (2009)

- however, concluded that
- this method is **not always the most efficient** option.

VS

Wouda and Eggen (2009)

- with two cutting points
- maximization **at the middle of the cutting points** resulted in the most accurate and the longest tests.

Multiple categories?



four **item selection methods** were developed for this study and were compared with current methods

investigate the effect of **the size of the indifference region**

consider **content and exposure control**

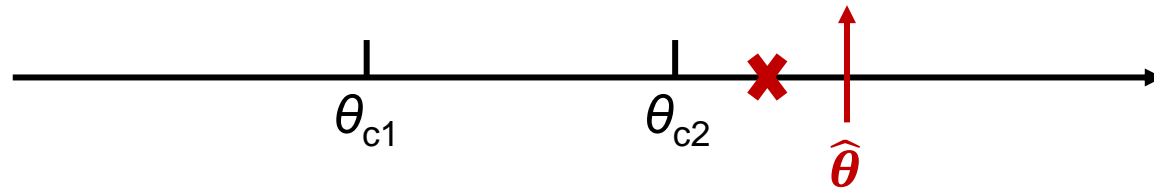
- Based on Fisher information

1. At the **current ability estimate**

$$\max_{i \in V_i} I_i(\theta) \quad I_i(\theta) = a_i^2 p_i(\theta) [1 - p_i(\theta)]$$

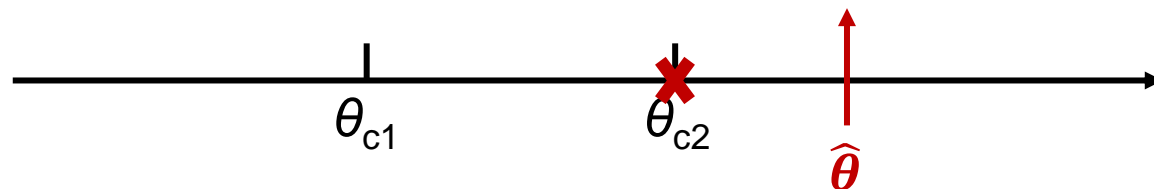
(V_i denotes the set of items still available for administration)

2. At **the middle of** the cutting points nearest to the current estimate



at **one point** on the latent scale
using the ability estimate

3. At the cutting point located **nearest** to the ability estimate



maximize information on all cutting points?

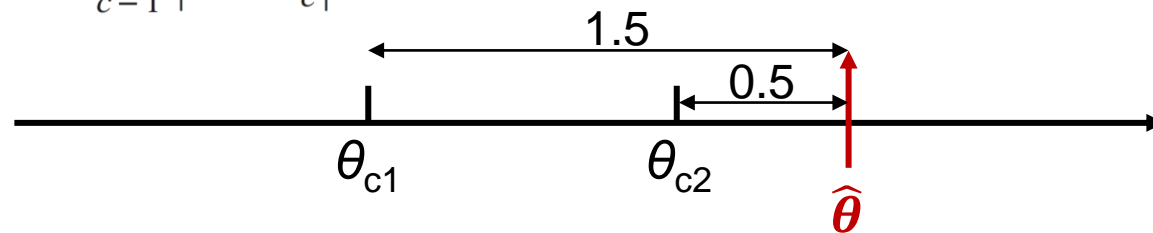


- Combine several objectives into one objective function

1. Weighting Methods

- the weight for a specific cutting point increases if the ability estimate is **closer to the cutting point**

$$\max \sum_{c=1}^C \frac{1}{|\hat{\theta} - \theta_c|} I_i(\theta_c), \text{ for } i \in V_i$$



e.g. $\max[\frac{1}{1.5} I_i(\theta_{c1}) + \frac{1}{0.5} I_i(\theta_{c2})]$

2. Goal Programming

- compute **the sum of the information** each available item can provide at each cutting point and at the current ability estimate (the item with the largest sum is selected)

$$\max \sum_{s \in V_s \rightarrow \{\theta_1, \dots, \theta_C, \hat{\theta}\}} w_s I_i(\theta_s), \text{ for } i \in V_i$$

e.g. $\max[w_1 I_i(\theta_{c1}) + w_2 I_i(\theta_{c2}) + w_3 I_i(\hat{\theta})]$

(In this study, all weights were set equal)

3. Global-Criterion Methods

- optimize all objectives **separately** and **combine** the results into one global criterion

$$\max \sum_{c=1}^C I_i(\theta_c), \text{ for } i \in V_{max}$$

Step1: optimize the objectives for each cutting point separately

Step2: calculating the sum of the information (global criterion)

	θ_{c1}	θ_{c2}	θ_{c3}
item 5	2	1	0.5
item 23	1	2	1
item 16	0.5	0.5	2

4. Maximin Methods

- **maximize the minimum** amount of information for each of the cut scores

- a lower boundary: should be low enough to ensure feasibility and high enough to ensure that the calculations do not consume unreasonable amounts of time
- the item was selected that maximized the boundary

item 6	0.5	0.5	0.5
item 4	0.9	0.3	0.4

Study 1

- Item Pools (2)

1. a Simulated Item Pool

- $a \sim N(1.50, 0.50)$ with $a > 0$
- $b \sim U(-3.00, 3.00)$
- 1000 items were generated for the item pool
- maximum test length = 40 items

- 1000 examinees with $\theta \sim N(0.00, 1.00)$

- cutting points:
 - 2 (33th, 66th)
 - 3 (25th, 50th, and 75th)
 - 4 (20th, 40th, 60th, and 80th)

- $\delta = 0.1; \alpha = \beta = 0.05$

2. the Mathematics Item Pool

- $\bar{a} = 3.09$
- $\bar{b} = 0$
- 250 items from a real test
- maximum test length = 40 items

- 1000 examinees with $\theta \sim N(0.294, 0.522)$


- cutting points:
 - 2 (-0.13, 0.33)

- $\delta = 0.1; \alpha = \beta = 0.05$

- Methods (8)

1. weighting methods (WM)
2. goal programming (GP) methods
3. global-criterion (GC) methods
4. maximin methods (MA)



1. maximizes information at the current ability estimate (AE)
2. at the middle of the nearest set of cutting points (MC)
3. at the nearest cutting point (NC)
4. random item selection (RA)  **serve as a baseline**

- Evaluation indexes (2)

1. average test length (ATL)
2. classification accuracy was defined as the proportion of correct decisions (PCD)



The simulations were executed for the eight item selection methods and were **replicated 100 times.**

Table 1. Results From Simulations With a Simulated Item Pool.

Item selection method	Two CP		Three CP		Four CP	
	ATL	PCD	ATL	PCD	ATL	PCD
RA	39.533	0.820	39.776	0.745	39.868	0.676
AE	32.646	0.906	34.861	0.866	35.938	0.826
MC	32.694	0.902	34.989	0.862	36.038	0.827
NC	32.721	0.908	35.009	0.867	36.170	0.828
WM	34.153	0.907	37.201	0.867	38.359	0.830
GP	33.065	0.907	37.296	0.863	39.364	0.818
GC	35.602	0.902	39.275	0.855	39.961	0.809
MA	33.259	0.902	36.856	0.853	38.444	0.798

Note. CP = cutting points; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion; MA = maximum.

Table 2. Results From Simulations With a Mathematics Item Pool.

Item selection method	ATL	PCD
RA	31.599	0.875
AE	20.213	0.915
MC	20.666	0.912
NC	20.593	0.916
WM	20.923	0.917
GP	20.831	0.914
GC	22.571	0.909
MA	22.514	0.908

Note. ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion.

- Various Delta Values

1. a Simulated Item Pool

- $a \sim N(1.50, 0.50)$ with $a > 0$
- $b \sim U(-3.00, 3.00)$
- 1000 items were generated for the item pool
- maximum test length = 40 items

- 1000 examinees with $\theta \sim N(0.00, 1.00)$

- cutting points:
 - 2 (33th, 66th)
 - ~~3 (25th, 50th, and 75th)~~
 - ~~4 (20th, 40th, 60th, and 80th)~~
- $\delta = 0.1; \alpha = \beta = 0.05$
↓
 $\delta \in (0.050, 0.400)$

2. the Mathematics Item Pool

- $\bar{a} = 3.09$
- $\bar{b} = 0$
- 250 items from a real test
- maximum test length = 40 items

- 1000 examinees with $\theta \sim N(0.294, 0.522)$

- cutting points:
 - 2 (-0.13, 0.33)
- $\delta = 0.1; \alpha = \beta = 0.05$
↓
 $\delta \in (0.025, 0.225)$

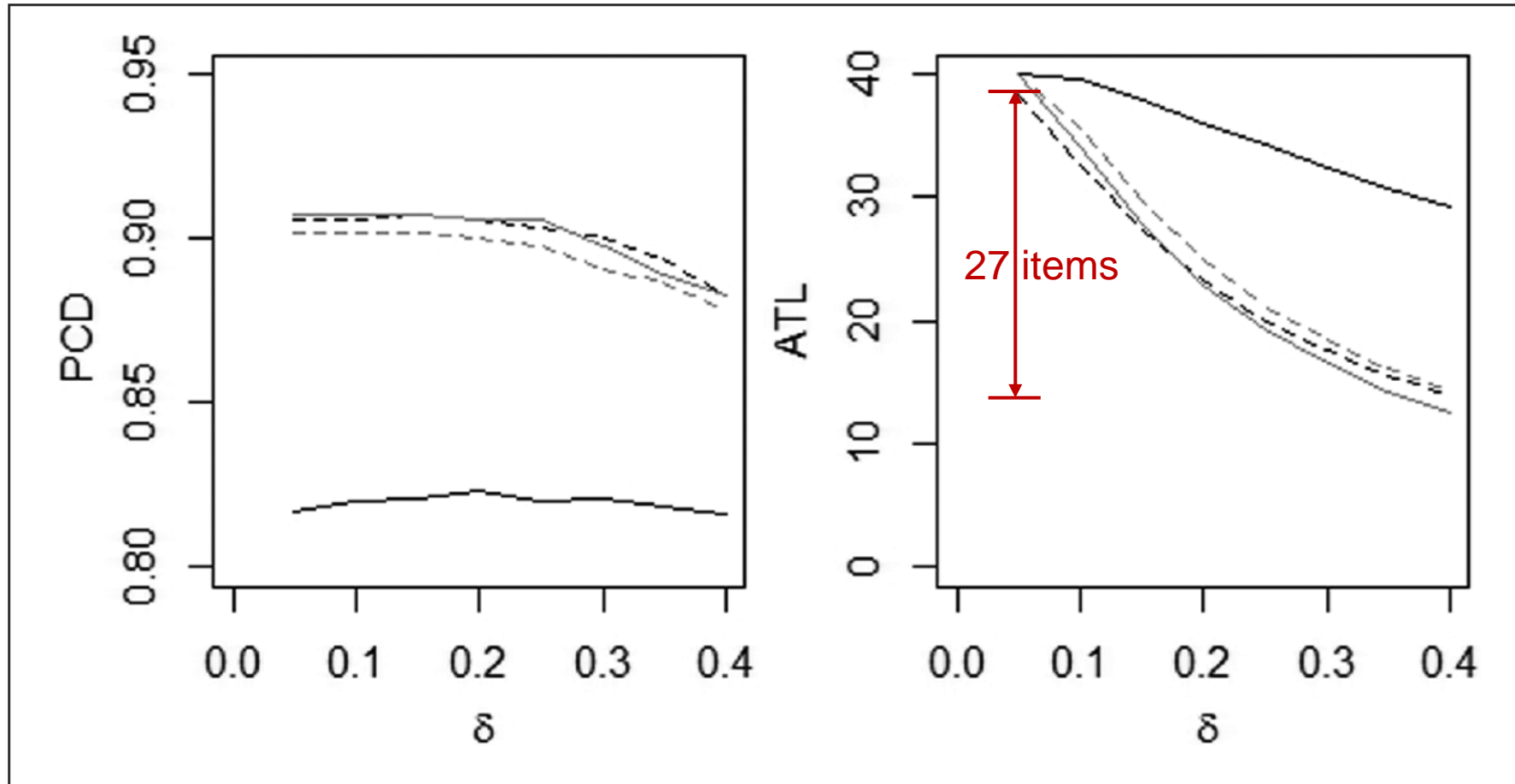


Figure I. Results from simulations with a simulated item pool for different sizes of the indifference region.

Note. The solid black line denotes RA, the solid gray line WM, the dotted black line AE, and the dotted gray line GC. RA = Random item selection; WM = Methods based on weighting; AE = ability estimate; GC = global criterion; PCD = proportion of correct decisions; ATL = average test length.

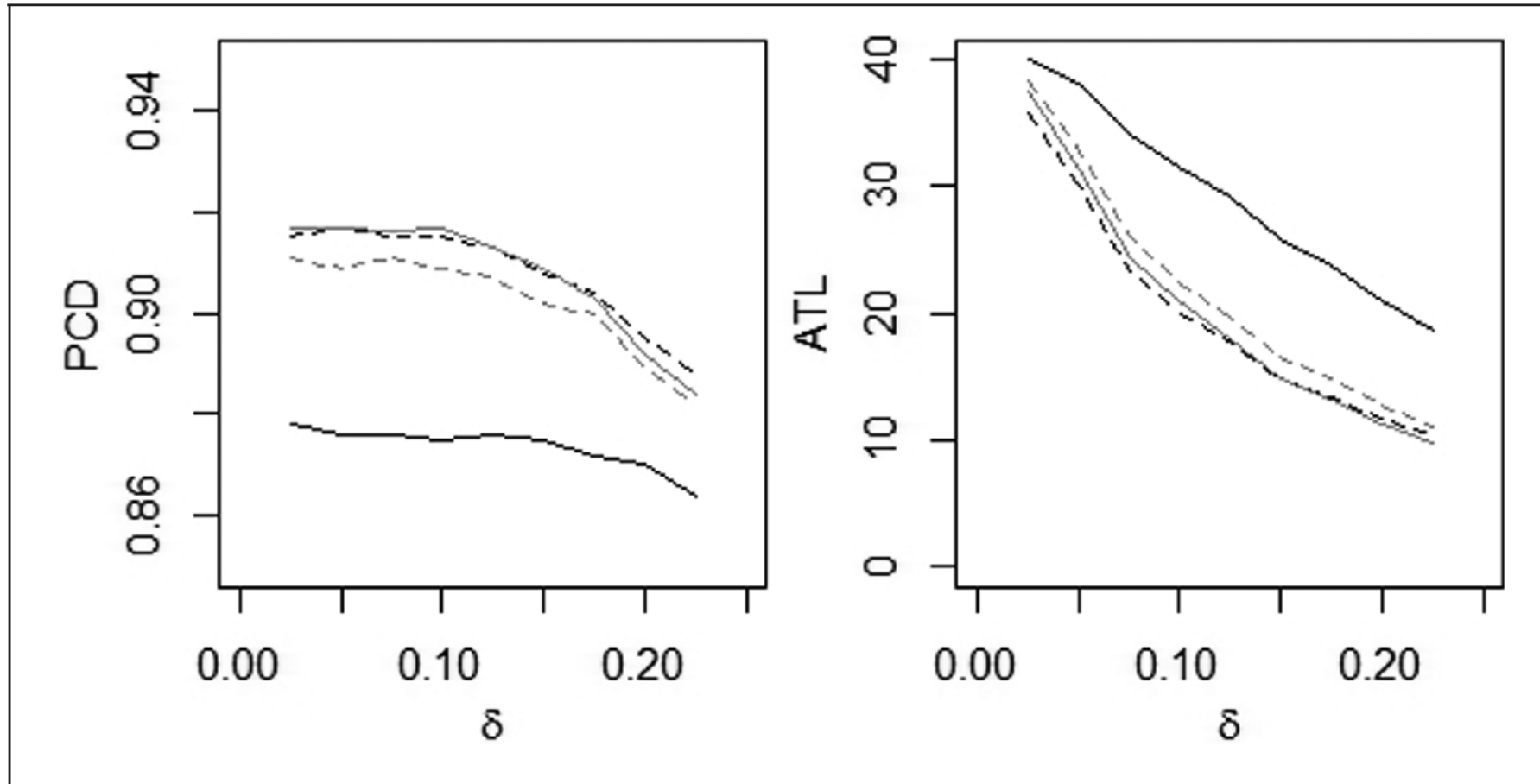


Figure 2. Results from simulations with the mathematics item pool for different sizes of the indifference region.

Note. The solid black line denotes RA, the solid gray line WM, the dotted black line AE, and the dotted gray line GC. RA = Random item selection; WM = methods based on weighting; AE = ability estimate; GC = global criterion; PCD = proportion of correct decisions; ATL = average test length.

- Content and Exposure Control (the mathematics item pool)

- Content control (C)

- 16% of the items from subdomain mental arithmetic/estimation
- 20% from measuring/geometry
- the other items from the other domains in the curriculum

- Exposure control (E)

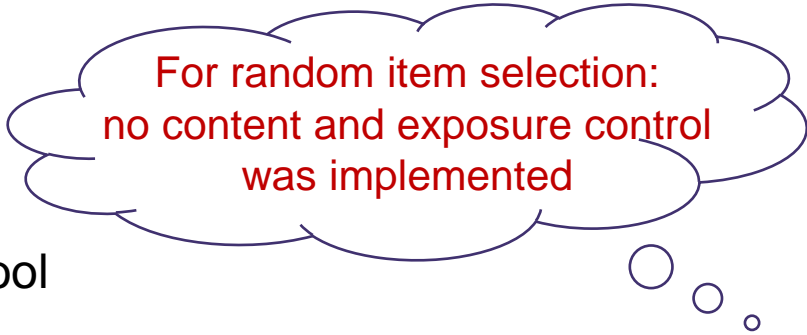
- When an item was selected, a **random number g** was drawn from the interval (0, 1).

{ if $g > 0.5$, administer
if not, reject

- The first 3 items

- An examinee was presented a **relatively easy item** from the item pool (54 items were denoted as easy items)

- Maximum test length = 25, $\delta = 0.10$, $\alpha = \beta = 0.05$



For random item selection:
no content and exposure control
was implemented

Table 3. Results From Simulations With and Without Content Constraints and Exposure Control.

Selection	No C, no E		C		E		C + E	
	ATL	PCD	ATL	PCD	ATL	PCD	ATL	PCD
RA	23.103	0.838						
AE	17.122	0.896	17.343	0.895	18.218	0.885	18.439	0.886
MC	17.330	0.890	17.685	0.891	18.314	0.883	18.494	0.883
NC	17.667	0.896	17.701	0.897	18.561	0.887	18.648	0.888
WM	17.987	0.897	17.969	0.896	18.768	0.889	18.897	0.889
GP	17.451	0.893	17.675	0.895	18.392	0.884	18.580	0.884
GC	18.985	0.885	18.970	0.889	19.779	0.881	20.196	0.881
MA	18.529	0.885	18.845	0.885	19.256	0.878	19.609	0.877

Note. C = content constraints; E = exposure control; ATL = average test length; PCD = proportion of correct decisions; RA = Random item selection; AE = ability estimate; MC = middle of the nearest set of cutting points; NC = nearest cutting point; WM = methods based on weighting; GP = goal programming; GC = global criterion.

- currently available item & multiple objective methods
 - use a multiple objective approach **in the starting phase of the test** and then switch to one of the currently available methods.
- a simulated pool & the mathematics item pool
 - **characteristics** of the item pool, **distribution** of ability, **settings** of the classification method, and the number of cutting points all influence test length and accuracy
- indifference regions & content and exposure control

- different **item pools**
- different **SPRT settings**
- different **examinee characteristics**

Thanks for listening!

Reporter: Yingshi Huang

谢谢大家

多谢晒~

ありがとう

Danke

Merci

The End © HYS 2020