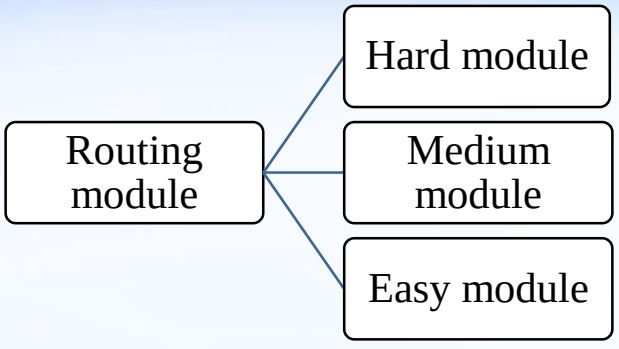
Listening. Learning. Leading.®

Computerized Multistage Testing and Applications

Duanli Yan
Educational Testing Service
The Seventeenth Annual Maryland Assessment Conference
November 3, 2017

Listening. Learning. Leading.®

Hypothetical Multistage Test (two stages)



Computerized Multistage Testing: Theory and Applications

Editors: Duanli Yan, Alina A. von Davier, Charles Lewis

Why MST?

- **Linear tests**: most popular; test takers all take every item (easy or difficult).

  A large number of items needed; no adaptation; need to design forms as parallel as possible.

- **Computerized adaptive tests (CAT):** adapts to the test taker's ability level; gets estimate after each item; guides selection of subsequent items.

  Test takers have different forms with different difficulties; complicated in design, assembly and estimation.

- **Multi-stage testing (MST):** similar to CAT, but by groups of items, or modules; an elegant compromise between Linear and CAT, flexibility.

  Shorter in test length, but as efficient as CAT for measurement.

  ➢ IRT-based and CTT-based methodologies.

  Development and use of MST are rapidly increasing.

Listening. Learning. Leading.®

Computerized Adaptive and Multistage Testing and R

- **Part I: Test Design, Item Pool, and Maintenance**
- **Part II: Test Assembly**
- **Part III: Routing, Scoring, and Equating**
- **Part IV: Test Reliability, Validity, Fairness, and Security**
- **Part V: Applications in Large –Scale Assessment**
- **Part VI: Past and Future**
- **CAT and MST with R**

Comparison of Linear Tests, CAT and MST

Table 1.1 A comparison of linear, CAT, and MST designs

Type of test	Advantages	Disadvantages
Linear test	Ease of assembly	Full length test
	Ease of administration	Inefficient for measurement
	Least effort for test development (TD)	Inflexible test schedule for test takers
		Prone to test copying
CAT	Shorter test length	Complicated to implement
	Efficient for measurement	Depends on strong model assumptions
	Flexible test schedule for test takers	Requires a large calibration data set
	Avoids test copying	Greatest effort for TD
		Item exposure more difficult to control
		Costly to administer via computer
		Robustness concerns
MST	Intermediate test length	Depends on model assumptions
	Efficient for measurement	Longer than CAT but shorter than linear test
	Allows test taker item review	Item exposure concerns (no more than CAT)
	Easier to implement	Costly to administer via computer
	Easier to assemble	(no more than CAT)
	Moderate effort for TD	
	Flexible test schedule for test takers	
	Reduces test copying	

Listening. Learning. Leading.®

MST Test Design and Implementation Considerations

- Test Design for Different Purposes
- Item Pool Design and Maintenance
- Content Balance and Test Assembly
- Routing and Scoring and Equating
- Psychometric models
- Reliability and Validity
- Security and Exposure Control
- Different types of adaptability
- Simulations for optimal design and implemantation

Listening. Learning. Leading.®

Examples of MSTs

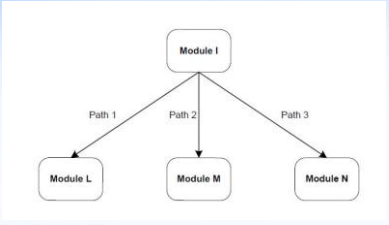


Figure 1.1. An Example of a Two-stage Multistage Testing Structure

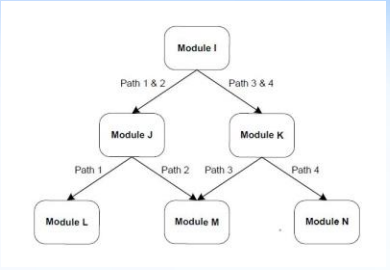


Figure 1.2. An Example of a Three-stage Multistage Testing Structure

Listening. Learning. Leading.®

An Example of an Item Pool

Table 5: The Crosstabulation of Items Based on P+ and r_{bi} for Calibration Sample (n=250)

	r_{bi}				
P+	0-.20	.20-.40	.40-.60	.60-1.00	Marginal
.84-.96	0	3	6	1	10
.70-.84	0	3	10	7	20
.51-.70	1	4	28	7	40
.36-.50	1	8	11	0	20
.20-.35	2	2	6	0	10

Listening. Learning. Leading.®

An Example of Module Specifications

Table 6: The Target Specifications for All the Modules in Design 1 (All Equal Length) for Typical Module Designs

Item Difficulty Category	Number in Pool	Lower r_{bi} I	Medium r_{bi} J	Medium r_{bi} K	Higher r_{bi} L	Higher r_{bi} M	Higher r_{bi} N	Lowest r_{bi} Not used
1	10	1	3	0	5	0	0	1
2	20	3	4	2	7	4	0	0
3	40	7	6	6	3	7	3	8
4	20	3	2	4	0	4	7	0
5	10	1	0	3	0	0	5	1
	100	15	15	15	15	15	15	10

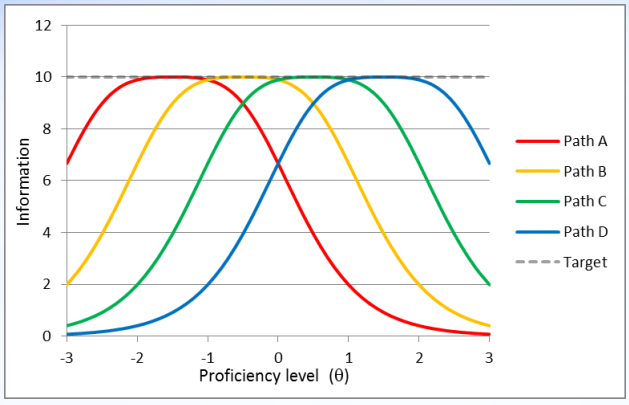
Listening. Learning. Leading.®

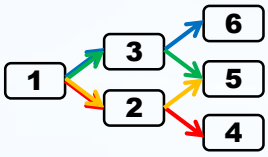
Routing, Scoring, and Equating

- IRT-Based MST
- Tree-Based MST
- MST for Categorical Decisions
- MST using Multidimensional Model
- MST using Diagnostic Models
- Considerations on Parameter Estimation, Scoring, and Linking

Listening. Learning. Leading.®

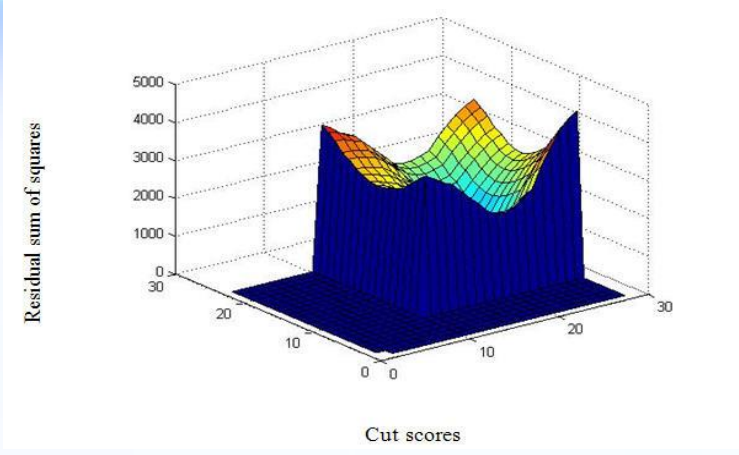
Path Information Functions (Weissman, 2013)





Listening. Learning. Leading.®

Identifying Cut Scores at Middle Stage



Residual Sum of Squares on Criterion as Function of Middle Stage Cut Scores

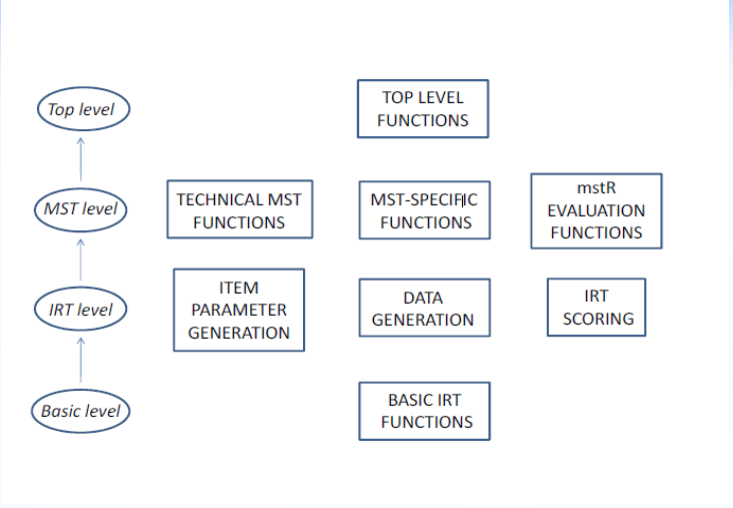
Listening. Learning. Leading.®

R Packages (Magis, 2017)

- catR: https://cran.r-project.org/web/packages/catR/catR.pdf
- mstR: https://cran.r-project.org/web/packages/mstR/mstR.pdf
- Designed to generate IRT-based MST scenarios under various options

Listening. Learning. Leading.®

R package *mstR*



Listening. Learning. Leading.®

mstR: Item bank, modules and paths

Three input arguments are mandatory to run MST simulation:

- an item bank
- a sorting of items into modules
- a MST structure describing stages and paths between stages

- item bank is a matrix with one row per item and as many columns as needed for item parameters
- Most common dichotomous IRT models (1PL, 2PL...) and polytomous IRT models (GRM, PCM...) available

Listening. Learning. Leading.®

mstR: Item bank

Illustration with 2PL model:

```
        a     b c d
[1,] 0.8   0.0 0 1
[2,] 0.9  -0.5 0 1
[3,] 1.0   1.1 0 1
[4,] 1.1  -0.8 0 1
[5,] 1.2   0.7 0 1
       .     .  .  .
       .     .  .  .
       .     .  .  .
```

Listening. Learning. Leading.®

mstR: modules

Example: modules matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

indicates that

- module 1 holds items 1, 2, 3, ...
- module 2 holds items 4, 6, ...
- module 3 holds items 5, ...
- ...

Listening. Learning. Leading.®

mstR: paths

Example:

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

leads to the following structure: paths can go

- from modules 1 to modules 2 or 3
- from module 2 to modules 4 or 5
- from module 3 to modules 5 or 6

mstR: randomMST()

The main meta-function, randomMST(), has many input arguments:

- trueTheta: the value of the true ability level (to be estimated)
- itemBank: the item bank for MST administration
- modules: the binary matrix that sets the modules from the item bank
- transMatrix: the transition matrix that sets the MST structure
- model: the type of IRT model used to calibrate the item bank

Listening. Learning. Leading.®

mstR: randomMST()

- responses: a possible vector of item responses for post-hoc simulations
- start: the list of options to select the first module
- test: the list of options for ad-interim ability estimation and next module selection
- final: the list of options for final ability estimation
- options to fix the random seed generation and save the output into external text files

Listening. Learning. Leading.®

mstR: randomMST()

The start list species options to select and administer the first module. Several modules are available at stage 1

- random selection
- selection forced by the test developer
- selection of the most informative module

Listening. Learning. Leading.®

mstR: randomMST()

The test list species options to select the next module and estimate ability at the end of module administration

- Ability estimators: ML, BM, EAP, WL, or total test score
- Module selection: extensions of MFI, MLWI, MPWI, KL, KLP to modules
- using predefined cut-scores between modules

mstR: randomMST()

The final list species options for final scoring and reporting:

- Final ability estimators: ML, BM, EAP, WL, or total test score
- For IRT estimates: final confidence interval

An Example

- 100 items
- Binary responses calibrated under 2PL model

First rows of bank matrix:

	a	b	c	d
[1,]	1.363	-1.623	0	1
[2,]	0.896	-1.008	0	1
[3,]	1.920	-1.397	0	1
[4,]	1.059	-1.772	0	1
[5,]	1.378	-0.687	0	1
[6,]	1.255	-1.526	0	1
	.	.	.	.

An Example, cont

- 1-2-3 design with 15 items in each module

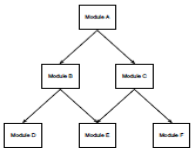
First rows of `modules` matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0	0	0	1	0	0
[2,]	0	0	0	1	0	0
[3,]	0	0	0	1	0	0
[4,]	0	1	0	0	0	0
[5,]	0	0	0	0	1	0
[6,]	0	0	0	1	0	0
	.	.	.	.	.	.
	.	.	.	.	.	.
	.	.	.	.	.	.

An Example, cont

Transition matrix `trans`:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0	1	1	0	0	0
[2,]	0	0	0	1	1	0
[3,]	0	0	0	0	1	1
[4,]	0	0	0	0	0	0
[5,]	0	0	0	0	0	0
[6,]	0	0	0	0	0	0



An Example, cont

MST options:

- First module chosen as most informative for starting ability level zero (by-default)
- Ability estimation: EAP
- Next module selection according to pre-specified cut-scores:
  - at stage 2, module 2 if $\hat{\theta} \leq 0$ and module 3 if $\hat{\theta} > 0$
  - at stage 3, module 4 if $\hat{\theta} \leq -1$, module 6 if $\hat{\theta} > 1$, module 5 otherwise
- Final ability estimation by ML

An Example, cont

Cut-off matrix `cut`:

```
      [,1] [,2]
[1,]   NA   NA
[2,] -Inf    0
[3,]    0  Inf
[4,] -Inf   -1
[5,]   -1    1
[6,]    1  Inf
```

- First row (for stage 1 module) mandatory (for compatibility) -Inf and Inf can be replaced by finite values

An Example, cont

R code for generating MST with prespecified options and for test taker with true ability level 0.5:

```
R> start <- list(theta = 0)
R> test <- list(method = "EAP", cutoff = cut)
R> final <- list(method = "ML")

R> ex <- randomMST(trueTheta = 0.5,
          itemBank = bank, modules = modules,
          transMatrix = trans,start = start,
          test = test, final = final)
```


An Example, cont

Selected output:

```
Random generation of a MST response pattern
  without fixing the random seed

 Item bank calibrated under Two-Parameter Logistic model

 True ability level: 0.5

 MST structure:
   Number of stages: 3
   Structure (number of modules per stage): 1-2-3

(...)
```

An Example, cont

```
Multistage test details:

 Stage 1:
   Module administered: 1
   Number of items in module 1: 15 items
   Items and responses:

Nr     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
Item   7 13 14 15 22 28 34 38 51 59 70 84 85 87 89
Resp.  1  0  1  0  1  0  1  1  1  1  1  1  0  1  0

    Provisional ability estimate (SE) after stage 1: 0.236 (0.545)
```


An Example, cont

```
  Stage 2:
    Module administered: 3
    Number of items in module 3: 15 items
    Items and responses:

Nr      1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
Item   32  35  36  45  52  57  61  63  66  78  81  82  88  92 100
Resp.   1   0   1   1   0   0   0   0   1   1   1   1   0   1   1

    Provisional ability estimate (SE) after stage 2: 0.29 (0.407)
```

An Example, cont

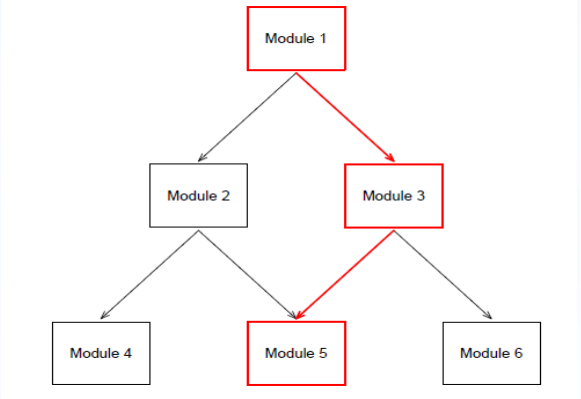
```
  Stage 3:
    Module administered: 5
    Number of items in module 5: 15 items
    Items and responses:

Nr     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15
Item   5 16 24 25 26 37 40 44 53 68 74 80 86 91 93
Resp.  1  1  1  1  1  0  0  0  1  1  1  1  1  1  0

    Provisional ability estimate (SE) after stage 3: 0.376 (0.336)
```

```
  Final results:
    Total length of multistage test: 45 items
    Final ability estimator: Maximum likelihood estimator
    Final range of ability values: [-4,4]
    Final ability estimate (SE): 0.398 (0.356)
    95% confidence interval: [-0.299,1.096]
```


An Example, cont

Visual representation of the generated path:

- R> plot(ex)

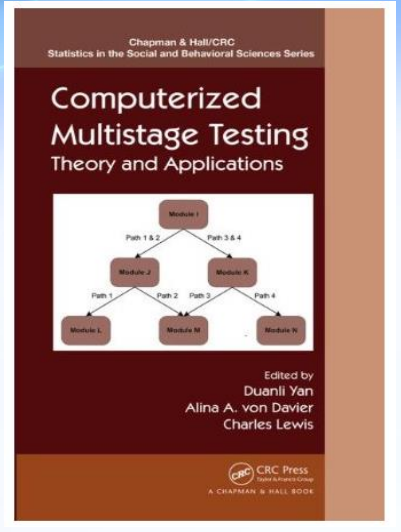


Listening. Learning. Leading.®

Applications in Large-Scale Assessment

- GRE Revised Test
- AICPA Examination
- K-12 Assessment
- NAEP Assessment
- International Survey Assessment
- MST with small samples
- PISA
- State and international Assessments
- More ….

Listening. Learning. Leading.®



Book Reviews in

- ***Psychometrika*** by Howard Wainer
- ***Journal of Educational and Behavioral Statistics (JEBS)*** by Robert J. Mislevy
- ***Applied Psychological Measurement (APM)*** by Mariana Curi

Listening. Learning. Leading.®

Computerized Adaptive and Multistage Testing with R (2017, Springer) by Magis, Yan and von Davier

AVAILABLE DECEMBER 2017

D. **Magis**, D. **Yan**, A. **von Davier**

Computerized Adaptive and Multistage Testing with R

Using Packages catR and mstR

Series: Use R!

- Provides exhaustive descriptions of CAT and MST processes in an R environment
- Guides users to simulate and implement CAT and MST using R for their applications
- Summarizes the latest developments and challenges of packages catR and mstR
- Provides R packages catR and mstR and illustrates to users how to do CAT and MST simulations and implementations using R

Listening. Learning. Leading.®

Thank you!

- Acknowledgment:
  - Contributors to the MST volume
  - AERA Division D Committee
  - ETS

- Contact:
  - dyan@ets.org