Use R!

Series Editors

Robert Gentleman
Kurt Hornik
Giovanni Parmigiani

More information about this series at http://www.springer.com/series/6991

David Magis • Duanli Yan • Alina A. von Davier

Computerized Adaptive and Multistage Testing with R

Using Packages *catR* and *mstR*

Springer

David Magis
Department of Education
University of Liege
Liege, Belgium

Duanli Yan
Educational Testing Service
Princeton, NJ, USA

Alina A. von Davier
ACTNext by ACT
Iowa City, IA, USA

ISSN 2197-5736 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-319-69217-3 ISBN 978-3-319-69218-0 (eBook)
https://doi.org/10.1007/978-3-319-69218-0

Library of Congress Control Number: 2017956343

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Maxime, Éléonore, and Anaëlle,
probably future R users

David Magis

To my daughter Victoria Song, an avid
R programmer

Duanli Yan

To my son, Thomas, whose time has come
to learn R

Alina A. von Davier

Foreword

Adaptive testing is, by now, a well-established procedure in educational and psychological assessment. Early adopters in the educational field in the USA were the licensing exam (NCLEX/CAT) of the National Council of State Boards of Nursing and the Graduate Record Examination (GRE). But at the time of this writing, the trend of using adaptive testing is worldwide. For instance, in the Netherlands, several adaptive tests are available for the compulsory nationwide final assessments in primary education. Further, also large-scale international educational surveys have implemented adaptive assessments; the Organisation for Economic Co-operation and Development (OECD) Program for International Student Assessment (PISA) and Program for the International Assessment for Adult Competencies (PIAAC) are prominent examples. Adaptive testing has also reached fields beyond educational assessment, such as epidemiology (the Patient-Reported Outcomes Measurement Information System (PROMIS) project for health assessment is an example), and organizational assessment (for supporting selection and promotion decisions). Most of these applications are based on adaptive selection of individual items. A more recent development is towards adaptive selection of testlets. This development is mainly motivated by the ease with which content control can be supported and generally goes under the name of multistage testing.

Given its widespread use, it may come as no surprise that good introductions to adaptive testing are available. I mention two edited volumes: *Computerized Adaptive Testing: A Primer* edited by Howard Wainer, which gives an excellent relatively nontechnical introduction to the key ingredients and practical implementation issues of adaptive testing, and *Elements of Adaptive Testing* edited by van der Linden and Glas, which delves a bit deeper into the statistical issues involved. An excellent introduction to multistage testing can be found in *Computerized Multistage Testing: Theory and Applications* edited by Yan, von Davier, and Lewis.

So where is the present volume and the software that goes with it to be positioned? First of all, the book provides a general overview of the statistical technology underlying adaptive testing, together with R packages that can provide illustrations for the theory. But there is more. When developing an adaptive

test, many questions must be answered and many decisions must be made. For instance, is my item bank big enough given the foreseen number of respondents and lifespan of the test, is the test length appropriate for the targeted reliability, and are the item parameters adequate for the targeted population, to mention a few questions. The website of the International Association for Computerized Adaptive Testing (IACAT) adds some more issues: content balancing (to create a test where the content matter is appropriately balanced), item exposure control (to prevent compromising of the test), combining multiple scales, and proctoring of administration via the Internet. To complicate things, all these aspects usually interact. The simulation studies such as those supported by the R packages can help answering such questions and support the decisions. The choice of R is a good one, motivated by the fact that it has become the most versatile and flexible open-source platform for statistical analyses and creation of graphical output. For most developers and users of statistical tools, R has become the standard of industry. It offers practitioners the opportunity to add their own functionality to existing applications and provides a common ground for exchanging their software. Therefore, I am sure that this book and the simulation software accompanying it will be extremely helpful for designing adaptive tests.

Professor of Social Science Research Methodology Cees Glas
University of Twente, The Netherlands

Acknowledgments

This book takes its origin from an inspiring discussion between the authors during the meeting of the International Association for Computerized Adaptive Testing (IACAT) in Cambridge (UK), October 2015. From this early project, several drafts were written, re-organized, reviewed, and updated to end up with the current version of the book.

The authors wish to express their sincere acknowledgments to the many people who took some of their precious time to read the provisional chapters and make insightful comments: Hong Jiao (University of Maryland), Kim Fryer, Lixiong Gu, Sooyeon Kim, Yanming Jiang, Longjuan Liang, Guangming Ling, Yuming Liu, Manfred Steffen, Fred Robin, Jonathan Weeks, Meng Wu (Educational Testing Service). The authors want to express their gratitude to Cees Glas who accepted to preface this book long before it came out in its final version. Eventually, special thanks to Jim Carlson from Educational Testing Service for his comments and suggestions throughout the whole book, and to Andrew Cantine (technical editor, ACTNext) for carefully editing the writing of this book.

Contents

1 Overview of Adaptive Testing 1
1.1 Linear Test, CAT and MST 1
1.1.1 Linear Test 1
1.1.2 CAT 2
1.1.3 MST 3
1.2 Organization of This Book 3

2 An Overview of Item Response Theory 7
2.1 Principles and Assumptions of Item Response Theory 7
2.2 Commonly Used IRT Models 9
2.2.1 Unidimensional Dichotomous IRT Models 9
2.2.2 Unidimensional Polytomous IRT Models 12
2.2.3 Multidimensional IRT Models 17
2.2.4 Other IRT Models 19
2.3 Parameter Estimation 20
2.3.1 Model Calibration 21
2.3.2 Ability Estimation 22
2.3.3 Information and Precision 25
2.4 Further Topics 27
2.4.1 Dimensionality 27
2.4.2 Local Item Independence 28
2.4.3 Model and Person Fit 28
2.4.4 Differential Item Functioning 29
2.4.5 IRT Linking and Equating 29

Part I Item-Level Computerized Adaptive Testing

3 An Overview of Computerized Adaptive Testing 35
3.1 Introduction and Background 35
3.2 CAT Basics 36

3.3 Test Design and Implementation 37
3.4 Test Assembly 38
3.5 The Item Bank 38
3.6 IRT-Based CAT 40
3.6.1 Initial Step 40
3.6.2 Test Step 41
3.6.3 Item Selection Method 42
3.6.4 Stopping Step 47
3.6.5 Final Step 48
3.7 Content Balance, Exposure and Security 48
3.8 CAT with Regression Trees 50
3.9 Final Comments 51

4 Simulations of Computerized Adaptive Tests 53
4.1 The R Package catR 53
4.2 Item Bank and Structure 54
4.3 General Architecture of catR 56
4.4 Basic IRT Functions 57
4.5 IRT-Level Functions 59
4.5.1 Item Parameter Generation 59
4.5.2 Data Generation 62
4.5.3 IRT Scoring 63
4.6 CAT-Level Functions 66
4.6.1 Technical CAT Functions 67
4.6.2 CAT-Specific Functions 67
4.7 Top-Level Function: `randomCAT()` 73
4.7.1 Input Information 74
4.7.2 The `Start` List 75
4.7.3 The `Test` List 77
4.7.4 The `Stop` List 79
4.7.5 The `Final` List 80
4.7.6 Output Information 80
4.8 Top-Level Function: `simulateRespondents()` 81
4.8.1 Input Arguments 81
4.8.2 Output Information 83

5 Examples of Simulations Using catR 87
5.1 Item Banks 87
5.1.1 The Dichotomous Item Bank 87
5.1.2 The Polytomous Item Bank 89
5.2 Example 1a: CAT with Dichotomous Item Bank 90
5.3 Example 1b: CAT with Polytomous Item Bank 94

5.4 Example 2: CAT for Placement Tests 97
5.4.1 Data Generation and Linear Design 97
5.4.2 CAT Design and Implementation 98
5.4.3 Output Analysis 99
5.5 Example 3: CAT with Unsuitable Data 102
5.5.1 Data Generation 102
5.5.2 CAT Design and Implementation 103
5.5.3 Results 104
5.6 Example 4: `simulateRespondents()` Function 105

Part II Computerized Multistage Testing

6 An Overview of Computerized Multistage Testing 113
6.1 Introduction and Background 113
6.2 MST Basics 114
6.3 Test Design and Implementation 116
6.4 Test Assembly 117
6.5 The Item Bank 117
6.6 IRT-Based MST 118
6.6.1 Module Selection 118
6.6.2 Routing 118
6.6.3 Latent Trait Estimation 119
6.7 IRT Linking 120
6.8 MST with Regression Trees 121
6.9 Final Comments 121

7 Simulations of Computerized Multistage Tests 123
7.1 The R Package **mstR** 123
7.2 Multistage Structure for Item Banks 124
7.3 MST Functions 128
7.3.1 The `startModule()` Function 128
7.3.2 The `nextModule()` Function 130
7.4 The `randomMST()` Function 136
7.4.1 Input Information 136
7.4.2 The `start` List 137
7.4.3 The `test` List 138
7.4.4 The `final` List 139
7.4.5 Output Information 139

8 Examples of Simulations Using *mstR* 141
8.1 Introduction 141
8.2 Example 1: MST Using `randomMST()` 144
8.3 Example 2: MST with Cut-Scores 148
8.3.1 Thresholds for Ability Estimation 148
8.3.2 Score-Based Thresholds 151

8.4 Example 3: Comparing MST Designs 153
8.4.1 Designs 153
8.4.2 Implementation 154
8.4.3 Results 155
8.5 Example 4: MST Versus CAT 156
8.5.1 Design and Code 157
8.5.2 Results 158

References 161

Index 171

Acronyms

CAT	Computerized Adaptive Testing
IRT	Item Response Theory
MST	Multistage Testing

List of Figures

Fig. 2.1 Item response functions of various unidimensional logistic IRT models. For the 2PL, 3PL and 4PL models, the difficulty b_j is fixed to zero. For the 3PL and 4PL models, the discrimination a_j is fixed to one. For the 4PL model, the lower asymptote c_j is fixed to zero 11

Fig. 2.2 Cumulative probabilities (left) and response category probabilities (right) of an artificial four-response item under GRM, with parameters $(\alpha_j, \beta_{j1}, \beta_{j2}, \beta_{j3}) = (1.2, -1.8, 0.3, 1.4)$ 13

Fig. 2.3 Response category probabilities of an artificial five-response item under PCM (left) and GPCM (right). The threshold parameters are $(\delta_{j1}, \delta_{j2}, \delta_{j3}, \delta_{j4}) = (-1.5, -0.8, 0.5, 1.2)$ for both panels, and the slope parameter of the GPCM is $\alpha_j = 1.6$ 16

Fig. 2.4 Response category probabilities of two artificial four-response items under NRM with $(c_{j1}, c_{j2}, c_{j3}) = (1.5, 2, 0.8)$ and different α_{jk} parameters. Left panel: $(\alpha_{j1}, \alpha_{j2}, \alpha_{j3}) = (1, 2, 3)$. Right panel: $(\alpha_{j1}, \alpha_{j2}, \alpha_{j3}) = (1, 1.5, 2)$ 17

Fig. 3.1 Schematic illustration of a CAT process 37

Fig. 3.2 An example of a tree-based CAT 51

Fig. 4.1 General structure of the **catR** package 56

Fig. 5.1 Scatterplot of discrimination and difficulty coefficients (left) and test information function (right) of the dichotomous (2PL) item bank 88

Fig. 5.2 Test information function of the polytomous (GPCM) item bank 90

Fig. 5.3 Sequences of successive ability estimates and related 95% confidence intervals with the 2PL item bank 94

Fig. 5.4 Proportions of test taker correctly classified (upper left), incorrectly classified (upper right) and without final classification (lower left) for linear and CAT assessment based on post-hoc simulations with 2PL item bank. Boxplots of CAT test lengths (lower right) are displayed for each true ability level. Classification threshold (−1.048) is displayed by a vertical solid line 100
Fig. 5.5 Bias (left) and root mean squared error (right) curves of EAP ability estimator for the 2PL (true) data and the 4PL (corrupted) data sets 105
Fig. 5.6 Output of the `simulateRespondents()` function 109
Fig. 5.7 Output of the `"expRatePara"` panel from `simulateRespondents()` function 110

Fig. 6.1 Example of an MST design 115

Fig. 7.1 General structure of the `mstR` package 124
Fig. 7.2 An example of an 1-2-3 MST design 125

Fig. 8.1 Graphical output of the first simulated MST example 148
Fig. 8.2 ASB (left) and RMSE (right) values for each MST design and true ability level generated in Example 3 156
Fig. 8.3 Scatter plot of CAT versus MST ability estimates (upper left), ASB (upper right), average SE (lower left) and RMSE (lower right) values for CAT and MST designs at each true ability level generated in Example 4 159

List of Tables

Table 1.1 A comparison of linear, CAT, and MST designs 4

Table 4.1 Maximum number of columns and order of item parameters for all possible polytomous IRT models with $K_j + 1$ (or $K + 1$) response categories 55

Table 4.2 List of technical CAT functions of **catR** for next item selection 67

Table 4.3 Input arguments for `randomCAT()` function 74

Table 4.4 Elements of the `start` list of function `randomCAT()` 76

Table 4.5 Elements of the `test` list of function `randomCAT()` 77

Table 4.6 Elements of the `stop` list of function `randomCAT()` 79

Table 4.7 Elements of the `plot.cat()` function 81

Table 4.8 Input arguments for `simulateRespondents()` function 82

Table 4.9 Selected output arguments of `simulateRespondents()` function 84

Table 4.10 Values for `type` argument of `plot.catResult()` function 85

Table 5.1 Contingency table of number of final classification outcomes (incorrect, undetermined, correct) for both linear and CAT designs 101

Table 7.1 Input arguments for `randomMST()` function 137

Table 7.2 Elements of the `start` list of function `randomMST()` 137

Table 7.3 Elements of the `test` list of function `randomMST()` 138

Table 7.4 Elements of the `plot.mst()` function 140

Table 8.1 Module sizes by stage for the six possible MST designs using the 2PL item bank 142
Table 8.2 Means and variances of item discrimination parameters in each module and for each MST design 142
Table 8.3 Means and variances of item difficulty parameters in each module and for each MST design 142