Statistics for Social and Behavioral Sciences

Wim J. van der Linden
Cees A.W. Glas

Elements of Adaptive Testing

Springer

Statistics for Social and Behavioral Sciences

Advisors:

S.E. Fienberg
W.J. van der Linden

For other titles published in this series, go to
http://www.springer.com/3463

Wim J. van der Linden • Cees A.W. Glas (Eds.)

Elements of Adaptive Testing

Springer

Wim J. van der Linden
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93940
USA
wim_vanderlinden@ctb.com

Cees A.W. Glas
Twente University
Fac. Behavioural Sciences
Dept. Research Methodology
7500 AE Enschede
The Netherlands
c.a.w.glas@utwente.nl

Series Editors

Stephen Fienberg
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA

Wim J. van der Linden
CTB/McGraw-Hill
20 Ryan Ranch Road
Monterey, CA 93940
USA

ISBN 978-0-387-85459-5 e-ISBN 978-0-387-85461-8
DOI 10.1007/978-0-387-85461-8
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010921197

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

For a long time, educational testing has focused mainly on paper-and-pencil tests and performance assessments. Since the late 1980s, when the rapid dissemination of personal computers in education began, these testing formats have been extended to formats suitable for delivery by computer. Such delivery of tests has several advantages. For example, it offers the possibility of testing on demand, that is, whenever and wherever an examinee is ready to take the test. Also, both the power of modern PCs and their ability to integrate multiple media can be used to create innovative item formats and more realistic testing environments. Furthermore, computers can be used to increase the statistical accuracy of test scores using computerized adaptive testing (CAT). Instead of giving each examinee the same fixed test, in adaptive testing after each new response the individual examinee's ability estimate is updated and the subsequent item is selected to have optimal properties at the new estimate.

The idea of adaptive item selection is certainly not new. In the Binet–Simon (1905) intelligence test, the items were classified according to mental age, and the classification was used to adapt the selection of the items to an estimate of the mental age of the examinee derived from the responses to the earlier items until the correct age could be identified with sufficient certainty. In fact, the idea of adaptive testing is as old as the practice of oral examinations. Any sensitive oral examiner knows how to tailor the questions to his or her impression of the examinee's knowledge level.

The development of item response theory (IRT) in the middle of the last century has provided a sound psychometric footing for adaptive testing. The key feature of IRT is its modeling of the response probabilities for an item with distinct parameters for the examinee's ability and the characteristics of the items. Due to this parameter separation, the statistical question of optimal item parameter values for the estimation of examinee ability could be addressed. The main answer to the question was given by Birnbaum (1968), who, for Fisher's information measure, showed that, unless guessing is possible, the optimal item is the one with the highest value for its discrimination parameter and a value for the difficulty parameter equal to the ability of the examinee.

The further development and fine-tuning of the psychometric techniques needed to implement adaptive testing took several decades. Because the first computers were slow and did not allow for statistically sound real-time ability estimation, early research was almost exclusively directed at finding approximate estimation methods and alternative adaptive formats that could be implemented in a traditional paper-and-pencil environment. Examples include the two-stage testing format (Cronbach & Gleser, 1965), Bayesian item selection with an approximation to the posterior distribution of the ability parameter (Owen, 1969), the up-and-down method of item selection (Lord, 1970), the Robbins–Monro algorithm (Lord, 1971a), the flexilevel test (Lord, 1971b), the stradaptive test (Weiss, 1973), and pyramidal adaptive testing (Larkin & Weiss, 1975).

With the advent of more powerful computers, the use of adaptive testing in large-scale, high-stakes testing programs became feasible. A pioneer in this field was the U.S. Department of Defense, with its Armed Services Vocational Aptitude Battery (ASVAB). After a developmental phase, which began in 1979, the first CAT version of the ASVAB became operational in the mid-1980s. An informative account of the development of the CAT-ASVAB is given in Sands, Waters, and McBride (1997). However, the migration from paper-and-pencil testing to computerized adaptive testing truly began when the National Council of State Boards of Nursing launched a CAT version of its licensing exam (NCLEX/CAT) and was followed with a CAT version of the Graduate Record Examination (GRE). Several other programs followed suit. After a temporary setback due to security problems for the GRE, large numbers of testing programs are now adaptive, not only in education but also in psychology and, more recently, areas such as marketing and health-outcome research.

Some of the early reasons to switch to computerized test administration were (1) the possibility for examinees to schedule tests at their convenience; (2) tests are taken in a more comfortable setting and with fewer people around than in large-scale paper-and-pencil administrations; (3) electronic processing of test data and reporting of scores are faster; and (4) wider ranges of questions and test content can be put to use (Educational Testing Service, 1994). In the current programs, these advantages have certainly been realized and are appreciated by the examinees. When offered the choice between a paper-and-pencil and a CAT version of the same test, typically nearly all examinees choose the latter.

But the first experiences with real-world CAT programs have also given rise to a host of new questions. For example, in programs with high-stakes tests, item security quickly became a problem. The capability of examinees to memorize test items as well as their tendency to share them with future examinees appeared to be much higher than anticipated. As a consequence, the need arose for effective methods to control the exposure of the items as well as to detect items that have been compromised. Also, the question of how to align test content with the test specifications and balance content across test administrations appeared to be more complicated than anticipated. This question has led to a variety of new testing algorithms. Furthermore, items can now be calibrated online during operational testing, and the feasibility of efficient methods of item calibration, using collateral information about the examinee and employing optimal design techniques, is currently being investigated. These examples highlight only a few practical issues met when the first CAT programs were implemented in practice. A more comprehensive review of such issues is given in Mills and Stocking (1996).

This volume is a completely revised and updated version of *Computerized Adaptive Testing: Theory and Practice* edited by the same authors and published by Kluwer, now part of the same company as Springer (van der Linden & Glas, 2000). Much has changed in the area of adaptive testing research and practice over the nearly 10 years that have passed since the publication of this volume, and the editors have appreciated the opportunity to change the composition of the volume, add new chapters, and update the chapters that have remained. The goal of the volume, however, has remained the same—not to provide a textbook with a basic introduction to adaptive testing but to present a snapshot of the latest exciting results from research and development efforts in the area. For a more comprehensive introduction to adaptive testing, the student or test specialist should therefore complement the volume with other books, such as Parshall, Spray, Kalohn and Davey (1969), Sands, Waters, and McBride (1997), Wainer (1990), and Weiss (1983). As the developments in adaptive testing are intricately related to those in computerized testing at large, reference to volumes on this topic edited by Bartram and Hambleton (2006), Drasgow and Olson-Buchanan (1999), and Mills, Potenza, Fremer and Ward (2002) are also appropriate.

As always, the book is the result of contributions by many people whose roles we gratefully acknowledge. First, we would like to express our gratitude to the contributing authors. Their cooperation and willingness to report on their research and developmental work in this volume are greatly appreciated. In spite of the current tendency to use journals rather than books as a primary outlet for new research, these contributors have allowed us to edit a volume with chapters that are based on original work. We would also like to thank John Kimmel for his support during the production of this volume. His way of asking us about our progress was always subtle and timely. Our thanks are also due to *Applied Measurement in Education*, *Applied Psychological Measurement*, *Psychometrika,* and the *Journal of Educational and Behavioral Statistics* for their permission to reproduce portions of figures in Chapters 1 and 2 as well as the Defense Manpower Data Center for the opportunity to use itsr data in one of the empirical examples in Chapter 2. Finally, the Law School Admission Council generously supported the research in Chapters 1, 2, 5, 6, 13, 20, and 21 of this volume. The research in Chapter 15 was funded by Deutsche Forschungsgemeinschaft (DFG), Schwerpunktprogramm "Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Kompetenzsprozessen" (SP 1293), Project "Rule-Based Item Generation of Statistical word Problems Based upon Linear Logistic Test Models for Item Cloning and Optimal Design." Without this support, the volume would not have been possible.

CTB/McGraw-Hill *Wim J. van der Linden*
University of Twente *Cees A. W. Glas*

References

Bartram, D. & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the internet: Issues and advances*. Chichester, UK: Wiley.

Binet, A. & Simon, Th. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. *l'Anneé Psychologie, 11*, 191–336.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological test and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Drasgow, F. & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Educational Testing Service. (1994). *Computer-based tests: Can they be fair to everyone?* Princeton, NJ: Educational Testing Service.

Larkin, K. C. & Weiss, D. J. (1975). *An empirical comparison of two-stage and pyramidal adaptive ability testing* (Research Report, 75-1). Minneapolis: Psychometrics Methods Program, Department of Psychology, University of Minnesota.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer assisted instruction, testing, and guidance* (pp. 139–183). New York: Harper and Row.

Lord, F. M. (1971a). Robbins–Monro procedures for tailored testing. *Educational and Psychological Measurement*, *31*, 2–31.

Lord, F. M. (1971b). The self-scoring flexilevel test. *Journal of Educational Measurement*, *8*, 147–151.

Mills, C. N., Potenza, M. T., Fremer, J. J. & Ward, W. C. (2002). *Computer-based testing: Building the foundation for future assessments*. Mahwah, NJ: Erlbaum.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in computerized adaptive testing. *Applied Psychological Measurement, 9,* 287–304.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Parshall, C. A., Spray, J. A., Kalohn. J. C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer.

Sands, W. A., Waters, B. K. & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

van der Linden, W. J. & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston: Kluwer.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hilsdale, NJ: Lawrence Erlbaum Associates.

Contributors

Adelaide A. Ariel Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Krista Breithaupt American Institute of Certified Public Accountants, 1230 Corporate Parkway Avenue, Ewing, NJ 08628–3018, USA

Tim Davey Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

Theo J.H.M. Eggen Cito Institute for Educational Measurement, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

Hanneke Geerlings Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Cees A.W. Glas Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

Ronald K. Hambleton Center for Educational Assessment, University of Massachusetts, Amherts, MA 01002, USA

Donovan R. Hare Department of Mathematics & Statistics, University of British Columbia Okanagan, 3333 University Way, Kelowna, BC V1V 1V7, Canada

J. Christine Harmes The Center for Assessment and Research Studies, James Madison University, 821 S. Main Street, MSC 6806, Harrisonburg, VA 22807, USA

Norio Hayashi Japan Institute for Educational Measurement, Inc., 162–8680 Tokyo, 55 Yokodera-cho Shinjuku-ku, Japan

Richard M. Luecht ERM Department, University of North Carolina at Greensboro, Greensboro, NC 26170, USA

Gerald J. Melican The College Board, 45 Columbus Avenue, New York, NY 10023–6992, USA

Rob R. Meijer Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Joris Mulder Department of Methodology and Statistics, Utrecht University,
P.O. Box 80140, 3508 TC Utrecht, The Netherlands

Yasuko Nogami Japan Institute for Educational Measurement, Inc., 162–8680
Tokyo, 55 Yokodera-cho Shinjuku-ku, Japan

Cynthia G. Parshall Measurement Consultant, 415 Dunedin Avenue, Temple
Terrace, FL 33617, USA

Peter J. Pashley Law School Admission Council, P.O. Box 40, Newtown,
PA 18940–0040, USA

Lawrence M. Rudner Graduate Management Admission Council, 1600 Tysons
Boulevard, Ste. 1400, McLean, VA 22102, USA

Daniel O. Segall Defence Manpower Data Center, 400 Gigling Road, Seaside,
CA 93955–6771, USA

Gerard J.J.M. Straetmans Cito Institute for Educational Measurement,
P.O. Box 1034, 6801 MG Arnhem, The Netherlands

Wim J. van der Linden CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey,
CA 93940, USA

Edith M.L.A. van Krimpen-Stoop Heymans Institute, University of Groningen,
Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

Bernard P. Veldkamp Department of Research Methodology, Measurement,
and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede,
The Netherlands

Angela J. Verschoor Cito Institute for Educational Measurement, P.O. Box 1034,
6801 MG Arnhem, The Netherlands

Hans J. Vos Department of Research Methodology, Measurement, and Data
Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands

Otto B. Walter Institut für Psychologie, RWTH Aachen University, Jägerstrasse
17/19, 52066 Aachen, Germany

April Zenisky Center for Educational Assessment, University of Massachusetts,
Amherst, MA 01002, USA

Yanwei Zhang American Institute of Certified Public Accountants, 1230
Corporate Parkway Avenue, Ewing, NJ 08628–3018, USA

Rebecca Zwick Department of Education, University of California, 2216 Phelps
Hall, Santa Barbara, CA 93106–9490, USA

Contents

Part I Item Selection and Ability Estimation

Part II Applications in Large-Scale Testing Programs

Part III Item Pool Development and Maintenance

Part IV Item Calibration and Model Fit

Part V Multistage and Mastery Testing

Part I
Item Selection and Ability Estimation

Chapter 1
Item Selection and Ability Estimation in Adaptive Testing

Wim J. van der Linden and Peter J. Pashley

1.1 Introduction

The last century saw a tremendous progression in the refinement and use of standardized linear tests. The first administered College Board exam occurred in 1901 and the first Scholastic Assessment Test (SAT) was given in 1926. Since then, progressively more sophisticated standardized linear tests have been developed for a multitude of assessment purposes, such as college placement, professional licensure, higher-education admissions, and tracking educational standing or progress. Standardized linear tests are now administered around the world. For example, the Test of English as a Foreign Language (TOEFL) has been delivered in approximately 88 countries.

Seminal psychometric texts, such as those authored by Gulliksen (1950), Lord (1980), Lord and Novick (1968), and Rasch (1960), have provided increasingly sophisticated means for selecting items for linear test forms, evaluating them, and deriving ability estimates using them. While there are still some unknowns and controversies in the realm of assessment using linear test forms, tried-and-true prescriptions for quality item selection and ability estimation abound. The same cannot yet be said for adaptive testing. To the contrary, the theory and practice of item selection and ability estimation for computerized adaptive testing (CAT) are still evolving.

Why has the science of item selection and ability estimation for CAT environments lagged behind that for linear testing? First of all, the basic statistical theory underlying adapting a test to an examinee's ability was only developed relatively recently. (Lord's 1971 investigation of flexilevel testing is often credited as one of the pioneering works in this field.) But more importantly, a CAT environment involves many more delivery and measurement complexities as compared to a linear testing format.

W.J. van der Linden (✉)
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

P.J. Pashley
Law School Admission Council, P.O. Box 40, Newtown, PA 18940–0040, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_1,

To illustrate these differences, consider the current development and scoring of one paper-and-pencil Law School Admission Test (LSAT). To begin, newly written items are subjectively rated for difficulty and placed on pretest sections by test specialists. Items that statistically survive the pretest stage are eligible for final form assembly. A preliminary test form is assembled using automated test assembly algorithms, and is then checked and typically modified by test specialists. The form is then pre-equated. Finally, the form is given operationally, to about 25,000 examinees on average, and most likely disclosed. Resulting number-right scores are then placed on a common LSAT scale by psychometricians using IRT scaling and true-score equating. The time lag between operational administrations and score reporting is usually about three weeks.

In contrast, within a CAT environment item selection and ability estimation occur in real time. As a result, computer algorithms must perform the roles of both test specialists and psychometricians. Because the test adapts to the examinee, the task of item selection and ability estimation is significantly harder. In other words, procedures are needed to solve a very complex measurement problem. These procedures must at the same time be robust enough to be relied upon with little or no human intervention.

Consider another, perhaps more subtle, difference between linear and CAT formats. As indicated above with the LSAT example, item selection and ability estimation associated with linear tests are usually conducted separately, though sometimes using similar technology, such as item response theory. Within a CAT format, item selection and ability estimation proceed hand in hand. Efficiencies in ability estimation are heavily related to the selection of appropriate items for an individual. In a circular fashion, the appropriateness of items for an individual depends in large part on the quality of interim ability estimates.

To start the exposition of these interrelated technologies, this chapter discusses what could be thought of as baseline procedures for the selection of items and the estimation of abilities within a CAT environment. In other words, it discusses basic procedures appropriate for unconstrained, unidimensional CATs that adapt to an examinee's ability level one item at a time for the purposes of efficiently obtaining an accurate ability estimate. Constrained, multidimensional, and testlet-based CATs, and CATs appropriate for mastery testing, are discussed in other chapters in this volume (Eggen, chap. 19; Glas & Vos, chap. 21; Mulder & van der Linden, chap. 4; Segall, chap.3; van der Linden, chap. 2; Vos & Glas, chap,. 20). Also, the focus in this chapter is on adaptive testing with dichotomously scored items. But adaptive testing with polytomous models has already been explored for such models as the nominal response model (e.g., De Ayala, 1992), graded response model (e.g., De Ayala, Dodd & Koch, 1992), partial credit model (Chen, Hou & Dodd, 1998), generalized partial credit model (van Rijn, Eggen, Hemker & Sanders, 2002), and an unfolding model (Roberts, Lin & Laughlin, 2001). Finally, in the current chapter, item parameters are assumed to have been estimated, with or without significant estimation error. A discussion of item parameter estimation for adaptive testing is given elsewhere in this volume (Glas, chap. 14; Glas, van der Linden & Geerlings, chap. 15).

Classical procedures are covered first. Often these procedures were strongly influenced by a common assumption or a specific circumstance. The common assumption was that what works well for linear tests probably works well for CATs. Selecting items based on maximal information is an example of this early thinking. The specific circumstance was that these procedures were developed during a time when fast PCs were not available. For example, approximations, such as Owen's (1969) approximate Bayes procedure, were often advocated to make CATs feasible to administer with slow PCs.

More modern procedures, better suited to adaptive testing using fast PCs, are then discussed. Most of these procedures have a Bayesian flavor to them. Indeed, adaptive testing seems to naturally fit into an empirical or sequential Bayesian framework. For example, the posterior distribution of θ estimated from $k-1$ items can readily be used both to select the kth item and as the prior for the derivation of the next posterior distribution.

When designing a CAT, a test developer must decide how initial and interim ability estimates will be calculated, how items will be selected based on those estimates, and how the final ability estimate will be derived. This chapter provides state-of-the-art alternatives that could guide the development of these core procedures for efficient and robust item selection and ability estimation.

1.2 Classical Procedures

1.2.1 Notation and Some Statistical Concepts

The following notation and concepts are needed. The items in the pool are denoted by $i = 1, ..., I$, whereas the rank of the items in the adaptive test is denoted by $k = 1, \ldots, K$. Thus, i_k is the index of the item in the pool administered as the kth item in the test. The theory in this chapter will be presented for the case of selecting the kth item in the test. The previous $k-1$ items form the set $S_k = \{i_i, \ldots, i_{k-1}\}$; they have responses that are represented by realizations of the response variables $U_{i_1} = u_{i_1}, \ldots, U_{i_{k-1}} = u_{i_{k-1}}$. The set of items in the pool remaining after $k-1$ items have been selected is $R_k = \{1, \ldots, I\} \backslash S_{k-1}$. Item k is selected from this set.

For the sake of generality, the item pool is assumed to be calibrated by the three-parameter logistic (3PL) model. That is, the probability of a correct response on item i is given as

$$p_i(\theta) \equiv \Pr(U_i = 1 \mid \theta) \equiv c_i + (1 - c_i) \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{1.1}$$

where $\theta \in (-\infty, \infty)$ is the parameter representing the ability of the examinee and $b_i \in (-\infty, \infty)$, $a_i \in [o, \infty)$, and $c_i \in [0, 1]$ represent the difficulty, discriminating power, and the guessing probability on item i, respectively. One of the classical item-selection criteria discussed below is based on the three-parameter normal-ogive model,

$$p_i(\theta) \equiv c_i + (1 - c_i)\Phi[a_i(\theta - b_i)], \tag{1.2}$$

where Φ is the normal cumulative distribution function.

The likelihood function associated with the responses on the first $k-1$ items is

$$L(\theta \mid u_{i_1} \dots u_{i_{k-1}}) \equiv \prod_{j=1}^{k-1} \frac{\{\exp[a_{i_j}(\theta - b_{i_j})]\}^{u_{i_j}}}{1 + \exp[a_{i_j}(\theta - b_{i_j})]}. \tag{1.3}$$

The second-order derivative of the loglikelihood reflects the curvature of the observed likelihood function at θ relative to the scale chosen for this parameter. The negative of this derivative is generally known as the observed information measure:

$$J_{u_{i_1} \dots u_{i_{k-1}}}(\theta) \equiv -\frac{\partial}{\partial\theta^2} \ln L(\theta \mid u_{i_1}, \dots, u_{i_{k-1}}). \tag{1.4}$$

The expected value of the observed information measure over the response variables is Fisher's expected information measure:

$$I_{U_{i_1} \dots U_{i_{k-1}}}(\theta) \equiv E[J_{U_{i_1} \dots U_{i_{k-1}}}(\theta)]. \tag{1.5}$$

For the response model in (1.1), the expected information measure reduces to

$$I_{U_{i_1} \dots U_{i_{k-1}}}(\theta) = \sum_{j=1}^{k-1} \frac{[p'_{i_j}(\theta)]^2}{p_{i_j}(\theta)[1 - p_{i_j}(\theta)]}, \tag{1.6}$$

with

$$p'_{i_j}(\theta) \equiv \frac{\partial}{\partial\theta} p_{i_j}(\theta). \tag{1.7}$$

In a Bayesian approach, a prior for the unknown value of the ability parameter, $g(\theta)$, is assumed. Together, the likelihood and prior yield the posterior distribution of θ:

$$g(\theta \mid u_{i_1} \dots u_{i_{k-1}}) = \frac{L(\theta \mid u_{i_1} \dots u_{i_{k-1}})g(\theta)}{\int L(\theta \mid u_{i_1} \dots u_{i_{k-1}})g(\theta)d\theta}. \tag{1.8}$$

Typically, this density is assumed to be uniform or, if the examinees can be taken to be exchangeable, to be an empirical estimate of the ability distribution in the population of examinees. The population distribution is often modeled to be normal. For the response models in (1.1) and (1.2), a normal prior distribution does not yield a normal small-sample posterior distribution, but the distribution is known to converge to normality (Chang & Stout, 1993).

It is common practice in adaptive testing to assume that the values of the item parameters have been estimated with enough precision to treat the estimates as the true parameter values. Under this assumption, the two-parameter logistic (2PL) and one-parameter logistic (1PL) or Rasch models, obtained from (1.1) by setting $c_i = 1$ and $a_i = 0$, subsequently, belong to the exponential family. Because for this family the information measures in (1.4) and (1.5) are identical (e.g., Andersen, 1980, sect. 3.3), the distinction between the two measures has only practical meaning for the 3PL model. This fact will be relevant for some of the Bayesian criteria later in this chapter.

1.2.2 Ability Estimators

The ability estimator after the responses to the first $k-1$ items is denoted as $\widehat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$, but for brevity we will sometimes use $\widehat{\theta}_{k-1}$. Several ability estimators have been used in CAT. In the past, the maximum-likelihood (ML) estimator was the most popular choice. The estimator is defined as the maximizer of the likelihood function in (1.3) over the range of possible θ values:

$$\widehat{\theta}^{\mathrm{ML}}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \arg\max_{\theta} \left\{L(\theta \mid u_{i_1}\ldots u_{i_{k-1}}) : \theta \in (-\infty,\infty)\right\}. \tag{1.9}$$

An alternative is Warm's (1989) weighted likelihood estimator (WLE), which is the maximizer of the likelihood in (1.3) weighted by a function $w_{k-1}(\theta)$:

$$\widehat{\theta}^{\mathrm{WLE}}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \arg\max_{\theta} \left\{w_{k-1}(\theta) L(\theta \mid u_{i_1}\ldots u_{i_{k-1}}) : \theta \in (-\infty,\infty)\right\}, \tag{1.10}$$

where the weight function $w_{k-1}(\theta)$ is defined to satisfy

$$\frac{\partial w_{k-1}(\theta)}{\partial\theta^2} \equiv \frac{H_{k-1}(\theta)}{2I_{k-1}(\theta)}, \tag{1.11}$$

with

$$H_{k-1}(\theta) \equiv \sum_{j=1}^{k-1} \frac{[p'_{i_j}(\theta)][p''_{i_j}(\theta)]}{p_{i_j}(\theta)[1-p_{i_j}(\theta)]}, \tag{1.12}$$

$$p''_{i_j}(\theta) \equiv \frac{\partial^2 p_{i_j}(\theta)}{\partial\theta^2}, \tag{1.13}$$

and $I_{k-1}(\theta) \equiv I_{U_{i_1}\ldots U_{i_{k-1}}}(\theta)$ as defined in (1.5). For a linear test, the WLE is attractive because it has been shown to be unbiased to order n^{-1} (Warm, 1989).

In a more Bayesian fashion, a point estimator of θ can be based on its posterior distribution in (1.8). Posterior-based estimators used in adaptive testing are the Bayes modal (BM) or maximum a posteriori (MAP) estimator and the expected a posteriori (EAP) estimator. The former is defined as the maximizer of the posterior of θ,

$$\widehat{\theta}^{\mathrm{MAP}}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \arg\max_{\theta} \left\{g(\theta \mid u_{i_1}\ldots u_{i_{k-1}}) : \theta \in (-\infty,\infty)\right\}; \tag{1.14}$$

the latter as its expected value:

$$\widehat{\theta}^{\mathrm{EAP}}_{u_{i_1}\ldots u_{i_{k-1}}} \equiv \int \theta g(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) d\theta. \tag{1.15}$$

The MAP estimator was introduced in IRT in Lord (1986) and Mislevy (1986). Use of the EAP estimator in adaptive testing is discussed extensively in Bock and Mislevy (1988).

A more principled Bayesian approach is to refrain from point estimates at all, and use the full posterior of θ as the ability estimator for the examinee. This estimator not only reveals the most plausible value of θ but shows the plausibility of any other value as well. It is common to summarize this uncertainty about θ in the form of the variance of the posterior distribution of θ:

$$\mathrm{Var}(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) \equiv \int [\theta - E(\theta \mid u_{i_1} \ldots u_{i_{k-1}})]^2 g(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) d\theta. \tag{1.16}$$

For the 3PL model, a unique maximum for the likelihood function in (1.3) does not always exist (Samejima, 1973). Also, for response patterns with all items correct or all incorrect, no finite ML estimates exist. However, for linear tests, the ML estimator is consistent and asymptotically efficient. For adaptive tests, the small-sample properties of the ML estimator depend on such factors as the distribution of the items in the pool and the item-selection criterion used. Large-sample theory for the ML estimator for an infinite item pool and one of the popular item-selection criteria will be reviewed later in this chapter.

For a uniform prior, the posterior distribution in (1.8) becomes proportional to the likelihood function over the support of the prior, and the maximizers in (1.9) and (1.14) are equal. Hence, for this case, the MAP estimator shares all the above properties of the ML estimator. For nonuniform prior distributions, the small-sample properties of the MAP estimator depend not only on the likelihood but also on the shape of the prior distribution. Depending on the choice of prior distribution, the posterior distribution may be multimodal. If so, unless precaution is taken, MAP estimation may result in a local maximum.

For a proper prior distribution, the EAP estimator always exists. Also, unlike the previous estimators, it is easy to calculate. No iterative procedures are required; one round of numerical integration generally suffices. This feature used to be important in the early days of computerized adaptive testing but has become less critical now that the typical adaptive testing platform has become much more powerful.

1.2.3 Choice of Estimator

The practice of ability estimation in linear testing has been molded by the availability of a popular computer program (e.g., BILOG, see Zimoski, Muraki, Mislevy & Bock, 2006; MULTILOG, see Thissen, Chen & Bock, 2002). In adaptive testing, such a de facto standard is missing. Most testing programs run their operations using their own software. In developing their software, most of them have taken an eclectic approach to ability estimation. The reason for this practice is that, unlike linear testing, in adaptive testing three different stages of ability estimation can be distinguished: (1) ability estimation to start the item-selection procedure; (2) ability estimation during the test to adapt the selection of the items to the examinee's ability; and (3) ability estimation at the end of the test to report a score for the examinee. Each of these stages involves its own requirements and problems.

Initial Ability Estimation

As already noted, the method of ML estimation does not produce finite estimates for response patterns with all items correct or all incorrect. Because such patterns are likely for the first few items, ML estimation cannot be used for ability estimation at the beginning of the test. Several measures have been proposed to resolve this problem. First, it has been proposed to fix the ability estimate at a small (incorrect items) or large value (correct items) until finite estimates are obtained. Second, ability estimation is sometimes postponed until a larger set of items has been answered. Third, the problem has been an important motive to use Bayesian methods such as the EAP estimator. Fourth, if relevant empirical information on the examinees is available, such as scores on earlier related tests, initial ability estimates can be inferred from this collateral information. A method for calculating such estimates is discussed later in this chapter.

None of these solutions is entirely satisfactory, though. The first two solutions involve an arbitrary choice of ability values and items, respectively. The third solution involves the choice of a prior distribution, which, in the absence of response data, completely dominates the choice of the first item. If the prior distribution is located away from the true ability of the examinee, it becomes counterproductive and can easily produce a longer initial string of correct or incorrect responses than necessary. (Bayesian methods are often said to produce a smaller posterior variance after each new datum, but this statement is not true; see Gelman, Carlin, Stern & Rubin, 1995, sect. 2.2. Initial ability estimation in adaptive testing with a prior at the wrong location is a good counterexample.) As for the fourth solution, although there are no technical objections to using empirical priors (see the discussion later in this chapter), the choice of them should be careful. For example, the use of general background variables easily leads to social bias and should be avoided.

Fortunately, the problem of inferring an initial ability estimate is only acute for short tests, for example, 10-item tests in a battery. For longer tests, of more than 20 to 30 items, say, the ability estimator generally does have enough time to recover from a bad start.

Interim Ability Estimation

Ideally, the next estimates should converge quickly to the true value of the ability parameter. In principle, any combination of ability estimator and item-selection criterion that does this job for the item pool could be used. Although some of these combinations look more "natural" than others (e.g., ML estimation with maximum-information item selection and Bayesian estimation with item selection based on the posterior distribution), practice of CAT has not been impressed by this argument and has often taken a more eclectic approach. For example, a popular choice has been the EAP estimator in combination with maximum-information item selection.

As already noted, in the early days of adaptive testing, the numerical aspects of these estimators used to be important. For example, in the 1970s, Owen's item-selection procedure was an important practical alternative to a fully Bayesian procedure because it did not involve any time-consuming, iterative calculations. However, for modern PCs, computational limitations to CAT no longer exist.

Final Ability Estimation

Although final ability estimates should have optimal statistical properties, their primary function is no longer to guide item selection but to provide the examinee with a meaningful summary of his or her performance in the form of the best possible score. For this reason, final estimates are sometimes transformed to an equated number-correct score on a reference test, that is, a released linear version of the test. The equations typically used for this procedure are the test characteristic function (e.g., Lord, 1980, sect. 4.4) and the equipercentile transformation that equates the ability estimates on the CAT into number-correct scores on a paper-and-pencil version of the test (Segall, 1997). The former is known once the items are calibrated; the latter has to be estimated in a separate empirical study. To avoid the necessity of explaining complicated ML scoring methods to examinees, Stocking (1966) proposed a modification to the likelihood equation such that its solution is a monotonic function of the number-correct score. However, the necessity to adjust the scores afterward can be entirely prevented by imposing appropriate constraints on the item selection that automatically equate the number-correct scores on an adaptive test to reference test (van der Linden, this volume, chap. 2).

The answer to the question of what method of ability estimation is best is intricately related to other aspects of the CAT. First of all, the choice of item-selection criterion is critical. Other aspects that have an impact on ability estimates are the composition of the item pool, whether or not the estimation procedure uses collateral information on the examinees, the choice of the method to control the exposure rates of items, and the presence of content constraints on item selection. The issue will be returned to at the end of this chapter where some of these aspects are discussed in more detail.

1.2.4 Classical Item-Selection Criteria

Maximum-Information Criterion

Birnbaum (1968) introduced the test information function as the main criterion for linear test assembly. The test information function is the expected information measure in (1.5) taken as a function of the ability parameter. Birnbaum's motivation for this function was the fact that, for increasing test length, the variance of the ML estimator is known to converge to the reciprocal of (1.5). In addition, the measure in (1.5) is easy to calculate and additive in the items. In adaptive testing, the maximum-information criterion was immediately adopted as a popular choice. The criterion selects the kth item to maximize (1.5) at $\theta = \widehat{\theta}_{u_{i_1},\dots,u_{i_{k-1}}}$. Formally, it can be presented as

$$i_k \equiv \arg\max_j \left\{ I_{U_1,\dots,U_{k-1},U_j}\,(\widehat{\theta}_{u_{i_1},\dots,u_{i_{k-1}}}) : j \in R_k \right\}. \qquad (1.17)$$

Because of the additivity of the information function, the criterion boils down to

$$i_k \equiv \arg\max_j \left\{ I_{U_j}\,(\widehat{\theta}_{u_{i_1},\dots,u_{i_{k-1}}}) : j \in R_k \right\}. \qquad (1.18)$$

Observe that, though the ML estimator is often advocated as the natural choice, the choice of estimator of θ in (1.18) is open. Also, the maximum-information criterion is often used in the form of a previously calculated information table for a fine grid of θ values (for an example, see Thissen & Mislevy, 1990, Table 5.2).

For a long time, the use of ML estimation of θ in combination with (1.19) as item-selection criterion in CAT missed the asymptotic motivation that existed for linear tests. Recently, such a motivation has been provided by Chang and Ying (2009). These authors show that, for this criterion, the ML estimator of θ converges to the true value with a sampling variance approaching the reciprocal of (1.5). The result holds only for an (infinite) item pool with all possible values for the discrimination parameter in the item pool bounded away from 0 and ∞, and values for the guessing parameter bounded away from 1. Also, for the 3PL model, a slight modification of the likelihood equation is necessary to prevent multiple roots. Because these conditions are mild, the results are believed to provide a useful approximation to adaptive testing from a well-designed item pool. As shown in Warm (1989), the WLE in (1.10) outperforms the ML estimator in adaptive testing. The results by Chang and Ying are therefore expected to hold for the combination of (1.18) with the WLE as well.

Owen's Approximate Bayes Procedure

Owen (1969; see also 1975) was the first to use a Bayesian approach to adaptive testing. His method had the format of a sequential Bayes procedure in which at each stage the previous posterior distribution of the unknown parameter serves as its new prior distribution.

Owen's method was formulated for the three-parameter normal-ogive model in (1.2) rather than its logistic counterpart. His criterion was to choose the kth item such that

$$\left| b_{i_k} - E(\theta \mid u_{i_1} \ldots u_{i_{k-1}}) \right| < \delta \tag{1.19}$$

for a small value of $\delta \geq 0$, where $E(\theta \mid u_{i_1} \ldots u_{i_{k-1}})$ is the EAP estimator defined in (1.15). After the item is administered, the likelihood is updated and combined with the previous posterior to calculate a new posterior. The same criterion is then applied to select a new item. The procedure is repeated until the posterior variance in (1.16) reaches the level of uncertainty about θ the test administrator is willing to tolerate. The last posterior mean is reported to the examinee as his or her final ability estimate.

In Owen's procedure, the selection of the first item is guided by the choice of a normal density for the prior, $g(\theta)$. However, the class of normal priors is not the conjugate for the normal-ogive model in (1.2); that is, they do not yield a normal posterior distribution. Because it was impossible to calculate the true posterior in real time, Owen provided closed-form approximations to the posterior mean and variance and suggested using these to normalize the posterior distribution. The approximation for the mean was motivated by its convergence to the true value of θ in mean square for $k \to \infty$ (Owen, 1975, Theorem 2).

Note that in (1.19), b_i is the only item parameter that determines the selection of the kth item. No further attempt is made to optimize item selection. However, Owen did make a reference to the criterion of minimal preposterior risk (see below) but refrained from pursuing this option because of its computational complexity.

1.3 Modern Procedures

Ideally, item-selection criteria in adaptive testing should allow for two different types of possible errors: (1) errors in the ability estimates and (2) errors in the estimates of the item parameter.

Because the errors in the first ability estimates in the test are generally large, item-selection criteria ignoring them tend to favor items with optimal measurement properties at the wrong value of θ. This problem, which was documented as the attenuation paradox in test theory a long time ago (Lord and Novick, 1968, sect. 16.5), has been largely ignored in adaptive testing. For the maximum-information criterion in (1.18), the "paradox" is illustrated in Figure 1.1, where the item that performs best at the current ability estimate, $\widehat{\theta}$, does worse at the true ability, θ^*. The classical solution for a linear test was to maintain high values for the discrimination parameter but space the values for the difficulty parameter (Birnbaum, 1968, sect. 20.5). This solution goes against the nature of adaptive testing.

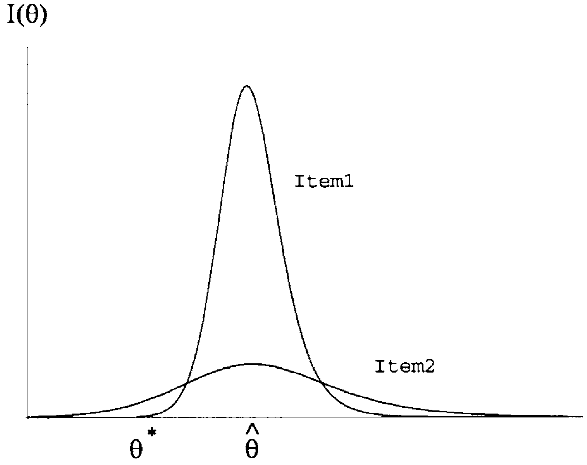


Fig. 1.1 Attenuation paradox in item selection in CAT

Ignoring errors in the estimates of the item parameter values is a strategy without serious consequences as long as the calibration sample is large. However, the first large-scale CAT applications showed that to maintain item pool integrity, the pools had to be replaced much more often than anticipated. Because the costs of replacement are high, the current trend is to minimize the size of the calibration sample. A potential problem for CAT from a pool of items with errors in their parameter values, however, is capitalization on chance. Because the items are selected to be optimal according to a criterion, the test will tend to have both items with optimal true values and less than optimal values with compensating errors in their parameter estimates. Figure 1.2 illustrates the effect of capitalization on chance on ability estimation for a simulation study of a 20-item adaptive test from item pools of varying sizes calibrated with samples of different sizes. For the smaller calibration samples, the error in the ability estimates at the lower-end scale goes up if the item pool becomes larger. This counterintuitive result is due only to capitalization on chance; for other examples of this phenomenon, see van der Linden and Glas (2000).

Recently, new item-selection criteria have been introduced to fix the above problems. These criteria have shown to have favorable statistical properties in extended computer simulation studies. Also, as for their numerical aspects, they can now easily be used in real time on the current generation of PCs.

1.3.1 Maximum Global-Information Criterion

To deal with large estimation error in the beginning of the test, Chang and Ying (1996) suggested replacing Fisher's information in (1.17) by a measure based on Kullback-Leibler information. The Kullback–Leibler information is a general

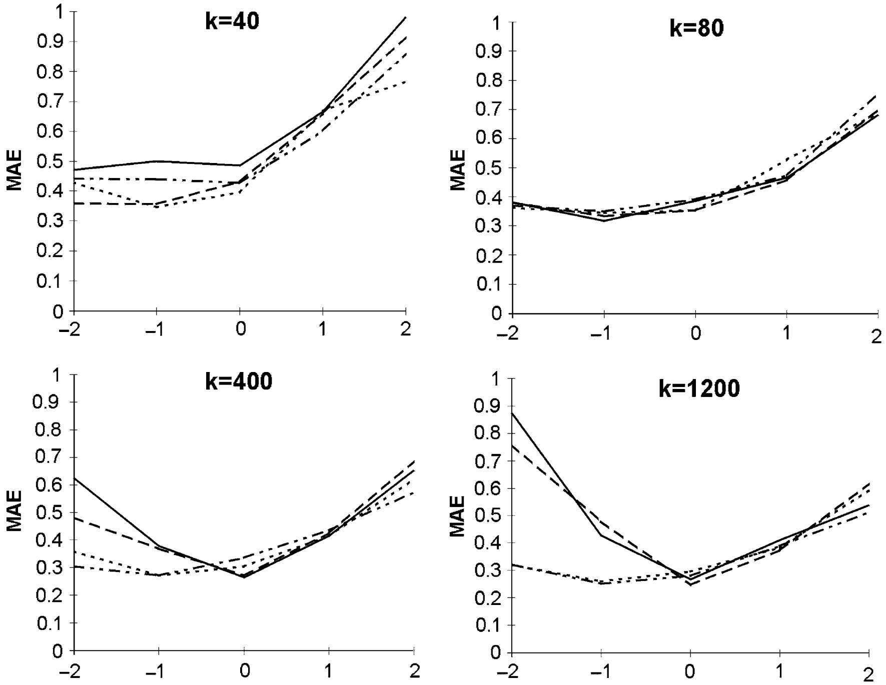


Fig. 1.2 Mean absolute error (MAE) in ability estimation from item pools with $k = 40$, 80, 400, and 1200 items (size of calibration samples: 250: solid; 500: dashed; 1200: dotted; 2500: dashed-dotted)

measure for the "distance" between two distributions. The larger the Kullback–Leibler information, the easier it is to discriminate between two distributions, or equivalently, between the values of the parameters that index them (Lehmann & Casella, 1998, sect. 1.7).

For the response model in (1.1), the Kullback–Leibler measure for the response distributions on the kth item in the test associated with the true ability value (θ_0) of the examinee and the current ability estimate ($\widehat{\theta}_{k-1}$) is

$$K_{i_k}\left(\widehat{\theta}_{k-1}, \theta_0\right) \equiv E\left[\log \frac{L(\theta_0 \mid U_{i_k})}{L(\widehat{\theta}_{k-1} \mid U_{i_k})}\right], \tag{1.20}$$

where the expectation is taken over response variable U_{i_k}. The measure can therefore be calculated as

$$\begin{aligned} K_{i_k}\left(\widehat{\theta}_{k-1}, \theta_0\right) &= p_{i_k}(\theta_0) \log \frac{p_{i_k}(\theta_0)}{p_{i_k}\left(\widehat{\theta}_{k-1}\right)} \\ &\quad + [1 - p_{i_k}(\theta_0)] \log \frac{1 - p_{i_k}(\theta_0)}{1 - p_{i_k}(\widehat{\theta}_{k-1})}. \end{aligned} \tag{1.21}$$

Because of conditional independence between the responses, information in the responses for the first k items in the test can be written as

$$K_k\left(\widehat{\theta}_{k-1}, \theta_0\right) \equiv E\left[\log \frac{L(\theta_0 \mid U_{i_1}, \ldots, U_{i_k})}{L(\widehat{\theta}_{k-1} \mid U_{i_1}, \ldots, U_{i_k})}\right] = \sum_{h=1}^{k} K_{i_h}\left(\widehat{\theta}_{k-1}, \theta_0\right). \tag{1.22}$$

Kullback–Leibler information tells us how well the response variable discriminates between the current ability estimate, $\widehat{\theta}_{k-1}$, and the true ability value, θ_0. Because the true value θ_0 is unknown, Chang and Ying propose replacing (1.20) by its integral over an interval about the current ability estimate, $[\widehat{\theta}_{k-1} - \delta_k, \widehat{\theta}_{k-1} + \delta_k]$, with δ_k a decreasing function of the rank number of the item in the adaptive test. The kth item in the test is then selected according to

$$i_k \equiv \arg\max_j \left\{ \int_{\widehat{\theta}_{k-1}-\delta_k}^{\widehat{\theta}_{k-1}+\delta_k} K_j(\widehat{\theta}_{k-1}, \theta) d\theta : j \in R_k \right\}. \tag{1.23}$$

Evaluation of the criterion will be postponed until all further criteria in this section have been reviewed.

1.3.2 Likelihood-Weighted Information Criterion

Rather than integrating the unknown parameter θ out, as in (1.23), the integral could have been taken over a measure of the plausibility of the possible values of θ. This idea has been advocated by Veerkamp and Berger (1997). Although they presented it for the Fisher information measure, it can easily be extended to the Kullback–Leibler measure.

In a frequentistic framework, the likelihood function associated with the responses $U_{i_1}{=}u_{i_1}, \ldots, U_{i_{k-1}} = u_{i_{k-1}}$ expresses the plausibility of the various values of θ given the data. Veerkamp and Berger proposed weighing Fisher's information with the likelihood function and selecting the kth item according to

$$i_k \equiv \arg\max_j \left\{ \int_{-\infty}^{\infty} L(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}) I_{i_k}(\theta) d\theta : j \in R_k \right\}. \tag{1.24}$$

If maximum-likelihood estimation of ability is used, the criterion in (1.24) places most weight on θ values close to the current ability estimate. In the beginning of the test, the likelihood function is flat, and values away from $\widehat{\theta}_{k-1}$ receive substantial weight. Toward the end of the test the likelihood function tends to become peaked, and nearly all of the weight will go to values close to $\widehat{\theta}_{k-1}$.

Veerkamp and Berger (1997) also specified an interval information criterion that, like (1.23), assumes integration over a finite interval of θ values about the current ability estimate. However, rather than defining an interval with the size of δ_k, they suggested using a confidence interval for θ. The same suggestion would be possible for the criterion in (1.23).

1.3.3 Fully Bayesian Criteria

All Bayesian criteria for item selection involve the use of a posterior distribution of θ. Because a posterior distribution is a combination of a likelihood function and a prior distribution, the basic difference with the previous criterion is the assumption of the latter. Generally, unless reliable collateral information about the examinee is available, the prior distribution of θ should be chosen to be low informative. The question of how to estimate an empirical prior from collateral information is answered in the next section. The purpose of the current section is to review several of the Bayesian criteria for item selection proposed in van der Linden (1998). For a more technical review, see van der Linden and Glas (2007).

Analogous to (1.24), a posterior-weighted information criterion can be defined as

$$i_k \equiv \arg\max_j \left\{ \int I_{U_j}(\theta) g(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}) d\theta : j \in R_k \right\}. \tag{1.25}$$

Generally, the criterion puts more weight on items with their information near the location of the posterior distribution. However, the specific shape of the posterior distribution determines precisely how the criterion discriminates between the information functions of the candidate items.

Note that the criterion in (1.25) is still based on Fisher's expected information in (1.5). Though the distinction between expected and observed information makes practical sense only for the 3PL model, a more Bayesian choice would be to use observed information in (1.4). Also, note that it is possible to combine (1.25) with the earlier Kullback–Leibler measure.

All of the next criteria are based on preposterior analysis. They predict the response distributions on the remaining items in the pool, $i \in R_k$, after $k-1$ items have been administered and then choose the kth item according to the update of a posterior quantity for these distributions. A key element in this analysis is the predictive posterior distribution for the response on item i, which has probability function

$$p(u_i \mid u_{i_1}, \ldots, u_{i_{k-1}}) = \int p(u_i \mid \theta) g(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}) d\theta. \tag{1.26}$$

Suppose item $i \in R_k$ were selected. The examinee would respond correctly to this item with probability $p_i(1 \mid u_{i_1}, \ldots, u_{i_{k-1}})$. A correct response would enable us to update any of the following quantities:

1. the full posterior distribution of θ;
2. any point estimate of the ability value of the examinee, $\widehat{\theta}_k$;

3. the observed information at $\widehat{\theta}_k$; and
4. the posterior variance of θ.

An incorrect response has probability $p_i(0 \mid u_{i_1}, \dots, u_{i_{k-1}})$ and could be used for similar updates. It should be noticed that the update of the observed information at $\widehat{\theta}_k$ involves an update from $\widehat{\theta}_{k-1}$ to $\widehat{\theta}_k$. Because of this, the information measure must be reevaluated at the latter not only for the predicted response to candidate item k but for all previous $k-1$ responses as well.

The first item-selection criterion based on preposterior analysis is the maximum expected information criterion. The criterion maximizes observed information over the predicted responses on the kth item. Formally, it can be represented as

$$\begin{aligned} i_k \equiv \arg\max_j \Big\{ & p_j(0 \mid u_{i_1}, \dots, u_{i_{k-1}}) J_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}(\widehat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}) \\ & + p_j(1 \mid u_{i_1}, \dots, u_{i_{k-1}}) J_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}(\widehat{\theta}_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}) \\ & : j \in R_k \Big\}. \end{aligned} \tag{1.27}$$

If in (1.27) observed information is replaced by the posterior variance of θ, the minimum expected posterior variance criterion is obtained:

$$\begin{aligned} i_k \equiv \arg\min_j \Big\{ & p_j(0 \mid u_{i_1}, \dots, u_{i_{k-1}}) \mathrm{Var}(\theta \mid u_{i_1}, \dots, u_{i_{k-1}}, U_j = 0) \\ & + p_j(1 \mid u_{i_1}, \dots, u_{i_{k-1}}) \mathrm{Var}(\theta \mid u_{i_1}, \dots, u_{i_{k-1}}, U_j = 1) \\ & : j \in R_k \Big\}. \end{aligned} \tag{1.28}$$

The expression in (1.28) is known as the preposterior risk associated with a quadratic loss function for the estimator. Owen (1975) referred to this criterion as a numerically more complicated alternative to his criterion in (1.19).

It is possible to combine the best elements of the ideas underlying the criteria in (1.25) and (1.28) by first weighting observed information using the posterior distribution of θ and then taking the expectation over the predicted responses. The new criterion is

$$\begin{aligned} i_k \equiv \arg\max_j \Bigg\{ & p_j(0 \mid u_{i_1}, \dots, u_{i_{k-1}}) \\ & \cdot \int J_{u_{i_1},\dots,u_{i_{k-1}},U_j=0}(\theta) g(\theta \mid u_{i_1}, \dots, u_{i_{k-1}}, U_j = 0) d\theta \\ & \cdot \int J_{u_{i_1},\dots,u_{i_{k-1}},U_j=1}(\theta) g(\theta \mid u_{i_1}, \dots, u_{i_{k-1}}, U_j = 1) d\theta : j \in R_k \Bigg\}. \end{aligned} \tag{1.29}$$

It is also possible to generalize the criteria in (1.26)–(1.28) to a larger span of prediction. For example, when predicting the responses for the next two items, $(i_k, i_{k'})$, the generalization involves the replacement of the posterior predictive probability function in the above criteria by

$$p(u_{i_k} \mid u_{i_1}, \ldots, u_{i_{k-1}}) p(u_{i_{k'}} \mid u_{i_1}, \ldots, u_{i_k}), \tag{1.30}$$

as well as a similar modification of the other posterior updates. Although the optimization is over pairs of candidates for items k and $k+1$, better adaptation is obtained if the candidate for item k is actually administered but the other item is returned to the pool, whereupon the procedure is repeated. Combinatorial problems inherent in the application of the procedure with larger item pools and spans of prediction can be avoided by using a trimmed version of the pool with unlikely candidate items left out.

1.3.4 Bayesian Criteria with Collateral Information

As indicated earlier, an informative prior located at the true value of θ would give Bayesian ability estimation its edge. For a large variety of item-selection criteria, such a prior would not only yield finite initial ability estimates but also improve item selection and speed up convergence of the estimates during the test. If useful collateral information on the examinee exists, for example, in the form of previous achievements or performances on a recent related test, an obvious idea is to infer the initial prior from this information. An attractive source of collateral information during the test is the response times (RTs) on the items. They can be used for a more effective update of the posterior distribution of θ during the rest of the test. This section deals with the use of both types of collateral information.

Statistically, no objections whatsoever exist against this idea; when the interest is only in ML or Bayesian estimation of θ, item-selection criteria based on collateral information are known to be ignorable (Mislevy & Wu, 1988). Nevertheless, if policy considerations preclude the use of collateral information in test scores, a practical strategy is to still use the information to improve the design of the test but to calculate the final ability estimate only from the last likelihood function for the examinee.

Initial Empirical Prior Distribution

Procedures for adaptive testing with the 2PL model with the initial prior distribution regressed on predictor variables are described in van der Linden (1999). Let the predictor variables be denoted by X_p, $p = 0, \ldots, P$. The regression of θ on the predictor variables can be modeled as

$$\theta = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon, \tag{1.31}$$

with

$$\varepsilon \sim N(0, \sigma^2). \tag{1.32}$$

Substitution of (1.30) into the response model gives

$$p_i(\theta) = \frac{\exp[a_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon - b_i)]}{1 + \exp[a_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon - b_i)]}. \tag{1.33}$$

For known values for the item parameters, the model amounts to logistic regression with examinees' values of ε missing. The values of the parameters $\beta_1, \ldots, \beta_P$ and σ can be estimated from data using the EM algorithm. The estimation procedure boils down to iteratively solving two recursive relationships given in van der Linden (1999, Eqs. 16–17). These equations are easily solved for a set of pretest data. They also allow for an easy periodical update of the parameter estimates from response data when the adaptive test is operational.

If the item selection is based on point estimates of ability, the regressed value of θ on the predictor variables,

$$\widehat{\theta}_0 = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P, \tag{1.34}$$

can be used as the prior ability estimate for which the initial item is selected. If the items are selected using a full prior distribution for θ, the choice of prior following (1.32)–(1.33) is

$$g(\theta) \equiv N\left(\widehat{\theta}_0, \sigma\right). \tag{1.35}$$

Observe that both (1.34) and (1.35) provide an individualized initialization for the adaptive test: Different examinees will start at different initial ability estimates. The procedure therefore offers more than statistical advantages. Initialization at the same ability estimate for all examinees leads to first items in the test that are always chosen from the same subset in the pool. Hence, they become quickly overexposed, and the testing program becomes vulnerable to security breaches. On the other hand, the empirical initialization of the test above entails a variable entry point to the pool, and hence offers a more even exposure of its items.

Item Selection with RTs as Collateral Information

RTs on test items are recorded automatically during adaptive testing, They are also a potentially rich source of collateral information about the examinee's ability. One possible use of RTs is as an additional source of information for the update of the posterior distribution of θ during testing. This procedure becomes possible as soon as we have a model for the RT distributions on the items in the pool that is statistically linked to the response model.

The modeling framework used in this demonstration of the procedure is a hierarchical framework with (i) the 3PL model and a lognormal model for the RT distribution as distinct first-level models and (ii) a bivariate normal model for the distribution of the person parameters in these models as a second-level model. The lognormal model is a normal model for the log of the RTs with $\tau_j \in (-\infty, \infty)$ as the speed for examinee j and $\beta_i \in (-\infty, \infty)$ and $\alpha_i \in (0, \infty)$ are the time intensity and discrimination parameters for item i. The model equation is

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))\right]^2\right\}. \tag{1.36}$$

At the second level,

$$(\theta, \tau) \sim \mathrm{MVN}(\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\Sigma}_{\mathcal{P}}), \tag{1.37}$$

with mean vector $\boldsymbol{\mu}_{\mathcal{P}} = (\mu_\theta, \mu_\tau)$ and covariance matrix $\boldsymbol{\Sigma}_{\mathcal{P}}$ for the person parameters in the population of examinees. More details on the model and the estimation of its parameters are given in Klein Entink, Fox, and van der Linden (2009) and van der Linden (2007).

The idea is to adjust the posterior distribution of θ in (1.8) using simultaneous updates of its two components:

1. An update of the likelihood $L(\theta \mid u_{i_1} \ldots u_{i_{k-1}})$ using the response on the item. This is the regular Bayesian update of a posterior distribution.
2. The retrofitting of the original prior $g(\theta)$ in (1.8) using the RTs on the items. The new prior distribution is the posterior predictive density of θ given the RTs, that is,

$$f(\widetilde{\theta} \mid \mathbf{t}_{k-1}) = \int f(\theta \mid \tau) f(\tau \mid \mathbf{t}_{k-1}) d\tau. \tag{1.38}$$

For the models in (1.36)–(1.37), use of the log RTs leads to a normal density for (1.38) with closed-form expressions for the mean and standard deviation that are easily calculated from the known item parameters and RTs on the previous items.

Observe that (1.38) leads to an individualized prior that is continuously improved during the test using additional information obtained from the individual test taker. The result is faster convergence of the posterior distribution of θ as well as the improved item exposure mentioned above relative to the case of a common fixed prior distribution for all examinees.

The procedure is demonstrated empirically in van der Linden (2008). Figure 1.3 shows the results from this study for adaptive tests of $n = 10$ and 20 items for various degrees of correlation between θ and τ. Even for a modest correlation of $\rho_{\theta\tau} = 0.2$, the improvement for the EAP estimator used as final estimate in this study is already conspicuous. In fact, a comparison between the two panels shows that for $\rho_{\theta\tau} = 0.2$ the MSE function for $n = 10$ already has a similar shape as the MSE function for $n = 20$ without the use of RTs. Also, observe that the curves for the conditions with RTs are generally flatter than the one for the case without. The empirical item pool used in this study was relatively scarce at the lower end of the scale (fewer easy items). The use of the RTs nicely compensated for this scarcity.

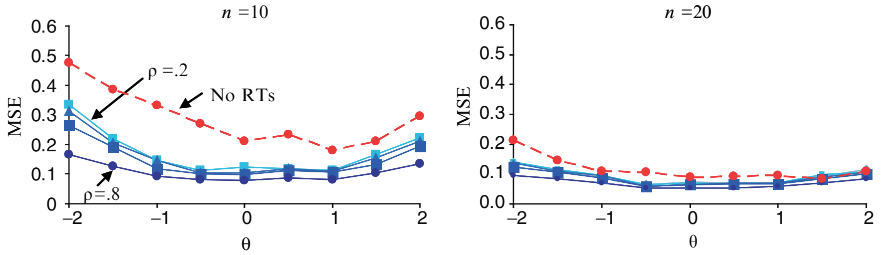


Fig. 1.3 MSE functions of EAP estimator of θ for item selection without RTs (dashed line) and with RTs with $\rho_{\theta\tau} = 0.2, 0.4, 0.6$, and 0.8 (solid lines; the darker the line, the higher the correlation) for tests of $n = 10$ and 20 items. [Reproduced with permission from W. J. van der Linden (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5-20.]

1.3.5 Bayesian Criteria with Random Item Parameters

If the calibration sample is small, errors in the estimates of the values of the item parameters should not be ignored but dealt with explicitly when estimating θ in adaptive testing. A Bayesian approach would not fix the item parameters at point estimates but leave them random, using their posterior distribution given all previous responses in the ability estimation procedure. Tsutakawa and Johnson (1990) describe this empirical Bayes approach to ability estimation for responses to linear tests. Their procedure can easily be modified for application in adaptive testing.

The modification is as follows: Let $\mathbf{y}$ be the matrix with response data from all previous examinees. For brevity, the parameters (a_i, b_i, c_i) for the items in the pool are collected into a vector $\boldsymbol{\xi}$. Suppose a new examinee has answered $k-1$ items, and we need the update of his or her posterior distribution for the selection of item k. Given a prior for $\boldsymbol{\xi}$, the derivation of the posterior distribution of this vector of item parameters is standard. The result is the posterior density $g(\boldsymbol{\xi} \mid u_{i_1}, \ldots, u_{i_{k-1}}, \mathbf{y})$.

Using the assumptions in Tsutakawa and Johnson (1990), the posterior distribution of θ after item $k-1$ can be updated as

$$g(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}, \mathbf{y}) = \frac{g(\theta) \int p(u_{i_{k-1}} \mid \theta, \boldsymbol{\xi}) g(\boldsymbol{\xi} \mid u_{i_1}, \ldots, u_{i_{k-2}}, \mathbf{y}) d\boldsymbol{\xi}}{p(u_{i_{k-1}} \mid u_{i_1}, \ldots, u_{i_{k-2}}, \mathbf{y})}. \tag{1.39}$$

Key in this expression is the replacement of the likelihood associated with the response to the last item, i_{k-1}, by its average over the posterior distribution of the item parameters given all previous data, $g(\boldsymbol{\xi} \mid u_{i_1}, \ldots, u_{i_{k-2}}, \mathbf{y})$. Such averaging is the Bayesian way of accounting for posterior uncertainty in unknown parameters. Given the posterior distribution of θ, the posterior predictive probability function for the response on item i_k can be derived as

$$p(u_{i_k} \mid u_{i_1}, \ldots, u_{i_{k-1}}, \mathbf{y}) \equiv \int p(u_{i_k} \mid \theta) g(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}}, \mathbf{y}) d\theta. \tag{1.40}$$

Once (1.40) is calculated, it can be used in one of the criteria in (1.25) or (1.27)–(1.29).

In spite of all our current computational power, a real-time update of the posterior distribution of the item parameters, $g(\boldsymbol{\xi} \mid u_{i_1}, \ldots, u_{i_{k-1}}, \mathbf{y})$, is prohibitive, due to the evaluation of complex multiple integrals. However, in practice, it makes sense to update the posterior only periodically, after prior screening of the new set of response patterns for possible aberrant behavior by some of the examinees or compromise of the items. When testing the next examinees, the posterior distribution of $\boldsymbol{\xi}$ then remains fixed until the next update. The resulting expression in (1.39)–(1.40) can easily be calculated in real time using appropriate numerical integration. Alternatively, we could use the simplifying assumptions for the update of $g(\boldsymbol{\xi} \mid \mathbf{y})$ given in Tsutakawa and Johnson (1990).

A different need for item-selection criteria to deal with random item parameters arises in adaptive testing with rule-based item generation. In this application, the traditional pool of discrete items is replaced by a pool of computer-generated items, or, more challenging, the items are generated by computer algorithms in real time. The first experiments with rule-based item generation typically involve two different types of rules. One type is based on the structural aspects of the items (generally referred to as "radicals") found in a cognitive analysis of the content domain. The second type is rules for item cloning, that is, for generating a family of items that look different but are based on the same combination of radicals. Within the families, the items thus differ only in their surface features (generally referred to as "incidentals"). Recent examples of the use of such types of rules are given in Freund, Hofer, and Holling (2008) and Holling, Bertling, and Zeuch (in press).

The structure of an item pool with items nested in families with the same combinations lends itself nicely to hierarchical response modeling with a regular response model for each individual item, such as the one in (1.1), as first-level models and a separate second-level model for each family to describe the distribution of its item parameters. Generally, the differences in item parameters between families will be much larger than within families. Nevertheless, explicit modeling of the within-family differences is much better than ignoring them and treating all items within a family as psychometrically equivalent. Hierarchical response models for this purpose have been proposed by Glas and van der Linden (2001, 2003; see also Sinharay, Johnson & Williamson, 2003) and Geerlings, van der Linden and Glas (2009). The first model is treated more in detail elsewhere in this volume (Glas, van der Linden & Geerlings, chap. 15); this chapter should be consulted for item calibration and model fit issues.

Let the pool be generated to have item families $p = 1, \ldots, P$, each with distribution $p(\boldsymbol{\xi} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ of its item parameters $\boldsymbol{\xi} = (a, b, c)$. In the hierarchical model by van der Linden and Glas, each family has a distinct normal distribution for its item parameters. The item pool is assumed to be calibrated using samples of items from each family to estimate its mean $\boldsymbol{\mu}_p$ and covariance $\boldsymbol{\Sigma}_p$.

Item selection from a pool of calibrated items proceeds along the following two steps:

1. adaptive selection of a family; i.e., identification of the family with the best match of its $\boldsymbol{\mu}_p$and $\boldsymbol{\Sigma}_p$ with the current θ
2. estimate; and
3. random selection of an item from the family.

More formally, in a Bayesian framework, the procedure is as follows. The update of the posterior distribution of θ after these $k-1$ items is given by

$$p(\theta \mid \mathbf{u}_{k-1}) \propto g(\theta) \prod_{p=1}^{k-1} \int p(u_p \mid \theta, \boldsymbol{\xi}_p) p(\boldsymbol{\xi}_p | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) d\boldsymbol{\xi}_p. \tag{1.41}$$

The first step is to select the kth family to be optimal at this posterior distribution. As an example, item selection based on the minimum expected posterior variance criterion in (1.32) is proposed. The only necessary change in this criterion is an adjustment of the posterior predicted distribution of the responses on the candidate item in (1.32) to those for a random item from a candidate family. Consider family p as the candidate for the kth family in the test; this candidate is denoted as p_k. The posterior predicted distribution for the response on a random item from this family has probability function

$$p(u_{p_k} \mid \mathbf{u}_{k-1}) = \int \left[\int p(u_{p_k} \mid \theta, \boldsymbol{\xi}_{p_k}) p(\boldsymbol{\xi}_{p_k} | \boldsymbol{\mu}_{p_k}, \boldsymbol{\Sigma}_{p_k}) d\boldsymbol{\xi}_{p_k} \right] p(\theta \mid \mathbf{u}_{k-1}) d\theta. \tag{1.42}$$

Observe that in this expression we first average the response probability over the distribution of the item parameters for family p_k to allow for the random sampling of an item from it, and then average the result over the posterior distribution of the ability of the examinee. This expression is used in (1.32) to identify the best family in the pool. The second step is to randomly sample an item from this family.

For an exploration of the behavior of this criterion using simulated adaptive testing, see Glas and van der Linden (2003).

1.3.6 Miscellaneous Criteria

The item-selection criteria presented thus far were statistically motivated. An item-selection procedure that addresses both a statistical and a more practical goal is the method of multistage α-stratified adaptive testing proposed in Chang and Ying (1999). The method was introduced primarily to reduce the effect of ability estimation error on item selection. As illustrated in Figure 1.1, if the errors are large, an item with a lower discrimination parameter value is likely to be more efficient over a larger range of θ values than one with a higher value.

These authors therefore propose stratifying the pool according to the values of the discrimination parameter for the items and restricting item selection to strata with increasing values during the test. In each stratum, items are selected according to the criterion of minimum distance between the value of the difficulty parameter and the current ability estimate. In a recent theoretical study, the authors showed why early selection of highly discriminating items after a few initial incorrect responses is detrimental to the estimation of θ (Chang & Ying, 2008). The procedure also provides a remedy to the problem of uneven item exposure in CAT. Because items with a lower discrimination parameter have an equal chance of being chosen, uneven exposure of the higher parameters is prevented.

To deal with capitalization on calibration error (see Figure 1.2), it may be effective to cross-validate item parameter estimation during adaptive testing. A practical way of doing so is to split the calibration sample into two parts, and estimate the item parameters separately for each part. One set of estimates can be used to select the items; the other to update the ability estimate after the examinee has taken them. Item selection then still tends to capitalize on the errors in the estimates in the first set, but the effects on ability estimation are neutralized by using the second set of estimates. Conditions under which this neutralization offsets the loss in precision due to calibration from a smaller sample were studied in van der Linden and Glas (2001).

Most of the item-selection criteria in this chapter select items for which the examinee has a probability of a correct response close to 0.5. For some educational applications, for instance, formative assessment to monitor the achievements of students during class work, such response probabilities may be less motivating. Eggen and Verschoor (2006) examined the effects of modifying item selection to produce higher or lower response probabilities. Direct selection on such probabilities worked well for the 1PL model but not for models with varying discrimination parameters, for which selection at a deliberate shift in the ability estimate worked better.

A final suggestion for item selection in adaptive testing was offered in Wainer, Lewis, Kaplan, and Braswell (1992). As selection criterion they used the posterior variance between the subgroups that scored the item in the pretest correctly and incorrectly. Results from an empirical study of this criterion are given in Schnipke and Green (1995).

1.3.7 Evaluation of Item-Selection Criteria and Ability Estimators

The question of which combination of item-selection criterion and ability estimation is best is too complicated for analytic treatment. Current statistical theory provides us only with asymptotic conclusions.

A well-known result from Bayesian statistics is that for $k \to \infty$, the posterior distribution $g(\theta \mid u_{i_1}, \ldots, u_{i_{k-1}})$ converges to degeneration at the true value of θ. Hence, it can be concluded that all posterior-based ability estimation and item-selection procedures reviewed in this chapter produce identical asymptotic results. Also, the result by Chang and Ying (2009) referred to earlier shows that for maximum-information item selection, the ML estimator converges to the true value of θ as well. The WLE in (1.10) is expected to show the same behavior.

However, particularly for adaptive testing with its much shorter test length, small-sample comparisons of estimators and criteria are more relevant. For such comparisons we have to resort to simulation studies.

Relevant studies have been reported in Chang and Ying (1999), van der Linden (1998), Veerkamp and Berger (1997), Wang, Hanson, and Lau (1999), Wang and Vispoel (1998), Weiss (1982), Weiss and McBride (1984) and Warm (1989), among others. Sample results for the bias and mean-square error (MSE) functions for five different combinations of ability estimators and item-selection criteria are given in Figures 1.4 and 1.5. All five combinations show the same slight

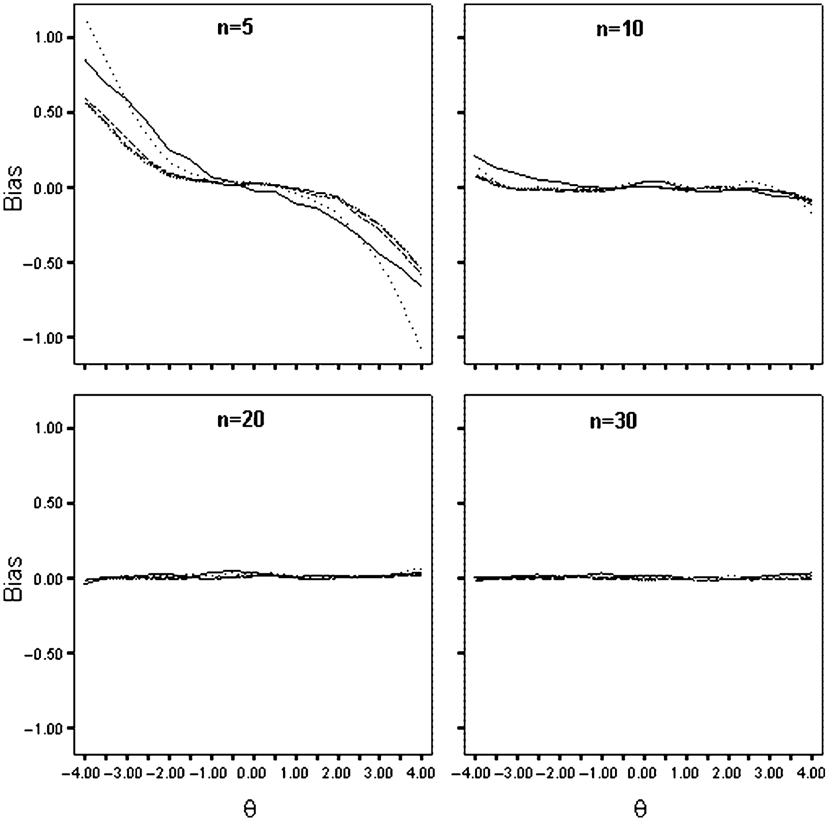


Fig. 1.4 Bias functions for five item-selection criteria after $n = 5, 10, 20, 30$ items (maximum-information with MLE: solid; maximum-posterior weighted Information: dotted; maximum expected information: dashed-dotted; maximum expected posterior variance: dashed; maximum expected posterior weighted information: finely dotted). [Reproduced with permission from W. J. van der Linden (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216.]

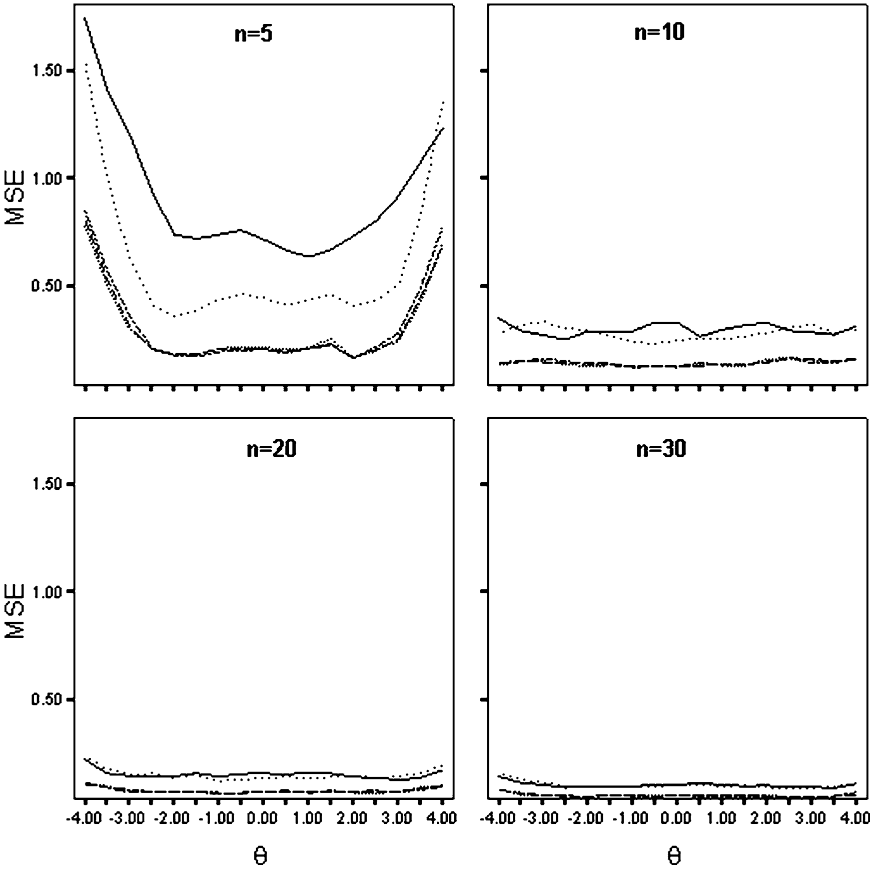


Fig. 1.5 MSE functions for five item-selection criteria after $n = 5$, 10, 20, 30 items (maximum-information with MLE: solid; maximum-posterior weighted information: dotted; maximum expected information: dashed-dotted; maximum expected posterior variance: dashed; maximum expected posterior weighted information: finely dotted). [Reproduced with permission from W. J. van der Linden (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216.]

inward bias for $n = 10$, which disappears completely for $n = 20$ and 30. Note that the bias for the ML estimators in Figure 1.4 has a direction opposite the one in the estimator for a linear test (e.g., Warm, 1989). This result is due to a feedback mechanism created by the combination of the contributions of the items to the bias in the estimator and the maximum-information criterion (van der Linden, 1998).

MSE functions for linear tests are typically U-shaped with the dip at the θ values where the items are located. However, as Figure 1.5 shows, for the same item-selection criteria as in Figure 1.4, after $n = 10$ items all MSE functions are already flat. The best functions were obtained for the criteria in (1.27)–(1.29). Each of these criteria was based on preposterior analysis. Hence, a critical element in the success of an item-selection criterion seems to be its use of posterior predictive probability functions to predict the item responses on the remaining items in the pool. As revealed by the comparison between the MSE functions for the maximum-information and maximum posterior-weighted information criteria in Figure 1.5, simply using the posterior distribution of θ appears to have little effect.

Weiss (1982) reported analogous results for the maximum-information criterion and Owen's criterion in (1.19). In Wang and Vispoel's (1998) study, the behavior of the ML, EAP, and MAP estimators in combination with the maximum-information criterion were compared with Owen's criterion. For a 30-item test from a real-world item pool, the three Bayesian procedures behaved comparably, whereas the ML estimator produced a worse standard error but a better bias function. Wang, Hanson, and Lau (1999) reported several conclusions for modifications of the ML and Bayesian estimators intended to remove their bias. A sobering result was given by Sympson, Weiss, and Ree (see Weiss, 1982, p. 478) who, in a real-world application of the maximum-information and Owen's selection criterion, found that approximately 85% of the items selected by the two criteria were the same. However, the result may largely be due to the choice of a common initial item for all examinees.

1.4 Concluding Remarks

As noted in the introduction section of this chapter, methods for item selection and ability estimation within a CAT environment are not yet as refined as those currently employed for linear testing. Hopefully, though, this chapter has provided evidence that substantial progress has been made in this regard. Modern methods have begun to emerge that directly address the peculiarities of adaptive testing, rather than relying on simple modifications of rules used in linear testing situations. Recent analytical studies with theoretical frameworks to evaluate the different procedures have been especially good to see. In addition, the constraints on timely numerical computations imposed by older and slower PCs have all but disappeared.

The studies discussed in this chapter only relate to a small part of the conditions that may prevail in an adaptive testing program. Clearly, programs can differ in the type of item-selection criterion and ability estimator they use. However, they can also vary in numerous other ways, such as the length of the test and whether the length is fixed or variable; the size and composition of the item pools; the availability of useful collateral information about the examinees; the size and composition of the calibration samples; the ability to update item parameter estimates using operational test data; the use of measures to control item exposure rates; and the content constraints imposed on the item-selection process. Important trade-offs exist among several of these factors, which also interact in their effect on the statistical behavior of the final ability estimates.

Given the complexities of a CAT environment and the variety of approaches (some untested) that are available, how should one proceed? One method would be to delineate all the relevant factors that could be investigated and then undertake an extensive simulation study—a daunting task at best. A more practical strategy is to study a few feasible arrangements in order to identify a suitable, though not necessarily optimal, solution for a planned adaptive testing program.

References

Andersen, E. B. (1980). *Discrete statistical models with social sciences applications*. Amsterdam: North-Holland.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. & Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444.

Chang, H.-H. & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*, 37–52.

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.

Chang, H.-H. & Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Chang, H.-H. & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*, 441–450.

Chang, H.-H. & Ying, Z. (2009). Nonlinear sequential designs for logistic item response models with applications to computerized adaptive tests. *The Annals of Statistics, 37*, 1466–1488.

Chen, S., Hou, L. & Dodd, B. G. (1998). A comparison of maximum-likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*, 569–595.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16*, 327–343.

De Ayala, R. J., Dodd, B. G. & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5*, 17–34.

Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal testing with easy and difficult items in computerized adaptive testing. *Applied Psychological Measurement, 30*, 379–393.

Freund, P. A., Hofer, S. & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement, 32*, 195–210.

Geerlings, H., van der Linden, W. J. & Glas, C. A. W. (2009). *Modeling rule-based item generation*. Submitted for publication.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Glas, C. A. W. & van der Linden, W. J. (2001). *Modeling item variability in item parameters in item response models* (Research Report 01-11). Enschede, the Netherlands: Department of Educational Measurement and Data Analysis, University of Twente.

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item clones. *Applied Psychological Measurement, 27*, 247–261.

Gulliksen, H. (1950). *Theory of mental tests*. Hillsdale, NJ: Erlbaum.

Holling, H., Bertling, J. P. & Zeuch, N. (in press). Probability word problems: Automatic item generation and LLTM modelling. *Studies in Educational Evaluation.*

Klein Entink, R. H., Fox, J.-P. & van der Linden, W. J. (2009). A multivariate multilevel approach to simultaneous modeling of accuracy and speed on test items. *Psychometrika, 74*, 21–48.

Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation*. New York: Springer-Verlag.

Lord, F. M. (1971). The self-scoring flexilevel test. *Journal of Educational Measurement, 8,* 147–151.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Mislevy, R. J. & Wu, P.-K. (1988). *Inferring examinee ability when some items response are missing* (Research Report 88-48-ONR). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedogogiske Institut.

Roberts, J. S., Lin, Y. & Laughlin, J. E. (2001). Computerized adaptive testing with the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 177–192.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in latent trait theory. *Psychometrika*, *38*, 221–233.

Samejima, F. (1993). The bias function of the maximum-likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58*, 195–210.

Schnipke, D. L. & Green, B. F. (1995). A comparison of item selection routines in linear and adaptive testing. *Journal of Educational Measurement*, *32*, 227–242.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 181–198). Washington, DC: American Psychological Association.

Sinharay, S., Johnson, M. S. & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*, 295–313.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21*, 365–389.

Thissen, D., Chen, W.-H. & Bock, R. D. (2002). *Multilog 7: Analysis of multi-category response data* [Computer program and manual]. Lincolnwood, IL: Scientific Software International.

Thissen, D. & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103–134). Hillsdale, NJ: Lawrence Erlbaum.

Tsutakawa, R. K. & Johnson, C. (1990). The effect of uncertainty on item parameter estimation on ability estimates. *Psychometrika*, *55*, 371–390.

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216.

van der Linden, W. J. (1999). A procedure for empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21–29.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287–308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5–20.

van der Linden, W. J. & Glas, C. A. W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 13*, 35–53.

van der Linden, W. J. & Glas, C. A. W. (2001). Cross-validating item parameter estimation in computerized adaptive testing. In A. Boomsma, M. A. J. van Duijn & T. A. M. Snijders (Eds.), *Essays on item response theory* (pp. 205–219). New York: Springer-Verlag.

van der Linden, W. J. & Glas, C. A. W. (2007). Statistical aspects of adaptive testing. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 27: Psychometrics) (pp. 801–838). Amsterdam: North-Holland.

van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T. & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26*, 393–411.

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Item-selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203–226.

Wainer, H., Lewis, C., Kaplan, B. & Braswell, J. (1991). Building algebra testlets: A comparison of hierarchical and linear structures. *Journal of Educational Measurement*, *28*, 311–323.

Wang, T., Hanson, B. A. & Lau, C.-M. A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement*, *23*, 263–278.

Wang, T. & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109–135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory with tests of finite length. *Psychometrika*, *54*, 427–450.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *4*, 473–285.

Weiss, D. J. & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement*, *8*, 273–285.

Zimoski, M. F., Muraki, E., Mislevy, R. & Bock, D. R. (2006). *BILOG-MG 3 for Windows* [Computer program and manual]. Lincolnwood, IL: Scientific Software International.

Chapter 2
Constrained Adaptive Testing with Shadow Tests

Wim J. van der Linden

2.1 Introduction

The intuitive principle underlying adaptive testing is that a test has better measurement properties if the difficulties of its items match the ability of the examinee. Items that are too easy or difficult have predictable responses and cannot provide much information about the ability of the examinee. The first to formalize this principle was Birnbaum (1968). The information measure he used was Fisher's well-known information in the sample. For dichotomous response models, the measure is defined as

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \frac{(P'(\theta))^2}{P(\theta)\,[1 - P(\theta)]}, \tag{2.1}$$

where $P_i(\theta)$ is the probability of a correct response to item $i = 1, \ldots, n$ for an examinee with ability θ, $I_i(\theta)$ is the information in the examinee's response to item i, and $I(\theta)$ is the information in his or her joint responses to the test.

For the one-parameter logistic (1PL) model, the information measure is maximal when the value of the difficulty parameter b_i is equal to the examinee's θ. The same relation holds for the two-parameter (2PL) model, though the maximum is now monotonically increasing in the value of the discrimination parameter of the items, a_i. The empirical applications discussed later in this chapter are all based on response data fitting the three-parameter (3PL) model,

$$P_i(\theta_j) \equiv c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}. \tag{2.2}$$

For this model, the optimal value of the item-difficulty parameter is greater than the ability of the examinee due to the possibility of guessing on the items. The

W.J. van der Linden (✉)
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_2,

difference between the optimal value and the ability of the examinee is known to be monotonically increasing with the guessing parameter, c_i.

Both test theoreticians and practitioners immediately adopted the information measure in (2.1) as their favorite criterion for the assembly of fixed test forms. The fact that item information additively contributes to the test precision has greatly enhanced its popularity. Though other criteria of item selection have been introduced later (for a review, see van der Linden and Pashley, this volume, chap. 1), the most frequently used criterion in computerized adaptive testing (CAT) has also been the one based on the information measure in (2.1).

Though adaptive testing research was initially mainly motivated by the intention to make test scores statistically more informative, the first testing programs to make the transition to CAT quickly discovered that adaptive testing operating only on this principle would lead to unrealistic results. For example, if items are only selected to maximize the information in the ability estimator, test content may easily become unbalanced for some ability levels. If examinees happen to learn about this feature, they may change their test preparations and, as a result, item calibrations might cease to be valid. Furthermore, even a simple attribute such as the answer key becomes a problem if the adaptive test administrations produced highly disproportionate use of one of the keys. Lower-ability examinees might then start benefiting from patterned guessing, whereas some of the more able examinees might become anxious and begin second-guessing their answers to previous items. As examinees get different selections of items, and items differ greatly in the amount of time they take, without any further provisions, adaptive tests may have a tendency to become differentially speeded—a condition leading to negatively biased scores for examinees who happen to get more time-intensive items.

More examples of necessary nonstatistical specifications for adaptive tests are easy to provide. In fixed-form testing, violations of such specifications are generally caught when candidate test forms are reviewed, but in adaptive testing this safety net is not available and the item-selection algorithm has to guarantee automatic satisfaction of the set of specifications. In fact, what most testing programs want if they make the transition from linear to adaptive testing is test administrations that meet the same set of test specifications as their old test forms (i.e., have exactly the same "look and feel") but that are much shorter because of a better adaptation to the ability levels of the individual examinees.

Formally, each test specification an adaptive test has to meet imposes a constraint on the selection of the items from the pool. As a consequence, a CAT algorithm that combines maximization of statistical information with the realization of several nonstatistical specifications can be viewed as an algorithm for *constrained sequential optimization*. The *objective function* to be optimized is the statistical information in the test items at the current ability estimate. All other specifications are the *constraints* subject to which the optimization has to take place.

The goal of this chapter is to develop this point of view further and discuss a general method of constrained sequential optimization for application in adaptive testing. This method has proven to be successful in several applications. The basic principle underlying the method is to implement all constraints through a series of *shadow tests* assembled to be optimal at the updated ability estimates of the examinee. The items to be administered are selected from these shadow tests rather than directly from the item pool. Use of the method will be illustrated for item pools from well-known large-scale testing programs.

2.2 Review of Existing Methods for Constrained CAT

2.2.1 Item-Pool Partitioning

An adaptation of the maximum-information criterion to make item selection balanced with respect to test content was presented in Kingsbury and Zara (1991). Their proposal was to partition the item pool according to the item attributes. While testing, the numbers of items selected from each class in the partition are recorded. In order to maintain content balance, the algorithm follows a minimax principle, selecting the next item from the class for which the largest number of items is lacking. A further modification was proposed to prevent items from being readministered to examinees that have taken the same test earlier. Finally, to reduce the exposure of the most informative items in the pool, these authors suggested not to select the most informative item from the current class in the partition but to pick one at random from among the κ best items in the class. The last adaptation was used in an early version of the CAT-ASVAB (Hetter & Sympson, 1997).

2.2.2 Weighted-Deviation Method

A more general approach is the weighted deviation method (WDM) by Swanson and Stacking (1993). In their approach, all content specifications for the CAT are formulated as a series of upper and lower bounds on the numbers of items to be selected from the various content classes. Likewise, a target for the test information is formulated as a series of upper and lower bounds on its values. A weighted sum of the deviations from all bounds is taken as the objective function, with the weights reflecting the desirability of the individual specifications. The items in the adaptive test are selected one at a time to minimize the objective function.

2.2.3 Maximum Priority Index Method

The maximum priority index method (Cheng & Chang, 2009) is related to the WDM in that it also weighs deviations from bounds on the numbers of items to be selected from content classes, and then selects the item with the highest value for a weighted index. But a critical difference between the two methods is that all deviations are scaled relative to the size of their bound, and the weights no longer have to be set to allow for scaling differences between constraints in addition to their relative importance. Because the rescaling only works for upper bounds, constraints with lower bounds are dealt with through an application of a two-phase selection method introduced by Cheng, Chang and Yi (2007), in which the items are initially selected treating the lower bounds in the constraints as upper bound, but the method focuses on the upper bound in a constraint as soon as its lower bound is met.

2.2.4 Testlet-Based Adaptive Testing

Some of the first to address the necessity of combining content specification and statistical criteria in item selection for CAT were Wainer and Kiely (1987). Their solution was to change the size of the units in the item pool. Rather than discrete items, they proposed the use of "testlets", that is, bundles of items related to sets of content specifications that are to be selected as intact units. Testlets are preassembled to have a fixed item order. New psychometric theory for testlet-based adaptive testing is offered by Glas, Wainer and Bradlow (2000), Vos and Glas (this volume, chap. 20), Wainer, Bradlow and Du (2000), and Wainer, Bradlow, and Wang (2007).

2.2.5 Multistage Testing

The idea of testlet-based adaptive testing is closely related to the older format of multistage testing (Lord, 1980). In multistage testing, examinees proceed through a sequence of subtests, moving to a more difficult subtest if they do well but to an easier one if their previous performances are low. Though the earlier literature discussed a paper-and-pencil version of this format with nonstatistical scoring of the ability of the examinees after each subtest, the advent of computers in testing practice has made an implementation with statistical estimation of ability after each subtest possible. Adema (1990) and van der Linden & Adema (1998) offer 0–1 integer programming models for the design of multistage testing systems based on the maximum-information criterion that allow for a large variety of constraints on the composition of the subtests. The possibility to include such constraints into multistage testing systems has rekindled the interest in this testing format (Luecht & Nungester, 1998; Zenisky, Hambleton & Luecht, this volume, chap. 18). For a review of a testing program designed around a multistage testing format, see Melican, Breithaupt, and Zhang (this volume, chap. 9).

2.2.6 Evaluation of Existing Approaches

The above approaches differ in important ways. The first three approaches implement the constraints through a modification of the item-selection algorithm. The last two approaches build all constraints directly into the units in the pool from which the test is administered. This distinction has consequences with respect to

1. the degree of adaptation possible during the test;
2. the possibility of expert review of actual test content;
3. the nature of constraint realization; and
4. the possibility of constraint violation.

The item-pool partitioning, WDM, and priority index methods allow for an update of the ability estimate after each item. They thus offer the maximum degree of adaptation possible. The WDM and priority index methods optimize an objective function based on weighted deviations from the constraints rather than controlling these constraints directly. Some of their constraints can therefore be violated. In all three approaches, both the selection of the items and the realization of the constraints are sequential. Though sequential item selection allows for optimal adaptation, sequential realization of constraints is less than ideal. Algorithms with this feature tend to pick items with an attractive value for the objective function early in the test—a choice that may turn out to be inadequate later on. If so, the result is completion of the test with constraint violation and/or less than optimal adaptation to the ability estimates. An empirical comparison among the item-pool partitioning method, WDM approach, and adaptive testing with the shadow-test approach in this chapter is reported in van der Linden (2005b).

The testlet-based and multistage approaches have the option of expert review of all intact testing material prior to administration. Explicit coding of relevant item attributes is not always necessary, but it is dangerous to omit such coding because human review easily leads to overlooking of relevant item attributes. However, two approaches select larger sets of items at a time, and adaptation of item selection to the ability estimates within these sets is impossible. Also, the task of assembling a pool of testlets or a multistage testing system such that any path an examinee may take satisfies all constraints involves a huge combinatorial problem that can quickly become too complicated for intuitive methods. The result may be a suboptimal branching system and/or constraint violation. However, as already noted, when adequate item coding is available, formal methods for assembling multistage testing systems can be exploited. In order to realize all constraints when assembling a pool of testlets or a multistage testing system, their use is necessary.

This evaluation of the existing methods for constrained CAT thus reveals an important dilemma. An optimal algorithm should select its items *sequentially* to allow for optimal adaptation but must realize all constraints *simultaneously* to prevent violation of certain constraints or suboptimal adaptation later in the test. Possible solutions to the dilemma are (1) to allow the algorithm to work backwardly to improve on previous decisions or (2) to have the algorithm project forwardly to take future consequences of decisions into account. In adaptive testing, backtracking is impossible; earlier choices cannot be undone. Thus, the only possibility left is to have the algorithm project forwardly each time a new item is selected. This is exactly what the shadow-test approach to adaptive testing does.

2.3 Constrained CAT with Shadow Tests

The basic concept of a shadow-test approach is illustrated in Figure 2.1. The selection of each new item in the adaptive test is preceded by real-time assembly of a shadow test. A shadow test is a full-length test that (1) meets all the test constraints, (2) contains all items already administered to the examinee, and (3) has maximum information at the current ability estimate. The item to be administered is the one with maximum information among the unused items in the shadow test. The horizontal axis of the graph shows the position of the items in the adaptive test; the vertical axis represents the ability measured by the items. The higher the vertical position of the shadow tests, the higher the current estimate of θ. Toward the end of the test, the estimates of θ stabilize. The darker portion of the shadow tests represents the items that are actually taken by the examinee. The lighter portion represents the part of it that is reassembled after a new update of the ability estimate. The very last shadow test contains the complete selection of the items actually taken by the examinee. Because each of the shadow tests meets all of the constraints, this selection does.

The following pseudo-algorithm gives a more precise summary of the idea:

Step 1: Initialize the ability estimator;
Step 2: Assemble a shadow test that meets the constraints and has maximum information at the current ability estimate;
Step 3: Administer the item in the shadow test with maximum information at the ability estimate;

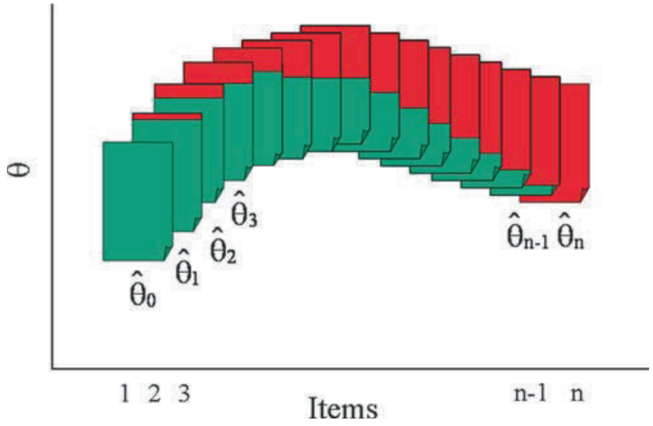


Fig. 2.1 Constrained adaptive testing with shadow tests

Step 4: Update the ability estimate;
Step 5: Update the test-assembly model to include the administered item in the next shadow test;
Step 6: Return all unused items to the pool;
Step 7: Repeat Steps 2-6 until n items have been administered.

Observe that the test length has been fixed in this algorithm. This choice is in agreement with practice in nearly all existing adaptive testing programs. Though a stopping rule based on a predetermined level of accuracy for the ability estimator is desirable from a statistical point of view, it is impossible to guarantee the same specifications for all examinees for a test with random length.

The ideal underlying all test assembly is a test that both is feasible (i.e., meets all specifications) and has maximal information at the examinee's true ability. But, as the true ability is always unknown, all one can hope for is item selection approximating this ideal as closely as possible. The shadow-test approach has this feature; it yields feasible adaptive tests converging to the optimal value for the information function at the true ability of the examinees.

This claim can be shown to hold as follows. The algorithm realizes all constraints simultaneously for each shadow test. Each next shadow test contains all items already administered to the examinee. Thus, the last shadow test is the actual adaptive test and always meets all constraints. Further, each shadow test is assembled to have a maximum value for the information function in (2.1), and the item selected from the shadow test has a maximum contribution to this function. For a consistent ability estimator, it follows from Slutsky's theorems (e.g., Ferguson, 1996) that the value for the function in (2.1) converges to the maximum value possible at the true ability of the examinee. Mild conditions for the case of maximum-information item selection to yield consistent maximum-likelihood estimation of θ are formulated in Chang and Ying (2009).

This argument assumes an infinitely large item pool with all possible combinations of values for the item parameters. However, the conclusion is expected to hold closely enough for all practical purposes for any well-designed finite item pool. Of course, the speed of convergence depends on the size and nature of the item pool as well as the set of constraints. For a severely constrained adaptive test from a small pool, convergence may be slower than for a test from a large pool involving only a few constraints. The empirical examples later in this chapter will shed some light on the question of how fast the ability estimator converges in typical applications of the procedure.

2.4 Technical Implementation

The idea of constrained adaptive testing with shadow tests was introduced in van der Linden and Reese (1998), who used the technique of 0–1 linear integer programming (IP) to assemble the shadow tests. The same idea was explored independently in Cordova (1997), whose test assembly work was based on the network-flow programming approach introduced in Armstrong and Jones (1992). A comprehensive review of approaches to automated test assembly, including adaptive assembly, is given in van der Linden (2005b).

In principle, any algorithm for automated test assembly that generates an optimal feasible solution and is fast enough for application in real time can be used to implement the above adaptive testing scheme. Even for test-assembly heuristics that tend to provide suboptimal solutions, considerable gain over the existing methods of constrained adaptive testing can be expected.

The examples later in this chapter are all based on the technique of 0–1 IP. This technique allows us to deal with virtually any type of constraint that can be met in test assembly and thus offers maximum flexibility when modeling the problem of shadow-test assembly (van de Linden, 2005b). In addition, a choice of powerful solvers for IP is available that can be used to solve such models in real time for adaptive testing programs.

2.4.1 Basic Notation and Definitions

In order to maintain generality, an IP model for the assembly of shadow tests from an item pool with some of its items organized as sets with a common stimulus is formulated. This testing format has become increasingly popular; several of the item pools used in the empirical examples later in this chapter involved this format. Typically, in testing with set-based items, the numbers of items per stimulus available in the pool are larger than the numbers to be selected in the test. The basic trick to use IP modeling for the assembly of set-based shadow tests is to introduce separate decision variables for the selection of the stimuli and items while using logical constraints to keep their values consistent.

The following notation is used throughout this chapter:

items in the pool	: $i = 1, \ldots, I$;
stimuli in the pool	: $s = 1, \ldots, S$;
set of items in the pool with stimulus s	: $U_s, s = 1, \ldots, S$;
items in the adaptive test	: $k = 1, \ldots, n$;
stimuli in the adaptive test	: $l = 1, \ldots, m$.

Thus, i_k and s_l are the indices of the kth item and lth stimulus in the adaptive test, respectively. Let $S_{k-1} \equiv \{i_1, \ldots, i_{k-1}\}$ be defined as the set of the first $k-1$ items administered. Consequently, $R_k \equiv \{1, \ldots, I\} \backslash S_{k-1}$ is the set of items remaining in the pool after $k-1$ items have been administered.

The kth shadow test is denoted as $T_k \equiv \{i_1, \ldots, i_{k-1}, i'_k, \ldots, i'_n\}$, where $i'_k, \ldots, i'_n$ are the free items in this test. Besides, $S_l \equiv (s_1, \ldots, s_l\}$ is defined as the set of the first l stimuli in the test. If the constraints on the number of items for the lth stimulus in the adaptive test have not yet been satisfied, s_l is called the *active stimulus* and U_{s_l} the *active item set* . If s_l is active, the next item is selected from $U_{s_l} \cap \{i_k^{'}, \dots, i_n^{'}\}$. As long as s_l is active, the constraints in the test-assembly model on the size of the item sets in the shadow test guarantee that U_{s_l} is not empty. Otherwise, the next item is selected from $\{i_k^{'}, \dots, i_n^{'}\}$. Therefore, the list of *eligible items* in the kth shadow test is defined as

$$A_k \equiv \begin{cases} U_{s_l} \cap \{i_k^{'}, \dots, i_n^{'}\}, & \text{if the } l\text{th stimulus is active;} \\ \{i_k^{'}, \dots, i_n^{'}\}, & \text{otherwise.} \end{cases} \tag{2.3}$$

Let $\widehat{\theta}_{k-1}$ denote the ability estimate updated after the first $k-1$ items in the adaptive test. It thus holds that the kth item in the adaptive test is

$$i_k \equiv \arg\max_i \{I_i(\widehat{\theta}_{k-1}); i \in A_k\}. \tag{2.4}$$

When assembling shadow tests, the objective function should be maximized only over the set of items eligible for administration. In particular, if the lth stimulus is active, it may be disadvantageous to maximize the information in the shadow test over items not in U_{s_l} (even though such items are needed to complete the shadow test). To implement this idea for the objective function in the model below, the following set is defined:

$$O_k \equiv \begin{cases} U_{s_l}, & \text{if the } l\text{th stimulus is active;} \\ R_k, & \text{otherwise.} \end{cases} \tag{2.5}$$

2.4.2 IP Model for Shadow Test

The model is an adapted version of the one for fixed-form test assembly with item sets presented in van der Linden (2005b, sect. 7.1). To formulate its objective function and constraints, 0–1 decision variables x_i and z_s are introduced. These variables take the value one if item i and stimulus s are selected in the shadow test, respectively; otherwise, they are equal to zero.

The following notation is needed to denote the various types of item and stimulus attributes that may play a role in the assembly of the shadow test. The set of items in the pool for stimulus s is denoted as V_s. Categorical item attributes, such as item content or format, partition the item pool into sets of items $V_c^{\text{item}}, c = 1, \dots, C$. Note that different attributes involve different partitions. For simplicity, however, only the case of one attribute is discussed; adding more constraints is straightforward. In addition, the items and stimuli are assumed to be described by quantitative attributes, such as a word count or an item difficulty parameter. For simplicity, the case of one quantitative attribute with value q_i for item i and q_s for stimulus s, respectively, is discussed. Finally, the use of logical constraints on the assembly of the shadow tests is illustrated through the presence of sets of items, $V_e^{\text{item}}, e = 1, \dots, E$, and sets of stimuli, $V_e^{\text{stim}}, e = 1, \ldots, E$, that clue each other; therefore, these items or stimuli cannot be selected for the same test. In sum, the notation is

set of items with categorical attribute c	: $V_c^{\text{item}}, c = 1, \ldots, C;$
set of stimuli with categorical attribute c	: $V_c^{\text{stim}}, c = 1, \ldots, C;$
quantitative item attribute	: $q_i^{\text{item}}, i = 1, \ldots, I;$
quantitative stimulus attribute	: $q_s^{\text{stim}}, s = 1, \ldots, S;$
sets of mutually exclusive items	: $V_e^{\text{item}}, e = 1, \ldots, E;$
sets of mutually exclusive stimuli	: $V_e^{\text{stim}}, e = 1, \ldots, E.$

The shadow-test model is

$$\text{maximize} \sum_{i \in O_k} I_i(\widehat{\theta}_{k-1})x_i \quad \text{(maximum information)} \tag{2.6}$$

subject to

$$\sum_{i=1}^{I} x_i = n; \text{ (test length)} \tag{2.7}$$

$$\sum_{s=1}^{S} z_s = m; \text{ (number of stimuli)} \tag{2.8}$$

$$\sum_{i \in S_{k-1}} x_i = k - 1; \text{ (items already administered)} \tag{2.9}$$

$$\sum_{i \in V_s} x_i \gtreqless n_s z_s, s = 1, .., S; \text{ (number of items per stimulus)} \tag{2.10}$$

$$\sum_{i \in V_c^{\text{item}}} x_i \gtreqless n_c^{\text{item}}, c = 1, \ldots, C; \text{ (categorical item attribute)} \tag{2.11}$$

$$\sum_{i=1}^{I} q_i x_i \gtreqless b_q^{\text{item}}; \text{ (quantitative item attribute)} \tag{2.12}$$

$$\sum_{i \in V_c^{\text{stim}}} z_s \gtreqless n_c^{\text{stim}}, c = 1, \ldots, C; \text{ (categorical stimulus attribute)} \tag{2.13}$$

$$\sum_{i=1}^{I} q_s z_s \gtreqless b_q^{\text{stim}}; \text{ (quantitative stimulus attribute)} \tag{2.14}$$

$$\sum_{i \in V_e^{\text{item}}} x_i \leq 1, e = 1, \ldots, E; \text{ (mutually exclusive items)} \tag{2.15}$$

$$\sum_{s \in V_e^{\text{stim}}} z_s \leq 1, e = 1, \ldots, E; \text{ (mutually exclusive stimuli)} \tag{2.16}$$

$$x_i \in \{0, 1\}, i = 1, \ldots, I; \text{ (domain of variables)} \tag{2.17}$$

$$z_s = \{0, 1\}, s = 1, \ldots, S. \text{ (domain of variables)} \tag{2.18}$$

The first two constraints set the numbers of items and stimuli in the test. The constraint in (2.9) forces all $k - 1$ items already administered to be in the test. In doing so, the model automatically accounts for the attributes of all these items. In the next constraints, $\gtreqless$ indicates the choice of an equality or inequality symbol. Besides, bounds on the number of items in a set are denoted as n and bounds on quantitative attributes as b, with appropriate subscripts and superscripts to denote the nature of the set or attribute. The constraints in (2.10) serve a double goal; not only do they set a bound on the number of items per stimuli, but the presence of the stimulus variables on their right-hand sides also keeps the selection of the stimuli and items consistent. It is necessary to use these constraints always for a combination of a lower and upper bound. If only one bound is needed, the other should be a dummy chosen to be large or small enough to remain inactive (for an explanation, see van de Linden, 2005b, sect. 7.1).

2.4.3 Numerical Aspects

A solution for an IP model as in (2.6)–(2.18) is a set of optimal values for the variables $x_i, i = 1, \ldots, I$, and $z_s, s = 1, \ldots, S$. Such solutions can only be obtained through implicit enumeration in the form of a well-implemented branch-and-bound (BAB) method. Although IP problems are known to be NP-hard (that is, have a running time for their worst cases not bounded by a polynomial in the size of the problem), we now have powerful solvers in the form of commercial software that preprocess the problem and automatically avoid such cases. Nevertheless, the following ideas are still helpful (cf. van der Linden, 2005b, sects. 4.2 and 9.1.5).

First, note that the constraints in (2.7)–(2.18) do not depend on the value of $\widehat{\theta}$. The update of this estimate only affects the objective function in (2.6). Repeated application of the model for $k = 1, \ldots, n$ can thus be described as a series of problems for which the space of feasible solutions remains the same but the coefficients in the objective function (i.e., the values for the item information function) change. The changes become generally small when the ability estimates stabilize. As a start from a good initial feasible solution is essential, the obvious choice is to use the $(k - 1)$th shadow test as the initial solution for the kth test. This measure has been proven to improve the speed of the solution processes dramatically. Also, the first shadow test need not be calculated during operational testing. If necessary, it can be preassembled for the initial estimate of θ in advance.

Second, additional improvement is gained by deliberate choice of the order in which the solver branches on the decision variables. The variables for the items, x_i, determine the selection of individual items, but those for the stimuli, z_s, have an impact on larger sets of items. It always pays off to branch first on the variables with the largest impact. In the branching order, the stimulus variables should thus precede the item variables.

Also, forcing the slack variables for the constraints in the model to be integers has proven to be efficient. In the branching order, slack variables for constraints with stimuli should have higher priority than the decision variables for these stimuli, and the same should hold for the items.

Finally, it is common to find values close to optimality for BAB processes long before the end of the process. It therefore makes sense to stop the process as soon as the objective function approaches a well-chosen bound satisfactorily closely. Good results have been found for the objective function values for the relaxed version of the model as upper bound with tolerances as small as 1–2% of the value.

All applications later in this chapter used CAT software developed at the University of Twente. To calculate the shadow tests, the software made calls to the solver in the CPLEX package. For a recent version of the solver (e.g., CPLEX 9.0; ILOG, 2003), the running time needed for one cycle of ability estimation and item selection on a current PC is less than a second for item pools much larger than the typical real-world pool. In fact, much larger times would not have involved any problem since it is always possible to calculate ahead; that is, calculate two solutions for the kth shadow test, one for the update of $\widehat{\theta}$ after a correct and the other after an incorrect response, while the examinee works on item $k-1$.

2.5 Four Applications to Adaptive Testing Problems

As shadow tests are full-size linear tests, any feature possible for regular fixed-form test assembly can also be realized for an adaptive test. The only thing needed is inserting the appropriate constraints into the shadow-test model in (2.6)–(2.18). The shadow-test approach thus immediately accommodates the earlier-discussed wish of testing programs that want to go adaptive but keep the same "look and feel" for its tests.

For a testing program that already assembles its fixed forms using IP, the only change required is to halve the right-hand-side bounds in the constraints in its models. The result is a shadow-test model for an adaptive test of half the length of its fixed test form but with the same relative composition. In fact, as discussed at the end of this chapter, an adaptive testing algorithm based on the shadow-test approach can used to assemble any of the existing test formats: adaptive, linear, multistage, etc., in real time.

The flexibility of the shadow-test approach is illustrated with four applications each addressing a different aspect of adaptive testing. In the first application, the practicality of the shadow-test approach is demonstrated for an adaptive testing program with an extremely large number of content constraints of varying nature. The second application deals with the problem of differential speededness in adaptive testing. As each examinee gets a different selection of items, and items differ greatly in their time intensity, some of them may have trouble completing the test. It is shown how the problem can be resolved by inserting a response-time constraint in the shadow-test model. The question of how to deal with item-exposure control in constrained adaptive testing with shadow tests is addressed in the third example. The control is based on the use of random constraints in the test-assembly model that determine which items are eligible for the examinees with probabilities that guarantee predetermined exposure rates. The last example addresses the case of a testing program that uses a released fixed form of the test as a reference test to help its examinees interpret their scores. It is shown how the observed scores on the two versions of the test can be automatically equated by inserting a few additional constraints into the model for the shadow test.

2.5.1 CAT with Large Numbers of Nonstatistical Constraints

In order to check the practicability of the shadow-test approach for a CAT program with a large number of nonstatistical constraints, a simulation study was conducted for a pool of 753 items from the Law School Admission Test (LSAT). A 50-item adaptive version of the LSAT was simulated. The current linear version of the LSAT is twice as long; all its specifications were reduced to half their size. The specifications dealt with such item and stimulus attributes as item and stimulus content, gender and minority orientation, word counts, and answer-key distributions. The set of content attributes defined an elaborate classification system for the items and stimuli for which the test had to meet a large number of specifications. In all, the IP model for the shadow test had 804 variables and 433 constraints.

Three conditions were simulated: (1) unconstrained CAT (adaptive version of the LSAT ignoring all current specifications); (2) constrained CAT with the least severely constrained section of the LSAT first; and (3) constrained CAT with the most severely constrained section first. Mean-square error (MSE) and bias functions were calculated for each of the three conditions after $n = 10, 20, \dots, 40$ items.

The results are shown in Figures 2.2 and 2.3. For all test lengths, the results for the conditions of unconstrained CAT and constrained CAT with the least severely constrained section first were practically indistinguishable. The MSE and bias functions for the condition of constrained CAT with the most severely constrained first were less favorable for the shorter test lengths but matched those of the other two conditions for $n > 20$. The results are discussed in more detail in van der Linden and Reese (1998). The main conclusion from the study is that adding large numbers of constraints to an adaptive test without substantial loss in statistical precision is possible.

2.5.2 CAT with Response-Time Constraints

A problem not anticipated before adaptive testing became operational was differential speededness of the test. Most adaptive testing programs offer fixed-length tests administered with fixed-size time slots. Test questions, however, vary considerably in the amount of time needed to complete them due to the amount of reading involved in the item, the nature of the problem formulated in it, etc. Because each

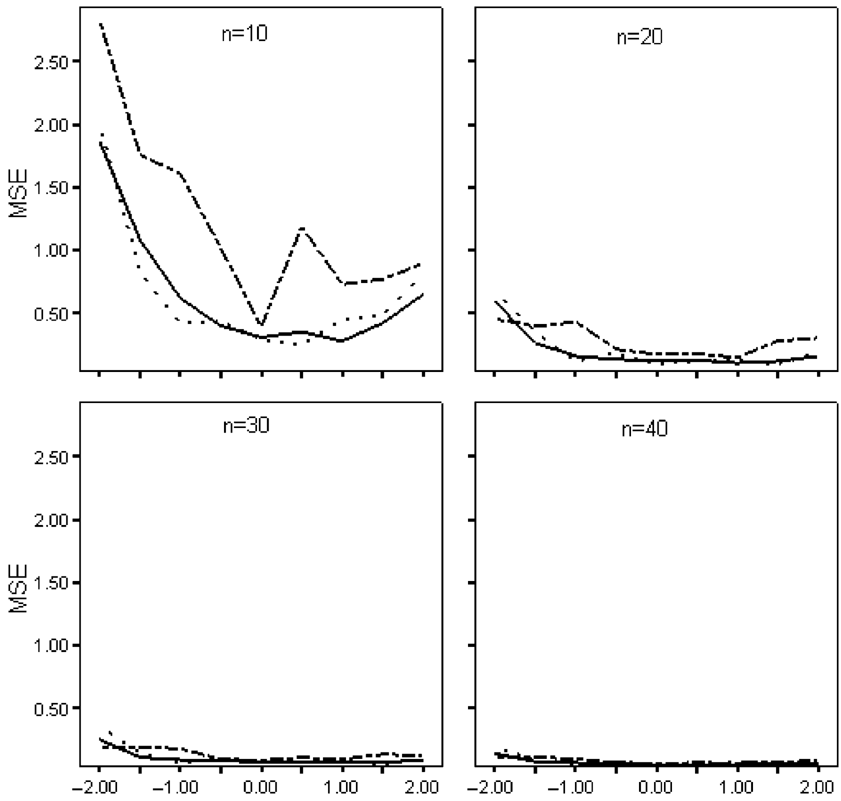


Fig. 2.2 MSE functions after $n = 10, 20, 30,$ and 40 items (unconstrained CAT: dotted; constrained CAT with least severely constrained section first: dashed; constrained CAT with most severely constrained section first: solid)

examinee gets an individual selection of items, some examinees may run out of time whereas others are able to finish easily.

The logic of adaptive testing with the shadow-test approach suggests resolving the issue of differential speededness by adding a constraint to the shadow-test model that guarantees the same time pressure for all examinees. Such a constraint is possible provided we have a model for the distribution of the response times (RTs) of an examinee responding to an item with separate parameters for the examinee and the item. The following approach is based on a lognormal response-time model (van der Linden, 2006). For an examinee j operating at a speed $\tau_j \in (-\infty, \infty)$ on the items in the test, the model assumes a normal density for the distribution of the logarithm of his or her RT on item i, $\ln T_{ij}$, with item parameters $\beta_i \in (-\infty, \infty)$ for the time intensity of the item and discrimination parameters $\alpha_i > 0$. The model equation is

$$f(t_{ij}; \tau_j, \alpha_i, \beta_i) = \frac{\alpha_i}{t_{ij}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}[\alpha_i(\ln t_{ij} - (\beta_i - \tau_j))]^2\right\}. \tag{2.19}$$

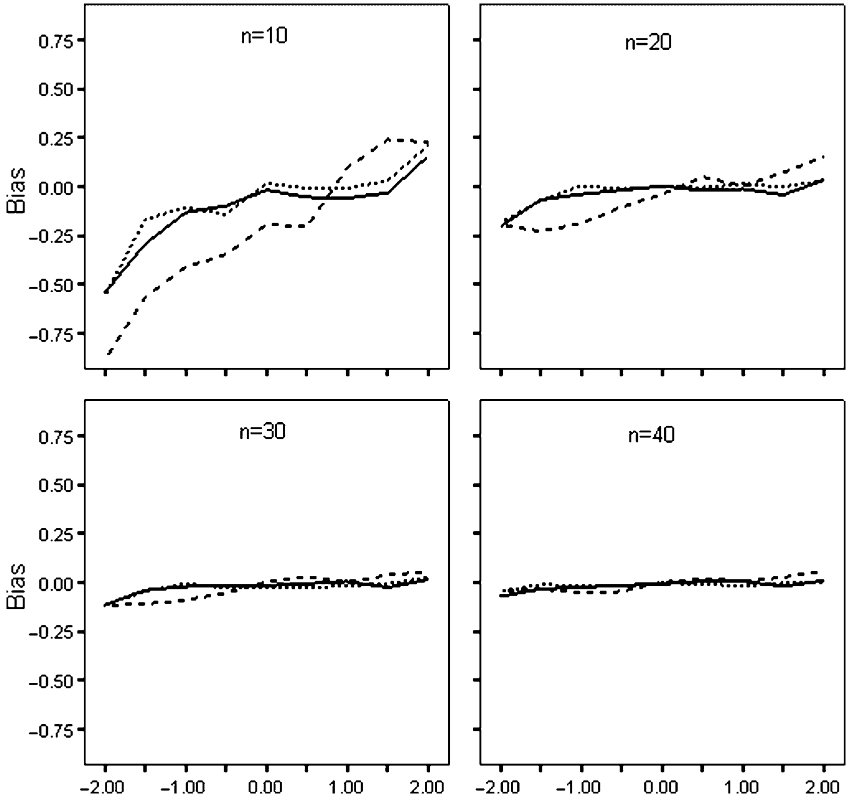


Fig. 2.3 Bias functions after $n = 10, 20, 30$, and 40 items (unconstrained CAT: dotted; constrained CAT with least severely constrained section first: dashed; constrained CAT with most severely constrained section first: solid)

The mean of the distribution is thus equal to $\beta_i - \tau_j$, which implies a tendency to a longer RT for an item that is more time-intensive and/or an examinee working at a higher speed. The relation between $\ln t_{ij}$ and its mean $\beta_i - \tau_j$ is modified by discrimination parameter α_i, which is the reciprocal of the standard deviation of the distribution. It is easy to estimate the item parameters as part of the standard item calibration required for adaptive testing, provided the RTs of the examinees during item pretesting have been saved (van der Linden, 2006).

The necessary constraint assumes a permanent update of an estimate of the examinee's speed parameter τ_j using his or her RTs during the test, just as the estimate of the ability parameter θ is updated using his or her responses. The estimate of τ_j is then used along with the known item parameters β_i and α_i to predict the RT distributions for the examinee on all remaining items in the pool from the model in (2.19). The constraint to be added to the shadow-test model requires the sum of the predicted RTs on the free items in the shadow test and the time actually used on the items already administered not to be larger than the time limit for the test.

The procedure combines excellently with a Bayesian update of τ during the test using the logRTs. For a normal prior distribution of θ, the posterior predictive densities of the logRTs on the remaining items are also normal with means and variances that are easily calculated. Details of the statistical aspects of this procedure are presented in van der Linden (2009a).

More formally,suppose $k-1$ items have already been administered. At this point, the total time spent on the test is equal to

$$\sum_{i \in S_{k-1}} t_{ij} x_i . \tag{2.20}$$

Also, the posterior predicted distributions of the RTs on the remaining items in R_k are known. Let $\widetilde{t}_{ij}$ be the predicted RT by examinee j on item i, whereas $\ln \widetilde{t}_{ij}^{\pi_k}$ is the πth percentile in the posterior distribution of $\ln T_{ij}$ to be used to select the kth item. It makes sense to choose more liberal values of π^k in the beginning of the test but become conservative toward the end of it. The predicted time on the free items in the shadow test is

$$\sum_{i \in R_k} \tilde{t}_{ij}^{\pi_k} x_i . \tag{2.21}$$

For time limit t_{lim} in use for the test, the constraint required to control the item selection for differential speededness is

$$\sum_{i \in S_{k-1}} t_{ij} x_i + \sum_{i \in R_k} \tilde{t}_{ij}^{\pi_k} x_i \leq t_{\text{lim}} . \tag{2.22}$$

The procedure was applied to an item pool for the adaptive version of the Arithmetic Reasoning Test in the Armed Services Vocational Aptitude Battery (ASVAB). The pool consisted of 186 items calibrated under the model in (2.1). Response times were recorded for 38,357 examinees who had taken the test previously. The test had a length of 15 items and the time limit was $t_{\text{lim}} = 39$ minutes (2,340 seconds). Percentile π_k was chosen to be the 50th percentile for $k = 1$ and moved up in equal steps to the 95th percentile for the last three items. In order to evaluate the effects of the constraint in (2.22) on the time needed by the examinees, versions of the test without and with the constraint were simulated. The range of the ability and speed parameters in the simulation was the same as for the empirical estimates of these parameters from the data set.

Some of the results are shown in Figure 2.4. The first panel shows the average time needed to complete the test as a function of speed parameter τ for the condition without the constraint. Different curves are displayed for the different values of θ used in the study. The faster examinees had considerable amounts of time left after completion of the test, but one of the ability groups among the slowest examinees ran out of time. The second panel shows the same information for the condition with the constraint. For this condition, none of the ability groups ran out of time.

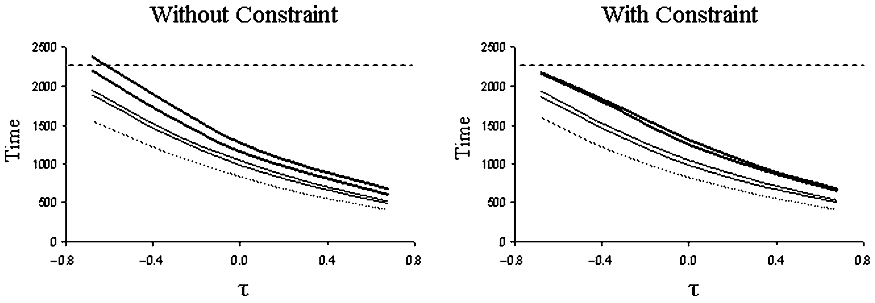


Fig. 2.4 Time needed to complete the test without (first panel) and with (second panel) the response time constraint in (2.22). Note: dotted line indicates the time limit. [Reproduced with permission from W. J. van der Linden (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25–41.]

Interestingly, the group that ran out of time for the condition without the constraint was the group with the highest ability. This finding may go against the reader's intuition but can be explained as follows. Additional analyses of the data set revealed a rather strong positive correlation between the difficulties and time intensities of the items in the pool ($r = 0.65$). Because the test was adaptive, the more able examinees received the more difficult items, particularly toward the end of the test, and thus needed more time to solve them. Consequently, the examinees in this group who operated more slowly had trouble completing the test on time. The correlation between speed and ability for the sample of ASVAB examinees was $r = 0.04$, suggesting that the combination of high ability with high speed was equally likely as with low speed.

The results indicate that the current time limit for the ASVAB test was quite generous. The effects of differential speededness were therefore minor. In fact, a much more conspicuous feature of the plots in Figure 2.4 is the large amount of unused time after completion of the test for the majority of the examinees. This finding suggests shortening the time limit and using the constraint in (2.22) to prevent differential speededness under the tighter limit. For results that show that the constraint remained fully effective when the time limit was shortened to $t_{\text{lim}} = 29$ minutes, see van der Linden (2009a).

2.5.3 CAT with Item-Exposure Control

For item selection based on (2.1), the probability of an item being selected for an examinee with a given estimate of θ depends on the size of its information measure at the estimate relative to the other items in the pool. As is well known, the information measure is predominantly determined by the discrimination parameter of the item. However, the fact that an item is selected frequently does not mean that all other items are necessarily much worse. Even a small difference in discriminating power can make one item be selected frequently and another rarely.

In order to avoid security problems due to capitalization on a small subset of items, CAT programs usually modify their item-selection process to yield exposure rates for their most popular items not larger than a well-chosen target value. As a result, the exposure rates of the other items in the pool go up, generally in the order of the value of their discrimination parameters.

Sympson and Hetter (1985) introduced the idea of having a probability experiment determine if an item selected should be actually administered or removed from the pool for the examinee. The experiment had to be repeated until an item was administered. Stocking and Lewis (1998, 2000) proposed a conditional version of this method. Let $P(S_i \mid \theta)$ denote the probability of selecting item i conditional on the ability θ, and A_i the event of administering the item. For a given maximum exposure rate, $r^{\max}$, it should thus hold that

$$P(A_i \mid \theta) = P(A_i, S_i \mid \theta) = P(A_i \mid S_i, \theta)P(S_i \mid \theta) \leq r^{\max}. \tag{2.23}$$

The probabilities of item selection $P(S_i \mid \theta)$ depend on the composition of the item pool and the algorithms in the CAT software, and are thus fixed by design. Hence, the upper bound on the exposure rates $P(A_i \mid \theta)$ should be realized by manipulating the values for the conditional probabilities $P(A_i \mid S_i, \theta)$. However, finding optimal values for these control parameters requires an extensive iterative process of computer simulations with cycles of (1) simulating the test, (2) estimating the probabilities of selection, and (3) adjusting the values for the control parameters.

Stocking and Lewis also suggested a more efficient probability experiment to implement Sympson–Hetter item-exposure control, which does not require separate experiments for each of the individual items selected by the CAT algorithm but enables us to pick the item from the result of a single experiment over the list of the κ most informative items at the ability estimate, where κ is a number to be selected by the testing program. The same experiment can easily be implemented for Sympson–Hetter item-exposure control in adaptive testing with shadow tests. The only necessary modification is the replacement of the list of the most informative items in the item pool by a list of the most informative free items in the shadow test.

A disadvantage of the implementation, however, is the decrease in the number of available free items toward the end of the test. An alternative without this problem is a multiple-shadow-test approach in which a set of shadow tests (two or three, say) is assembled prior to the selection of each item instead of a single test. Each of the shadow tests in the set has to meet the same set of test specifications. Also, the tests share the $k - 1$ items that have already been administered; otherwise, they are different but parallel. The list of the best κ items for the Stocking–Lewis experiment is then selected from the set of all free items in the shadow tests.

Simultaneous assembly of a set of shadow tests requires only a simple reformulation of the optimization model in (2.6)–(2.18), and the method does not involve any new technical or computational complications (van der Linden, 2005b, sect. 9.4.3; Veldkamp & van der Linden, 2008). As demonstrated by the empirical results in these references, the method works well. However, like any other method based on

Sympson–Hetter item-exposure control, it still requires the time-consuming iterative process of finding the optimal control parameters in (2.23).

An alternative approach within the shadow-test framework, which avoids this process, is to control the exposure rates through random selection of item-ineligibility constraints in the shadow test model rather than a probability experiment that eliminates items after they have been selected. The appropriate constraints are

$$x_1 = 0, \qquad i = 1, \ldots, I. \tag{2.24}$$

To derive the probabilities with which the constraints have to be imposed for each of the items in the pool, a relation analogous to (2.23) can be derived. Let E_i denote the event of item i being eligible for the current examinee whereas A_i still denotes the event of the item being administered. Because $A_i \subset E_i$, it should thus hold that

$$P(A_i \mid \theta) = P(A_i, E_i \mid \theta) = P(A_i \mid E_i, \theta)P(E_i \mid \theta) \leq r^{\max}. \tag{2.25}$$

As shown in van der Linden and Veldkamp (2007), the relation can be rewritten as

$$P(E_i \mid \theta) \leq \frac{r^{\max}}{P(A_i \mid \theta)} P(E_i \mid \theta). \tag{2.26}$$

Let j denote the examinees in the order in which they take the test. We can conceive of (2.26) as a recurrence relation across these examinees; that is, for examinee $j + 1$, the relation becomes

$$P^{(j+1)}(E_i \mid \theta) \leq \frac{r^{\max}}{P^{(j)}(A_i \mid \theta)} P^{(j)}(E_i \mid \theta). \tag{2.27}$$

The equation gives the update of the probability of item eligibility for each next examinee. The two right-hand-side probabilities can easily be estimated from continuously updated counts of the events E_i and A_i during all previous examinees. To determine if the constraint in (2.23) has to be imposed, a Bernoulli experiment for each of the items with the left-hand-sided probability in (2.27) suffices. The experiments have to be conducted only once, before the examinee begins the test.

Use of the equation in (2.27) results in a self-adaptive system of item-exposure control. As soon as $P^{(j)}(A_i \mid \theta) > r^{\max}$, the probability of eligibility $P^{(j+1)}(E_i \mid \theta)$ goes down, and so will $P^{(j)}(A_i \mid \theta)$. The reverse also holds. For these and other features of the method, see van der Linden and Veldkamp (2007).

Figure 2.5 shows a selection from the conditional exposure rates in a simulation study with a 25-item adaptive version of a section from the LSAT from a pool of 305 items. Six different levels of exposure control were simulated: no control and control with $r^{\max} = 0.20, 0.15, 0.10, 0.05$, and 0.025. The patterns of conditional exposure rates for the 305 items in Figure 2.3 were typical of all patterns in this study: First, except for some negligible remaining random noise, the maximum exposure rate was always lower than the target value $r^{\max}$. Second, the lower this target, the greater

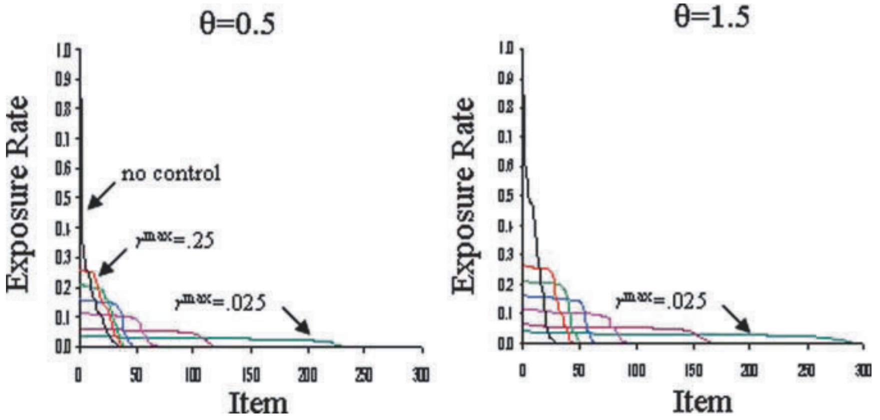


Fig. 2.5 Time needed to complete the test without (first panel) and with (second panel) the response time constraint in (2.22). Note: dotted line indicates the time limit. [Reproduced with permission from W. J. van der Linden and B. P. Veldkamp (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418.]

the set of items with positive exposure rates. Finally, for the lowest value for the target $r^{\max}$, nearly all items became active.

For programs with larger sets of content constraints, low target values, and smaller item pools, the Bernoulli experiments may occasionally result in too many ineligibility constraints and as a result the shadow-test model may become infeasible. Also, rather than using direct updates of the probability estimates in (2.27), it is advantageous to use a technique from Bayesian networks known as fading. For these and other implementation issues, see van der Linden and Veldkamp (2007).

2.5.4 CAT with Equated Number-Correct Scores

The necessity to equate scores on an adaptive test to number-correct scores on a fixed, linear test has at least two practical reasons. First, testing programs making the transition to an adaptive testing format may want to offer their examinees the choice between a former linear version with number-correct scoring and the new adaptive version of the test. However, this choice is only justified if the scores on both versions are comparable. To achieve comparable scores, the method of equipercentile equating has been applied to equate ability estimates to number-correct scores for this purpose (Segall, 1997). Second, as the items in an adaptive test cannot be released when an examinee takes the test, some testing programs release a linear version of the test to help the examinees with the interpretation of their CAT scores. This use of a linear test as a reference test requires the same type of score equating.

Instead of a separate equating study for each new reference test, the logic of constrained adaptive testing with shadow tests proposed in this chapter suggests the use of constraints that guarantee an adaptive test to have observed scores automatically equated to those on the linear test. Such constraints are possible using a condition derived in van der Linden and Luecht (1998). These authors show that, for any value of θ, the conditional distributions of observed number-correct scores on two test forms with items $i = 1, \ldots, n$ and $j = 1, \ldots, n$ are identical if and only if

$$\sum_{i=1}^{n} P_i^r(\theta) = \sum_{j=1}^{n} P_j^r(\theta), \quad r = 1, \ldots, n. \tag{2.28}$$

They also show that the conditions quickly vanish for $r \to n$, and report nearly perfect empirical results for $r = 2$ or 3.

Note that the conditions in (2.28) are linear in the items. They therefore lend themselves easily to insertion in an IP model for the assembly of shadow tests.

Let $j = 1, \ldots, n$ indicate the items in the reference test to which the adaptive test has to be equated. The following set of constraints should be used:

$$\sum_{i=1}^{n} P_i^r\left(\widehat{\theta}_{k-1}\right) x_i - \sum_{j=1}^{n} P_j^r\left(\widehat{\theta}_{k-1}\right) \leq c, \quad r = 1, \ldots, R, \leq n, \tag{2.29}$$

$$\sum_{i=1}^{n} P_i^r(\widehat{\theta}_{k-1}) x_i - \sum_{j=1}^{n} P_j^r\left(\widehat{\theta}_{k-1}\right) \geq -c, \quad r = 1, \ldots, R \leq n, \tag{2.30}$$

where c is a tolerance parameter with an arbitrarily small value and R need not be larger than 3 or 4. Note that these constraints thus require the difference between the sums of powers of the response functions at $\widehat{\theta}_{k-1}$ to be in an arbitrarily small interval about zero, $(-c, c)$. They do not require the two sets of response functions to be identical across the whole range of θ. Also, they only require sums of powers of their values at $\widehat{\theta}_{k-1}$ to be identical, not the powers of the response functions of the individual items. Thus, the algorithm does *not* build adaptive tests that are required to be item-by-item parallel to the alternative test.

In order to assess the effects of the constraints in (2.29)–(2.30) on the observed number-correct scores in the adaptive test, a simulation study was conducted for the same item pool from the LSAT as in the first application above. Two results for two different conditions were compared: (1) unconstrained CAT and (2) constrained CAT with the above conditions for $R = 1, 2$. In either condition, the true values of the examinees were sampled from $N(0, 1)$. The observed number-correct scores were recorded after $n = 20, \ldots, 50$ items. As a reference test, a previous form of the LSAT was used.

The results are given in Figure 2.6. As expected, the observed number-correct distribution for the unconstrained CAT was peaked with a mode slightly larger than $n/2$. After 20 items the observed number-correct distributions for the constrained condition had already moved away from this distribution toward the target

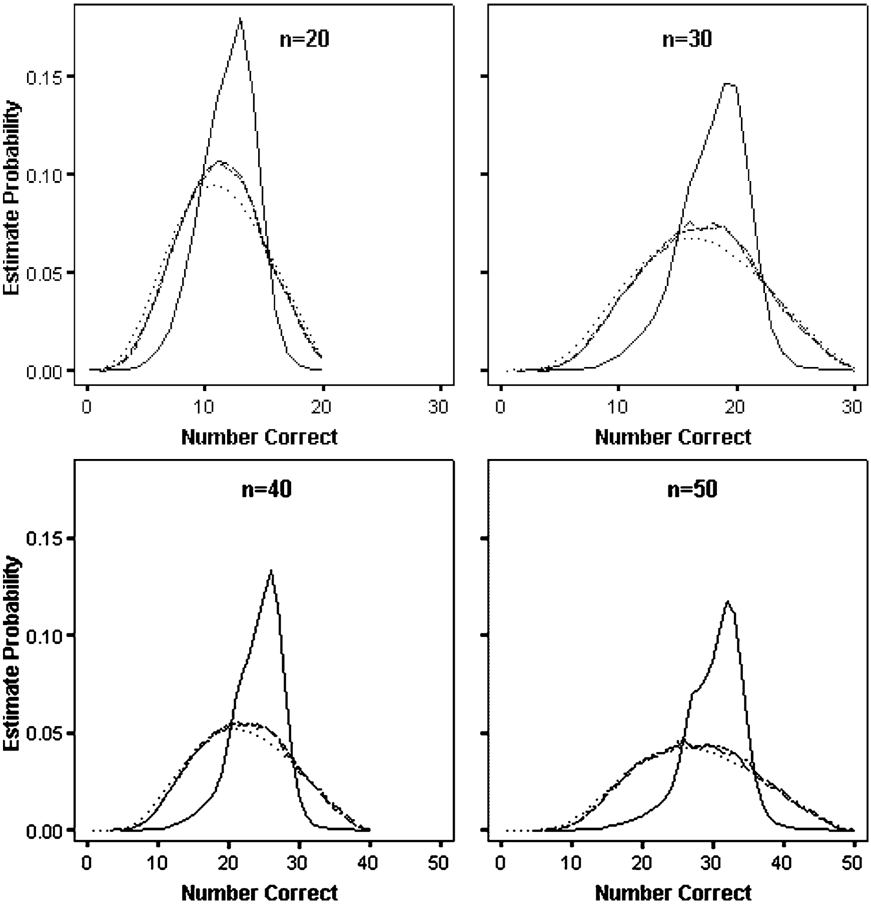


Fig. 2.6 Observed-score distributions for CAT with and without constraints for number-correct score equating (Target distribution: dotted; CAT without constraints: solid; CAT with constraint for $R = 1$: dashed-dotted; CAT with constraints for $R = 1, 2$: dashed). [Reproduced with permission from W. J. van der Linden (2001). Adaptive testing with equated number-correct scoring. *Applied Psychological Measurement*, *24*, 343–355.]

distribution on the reference test. After 30 items, the observed number-correct distributions for the constrained CAT and the reference test were indistinguishable for all practical purposes. The choice of value for R did not seem to matter much. Neither did the ability estimators show any differences in bias between the two conditions. On the other hand, the study revealed a loss in mean-square error for the constrained CAT condition comparable to that for the constrained CAT conditions in Figure 2.2.

Observe that the equating introduced by the constraints in (2.29)–(2.30) is actually much more powerful than traditional equipercentile equating. The goal of the latter is only to match the marginal distributions for the population in the study whereas the current constraints imply local equating, that is, matching of the conditional distributions given the ability level of the examinees as well (van der Linden, 2009b).

2.6 Concluding Remarks

The empirical examples above illustrate the application of several types of constraints. These examples do not exhaust all possibilities. A more recent example is the formulation of Chang and Ying's (1999) α-stratified multistage adaptive testing scheme in the current framework to allow for large numbers of content constraints on the adaptive test (Chang & van der Linden, 2003; van der Linden & Chang, 2003). Also, because the adaptive test is realized through a *series* of shadow tests, specifications can be imposed in ways that do not exist for the assembly of a single linear test. These new implementations include the possibility of alternating systematically between objective functions for successive shadow tests to deal with cases of multiple-objective test assembly and, as illustrated in the third application above, using random constraints. However, the full array of possible applications of such implementations to constrained adaptive testing still has to be explored.

It seems tempting to think of adaptive testing as a specific form of test assembly in which one item is selected at a time rather than all items simultaneously. However, as already hinted at, a software program for adaptive testing with the shadow-test approach is the most general test assembler possible. A change from one testing format to another can easily be realized through a change of the test-assembly model or the selection of the items from the shadow tests. Earlier in this chapter, we have already indicated how to change from a linear testing program to an adaptive program with the same content specifications. Conversely, a linear test form can be treated as a common first shadow test administered in full to all examinees. Likewise, linear-on-the-fly testing can be made adaptive by first assembling shadow tests at the initial ability estimates for the examinees, which then are also taken in full. A software program for adaptive testing with shadow tests can also serve as a real-time assembler of multistage tests. The only necessary change is administering more than one item from a shadow test before reassembling it at a new ability estimate.

References

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement*, *27*, 241–253.

Armstrong, R. D. & Jones, D. H. (1992). Polynomial algorithms for item matching. *Applied Psychological Measurement, 16*, 271–288.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Chang, H.-H. & van der Linden, W. J. (2003). Optimal stratification of item pools in alpha-stratified adaptive testing. *Applied Psychological Measurement, 27*, 262–274.

Chang, H. & Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Chang, H.-H. & Ying, Z. (2009). Nonlinear sequential designs for logistic item response models with applications to computerized adaptive tests. *Annals of Statistics*, *37*, 1466–1488.

Cheng, Y. & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369–383.

Cheng, Y., Chang, H.-H. & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*, 467–482.

Cordova, M. J. (1997). *Optimization methods in computerized adaptive testing*. Unpublished doctoral dissertation, Rutgers University, New Brunswick, NJ.

Ferguson, T. S. (1996). *A course in large-sample theory.* London: Chapman & Hall.

Glas, C. A. W., Wainer, H. & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston: Kluwer-Nijhof Publishing.

Hetter, R. D. & Sympson, J. B. (1997). Item exposure in CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.),*Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

ILOG, Inc. (2003). *CPLEX 9.0* [Computer Program and Manual]. Incline Village, NV: Author.

Kingsbury, G. G. & Zara, A. R. (1991). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359–375.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luecht, R. D. (1988). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224–236.

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229–249.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 81–198). Washington, DC: American Psychological Association.

Stocking, M. L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57–75.

Stocking, M. L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston: Kluwer-Nijhof Publishing.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.

Sympson, J. B. & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.

van der Linden, W. J. (1999). A procedure for empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 21–29.

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement*, *24*, 225–240.

van der Linden, W. J. (2001a). Adaptive testing with equated number-correct scoring. *Applied Psychological Measurement*, *24*, 343–355.

van der Linden, W. J. (2001b). On complexity in computer-based testing. In G. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum Associates.

van der Linden, W. J. (2005a). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*, 283–302.

van der Linden, W. J. (2005b). *Linear models for optimal test design.* New York: Springer-Verlag.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*, 181–204.

van der Linden, W. J. (2009a). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement, 33*, 25–41.

van der Linden, W. J. (2009b). Local observed-score equating. In A. A. von Davier (Ed.), *Statistical models for equating, scaling, and linking*. New York: Springer-Verlag. In press.

van der Linden, W. J. & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement*, *35*, 185–198 [Addendum in Vol. 36, 90–91].

van der Linden, W. J. & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*, 107–120.

van der Linden, W. J. & Luecht, R. M. (1998). Observed equating as a test assembly problem. *Psychometrika*, *62*, 401–418.

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

van der Linden, W. J. & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418.

Veldkamp, B. P. & van der Linden, W. J. (2008). A multiple-shadow-test approach to Sympson-Hetter item-exposure control in adaptive testing. *International Journal of Testing, 8*, 272–289.

Wainer, H., Bradlow, E. T. & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer-Nijh of Publishing.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H. & Kiely, G. L. (1987). Item clusters in computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Chapter 3
Principles of Multidimensional Adaptive Testing

Daniel O. Segall

3.1 Introduction

Tests used to measure individual differences are often designed to provide comprehensive information along several dimensions of knowledge, skill, or ability. For example, college entrance exams routinely provide separate scores on math and verbal dimensions. Some colleges may elect to base qualification on a compensatory model, where an applicant's total score (math plus verbal) must exceed some specified cutoff. In this instance, the individual math and verbal scores may provide useful feedback to students and schools about strengths and weaknesses in aptitudes and curriculum. In other instances, colleges may elect to base qualification on a multiple-hurdle model, where the applicant's scores on selected components must exceed separate cutoffs defined along each dimension. For example, a college may elect to have one qualification standard for math knowledge and another standard for verbal proficiency. Applicants may be required to meet one or the other, or both standards, to qualify for entrance. In all these instances, it is useful and important for the individual component scores to possess adequate psychometric properties, including sufficient precision and validity.

When the dimensions measured by a test or battery are correlated, responses to items measuring one dimension provide clues about the examinee's standing along other dimensions. An examinee exhibiting a high-level vocabulary proficiency is likely (although not assured) to exhibit a similar high level of reading comprehension, and vice-versa. Knowledge of the magnitude of the association between the dimensions in the population of interest, in addition to the individual's performance levels, can add a unique source of information, and if used properly can lead to a more precise estimate of proficiencies. This cross-information is ignored by conventional scoring methods and by unidimensional item selection and scoring methods used in computerized adaptive testing (CAT). The challenge discussed in this chapter is to increase the efficiency of adaptive item selection and scoring algorithms by extending unidimensional methods to the simultaneous measurement of multiple dimensions.

D.O. Segall (✉)
Defence Manpower Data Center, 400 Gigling Road, Seaside, CA 93955–6771, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_3,

The cross-information gathered from items of correlated dimensions can be effectively modeled by multidimensional item response theory. In the case of computerized adaptive testing, this information can aid measurement in two ways. First, it can aid in the selection of items, leading to the choice of more informative items. Second, it can aid in the estimation of ability, leading to test scores with added precision. In order to realize these benefits, two generalizations of unidimensional adaptive testing are necessary, one for item selection, and another for scoring. The benefit of this multidimensional generalization is increased measurement efficiency—manifested by either greater precision or reduced test lengths.

3.2 Literature Review

Bloxom and Vale (1987) were the first to formally consider the extension of unidimensional adaptive testing methods to multiple dimensions. They noted that the direct multivariate generalization of unidimensional IRT scoring procedures could easily exceed the computational power of personal computers of the time (mid-1980s). To avoid intensive calculations associated with iterative algorithms and numerical integration, they proposed an efficient scoring procedure based on a multivariate extension of Owen's (1975) sequential updating procedure. Through a series of normal approximations, the multivariate extension provides closed-form expressions for point estimates of ability. The issue of efficient item selection was not explicitly addressed by Bloxom and Vale.

Tam (1992) developed an iterative maximum likelihood (ML) ability estimation procedure for the two-dimensional normal-ogive model. Tam evaluated this procedure along with several others using such criteria as precision, test information, and computation time. Like Bloxom and Vale, the problem of item selection was not specifically addressed. Tam's item selection procedures assumed ideal item pools where the difficulty of the item was matched to the current ability level of the examinee.

Segall (1996) extended previous work (Bloxom and Vale, 1987; Tam, 1992) by providing a theory-based procedure for item selection that incorporates prior knowledge of the joint distribution of ability. Segall also presented maximum likelihood and Bayesian procedures for item selection and scoring of multidimensional adaptive tests for the general H-dimensional model. By this point in time (mid-1990s), the power of personal computers could support the computations associated with iterative numerical procedures, making approximate-scoring methods of the type suggested by Bloxom and Vale (1987) less desirable. The benefits of the Bayesian approach were evaluated from simulations based on a large-scale high-stakes test: the Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (Segall and Moreno, 1999). Segall demonstrated that for realistic item pools, multidimensional adaptive testing can provide equal or higher precision with about one third fewer items than required by unidimensional adaptive testing.

Luecht (1996) examined the benefits of applying multidimensional item selection and scoring techniques in a licensing/certification context, where mandatory complex content constraints were imposed. He compared the reliability of the ML multidimensional approach (Segall, 1996) to a unidimensional CAT. Results demonstrated that a shorter multidimensional CAT with content constraints could achieve about the same subscore reliability as its longer unidimensional counterpart. Estimated savings in test lengths were consistent with Segall's findings, ranging from about 25–40%.

van der Linden (1999) presents a multidimensional adaptive testing method intended to optimize the precision of a composite measure, where the composite of interest is a linear function of latent abilities. The composite weights associated with each dimension are specified a priori, based on external developer-defined criteria. Ability estimation proceeds according to ML, and item selection is based on a minimum error variance criterion. The error (or sampling) variance of the composite measure is obtained from a linear combination of elements from the inverse Fisher-information matrix. van der Linden demonstrates that for a two-dimensional item pool, a 50-item adaptive test provides nearly uniform measurement precision across the ability space. For shorter tests (of 10 and 30 items), the ML estimates tended to be considerably biased and inefficient.

As pointed out by van der Linden (1999), a Bayesian procedure such as the one proposed by Segall (1996) can lead to inferences that are more informative than those based on ML approaches. This is true when the dimensions underlying the item responses are correlated, and the joint distribution of latent ability is known or estimable with a fair degree of accuracy. The greatest benefit of a Bayesian approach is likely to occur for short to moderate test lengths, or when test length is short relative to the number of dimensions. Added information is provided by the prior distribution, which incorporates known dependencies among the ability variables. The remainder of this chapter provides an explication of Segall's (1996) Bayesian methodology for item selection and scoring of multidimensional adaptive tests.

3.3 Multidimensional Item Selection and Scoring

Two-unidimensional item selection and scoring approaches, based on maximum likelihood and Bayesian estimation, have direct multidimensional counterparts. Associated with each of the two approaches are adaptive item selection rules, and methods for estimating ability and quantifying the level of uncertainty in the ability estimates. However, the extension based on maximum-likelihood theory has serious deficiencies. Although successful unidimensional applications of maximum-likelihood procedures exist, at least two drawbacks are exacerbated in multidimensional CAT. First, toward the beginning of the adaptive test, item selection is hampered by noninformative likelihood functions that possess indeterminate or poorly defined maxima. Consequently, some *adhockery* is needed to bolster item selection procedures in the absence of sufficient data. Second, the ML item selection approach does not consider prior knowledge about the joint distribution of ability. These shortcomings are remedied by the Bayesian methodology.

The objective of the multidimensional adaptive testing algorithms is the efficient estimation of the H-dimensional vector of ability values $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_H\}$. The development of these algorithms is based on five principles and their associated steps drawn primarily from Bayesian theory. The first section below describes the specification of the *prior density* function which characterizes all usable information about the latent ability parameters before the data (item-responses) are observed. The second section describes the *likelihood function,* which provides a mathematical description of the process giving rise to the observed item responses in terms of the unknown ability parameters. The next section outlines the specification of the *posterior distribution*, which summarizes the current state of knowledge (arising from both the observed responses and the prior information). The fourth section casts the issue of *item selection* in terms of a Bayes decision problem for choosing among optimal experiments, and derives an expression for item-specific information measures to be used for choosing among candidate items. The final section derives specific *posterior inference* statements from the posterior distribution, which consist of point estimates of ability.

3.3.1 Prior Density

Here we consider a two-stage process that leads to an individual's item responses. First, an individual is sampled from a population (with a known distribution). That is, a value of $\boldsymbol{\theta}$ is sampled from a population with distribution $f(\boldsymbol{\theta})$. Second, this individual (with fixed $\boldsymbol{\theta}$) is administered multiple test items resulting in a set of binary (correct/incorrect) responses. Under this model, the ability parameters $\boldsymbol{\theta}$ are treated as random variables with distribution $f(\boldsymbol{\theta})$. We shall consider the case in which f is multivariate normal,

$$f(\boldsymbol{\theta}) = (2\pi)^{-H/2}|\boldsymbol{\Phi}|^{-1/2} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})'\boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right], \tag{3.1}$$

with mean vector $\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_H\}$ and $H \times H$ covariance matrix $\boldsymbol{\Phi}$. We further assume that $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ are known. The prior $f(\boldsymbol{\theta})$ encapsulates all usable knowledge about $\boldsymbol{\theta}$ before the item responses have been collected.

Figure 3.1 illustrates the joint density function for a bivariate normal prior with centroid $\mu_1 = \mu_2 = 0$, and covariance terms $\phi_{11} = \phi_{22} = 1$, and $\phi_{12} = 0.6$. Also displayed are the marginal distributions for each dimension, which are also normally distributed with means and variances equal to their corresponding values in the bivariate distribution (i.e., 0 and 1, respectively). From Figure 3.1 it is evident that prior information about θ_1 comes from two sources. First, the range of probable values is confined primarily to the interval $(-2, +2)$, as indicated by the

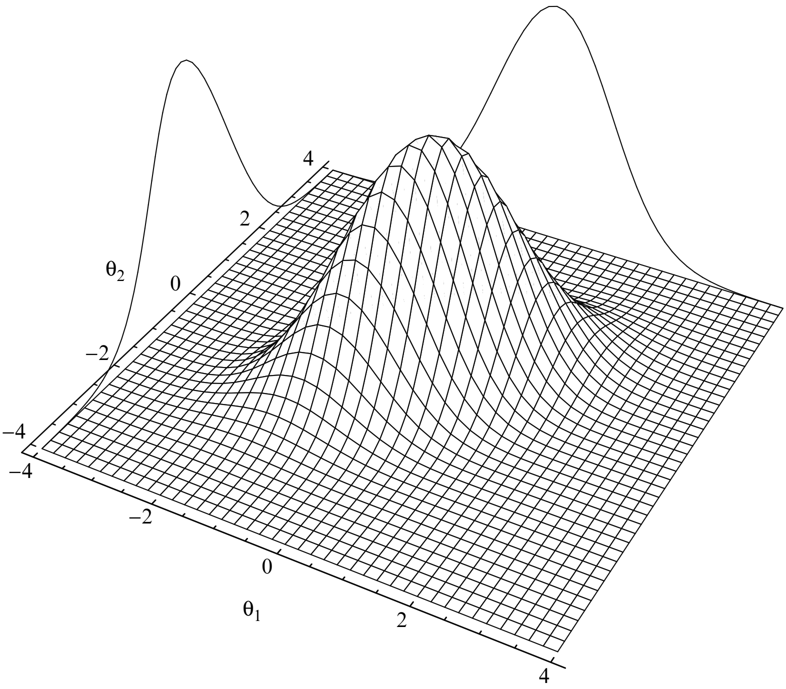


Fig. 3.1 Bivariate normal prior distribution

marginal distribution. Second, small values of θ_1 tend to be associated with small values of θ_2, and a similar association is observed for moderate and large values of θ_1 and θ_2. Thus, a second source of information about θ_1 comes from its association with θ_2. In the general H-dimensional case, prior information derived from correlated dimensions leads to additional precision for the estimation of individual ability parameters.

3.3.2 Likelihood Function

The modeled data consist of a vector of scored responses from an individual examinee $\mathbf{u}_n = \{u_{i_1}, u_{i_2}, \ldots, u_{i_n}\}$ to n adaptively administered items. The set of administered items is denoted by $S_n = \{i_1, i_2, \ldots, i_n\}$, whose elements uniquely identify the items, which are indexed in the pool according to $i = 1, 2, \ldots, I$. For example, if the first item administered was the 50th item in the pool, then $i_1 = 50$; if the second item administered was the 24th item in the pool, then $i_2 = 24$; and so forth. In this case $S_2 = \{50, 24\}$. If item 50 was answered correctly, and item 24 answered incorrectly, then $\mathbf{u}_2 = \{1, 0\}$. For notational simplicity, we shall assume that the nonsubscripted item-response vector $\mathbf{u}$ contains n elements (i.e. $\mathbf{u} \equiv \mathbf{u}_n$).

Furthermore, examinees can be characterized by their standing on H traits denoted by the vector $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_H\}$, where each trait affects performance on one or more test items. The item-response function for item i is given by

$$p_i(\boldsymbol{\theta}) \equiv \text{Prob}(U_i = 1|\boldsymbol{\theta}) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\mathbf{a}_i'(\boldsymbol{\theta} - b_i\mathbf{1})\right]}, \tag{3.2}$$

where U_i is the binary random variable containing the response to item $i(U_i = 1$, if item i is answered correctly; and $U_i = 0$ otherwise), c_i is the probability that a person with infinitely low ability will answer item i correctly, b_i is the difficulty parameter of item i, $\mathbf{1}$ is an $H \times 1$ vector of 1's, D is the constant 1.7, and $\mathbf{a}_i'$ is an $1 \times H$ vector of discrimination parameters for item i.

The form of the item-response function (3.2) is a generalization of the three-parameter logistic model proposed by Birnbaum (1968), with the addition of a linear-compensatory rule for multiple latent traits. For a one-dimensional model ($H = 1$), this function reduces to the standard three-parameter logistic model. Also note that the model possesses a single difficulty parameter b_i. Separate difficulty parameters for each dimension are indeterminate and thus cannot be estimated from observed response data.

Another basic model assumption is that of local or conditional independence (Lord & Novick, 1968). According to this assumption, the joint probability function of a set of n responses $\{u_{i_1}, u_{i_2}, \ldots, u_{i_n}\}$ for an examinee of ability $\boldsymbol{\theta}$ is equal to the product of the probabilities associated with the individual item responses,

$$f(U_{i_1} = u_{i_1}, U_{i_2} = u_{i_2}, \ldots, U_{i_n} = u_{i_n}|\boldsymbol{\theta}) = \prod_{i \in S_n} p_i(\boldsymbol{\theta})^{u_i} q_i(\boldsymbol{\theta})^{1-u_i}, \tag{3.3}$$

where the product runs over the set of administered (or selected) items $S_n = \{i_1, i_2, \ldots, i_n\}$, and $q_i(\boldsymbol{\theta}) = 1 - p_i(\boldsymbol{\theta})$. The ability to express $f(U_{i_1} = u_{i_1}, U_{i_2} = u_{i_2}, \ldots, U_{i_n} = u_{i_n}|\boldsymbol{\theta})$ as a product of terms that depend on individual item-response functions leads to computational simplifications in item selection and scoring. Without the assumption of local independence, expressions required by ML and Bayes methods would be intractable for all but very short tests.

The likelihood function given by

$$L(\mathbf{u}|\boldsymbol{\theta}) = \prod_{i \in S_n} p_i(\boldsymbol{\theta})^{u_i} q_i(\boldsymbol{\theta})^{1-u_i} \tag{3.4}$$

is algebraically equivalent to the joint probability function (3.3). The change in notation, however, reflects a shift in emphasis from the random variables $\mathbf{u}$ with $\boldsymbol{\theta}$ fixed, to the parameters $\boldsymbol{\theta}$, with $\mathbf{u}$ fixed. Since $\mathbf{u}$ are a set of sampled (observed) values of the item responses, the quantity $L(\mathbf{u}|\boldsymbol{\theta})$ is merely a function of the parameters $\boldsymbol{\theta}$.

Figure 3.2 illustrates the likelihood function for the pattern of responses to the eight-item test displayed in Table 3.1. As indicated, the region of highest likelihood

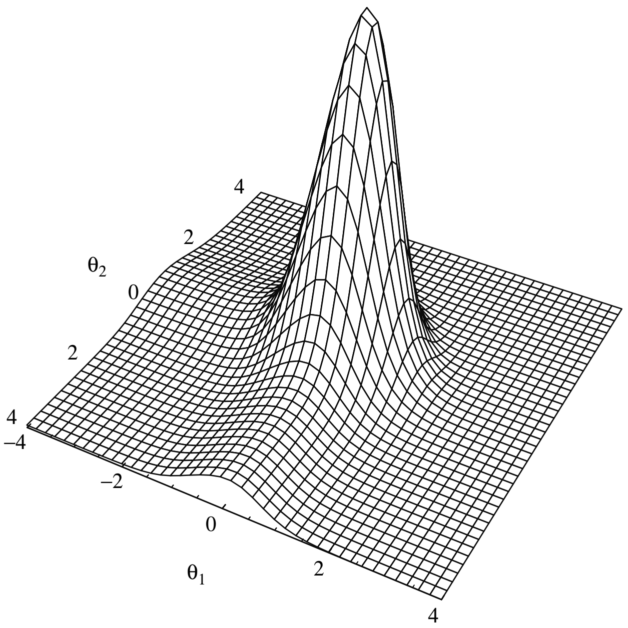


Fig. 3.2 Likelihood function $L(\mathbf{u}|\boldsymbol{\theta})$

Table 3.1 Example item parameters and responses

Item	a_1	a_2	b	c	u	Item	a_1	a_2	b	c	u
1	1.0	0.0	0.00	0.20	1	5	0.0	1.0	0.55	0.20	1
2	0.0	1.0	0.50	0.20	0	6	1.0	1.5	0.95	0.20	0
3	1.0	1.0	0.75	0.30	1	7	2.0	0.0	−1.00	0.20	1
4	2.0	0.0	0.60	0.25	0	8	0.0	1.7	−0.70	0.20	1

occurs near the point $(0.257, 0.516)$. However, this value as a point estimate of ability does not consider information contributed by the prior distribution. Both sources of information are combined, however, by the *posterior density*.

3.3.3 Posterior Density

Given specifications for the prior $f(\boldsymbol{\theta})$ and likelihood $L(\mathbf{u}|\boldsymbol{\theta})$, we are now in a position to make probability statements about $\boldsymbol{\theta}$ given $\mathbf{u}$. These can be made through an application of Bayes' rule that is used to construct the posterior density function

$$f(\boldsymbol{\theta}|\mathbf{u}) = \frac{L(\mathbf{u}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{u})}, \tag{3.5}$$

where $f(\boldsymbol{\theta})$ is the multivariate normal density function (3.1), $L(\mathbf{u}|\boldsymbol{\theta})$ is the likelihood function (3.4), and $f(\mathbf{u})$ is the marginal probability of $\mathbf{u}$ given by

$$f(\mathbf{u}) = \int_{-\infty}^{\infty} L(\mathbf{u}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The posterior density $f(\boldsymbol{\theta}|\mathbf{u})$ contains all existing information about $\boldsymbol{\theta}$ and is used as a basis to provide point and interval estimates of ability parameters $\boldsymbol{\theta}$. As implied by the notation, the posterior represents the distribution of $\boldsymbol{\theta}$ for fixed $\mathbf{u}$, where $\mathbf{u}$ is fixed at the observed response values for the n items.

Figure 3.3 displays the posterior density function for the pattern of responses displayed in Table 3.1. According to (3.5), the height of the posterior density function is proportional to the product of the prior times the likelihood. Since the posterior distribution incorporates the information from the item responses, it is generally less variable than the prior distribution (Figure 3.1). Note also that the posterior density function forms a compromise between the prior and likelihood (Figure 3.2) functions. In the example displayed in Figure 3.3, the mode of the posterior density $(0.233, 0.317)$ forms a compromise between the prior centered at $(0, 0)$ and the data, which suggest that the most likely value of $\boldsymbol{\theta}$ is $(0.257, 0.516)$. In this example, the centroid of the posterior distribution is more heavily influenced by the data than by the prior. In general, the role of the prior diminishes as the test length is increased.

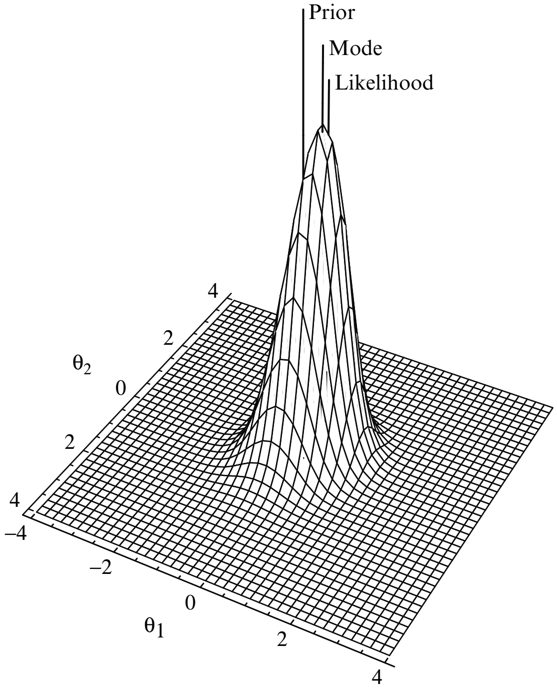


Fig. 3.3 Posterior distribution $f(\boldsymbol{\theta}|\mathbf{u})$

3.3.4 Item Selection

Item selection in adaptive testing can be framed in terms of a specialized area of Bayesian decision theory, namely the area pertaining to choice of experiments (Bernardo & Smith, 1994, p. 63). Suppose that to assist in the measurement of $\boldsymbol{\theta}$, we can choose among several experiments. In the adaptive testing context, the kth experiment would involve the administration of item i_k from the pool of remaining items, denoted by $R_k = \{1, 2, \ldots, I\} \setminus S_{k-1}$. Given that $k - 1$ items have already been administered, the task is to decide which item is to be administered as the next (kth) item from the set of remaining items R_k.

Bayesian decision theory considers the utility of administering each candidate item, and chooses the item with the highest expected utility. The item-specific utility can be expressed as a function of at least two sources: the cost of administering item i, and a loss for making inference $\hat{\boldsymbol{\theta}}$ when the true parameter is $\boldsymbol{\theta}$. When the cost of administration is the same for all items, then it can be shown (O'Hagan, 1994, p. 87) that the following item-selection strategy will maximize utility: *Choose the item that provides the largest decrement in the size of the posterior credibility region.*

This item-selection criterion can be illustrated with an example in which two dimensions are measured, and where the posterior distribution $f(\boldsymbol{\theta}|\mathbf{u}_k)$ is bivariate normal. Figure 3.4 displays two normal posterior distributions with centroids $(0, 0)$, $\rho = 0.6$, and standard deviations $\sigma_1 = \sigma_2 = 1$ (Figure 3.4a) and $\sigma_1 = \sigma_2 = 0.8$ (Figure 3.4b). Associated with each distribution is an isodensity contour—a cross-section of the surface made by a plane parallel to the (θ_1, θ_2)-plane. In general, these contours are elliptical and can be used to define multidimensional credible regions—regions of the posterior distribution containing 50%, 90%, 95%, or 99% of the probability under $f(\boldsymbol{\theta}|\mathbf{u}_k)$. The coverage proportions or percentages can be adjusted by raising or lowering the altitude of the intersecting parallel plane, which in turn influences the size of the elliptical region.

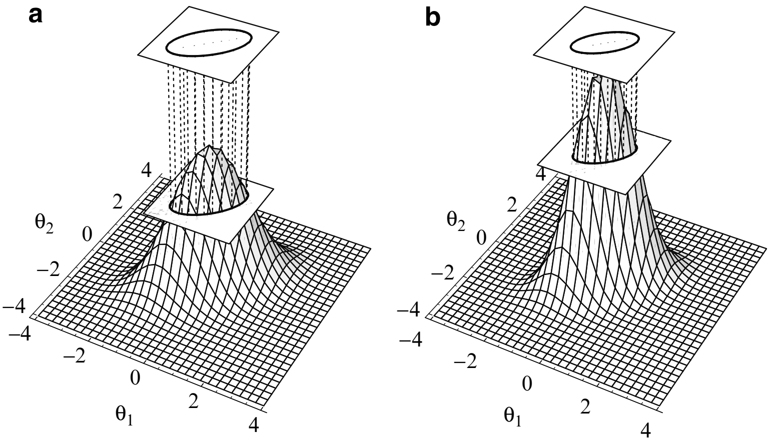


Fig. 3.4 Isodensity contours: (**a**) $N(\boldsymbol{\mu} = 0; \sigma_1 = \sigma_2 = 1.0)$; (**b**) $N(\boldsymbol{\mu} = 0; \sigma_1 = \sigma_2 = 0.8)$

The two elliptical regions displayed in Figure 3.4 each contain about 39% of the probability under their respective densities. Geometrically, this means that the volume of the three-dimensional region between the bivariate normal surface and the (θ_1, θ_2)-plane, bounded laterally by the right elliptic cylinder based on the pictured ellipse, is equal to 0.39 (Tatsuoka, 1971, p. 70). Note, however, that the size (area) of the elliptical credible region in Figure 3.4b is considerably smaller than the region in Figure 3.4a. If these distributions represented the expected posterior outcomes from administering two different items, we would prefer the outcome depicted in Figure 3.4b—the outcome that provides the smallest credible region.

For a normal posterior distribution, the size (length, area, volume) of the credible region is given by

$$V_i = |\boldsymbol{\Sigma}_i|^{1/2} \times g(H) \times \left[\chi^2_H(p)\right]^{H/2}, \tag{3.6}$$

where $\boldsymbol{\Sigma}_i$ is the posterior covariance matrix based on the administration of item i (for $i \in R_k$), H is the number of dimensions, $g(H)$ is a term based on the number of dimensions, and $\chi^2_H(p)$ is the χ^2-value $(df = H)$ located at the $p \times 100$ percentile (Anderson, 1984, p. 263). The coverage probability p can be altered by reference to the appropriate percentiles of the χ^2-distribution. When comparisons are made among the credible regions of different candidate items, all terms except the first remain constant in (3.6). Thus, the item with the smallest value of $|\boldsymbol{\Sigma}_i|$ will provide the largest decrement in the size of the posterior credibility region.

Two related issues hamper the direct application of $|\boldsymbol{\Sigma}_i|$ as an item selection criterion. First, the posterior density $f(\boldsymbol{\theta}|\mathrm{u}_k)$ as parameterized by (3.5) is not normal. Second, the posterior density of interest is based on responses u_{i_k} to candidate items $(i_k \in R_k)$, which have not yet been observed. Both these problems can be solved by approximating the nonnormal posterior with a multivariate normal density based on the curvature at the mode. Specifically, the posterior distribution $f(\boldsymbol{\theta}|\mathbf{u}_k)$ (obtained after the administration of item i_k and observation of the associated response u_{i_k}) can be approximated by a normal distribution having mean equal to the posterior mode $\hat{\boldsymbol{\theta}}^{k-1}$, and covariance matrix $\boldsymbol{\Sigma}_{i|S_{k-1}}$ equal to the inverse of the posterior information matrix evaluated at the mode $\hat{\boldsymbol{\theta}}^{k-1}$:

$$\boldsymbol{\Sigma}_{i|S_{k-1}} = \left[\mathbf{I}_{i|S_{k-1}}\right]^{-1},$$

where the information matrix $\mathbf{I}_{i|S_{k-1}}$ is minus the expected Hessian (second derivative matrix) of the log posterior

$$\mathbf{I}_{i|S_{k-1}} = -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} \ln f(\boldsymbol{\theta}|\mathbf{u}_k)\right], \tag{3.7}$$

and where the expectation is over the random item-response variables $\mathbf{u}_k$. As the conditional notation "$i|S_{k-1}$" implies, the posterior covariance and information matrices associated with the ith item depend on both (a) the characteristics of the candidate item i itself, and (b) the characteristics of the administered items S_{k-1}.

Note that in (3.7), the last element of $\mathbf{u}_k$, namely u_{i_k}, has not yet been observed. However, by taking the expectation of the matrix of second derivatives, the item-response terms $\mathbf{u}_k$ are replaced by their expected values $p_i(\boldsymbol{\theta})$, and the covariance matrix $\boldsymbol{\Sigma}_{i|S_{k-1}}$ can be calculated prior to the administration of the candidate item i. The required posterior mode $\hat{\boldsymbol{\theta}}^{k-1}$ is calculated from the $k-1$ administered items. The information matrix $\mathbf{I}_{i|S_{k-1}}$ is calculated from the item parameters of the $k-1$ administered items, and from the parameters of the ith candidate item. These calculations are detailed in Section 3.6.

One additional simplification can be made by noting that the determinant of the inverse of $\mathbf{I}_{i|S_{k-1}}$ is equal to the reciprocal of the determinant (Searle, 1982, p. 130). With this simplification the item selection criterion becomes

$$\begin{aligned} \left|\boldsymbol{\Sigma}_{i|S_{k-1}}\right| &= \left|[\mathbf{I}i\,|S_{k-1}]^{-1}\right| \\ &= \left|\mathbf{I}_{i|S_{k-1}}\right|^{-1}. \end{aligned} \tag{3.8}$$

Then from inspection of (3.8) we see that the candidate item that maximizes the determinant of the posterior information matrix $\mathbf{I}_{i|S_{k-1}}$ will provide the largest decrement in the size of the posterior credibility region.

The suitability of the item-selection criterion depends in part on how well the nonnormal posterior can be approximated by a normal distribution. Figure 3.5 displays the normal approximation to the posterior distribution based on the eight sets of item responses and parameters provided in Table 3.1. The centroid of the dis-

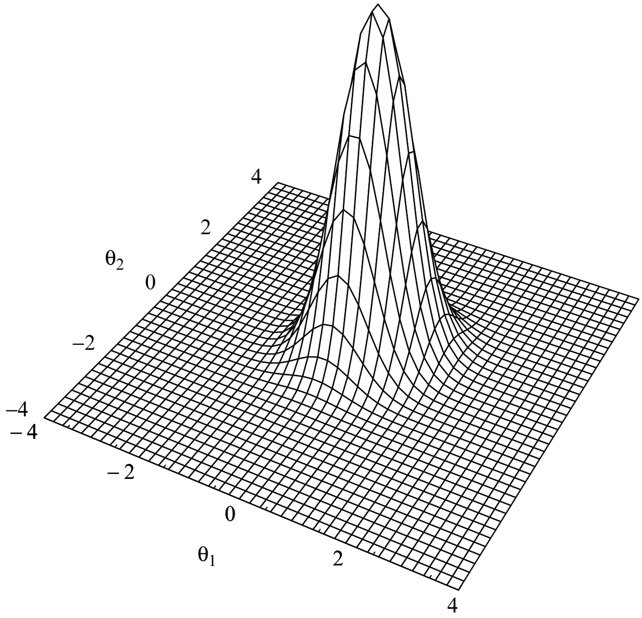


Fig. 3.5 Normal approximation to posterior distribution

tribution was set equal to the mode of the posterior $(0.233, 0.317)$. The covariance matrix $\boldsymbol{\Sigma}$ was computed from the inverse of the information matrix (3.7), which provides

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.343 & 0.092 \\ 0.092 & 0.367 \end{pmatrix}.$$

A comparison of Figures 3.3 and 3.5 suggests that with these data, the normal approximation provides a close representation of the nonnormal posterior. In general, close agreement of the sort displayed here provides support for the use of $\left|\mathbf{I}_{i|S_{k-1}}\right|$ as an inverse indicator of the credibility region's volume.

Selection of the kth adaptively administered item involves evaluation of the determinant of the posterior information matrix for candidate item i, denoted by $\left|\mathbf{I}_{i|S_{k-1}}\right|$. This quantity is computed for each of the unadministered (remaining) items contained in the pool, $i \in R_k$. The candidate item with the largest criterion value will be selected for administration. Computational details are provided in Section 3.6.

3.3.5 *Posterior Inference*

Because the information in the form given by (3.5) is not readily usable, various numerical summaries of the posterior distribution are used. In item-response theory, the posterior distribution is typically characterized by summary measures of central tendency and dispersion. Point estimates of ability are typically defined as the mean or mode of the posterior distribution. In many instances (for tests of moderate to long lengths), these will be nearly identical. However, the mode of the posterior distribution (modal estimate) is better suited than the mean for applications involving higher dimensionality, since far fewer calculations are required. In addition to providing a score (posterior summary-measure of ability), the mode is also required for item selection purposes, as described in the previous section. Accordingly, it is computed after each item response to aid in the selection of the next item, and can be computed at the end of the test to provide an overall or final point estimate of ability. Below we drop the subscripts k in $\mathbf{u}_k$ with the understanding that modal estimates can be computed for any set or super-set of responses by straightforward application of the following formulas.

The modal estimates of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, are those values that correspond to the maximum of the posterior density function: $\hat{\boldsymbol{\theta}} = \max_{\theta} f(\boldsymbol{\theta}|\mathbf{u})$. The estimate $\hat{\boldsymbol{\theta}}$ can be found by taking the H partial derivatives of the log-posterior density function, setting these equal to zero, and solving the H simultaneous nonlinear equations for $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta}|\mathbf{u}) = \mathbf{0}. \tag{3.9}$$

Since there is no closed-form solution to (3.9), an iterative method is required. Suppose we let $\boldsymbol{\theta}^{(m)}$ denote the mth approximation to the value of $\boldsymbol{\theta}$ that maximizes

$\ln f(\boldsymbol{\theta}|\mathbf{u})$; then a better approximation is generally given by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} - \boldsymbol{\delta}^{(m)}, \tag{3.10}$$

where $\boldsymbol{\delta}^{(m)}$ is the $H \times 1$ vector

$$\boldsymbol{\delta}^{(m)} = \left[\mathbf{M}\left(\boldsymbol{\theta}^{(m)}\right)\right]^{-1} \times \frac{\partial}{\partial \boldsymbol{\theta}} \ln f\left(\boldsymbol{\theta}^{(m)}|\mathbf{u}\right). \tag{3.11}$$

The matrix $\mathbf{M}(\boldsymbol{\theta}^{(m)})$ is either the matrix of second partial derivatives $\mathbf{J}(\boldsymbol{\theta})$ (Newton-Raphson method) or the negative posterior information matrix $-\mathbf{I}(\boldsymbol{\theta})$ (Fisher method of scoring)—evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)}$. Modal estimates can be obtained through successive approximations using (3.10) and (3.11). Additional approximations are obtained until the elements of $\boldsymbol{\theta}^{(m)}$ change very little from one iteration to the next. Explicit expressions for the required derivatives and information matrix are provided in Section 3.6.

3.4 Example

This section provides a detailed example of multidimensional item selection and scoring calculations based on the methodology presented in the previous section. The calculations presented below are for a single examinee administered a fixed-length, four-item adaptive test, where items are selected from a pool of eight items spanning two latent dimensions. Note that all computational formulas presented below can be applied to higher-dimensionality ($H > 2$) problems without modification. The matrix notation used enables the calculations to be presented in a way that is independent of the number of dimensions H. Item parameters and associated responses are displayed in Table 3.1. The prior distribution of ability is assumed to be multivariate normal with unit variances, zero means, and correlated dimensions ($\phi_{12} = 0.6$). The basic steps consisting of initialization, provisional ability estimation, item selection, and scoring are detailed below.

3.4.1 Initialization

First, the provisional ability estimate $\hat{\boldsymbol{\theta}}^k$ (where $k = 0$) is set equal to the mean of the prior distribution of ability. In this example the mean of the prior is $\boldsymbol{\mu} = (0, 0)$. The inverse of the prior covariance matrix is also calculated, since it is used in all subsequent item selection calculations:

$$\boldsymbol{\Phi}^{-1} = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1.563 & -0.938 \\ -0.938 & 1.563 \end{pmatrix}.$$

3.4.2 Item Selection

Item selection proceeds by computing the determinant of the posterior information matrix $\left|\mathbf{I}_{i|S_{k-1}}\right|$ for each candidate item ($i \in R_k$), where the information matrix is evaluated at the provisional ability estimate $\hat{\boldsymbol{\theta}}^{k-1}$. From (3.14) (Section 3.6), we see that the posterior information matrix consists of summands arising from three sources:

$$\mathbf{I}_{i|S_{k-1}} = \overbrace{\boldsymbol{\Phi}^{-1}}^{\text{Prior}} + \overbrace{\mathbf{W}_{S_{k-1}}}^{\text{Administered Items}} + \overbrace{\mathbf{W}_i}^{\text{Candidate Item}} . \tag{3.12}$$

The first source is the inverse prior covariance matrix (initialized in the first step). The second source consists of summed $\mathbf{W}$-matrices associated with previously administered items

$$\mathbf{W}_{S_{k-1}} = \sum_{j \in S_{k-1}} \mathbf{W}_j ,$$

where $\mathbf{W}_j$ for item j is defined by (3.13), and the sum $\sum_{j \in S_{k-1}}$ runs over those items already selected. The final term consists of the $\mathbf{W}$-matrix for the candidate item i, also defined by (3.13).

Table 3.2 displays values required to select the first item. These include the $\mathbf{W}_i = \{w_{i(11)}, w_{i(12)} = w_{i(21)}, w_{i(22)}\}$ and posterior information $\mathbf{I}_i = \{I_{i(11)}, I_{i(12)} = I_{i(21)}, I_{i(22)}\}$ matrices and their determinants for the eight candidate items. Since no items have been administered prior to the first item, the posterior information matrix consists of terms from two (rather than three) sources:

$$\mathbf{I}_i = \overbrace{\boldsymbol{\Phi}^{-1}}^{\text{Prior}} + \overbrace{\mathbf{W}_i}^{\text{Candidate Item}} .$$

From inspection of the last column in Table 3.2, item 8 is selected for administration, since it has the largest criterion value: $|\mathbf{I}_8| = 2.612$.

Table 3.2 Item selection calculations: First item

Item i	$w_{i(11)}$	$w_{i(12)}$	$w_{i(22)}$	$I_{i(11)}$	$I_{i(12)}$	$I_{i(22)}$	$\lvert\mathbf{I}_i\rvert$
1	0.482	0.000	0.000	2.044	−0.938	1.563	2.315
2	0.000	0.000	0.330	1.563	−0.938	1.893	2.079
3	0.028	0.028	0.028	1.591	−0.909	1.591	1.703
4	0.302	0.000	0.000	1.865	−0.938	1.563	2.034
5	0.000	0.000	0.310	1.563	−0.938	1.873	2.047
6	0.003	0.005	0.007	1.566	−0.933	1.570	1.588
7	0.287	0.000	0.000	1.850	−0.938	1.563	2.011
8	0.000	0.000	0.672	1.563	−0.938	2.234	2.612

3.4.3 Provisional Ability Estimation

Once the kth selected item has been administered and scored, the provisional ability estimate $\hat{\boldsymbol{\theta}}^{k-1}$ is updated using the full set of k observed responses and administered items to produce $\hat{\boldsymbol{\theta}}^{k}$. Unfortunately, there is no guarantee that the required iterative numerical procedures (Newton–Raphson or Fisher's scoring algorithms) will converge if the starting value for the ability parameter $\boldsymbol{\theta}^{(1)}$ in (3.10) and (3.11) is far from the maximum. However, satisfactory convergence behavior is generally obtained by setting the starting value $\boldsymbol{\theta}^{(1)}$ equal to the posterior mode obtained from the previous calculations (i.e. $\boldsymbol{\theta}^{(1)} = \hat{\boldsymbol{\theta}}^{k-1}$). The starting value $\boldsymbol{\theta}^{(1)}$ for the first provisional update is set equal to the mean of the prior. Typically if one method fails (Newton–Raphson or Fisher's scoring), the other will converge to the true maximum. In practice, it is useful to program both methods, using one as a backup in case the other fails to converge. Using the Newton–Raphson algorithm based on (3.10) and (3.11), a correct response to item 8 results in the posterior mode estimate $\hat{\boldsymbol{\theta}}^{1} = (0.102, 0.170)$.

3.4.4 Item Selection and Scoring Cycle

The previous two steps of item selection and provisional ability estimation are repeated until the test termination criterion has been satisfied—in this example, until four items have been administered. Tables 3.3 and 3.4 display key summary calculations used in item selection and scoring.

Table 3.3 provides a summary of the administered items i, responses u, and modal ability estimates ($\hat{\theta}_1$ and $\hat{\theta}_2$). As indicated, the first item selected was item 8. A correct response to this item ($u = 1$) resulted in a two-dimensional Bayes mode estimate of $\hat{\theta}_1 = 0.102$ and $\hat{\theta}_2 = 0.170$. The second, third, and fourth items selected were 1, 4, and 2, respectively. Note that a correct response to an item resulted in higher $\hat{\theta}$ values along *both* dimensions. Similarly, an incorrect response also influenced the provisional ability estimates of both dimensions—resulting in lower $\hat{\theta}$ scores. The final ability estimate after providing an incorrect response to the fourth item was $\hat{\boldsymbol{\theta}}^{4} = (0.034, -0.075)$.

Table 3.3 Item selection and scoring summary

			Posterior Mode	
Sequence k	Item i	u	$\hat{\theta}_1$	$\hat{\theta}_2$
0	–	–	0.000	0.000
1	8	1	0.102	0.170
2	1	1	0.457	0.343
3	4	0	0.103	0.171
4	2	0	0.034	−0.075

Table 3.4 Item selection indices $|\mathbf{I}_{i\,|\,S_{k-1}}|$

	Adaptive Test Sequence k			
Item i	1	2	3	4
1	2.315	⇒ 3.265	–	–
2	2.079	2.893	3.782	⇒ 5.247
3	1.703	2.591	3.920	4.617
4	2.034	3.247	⇒ 5.572	–
5	2.047	2.864	3.754	5.201
6	1.588	2.355	3.390	4.345
7	2.011	2.692	2.992	4.658
8	⇒ 2.612	–	–	–
$\hat{\boldsymbol{\theta}}^{k-1}$	$(0.000, 0.000)$	$(0.102, 0.170)$	$(0.457, 0.343)$	$(0.103, 0.171)$

Table 3.5 Item selection calculations: Fourth item

Item i	$w_{i(11)}$	$w_{i(12)}$	$w_{i(22)}$	$I_{i(11)\|S_3}$	$I_{i(12)\|S_3}$	$I_{i(22)\|S_3}$	$\|\mathbf{I}_{i\,\|\,S_3}\|$
1	0.491	0.000	0.000	–	–	–	–
2	0.000	0.000	0.396	2.538	−0.938	2.413	5.247
3	0.058	0.058	0.058	2.597	−0.879	2.075	4.617
4	0.485	0.000	0.000	–	–	–	–
5	0.000	0.000	0.378	2.538	−0.938	2.395	5.201
6	0.010	0.015	0.022	2.548	−0.923	2.039	4.345
7	0.206	0.000	0.000	2.745	−0.938	2.017	4.658
8	0.000	0.000	0.455	–	–	–	–

Table 3.4 provides the item selection criteria based on the $\left|\mathbf{I}_{i\,|\,S_{k-1}}\right|$ indices. The last row displays the provisional ability estimate used in the evaluation of the posterior information matrix $\mathbf{I}_{i\,|\,S_{k-1}}$. As indicated, the first item selected was Item 8, which had the maximum value of the criterion $|\mathbf{I}_8| = 2.612$. The second item selected was item 1, which had the largest criterion value among the remaining candidate items, and so forth.

Table 3.5 displays calculations associated with the selection of the last (fourth) item. The elements of the $\mathbf{W}_i$-matrices evaluated at the provisional ability estimate $\hat{\boldsymbol{\theta}}^3 = (0.103, 0.171)$ are displayed in columns 2–4. Columns 5–7 display elements of the posterior information matrices $\mathbf{I}_{i\,|\,S_3}$ for candidate items (those not previously administered). These matrices are computed from (3.14), which, after the third administered item, take the form

$$\mathbf{I}_{i\,|\,S_3} = \overbrace{\boldsymbol{\Phi}^{-1}}^{\text{Prior}} + \overbrace{\mathbf{W}_1 + \mathbf{W}_4 + \mathbf{W}_8}^{\text{Administered Items}} + \overbrace{\mathbf{W}_i}^{\text{Candidate Item}} \quad .$$

The item selection criteria computed from the determinant of the posterior information matrices are displayed in the last column of Table 3.5. The maximum value is associated with item 2, which was administered as the fourth item in the adaptive sequence.

3.5 Discussion

The multidimensional item-selection and scoring methods presented here provide an opportunity for increased measurement efficiency over unidimensional adaptive testing methods. However, before these benefits can be fully realized, several practical issues including item parameter specification and item exposure must be addressed. Segall (1996) provides a discussion of a straightforward approach for item-parameter specification based on unidimensional 3PL estimates. Also discussed is an approach to exposure control that places a ceiling on the administration rates of the pool's most informative items, while sacrificing only small to moderate amounts of precision.

By applying Bayesian principles to multidimensional IRT, item-selection and scoring algorithms can be specified that enhance the precision of adaptive test scores. This increase in precision or efficiency can be potentially large for test scores obtained from batteries that measure several highly correlated dimensions. However, the magnitude of the efficiency gain over unidimensional methods is likely to be test- or battery-specific. For specific applications, efficiency gains can be investigated through a direct comparison of unidimensional and multidimensional approaches. To this end, this chapter presents the underlying theoretical and computational bases for the multidimensional approach—increasing the accessibility of this new methodology to interested researchers and practitioners.

3.6 Appendix: Computational Formulas

First Partial Derivatives

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta} | \mathbf{u}) = D \sum_{i \in S} v_i \mathbf{a}_i - \boldsymbol{\Phi}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}),$$

where the sum runs over items contained in S, and

$$v_i = \frac{[p_i(\boldsymbol{\theta}) - c_i]\,[u_i - p_i(\boldsymbol{\theta})]}{(1 - c_i) p_i(\boldsymbol{\theta})}.$$

Second Partial Derivatives

$$\mathbf{J}_S(\boldsymbol{\theta}) \equiv \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \ln f(\boldsymbol{\theta} | \mathbf{u}) = D^2 \sum_{i \in S} \mathbf{a}_i \mathbf{a}_i' w_i - \boldsymbol{\Phi}^{-1},$$

where

$$w_i = \frac{q_i(\boldsymbol{\theta})[p_i(\boldsymbol{\theta}) - c_i \left[c_i u_i - p_i^2(\boldsymbol{\theta})\right]}{p_i^2(\boldsymbol{\theta})(1 - c_i)^2}.$$

Posterior Information Matrix

The information matrix for a set of items S is given by

$$\mathbf{I}_S = -\mathrm{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\ln f(\boldsymbol{\theta}|\mathbf{u})\right] = \boldsymbol{\Phi}^{-1} + \sum_{i\in S}\mathbf{W}_i,$$

where

$$\mathbf{W}_i = D^2\mathbf{a}_i\mathbf{a}_i' w_i^* \tag{3.13}$$

and

$$w_i^* = \frac{q_i(\boldsymbol{\theta})}{p_i(\boldsymbol{\theta})} \times \left[\frac{p_i(\boldsymbol{\theta}) - c_i}{1 - c_i}\right]^2.$$

The posterior information matrix associated with candidate item i,

$$\mathbf{I}_{i|S_{k-1}} = \boldsymbol{\Phi}^{-1} + \boldsymbol{W}_i + \sum_{j\in S_{k-1}}\mathbf{W}_j, \tag{3.14}$$

is formed from **W**-terms associated with previously administered items S_{k-1}, and from a **W**-term associated with candidate item i.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* New York: John Wiley & Sons.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian theory.* New York: John Wiley & Sons.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Bloxom, B. M. & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centroid and dispersion of theta.* Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404.

O'Hagan, A. (1994). *Kendall's advanced theory of statistics: Bayesian inference* (Vol. 2B). London: Edward Arnold.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

Searle, S. R. (1982). *Matrix algebra useful for statistics.* New York: John Wiley & Sons.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354.

Segall, D. O. & Moreno, K. E. (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait.* Unpublished doctoral dissertation, Columbia University, New York.

Tatsuoka, M. M. (1971). *Multivariate analysis: Techniques for educational and psychological research.* New York: John Wiley & Sons.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24,* 398–412.

Chapter 4
Multidimensional Adaptive Testing with Kullback–Leibler Information Item Selection

Joris Mulder and Wim J. van der Linden

Although multidimensional item response theory (IRT) (e.g., McDonald, 1962, 1997; Reckase, 1985, 1997; Samejima, 1974) has been available for some time, it has been applied much less frequently in adaptive testing than unidimensional IRT. The main reason for this was a lack of computational power. However, recently this condition has changed dramatically. Even for the more time-intensive Bayesian treatments of multidimensional models, regular PCs now have plenty of computational power to deal with them in a large variety of applications.

A promising area of multidimensional IRT is testing for diagnosis, with its goal of extracting as much information as possible about the multiple abilities required, for instance, to solve complex learning tasks (e.g., Boughton, Yoa & Lewis, 2006; Yao & Boughton, 2007). The test batteries used in this area are generally time-intensive and could profit greatly from the high efficiency of multidimensional adaptive testing.

Earlier explorations of multidimensional implementations of adaptive testing are offered in Bloxom and Vale (1987), Fan and Hsu (1996), Luecht (1996), Segall (1996, this volume, chap. 3), van der Linden (1999; 2005, chap. 9), Veldkamp and van der Linden (2002), and Mulder and van der Linden (2009). Some of these implementations used classical statistics; the others were Bayesian. The former explored the use of likelihood-based ability estimation and item selection based on a criterion of optimality defined on the Fisher information matrix. The latter used the responses to the items to update the (joint) posterior distribution of the ability parameters and based item selection on these updates. For a more general review of the differences between these two approaches in unidimensional adaptive testing, see van der Linden and Pashley (this volume, chap. 1).

J. Mulder
Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands

W.J. van der Linden (✉)
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_4,

The research reported in this chapter continues the Bayesian tradition. More specifically, the goal was to evaluate the behavior of Bayesian item selection based on the Kullback–Leibler (KL) information measure. KL information is a versatile measure for the "distance" between two distributions with many applications elsewhere in statistics (e.g., Lehmann and Casella, 1998, sect. 4.5). In the current research, it was combined with the posterior distribution of the ability parameters to define new multidimensional item-selection criteria, to analyze the preferences of these criteria for the statistical features of the items, and to evaluate the impact of these criteria on the change of the posterior distribution during the test. As our intention was to understand theoretically the complex interactions between these quantities, our approach was mainly analytical. In particular, we refrained from the more superficial evaluation of item selection based on simulating adaptive testing.

Since the statistical inference in multidimensional adaptive testing has to be optimized with respect to multiple ability components, attention should be paid to the status of each of them. We therefore also examined the behavior of KL-based item selection in the special case when not all parameters are intentional but some of them should be ignored as nuisance parameters. Nuisance parameters occur in educational testing when, in addition to the primary abilities that are tested, the test items appear to be sensitive to abilities related to their format as well. Examples of such abilities are language abilities or abilities to deal with graphical information. In addition, we examined KL-based item selection when the inference has to be optimized with respect to a weighted combination of ability parameters. The last case arises, for instance, when the test has to be scored by a scalar that summarizes the intentional abilities measured by the test.

4.1 Multidimensional IRT model

A multidimensional version of the common three-parameter logistic (3PL) model for dichotomous items is assumed throughout this chapter. The model gives the probability of a correct response to item i as a function of a p-dimensional ability $\boldsymbol{\theta}$ as

$$\begin{aligned} P_i(\boldsymbol{\theta}) &\equiv P(U_i = 1|\boldsymbol{\theta}, \mathbf{a}_i, d_i, c_i) \\ &\equiv c_i + \frac{1 - c_i}{1 + \exp\left(-\mathbf{a}_i^T \boldsymbol{\theta} + d_i\right)}, \end{aligned} \tag{4.1}$$

where $\mathbf{a}_i$ is the (column) vector with the discrimination parameters of item i for each of the component abilities $\theta_l, l = 1, \ldots, p$ in $\boldsymbol{\theta}$, d_i is the scalar parameter for the difficulty of the item, and c_i is the height of the lower asymptote for the probability of a correct response necessary to deal with the effects of random guessing. The probability of an incorrect response is given by $Q_i(\boldsymbol{\theta}) = 1 - P_i(\boldsymbol{\theta})$. As is customary in adaptive testing, the items are assumed to be calibrated with enough precision to consider their parameters as known.

The following notation will be used to describe the testing process:

N: size of the item pool;
n: length of the adaptive test;
$i = 1, \ldots, N$: index for the items in the pool;
$k = 1, \ldots, n$: index for the items in the test;
i_k: index of the item in the pool administered as the kth item in the test;
S_{k-1}: set of first $k-1$ items administered in the test;
$\mathbf{u}_{k-1}$: vector with the responses to the first $k-1$ items;
R_k: set of items in the pool from which item k is picked, i.e., $\{1, \ldots, N\} \backslash S_{k-1}$.

4.2 Bayesian Estimation of $\boldsymbol{\theta}$

In a Bayesian approach, the posterior distribution of $\boldsymbol{\theta}$ is updated after each observed response. Suppose $k-1$ responses have already been observed and the current posterior has density $f(\boldsymbol{\theta}|\mathbf{u}_{k-1})$. The update of the posterior distribution after the response to the kth item follows from Bayes' theorem as

$$f(\boldsymbol{\theta}|\mathbf{u}_k) = \frac{f(u_{i_k}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1})}{f(u_{i_k}|\mathbf{u}_{k-1})}, \tag{4.2}$$

where $f(u_{i_k}|\boldsymbol{\theta})$ is the model probability for the response $U_{i_k} = u_{i_k}$ to item i_k given by

$$f(u_{i_k}|\boldsymbol{\theta}) = P_{i_k}^{u_{i_k}}(\boldsymbol{\theta}) Q_{i_k}^{1-u_{i_k}}(\boldsymbol{\theta}) \tag{4.3}$$

and $f(u_{i_k}|\mathbf{u}_{k-1})$ is the posterior predictive probability function defined as

$$f(u_{i_k}|\mathbf{u}_{k-1}) = \int_{\boldsymbol{\theta}} f(u_{i_k}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) d\boldsymbol{\theta}. \tag{4.4}$$

Unfortunately, the model probability in (4.3) does not have a conjugate family from which the prior distribution of $\boldsymbol{\theta}$ could be chosen. This means that the posterior distribution cannot be in the same family as the prior distribution and the update of the latter never reduces to a simple update of its parameters. For this reason, Owen (1975) proposed a Bayesian method for adaptive testing more generally known as restricted Bayesian updating. His method assumes a normal prior distribution and replaces the subsequent posterior distributions by a normal with the same mean and variance as the true posterior. This normal approximation is not unreasonable; Chang and Stout 1993) have shown that, under mild nonparametric assumptions, the posterior in (4.2) is asymptotically normal with a mean equal to $\boldsymbol{\theta}$. We will use this approximation in our numerical examples and then assume a multivariate normal posterior distribution with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and density

$$f(\boldsymbol{\theta}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})\right]. \tag{4.5}$$

It is important to note that the off-diagonal elements of $\boldsymbol{\Sigma}$ are generally nonzero; that is, the abilities in the posterior distribution correlate. One reason for the correlation is the fact that the model probability in (4.3) does not factor according to the individual abilities. Another reason is possible correlation between the abilities in the assumed prior distribution. This makes sense, for instance, when the prior distribution is chosen to reflect empirical correlation in a population of interest.

Usual choices of point estimates of $\boldsymbol{\theta}$ are the mean and the mode of its posterior distribution known as the expected (EAP) and maximum a posteriori (MAP) estimates, respectively. The former requires numerical integration; in our examples, we will use the Gauss–Hermite formulas from Glas (1992). The latter can be determined using a Newton–Raphson procedure; see, for instance, Segall (1996).

Two features of Bayesian inference for the IRT model in adaptive testing are worth noting. First, the well-known statistical relations

$$E(\boldsymbol{\theta}) = E(E(\boldsymbol{\theta}|u)) \tag{4.6}$$

and

$$\text{var}(\boldsymbol{\theta}) = E(\text{var}(\boldsymbol{\theta}|u)) + \text{var}(E(\boldsymbol{\theta}|u)) \tag{4.7}$$

aqre used (e.g., Gelman et al., 1995, sect.1.8). The first equation implies that the average posterior mean of $\boldsymbol{\theta}$ over the distribution of possible response on an item is equal to its prior mean. The second equation is more interesting in that it implies an average posterior variance of $\boldsymbol{\theta}$ is smaller than its prior variance by an amount equal to the variance of the posterior mean over the distribution of possible responses. Thus, generally, from an item selection point of view, it is desirable to select items that result in posteriors with highly variable means.

Second, in item selection the discrimination parameters appear to be the critical parameters in the model probabilities in (4.3). For an analysis that leads to this conclusion, see Mulder and van der Linden (2009). Here, we only illustrate the role of the discrimination parameters graphically for a unidimensional ability space; the generalization to a multidimensional space is immediate. Figure 4.1 shows a standard normal prior distribution along with the posterior distributions for a correct response on an item with varying discrimination parameters a whereas d and c are assumed to be fixed at zero. (Note that the posterior distributions for an incorrect response reflect these curves at the line $\theta = 0$.) For $a \to \infty$, the probability of a correct response approaches the step function

$$f(U=1|\theta) = \begin{cases} 1 & \text{if } \theta > 0, \\ 0.5 & \text{if } \theta = 0, \\ 0 & \text{if } \theta < 0, \end{cases}$$

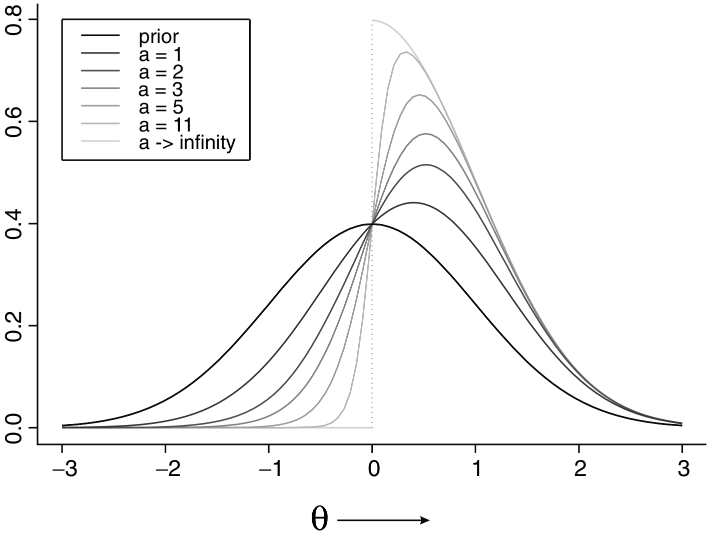


Fig. 4.1 Plots of standard normal prior density and posterior densities for a correct response on an item with $d = 0$ and $c = 0$ for an increasing discrimination parameter a. Note: For $a \to \infty$, the likelihood approaches a step function and the posterior distribution becomes discontinious

which explains the discontinuity of its posterior at $\theta = 0$. The figure clearly shows a decrease in posterior variance with the size of the discrimination parameter. Items with a larger discrimination parameter are therefore more informative about the ability parameter.

4.3 Kullback–Leibler Information

KL information for two alternative densities f and g for a continuous variable X is defined as

$$K(f, g) = E_f \left[\log \frac{f(X)}{g(X)} \right] \tag{4.8}$$

$$= \int f(x) \log \frac{f(x)}{g(x)} dx, \tag{4.9}$$

where E_f denotes expectation under f. For a discrete variable, KL information is defined as

$$K(p, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}, \tag{4.10}$$

where p and q are probability functions (Cover and Thomas, 1991). Note that these definitions admit any dimension of X and that the result is always a scalar.

Unlike item selection based on Fisher information, which becomes a full matrix for a multidimensional ability space, the use of KL information allows us to generalize immediately from unidimensional to multidimensional adaptive testing.

Lehmann and Casella (1998) point at a useful feature of KL information: The larger $K(f, g)$ is, the easier it is to discriminate between f and g. Chang and Ying's (1996) choice of a KL-based item-selection criterion of unidimensional adaptive testing (see below) was motivated by this feature.

The same feature explains the popular interpretation of KL information as a measure for the distance between two distributions. This interpretation is motivated by two of its features. Firstly, for any two distributions f and g, $K(f, g) \geq 0$. Using a similar argument as in Lehmann and Casella (1998), this can be proven as follows:

$$\begin{aligned} K(f, g) &= E_f\left[\log \frac{f(x)}{g(x)}\right] \\ &= -E_f\left[\log \frac{g(x)}{f(x)}\right] \\ &\geq -\log E_f\left[\frac{g(x)}{f(x)}\right] \\ &= -\log \int \left[f(x)\frac{g(x)}{f(x)}dx\right] \\ &= -\log \int [g(x)dx] \\ &= 0, \end{aligned}$$

where the third step is based on Jensen's inequality $Eh(X) \geq h(EX)$ for a convex function $h(x) = -\log(x)$. Secondly,

$$\begin{aligned} K(f, f) &= E_f\left[\log \frac{f(x)}{f(x)}\right] \\ &= E_f\,[\log 1] \\ &= 0, \end{aligned}$$

which makes intuitive sense because the distance between two identical densities should be equal to zero.

On the other hand, KL information is not symmetric, i.e., $K(f, g) \neq K(g, f)$. The lack of symmetry follows from the fact that the expectation in (4.8) is taken over the first argument. Also, neither does the triangular inequality hold.

We will nevertheless follow the interpretation of KL information as a distance measure in this chapter and refer to it as "L distance" but admit that the only correct interpretation of it is as the information about g at f.

The KL distance between two subsequent posterior distributions in adaptive testing quantifies the impact of the information in the response to the item on our uncertainty about $\boldsymbol{\theta}$. If we use the earlier normal approximation, the KL distance has to be calculated between two normal distributions. For normal densities f and g with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the expression for the KL distance between f and g is

$$\begin{aligned} K(f,g) &= \int f(x)\log\frac{f(x)}{g(x)}dx \\ &= \int \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \\ &\quad \times \log\left(\frac{\sigma_2}{\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}+\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)\right)dx. \end{aligned} \tag{4.11}$$

Writing out the second factor and distributing the first factor, we obtain

$$\begin{aligned} K(f,g) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \left(x^2\left(\frac{1}{2\sigma_2^2}-\frac{1}{2\sigma_1^2}\right) + x\left(\frac{\mu_1}{\sigma_1^2}-\frac{\mu_2}{\sigma_2^2}+\frac{\mu_2^2}{2\sigma_2^2}-\frac{\mu_1^2}{2\sigma_1^2}\right)\right) \\ \times \int \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)dx \end{aligned} \tag{4.12}$$

which simplifies to the closed form

$$K(f,g) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{\sigma_1^2}{2\sigma_2^2} + \frac{(\mu_2-\mu_1)^2}{2\sigma_2^2} - \frac{1}{2} \tag{4.13}$$

Two interesting properties follow. First, the KL distance depends only on the means μ_1 and μ_2 through their difference. If the variances are equal, that is, for $\sigma_1 = \sigma_2 = \sigma$,

$$K(f,g) = \frac{(\mu_2-\mu_1)^2}{2\sigma^2}. \tag{4.14}$$

Thus, when the variances are equal, the distance between two normal distributions is symmetric and proportional to the difference between their means and the reciprocal of their common variance. Both relations make intuitive sense.

Second, if the means are equal, the distance depends only on the variances through their ratio; that is, for $\mu_1 = \mu_2$,

$$K(f,g) = \log\left(\frac{\sigma_2}{\sigma_1}\right) + \frac{1}{2}\left(\frac{\sigma_1}{\sigma_2}\right)^2 - \frac{1}{2}. \tag{4.15}$$

The KL distance is then no longer symmetric. The distance is plotted as a function of $\sigma_2/\sigma_1 > 0$ in the first panel of Figure 4.2. The second panel of this figure shows that

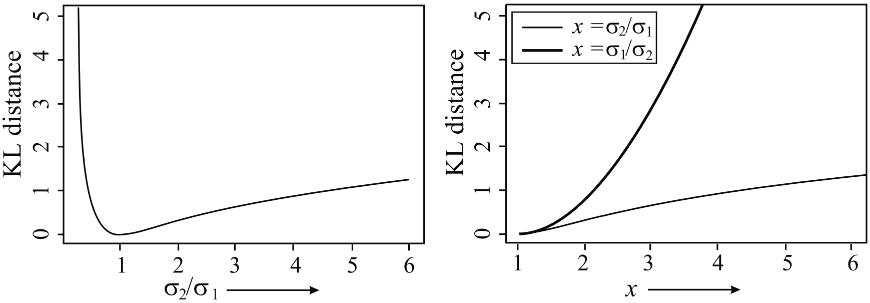


Fig. 4.2 Kullback–Leibler distance between two normal distributions with equal means as a function of the ratio of their standard deviations σ_2/σ_1

when $\mu_1 = \mu_2$ and $\sigma_1 > \sigma_2$, $K(f, g) > K(g, f)$. This property will be used later in this chapter when the KL distance between subsequent posterior distributions is calculated.

4.3.1 Mutual Information

An interesting generalization of KL information is the mutual information measure. For two continuous random variables X and Y, mutual information is defined as

$$I_M(X,Y) = \int_Y \int_X f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy. \tag{4.16}$$

For discrete variables, the integrals are replaced by sums.

Mutual information I_M is a measure of the amount of information X provides about Y. Expression (4.16) clearly shows symmetry; I_M is thus also a measure of the amount of information in X about Y. Further, when X and Y are independent, $I(X,Y) = 0$. This feature follows directly from the fact that under this condition $f(x,y) = f(x)f(y)$.

Mutual information between two identical random variables X is also known as the entropy of X. Entropy is a measure of how much information is contained in a variable X. For more on entropy and information theory, see Cover and Thomas (1991).

4.4 Item Selection Using KL Information

Three Bayesian criteria for item selection based on KL information are discussed. Each of the criteria involves a different, plausible summary of the information in a response to a test item. Interestingly, the criteria can be used in both unidimensional and multidimensional adaptive testing. The only modification involved in a change of dimensionality is in the dimension of $\boldsymbol{\theta}$; otherwise, the mathematical expressions for each of the criteria remain the same. We believe this to be an advantage over item selection based on Fisher information, which for the multidimensional case generalizes to a matrix. In order to use this matrix for a ranking of the items in the pool, an additional criterion of optimality has to be specified. For the use of criteria from the optimal design literature (e.g., Silvey, 1980), such as D-, A-, and E-optimality, for this purpose, see Mulder and van der Linden 2009).

4.4.1 Posterior Expected Kullback–Leibler Information

Chang and Ying (1996) were the first to propose the use of KL information for item selection in unidimensional adaptive testing. Their item-selection rule was based on the distance between the response distributions on the candidate item at the current ability estimate $\hat{\theta}$ and the true ability θ, with the expectation taken over the response.

For a multidimensional ability space, this measure generalizes to

$$K_i(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) = E\left[\log \frac{f\left(U_i|\hat{\boldsymbol{\theta}}\right)}{f(U_i|\boldsymbol{\theta})}\right] \tag{4.17}$$

$$= P_i\left(\hat{\boldsymbol{\theta}}\right)\log\frac{P_i\left(\hat{\boldsymbol{\theta}}\right)}{P_i\left(\boldsymbol{\theta}\right)} + Q_i\left(\hat{\boldsymbol{\theta}}\right)\log\frac{Q_i\left(\hat{\boldsymbol{\theta}}\right)}{Q_i\left(\boldsymbol{\theta}\right)}, \tag{4.18}$$

where $P_i(\cdot)$ is the response function for item i in (4.3) and $Q_i(\cdot) = 1 - P_i(\cdot)$. As already indicated, the larger the measure, the better the item discriminates between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

For a test of n items, KL information is equal to

$$\begin{aligned} K_n(\boldsymbol{\theta}\hat{;}\theta) &\equiv E\left[\log\frac{f(U_1,\ldots,U_n|\hat{\boldsymbol{\theta}})}{f(U_1,\ldots,U_n|\boldsymbol{\theta})}\right] \\ &= E\left[\log\prod_{i=1}^{n}\frac{f(U_i|\hat{\boldsymbol{\theta}})}{f(U_i|\boldsymbol{\theta})}\right] \\ &= \sum_{i=1}^{n} E\left[\log\frac{f(U_i|\hat{\boldsymbol{\theta}})}{f(U_i|\boldsymbol{\theta})}\right] \\ &= \sum_{i=1}^{n} K_i(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}), \end{aligned}$$

where the second step follows from the usual assumption of conditional independence between the responses given $\boldsymbol{\theta}$. The KL information measure is thus additive in the items.

Because the examinee's true ability $\boldsymbol{\theta}$ is unknown, Chang and Ying proposed to integrate (4.18) over a confidence interval for $\boldsymbol{\theta}$. For the multidimensional case, the proposal would generalize to integration over a confidence region.

Alternatively, Veldkamp and van der Linden (2002) proposed a Bayesian version of this item-selection criterion. This criterion, which we will refer to as posterior expected KL information, K^B, selects the item

$$\arg\max_{i_k \in R_k} K_{i_k}^{B}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) = \arg\max_{i_k \in R_k} \int_{\boldsymbol{\theta}} K_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}; \boldsymbol{\theta}) f(\boldsymbol{\theta} | \mathbf{u}_{k-1}) d\boldsymbol{\theta}, \tag{4.19}$$

where $\mathbf{u}_{k-1}$ is the vector of item responses of the previously $k-1$ administered items and $\hat{\boldsymbol{\theta}}_{k-1}$ is the EAP estimate of the ability

$$\hat{\boldsymbol{\theta}}_{k-1} = \int \boldsymbol{\theta} f(\boldsymbol{\theta} | \mathbf{u}_{k-1}) d\boldsymbol{\theta} \tag{4.20}$$

after $k-1$ items have been administered. This idea implies the selection of an item that maximally discriminates between the EAP estimate and the other abilities in the multidimensional ability space covered by the current posterior. Simulating adaptive testing, these authors showed that K^B results in ability estimates of approximately the same accuracy as the Bayesian item-selection criterion based on the information matrix by Segall (1996; this volume, chap. 3). They further note that K^B is easier to use in real-world adaptive testing than Segall's criterion because when the test has to satisfy content constraints, e.g., bounds on the numbers of items for certain topics or skill categories, use of the shadow-test approach (van der Linden, this volume, chap. 2) is straightforward.

The information surfaces of K^B for two items with parameters $\mathbf{a}_1^T = [1\ 0.5]$, $\mathbf{a}_2^T = [0\ 0.8]$, and all other parameters fixed at zero are plotted in Figure 4.3. Remember that we use a normal approximation for the current posterior distribution. The integration in (4.19) was performed using the Gauss–Hermite formulas in Glas (1992). The surfaces were calculated by varying the posterior mean $(\mu_{\theta_1}, \mu_{\theta_2})$ with a covariance matrix with fixed positive covariances between the abilities. The influence of the posterior covariances on the criterion will be discussed later in this chapter.

The same items were used in Mulder and van der Linden (2009) to check if items discriminating along one ability dimension are generally more informative than those discriminating along all dimensions. For Fisher information with D-optimality or A-optimality, this was found to be the case. However, the surface in Figure 4.3 shows that item 1 is generally more informative than item 2 and, therefore, that K^B has a preference for items that discriminate along all dimensions. One possible reason is the dependence between the abilities in the joint posterior of $\boldsymbol{\theta}$. Also, as will become clear later, when the posterior dependence is strong, item information tends

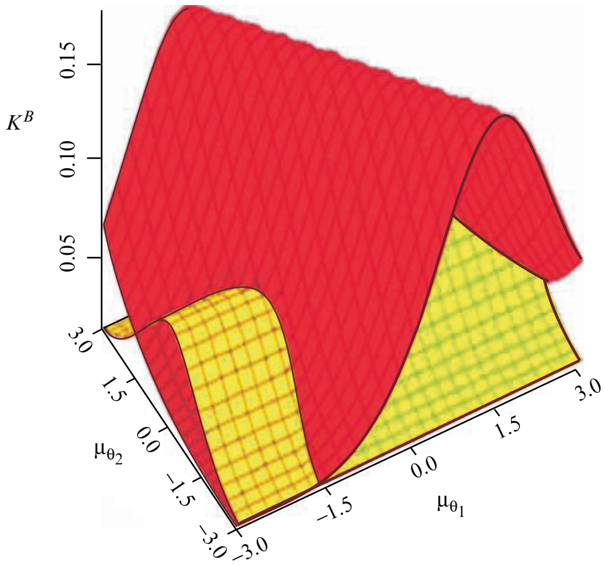


Fig. 4.3 Kullback–Leibler distance between subsequent posterior distributions, K^B, for two items with $\mathbf{a}_1^T = [1.5\ 0.5]$, $\mathbf{a}_2^T = [0\ 0.8]$ and $d_1 = d_2 = 0$ and $c_1 = c_2 = 0$ for a multivariate normal prior distribution with fixed covariance matrix [1 0.4; 0.4 1] and varying mean $\boldsymbol{\mu} = (\mu_{\theta_1}, \mu_{\theta_2})$. Note: The darker surface is for the item with $\mathbf{a}_1^T = [1.5\ 0.5]$

to depend on the discrimination parameters through their sum rather than their individual values. Finally, note that K^B is constant wherever $\mathbf{a}^T \cdot \boldsymbol{\theta}$ is. This was also the case for the common factor $g(\boldsymbol{\theta}; \mathbf{a}, d, c)$ in the elements in the Fisher information matrix in Mulder and van der Linden (2009). The constancy is a direct consequence of the presence of the same linear combination of the abilities parameters in the response function in (4.3).

4.4.2 KL Distance between Subsequent Posteriors

If the posterior distribution of $\boldsymbol{\theta}$ did not change much after administering an item, which is the case, for instance, when a hard item is administered to a low-ability test taker, the item should be avoided. This observation suggests selecting items with the largest expected distance between the current and new posterior distributions of $\boldsymbol{\theta}$. The KL distance can be used to formalize this argument. Because one of the possible responses to the candidate item would move the posterior toward the examinee's true ability and the other would move it away from it, the criterion is defined as the expected KL distance across the response distribution.

More formally, this item-selection criterion, denoted as K^P, selects the item

$$\begin{aligned}&\arg\max_{i_k \in R_k} K^P_{i_k}[f(\boldsymbol{\theta}|\mathbf{u}_{k-1})]\\&\qquad = \arg\max_{i_k \in R_k} \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) K(f(\boldsymbol{\theta}|\mathbf{u}_{k-1}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k})), \quad (4.21)\end{aligned}$$

where $K(f(\boldsymbol{\theta}|\mathbf{u}_{k-1}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_i))$ is the KL distance between the two posterior densities and $f(u_{i_k}|\mathbf{u}_{k-1})$, $u_{i_k} = 0, 1$, is the posterior predictive probability function in (4.4).

A plot of the information surfaces for K^P for the same two items as for the previous criterion in Figure 4.3 resulted in surfaces that were only slightly wider than the information surfaces of K^B but otherwise entirely similar. These plots are therefore omitted here.

4.4.3 Mutual Information

The third criterion is based on the definition of mutual information in (4.16). For unidimensional adaptive testing, Weissman (2007) suggested selecting items that maximize mutual information between the test taker's current posterior distribution and the response distribution on the candidate item, the idea being that these items are closest to $\boldsymbol{\theta}$ according to the posterior information in the previous items. Let I_M denote this criterion. Formally, the best item according to the criterion is

$$\begin{aligned}&\arg\max_{i_k \in R_k} I_M(\boldsymbol{\theta}; u_{i_k})\\&\quad = \arg\max_{i_k \in R_k} \sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1}) \log \frac{f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1})}{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) f(u_{i_k}|\mathbf{u}_{k-1})} d\boldsymbol{\theta}. \quad (4.22)\end{aligned}$$

Note that $f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1}) = f(u_{i_k}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1})$; hence, the factor with the logarithm in (4.22) simplifies to

$$\log \frac{f(u_{i_k}|\boldsymbol{\theta})}{f(u_{i_k}|\mathbf{u}_{k-1})},$$

which is the log of the ratio between the model and the posterior predictive probability of u_{i_k}.

An important interpretation of mutual information comes from the relation

$$I_M(\boldsymbol{\theta}; u_{i_k}) = H(\boldsymbol{\theta}) - H(\boldsymbol{\theta}|u_{i_k}),$$

with

$$H(\boldsymbol{\theta}) = -\sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, u_{i_k}) \log f(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

and

$$H(\boldsymbol{\theta}|u_{i_k}) = -\sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, u_{i_k}) \log f(\boldsymbol{\theta}|u_{i_k}) d\boldsymbol{\theta}.$$

$H(\boldsymbol{\theta})$ is the information in $\boldsymbol{\theta}$ and $H(\boldsymbol{\theta}|u)$ is the information in $\boldsymbol{\theta}$ upon the observation of u. Consequently, mutual information $I_M(\boldsymbol{\theta}; u_{i_k})$ can be interpreted as the reduction in uncertainty about $\boldsymbol{\theta}$ due to the response $U_{i_k} = u_{i_k}$.

Again, a plot with the surfaces of the mutual information measure for the same items as in Figure 4.3 yielded surfaces a little wider than those for both K^B and K^P but otherwise entirely similar.

4.5 Relationship between Selection Criteria

Rewriting (4.19) and (4.21), it can be shown that K^B and K^P only differ in their definitions of the probabilities of a correct and incorrect response. For the former, it holds that

$$K_{i_k}^B\left(\hat{\boldsymbol{\theta}}_{k-1}\right) = \int_{\boldsymbol{\theta}} K_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) d\boldsymbol{\theta} \tag{4.23}$$

$$\begin{aligned}
&= \int_{\boldsymbol{\theta}} \left(P_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \log \frac{P_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right)}{P_{i_k}(\boldsymbol{\theta})} + Q_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \log \frac{Q_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right)}{Q_{i_k}(\boldsymbol{\theta})} \right) \\
&\qquad \times f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) d\boldsymbol{\theta} \\
&= P_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \log P_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) + Q_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \log Q_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \\
&\quad - P_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log P_{i_k}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad - Q_{i_k}\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log Q_{i_k}(\boldsymbol{\theta}) d\boldsymbol{\theta},
\end{aligned} \tag{4.24}$$

where $P_{i_k}(\hat{\boldsymbol{\theta}}_{k-1})$ and $Q_{i_k}(\hat{\boldsymbol{\theta}}_{k-1})$ are the model probabilities for a correct and incorrect response evaluated at the point estimate of $\boldsymbol{\theta}$.

Likewise, substituting the definition of the KL distance in (4.13) and rewriting the result, the latter can be shown to be equal to

$$\begin{aligned}
K_{i_k}^P[f(\boldsymbol{\theta})|\mathbf{u}_{k-1}] = {} & f(0|\mathbf{u}_{k-1}) \log f(0|\mathbf{u}_{k-1}) + f(1|\mathbf{u}_{k-1}) \log f(1|\mathbf{u}_{k-1}) \\
& - f(0|\mathbf{u}_{k-1}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log Q_{i_k}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
& - f(1|\mathbf{u}_{k-1}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log P_{i_k}(\boldsymbol{\theta}) d\boldsymbol{\theta},
\end{aligned} \tag{4.25}$$

where $f(1|\mathbf{u}_{k-1})$ and $f(0|\mathbf{u}_{k-1})$ are the posterior predicted probabilities of a correct and incorrect response in (4.4).

Thus, K^P operates more in agreement with standard Bayesian methodology than K^B; it uses the posterior distribution to calculate an update of the response probabilities. Unlike K^B, K^P therefore takes the uncertainty in the ability estimate into account and is more robust with respect to ability estimation. This suggests K^P to be a better item-selection criterion than K^B.

For a multivariate normal posterior $f(\boldsymbol{\theta}|\mathbf{u}_{k-1})$ and guessing parameter $c_{i_k} = 0$,

$$K_{i_k}^B\left(\hat{\boldsymbol{\theta}}_{k-1}\right) \leq K_{i_k}^P[f(\boldsymbol{\theta}|\mathbf{u}_{k-1})], \tag{4.26}$$

with equality at $-\mathbf{a}^T\widehat{\boldsymbol{\theta}}_{k-1} + d = 0$. This relationship holds because, for this symmetric posterior, $P_{i_k}(\hat{\boldsymbol{\theta}}_{k-1}) = \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1})P_{i_k}(\boldsymbol{\theta})d\boldsymbol{\theta}$.

For $c_{i_k} = 0$, the two criteria appear to take their maximum value at $-\mathbf{a}_{i_k}^T\widehat{\boldsymbol{\theta}}_{k-1} + d_{i_k} = 0$. Although intuitively clear, due to the complexity of their expressions, it has not been possible to construct a formal proof of this feature.

Similarly as mutual information I_M in (4.22), KL information between subsequent posteriors, K^P, can also be written as the KL distance between a joint probability distribution and the product of its marginal distributions:

$$\begin{aligned} K_{i_k}^P[f(\boldsymbol{\theta}|\mathbf{u}_{k-1})] &= \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1})K(f(\boldsymbol{\theta}|\mathbf{u}_{k-1}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k})) \\ &= \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log \frac{f(\boldsymbol{\theta}|\mathbf{u}_{k-1})}{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k})} d\boldsymbol{\theta} \\ &= \sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(u_{i_k}|\mathbf{u}_{k-1}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \log \frac{f(u_{i_k}|\mathbf{u}_{k-1}) f(\boldsymbol{\theta}|\mathbf{u}_{k-1})}{f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1})} d\boldsymbol{\theta}. \end{aligned} \tag{4.27}$$

It can be concluded that K^P is the KL distance between the joint and product distributions of $\boldsymbol{\theta}$ and the response on the candidate item.

It can also be shown that mutual information is the average KL distance between the new and current posteriors. (Observe that K^P is defined as the average KL distance between the current and new posteriors; the measure is not symmetric!) This is proved by the following derivation:

$$\begin{aligned} I_M(\boldsymbol{\theta}; u_{i_k}) &= \sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1}) \log \frac{f(\boldsymbol{\theta}, u_{i_k}|\mathbf{u}_{k-1})}{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) f(u_{i_k}|\mathbf{u}_{k-1})} d\boldsymbol{\theta} \\ &= \sum_{u_{i_k}=0}^{1} \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k}) f(u_{i_k}|\mathbf{u}_{k-1}) \log \frac{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k}) f(u_{i_k}|\mathbf{u}_{k-1})}{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) f(u_{i_k}|\mathbf{u}_{k-1})} d\theta \end{aligned}$$

$$= \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k}) \log \frac{f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k})}{f(\boldsymbol{\theta}|\mathbf{u}_{k-1})} d\boldsymbol{\theta}$$

$$= \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) K[f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1})]. \tag{4.28}$$

In the previous section, we observed that, for equal means $\mu_1 = \mu_2$ and variances $\sigma_1 > \sigma_2$, the KL distance between two normal densities f and g is less robust with respect to the ratio of the variances than the distance between g and f. This property suggests that, because the expected posterior variance decreases upon the administration of an item, $K[f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1})]$ should be more robust than $K[f(\boldsymbol{\theta}|\mathbf{u}_{k-1}), f(\boldsymbol{\theta}|\mathbf{u}_{k-1}, u_{i_k})]$. If so, I_M would be more robust with respect to error in the ability estimate than K^P.

4.6 Special Status of Some of the Ability Parameters

As described in van der Linden (1996), five different cases of multidimensional testing can be distinguished depending on whether (i) the abilities measured by the test are intentional or a nuisance and (ii) the interest is in scoring separate abilities or a composite of them. The item-selection criteria above are particularly apt for the case of all abilities intentional. We now explore what modifications of K^B, K^P, and I_M are necessary when some of the abilities are nuisances or the interest is only in the estimation of a given linear combination of intentional abilities. In order to provide insight into the behavior of these modified criteria, a few numerical examples will be given.

4.6.1 Nuisance Abilities

Assume that the ability space consists of h intentional abilities and $p - h$ nuisance abilities. The following notation for the ability vector $\boldsymbol{\theta}$ will be used: $\boldsymbol{\theta}^I$, a vector of length h consisting of the intentional abilities,m and $\boldsymbol{\theta}^N$ a vector of length $p - h$ consisting of the nuisance abilities.

Veldkamp and van der Linden (2002) proposed a modification of K^B to deal with the presence of nuisance abilities following a profile-likelihood approach. For the selection of the kth item, the modification comprises of the following steps: First, the EAP estimate of the nuisance abilities from the first $k-1$ responses is substituted in the probability function for the response on the candidate item. Although this approximation of the function may be poor in the beginning of the test, the EAP estimates of the nuisance abilities do converge to their true values with an increase in the number of items. For candidate item i_k, the approximation is defined as

$$f\left(u_{i_k}|\boldsymbol{\theta}^I, \hat{\boldsymbol{\theta}}^N_{k-1}\right) = P_{i_k}\left(\boldsymbol{\theta}^I, \hat{\boldsymbol{\theta}}^N_{k-1}\right)^{u_{i_k}} Q_{i_k}\left(\boldsymbol{\theta}^I, \hat{\boldsymbol{\theta}}^N_{k-1}\right)^{1-u_{i_k}}. \tag{4.29}$$

Second, the approximate probability function is substituted into the KL distance $K_{i_k}(\hat{\boldsymbol{\theta}}^I_{k-1};\boldsymbol{\theta}^I)$. Third, integrating $K_{i_k}(\hat{\boldsymbol{\theta}}^I_{k-1};\boldsymbol{\theta}^I)$ over the marginal posterior of $\boldsymbol{\theta}^I$, which has density function

$$f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right) = \int \ldots \int f\left(\boldsymbol{\theta}|\mathbf{u}_{k-1}\right) d\boldsymbol{\theta}^N, \tag{4.30}$$

gives the selection criterion

$$\begin{aligned}\arg\max_{i_k\in R_k} K^B_{i_k}\left(\hat{\boldsymbol{\theta}}^I_{k-1}\right)\\ &= \arg\max_{i_k\in R_k} \int_{\boldsymbol{\theta}} K_{i_k}\left(\hat{\boldsymbol{\theta}}^I_{k-1};\boldsymbol{\theta}^I\right) f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right) d\boldsymbol{\theta}^I. \end{aligned}\tag{4.31}$$

As for the modification of the KL distance between subsequent posteriors, K^P, an obvious Bayesian choice is to use the largest expected distance between the subsequent marginal posterior distributions of the intentional abilities as a criterion. The measure is defined as

$$\begin{aligned}K\left(f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right), f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1},u_{i_k}\right)\right) &= \int f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right)\log\frac{f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right)}{f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1},u_{i_k}\right)}d\boldsymbol{\theta}^I\\ &= \int f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right)\log\frac{f\left(u_{i_k}|\mathbf{u}_{k-1}\right)}{f\left(u_{i_k}|\boldsymbol{\theta}^I\right)}d\boldsymbol{\theta}^I,\end{aligned}\tag{4.32}$$

where $f\left(u_{i_k}|\boldsymbol{\theta}^I\right)$ is likelihood for the intentional abilities obtained as

$$f\left(u_{i_k}|\boldsymbol{\theta}^I\right) = \int f\left(u_{i_k}|\boldsymbol{\theta}^I,\boldsymbol{\theta}^N\right) f\left(\boldsymbol{\theta}^N|\boldsymbol{\theta}^I,\mathbf{u}_{k-1}\right) d\boldsymbol{\theta}^N. \tag{4.33}$$

Consequently, the kth item is selected according to

$$\begin{aligned}\arg\max_{i_k\in R_k} K^P_{i_k}\left[f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right)\right]\\ = \arg\max_{i_k\in R_k}\sum_{u_{i_k}=0}^{1} f\left(u_{i_k}|\mathbf{u}_{k-1}\right) K\left(f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1}\right), f\left(\boldsymbol{\theta}^I|\mathbf{u}_{k-1},u_{i_k}\right)\right).\end{aligned}\tag{4.34}$$

A main difference between the two modifications is inserting the estimates of the nuisance abilities in the likelihood for K^B instead of marginalizing the posterior distribution in K^P. The former does not allow for any uncertainty about the nuisance abilities, but the latter does. Consequently, near the beginning of the test, when the uncertainty is high, the modified version of K^P is expected to be more appropriate for item selection than K^B.

When nuisance abilities are present, mutual information I_M can be modified analogously to K^P. As shown in (4.28), I_M is equal to the average KL distance between the new and current posteriors. This relationship suggests the replacement of the full posterior distributions for all abilities in (4.28) by the marginal distributions of the intentional abilities, i.e., the modified criterion

$$\begin{aligned}
&\arg\max_{i_k \in R_k} I_M\left(\boldsymbol{\theta}^N; u_{i_k}\right) \\
&\quad = \arg\max_{i_k \in R_k} \sum_{u_{i_k}=0}^{1} f\left(u_{i_k} | \mathbf{u}_{k-1}\right) K\left(f\left(\boldsymbol{\theta}^I | \mathbf{u}_{k-1}, u_{i_k}\right), f\left(\boldsymbol{\theta}^I | \mathbf{u}_{k-1}\right)\right). \quad (4.35)
\end{aligned}$$

Numerical Example

Again, as our interest is in the effects of the discrimination parameters, we fix the difficulty and guessing parameters at $d = 0$ and $c = 0$. Suppose we have a two-dimensional ability space with θ_1 intentional and θ_2 a nuisance ability and a current posterior distribution of θ given by

$$f(\boldsymbol{\theta} | \mathbf{u}_{k-1}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}\right). \quad (4.36)$$

Figure 4.4 shows the new posterior distribution of the intentional ability θ_1 upon a response to three different items with a_{i1} fixed at 1.5 but different values for a_{i2}.

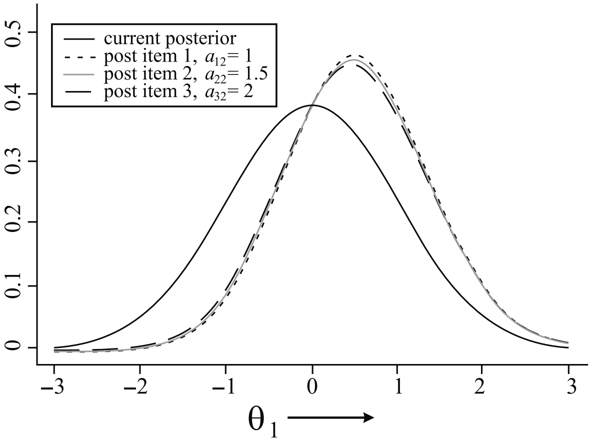


Fig. 4.4 Current and new (marginal) posterior distributions of θ_1 upon observation of a correct response for three different values for $a_i = (1.5, a_{i2}), i = 1, 2, 3$. Note: The posterior distributions for an incorrect response are mirrored about $\theta_1 = 0$

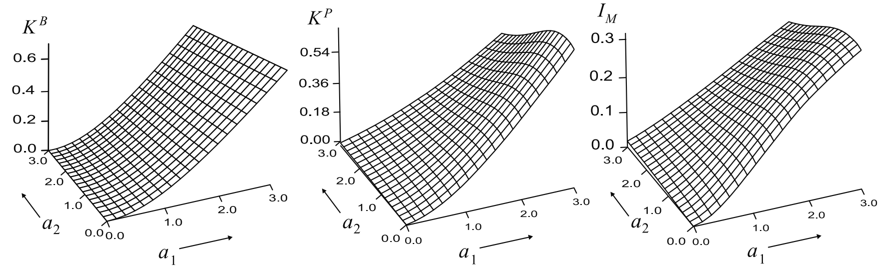


Fig. 4.5 The three modified versions of K^B, K^P, and I_M as a function of the discrimination parameters when θ_1 is intentional and θ_2 is a nuisance ability for a normal current posterior distribution with mean $(0, 0)$, covariance matrix $[10.3; 0.31]$, and the item parameters fixed at $d = 0$ and $c = 0$

The item with the smallest value for a_2 results in the smallest posterior variance of the intentional ability. It can be considered to be the most informative item. Obviously, an item that discriminates less along a nuisance ability causes less noise in the responses used for inference about an intentional ability.

We next illustrate the differences among the three modified versions of information measures K^B, K^P, and I_M for the same posterior as in (4.36). The surfaces of the three measures as a function of a_1 and a_2 are shown in Figure 4.5. These surfaces were also calculated using the Gauss–Hermite formulas in Glas (1992). As the measures are used to select the best item, only the ordinal features of their surfaces should be evaluated. Each of the three measures increases with the discrimination parameter for the intentional ability, a_1, which is what they should do. However, for the current posterior and $\theta_2 = 0$, the version of K^B appears to be independent of a_2, whereas both K^P and I^M decrease slightly with it for the more substantial values of a_1. This independence of K^B does not make much sense. As we just saw, a large value of a_2 actually results in a larger posterior variance of θ_1 and a plausible criterion should therefore decrease with it. We therefore conclude that the modified versions of K^P and I^M are the best criteria for item selection in the presence of a nuisance ability.

4.6.2 Composite Ability

Suppose the test has to be scored only for a linear combination of the abilities

$$\sum_{l=1}^{p} \lambda_l \theta_l, \tag{4.37}$$

with $0 < \lambda_l < 1$ for $l = 1, \ldots, p$ and $\sum_{l=1}^{p} \lambda_l = 1$.

Veldkamp and van der Linden (2002) suggest a change of variables so that one of the new ability variables, say the first, is precisely the composite in (4.37). This ability is then treated as an intentional ability and all other abilities are considered to be nuisance dimensions in the new ability space. We follow the suggestion for each of the three KL-based item-selection criteria.

The mapping of $\boldsymbol{\theta}$ to a new ability space $\boldsymbol{\xi}$ is

$$\boldsymbol{\xi} = \mathbf{B}\boldsymbol{\theta}, \tag{4.38}$$

where $\mathbf{B}$ is a $p \times p$ matrix with weights λ_l, $l = 1, \ldots, p$, in the first row so that $\xi_1 = \sum_{l=1}^{p} \lambda_l \theta_l$. The other elements in $\mathbf{B}$ can be chosen arbitrarily as long as $\mathbf{B}$ is invertible. We choose them to yield orthogonal new abilities ξ_l, which have the advantage of the other abilities ξ_l, $l = 2, \ldots, p$, not having any influence on the item-selection criterion. This choice implies

$$\mathbf{B} = \begin{bmatrix} \lambda_1 & \lambda_2 & \lambda_3 & \cdots & \lambda_p \\ -\lambda_2 & \lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_3 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\lambda_p & \lambda_{p-1} \end{bmatrix}. \tag{4.39}$$

As $\boldsymbol{\theta} = \mathbf{B}^{-1}\boldsymbol{\xi}$, the parameterization of the item response function in (4.3) becomes

$$\begin{aligned} \tilde{P}_i(\boldsymbol{\xi}) &= c_i + \frac{1 - c_i}{1 + \exp(-\mathbf{a}_i^T \mathbf{B}^{-1}\boldsymbol{\xi} + d_i)} \\ &= c_i + \frac{1 - c_i}{1 + \exp(-\tilde{\mathbf{a}}_i^T \boldsymbol{\xi} + d_i)}, \end{aligned} \tag{4.40}$$

with a new discrimination parameter for the intentional ability ξ_1 equal to

$$\tilde{a}_1 = \frac{\sum_{l=1}^{p} \lambda_l a_l}{\sum_{l=1}^{p} \lambda_l^2}.$$

Finally, the new model probability function is

$$f(u_{i_k}|\boldsymbol{\xi}) = \tilde{P}_{i_k}^{u_{i_k}}(\boldsymbol{\xi})\tilde{Q}_{i_k}^{1-u_{i_k}}(\boldsymbol{\xi}), \tag{4.41}$$

with $\tilde{Q}_{i_k}(\boldsymbol{\xi}) = 1 - \tilde{P}_{i_k}(\boldsymbol{\xi})$.

A useful property of the linear transformation in (4.38) is that when $\boldsymbol{\theta}$ is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the new ability vector $\boldsymbol{\xi}$ has a multivariate normal distribution with mean $\mathbf{B}\boldsymbol{\mu}$ and covariance matrix $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T$.

Applying the modification of K^B above to the case of ξ_1 an intentional and $\xi_2, \ldots, \xi_p$ nuisance abilities, the item-selection criterion changes into

$$\arg\max_{i_k \in R_k} K^B_{i_k}(\widehat{\xi}_{1,k-1}) = \arg\max_{i_k \in R_k} \int K_{i_k}(\widehat{\xi}_{1,k-1}; \xi_1) f(\xi_1 | \mathbf{u}_{k-1}) d\xi_1, \quad (4.42)$$

with $\widehat{\xi}_{1,k-1}$ the EAP estimate of ξ_1 and $f(\xi_1|\mathbf{u}_{k-1})$ the marginal posterior density obtained by integrating all nuisance abilities out of the joint posterior distribution of $\boldsymbol{\xi}$.

Likewise, the modified versions of K^P and I_M are

$$\begin{aligned}&\arg\max_{i_k \in R_k} K^P_{i_k}[f(\xi_1|\mathbf{u}_{k-1})] \\ &\quad = \arg\max_{i_k \in R_k} \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) K(f(\xi_1|\mathbf{u}_{k-1}), f(\xi_1|\mathbf{u}_{k-1}, u_{i_k})) \quad (4.43)\end{aligned}$$

and

$$\begin{aligned}&\arg\max_{i_k \in R_k} I_M(\xi_1; u_{i_k}) \\ &\quad = \arg\max_{i_k \in R_k} \sum_{u_{i_k}=0}^{1} f(u_{i_k}|\mathbf{u}_{k-1}) K[f(\xi_1|\mathbf{u}_{k-1}, u_{i_k}), f(\xi_1|\mathbf{u}_{k-1})], \quad (4.44)\end{aligned}$$

respectively.

Numerical Example

Suppose the interest is in the composite $0.5\theta_1 + 0.5\theta_2$. As before, the posterior distribution of $\boldsymbol{\theta}$ is the one in (4.36), the difficulty and guessing parameters are arbitrarily fixed at $d = 0$ and $c = 0$, and we analyze the new information measures as a function of the discrimination parameters. The orthogonal matrix in (4.39) is used to map $\boldsymbol{\theta}$ to $\boldsymbol{\xi}$; that is,

$$\mathbf{B} = \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & 0.5 \end{bmatrix}. \quad (4.45)$$

Hence,

$$\begin{aligned}\tilde{\mathbf{a}}_i &= \left(B^{-1}\right)^T \mathbf{a}_i \\ &= \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \mathbf{a}_i \\ &= \begin{bmatrix} a_1 + a_2 \\ -a_1 + a_2 \end{bmatrix}.\end{aligned}$$

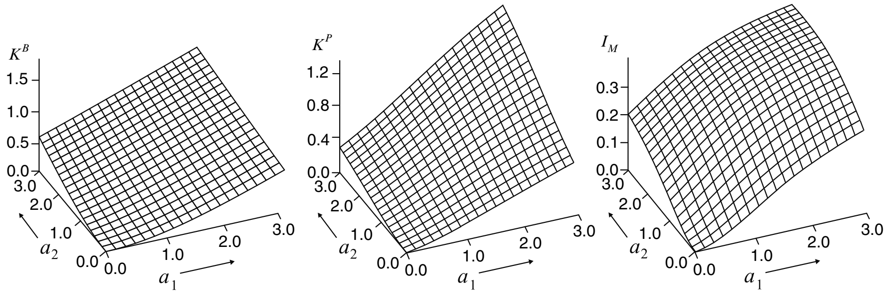


Fig. 4.6 The three modified versions of K^B, K^P, and I_M as a function of the discrimination parameters when the interest is in a linear combination of θ_1 and θ_2 for a normal current posterior distribution with mean $(0, 0)$, covariance matrix [10.3; 0.31], and the item parameters fixed at $d = 0$ and $c = 0$. Note: The transformation matrix is $B = [0.50.5; -0.50.5]$

Figure 4.6 shows the surfaces of K^B, K^P, and I_M as a function of the discrimination parameters. The surfaces reveal a similar general shape for all three criteria. Each of them increases with a_1 and a_2. The fact that some of them increase locally somewhat faster than the others will hardly influence item selection. Thus, these results do not suggest any differential performance of the three criteria for item selection with the interest in the estimation of an equally weighted linear composite of θ_1 and θ_2.

4.7 Posterior Covariance

In all previous examples, the posterior covariance between the ability parameters was fixed. In this section, the role of this covariance is explored. In particular, the interest is in whether the covariance modifies the role of the discrimination parameters in the selection of the items for the three criteria. For instance, as demonstrated earlier, unlike the criteria of D- and A-optimality for the Fisher information matrix (Mulder and van der Linden, 2009), item-selection criteria K^B, K^P, and I_M do not tend to select items that mainly test a single ability dimension. In this section, we explore whether this selection behavior changes as a function of the correlation between the abilities.

In each of the examples below, the case of two abilities was investigated with the sum of the two discrimination parameters fixed at $a_1 + a_2 = 3$. The choice allowed us to vary the relative size of the individual discrimination parameters without increasing the overall quality of the item. As before, the difficulty and guessing parameters were fixed at $d = 0$ and $c = 0$. Furthermore, the current posterior was assumed to be multivariate normal,

$$f(\boldsymbol{\theta}|\mathbf{u}_{k-1}) \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_{12} \\ \sigma_{12} & 1 \end{bmatrix}\right). \tag{4.46}$$

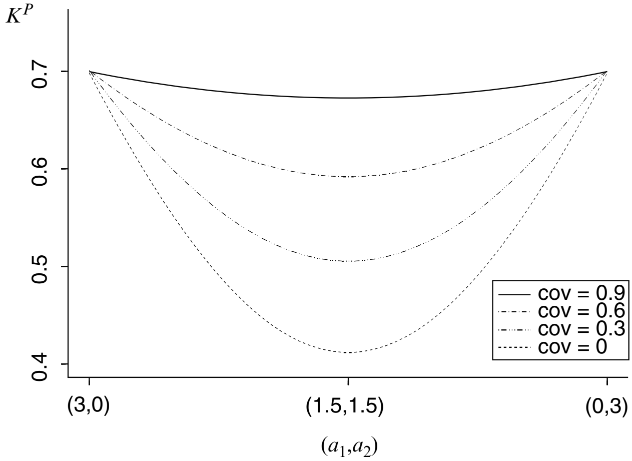


Fig. 4.7 Information measure K^P as a function of the combination of the two discrimination parameters $a_1 + a_2 = 3$ for different posterior covariances σ_{12} (both abilities intentional)

We varied the size of the posterior correlation σ_{12} and evaluated its impact on the three KL-based information measures.

First, the case of both θ_1 and θ_2 being intentional is reported. Figure 4.7 displays the KL distance between subsequent posterior distributions, K^P, for different values of σ_{12}. The larger the covariance, the higher K^P for any combination of the discrimination parameters. The Bayesian explanation for this relationship is that of borrowing information. Because θ_1 and θ_2 correlate, information in the response on either of them also means information on the other. Except for a difference in scale, the shapes of the curves for K^B and I_M were entirely similar to those in Figure 4.7; their figures are therefore omitted here.

The second case is that of θ_1 being intentional and θ_2 a nuisance ability. From (4.31), it can be seen that K^B is independent of the posterior covariance. The other two measures, K^P and I_M, do depend on the posterior covariance. The curves for I_M are shown in Figure 4.8a. The curves for K^P were similar and are therefore omitted. Obviously, an item that mainly tests the nuisance ability (i.e., high value of a_2) becomes more informative when this nuisance ability correlates highly with the intentional ability. The principle of borrowing information also explains this fact. The curve for K^B was similar to that in Figure 4.8a for $\sigma_{12} = 0$. The intuitively more attractive behavior of K^P and I_M for correlated abilities therefore does not hold for K^B.

The final case is the estimation of the composite ability $\lambda_c = \lambda_1\theta_1 + \lambda_2\theta_2$. The orthogonal transformation in (4.45) was used to map ability space $\boldsymbol{\theta}$ to $\boldsymbol{\xi}$, with $\xi_1 = \lambda_1\theta_1 + \lambda_2\theta_2$ as the new intentional ability. Figure 4.8b shows the curves for I_M. The curves for K^P were entirely similar and are omitted here. Although θ_1 and θ_2 have equal weights, the most informative combination of discrimination parameters was $\mathbf{a}^T = (1.5, 1.5)$. This is somewhat surprising because the case of

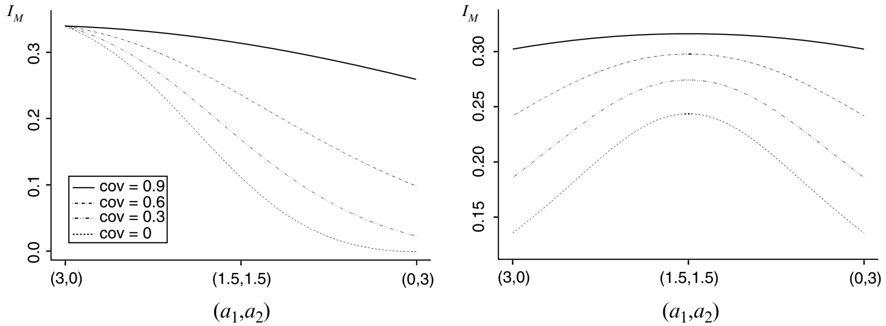


Fig. 4.8 Information measure I_M as a function of combination of the two discrimination parameters $a_1 + a_2 = 3$ for different posterior covariances σ_{12} when θ_1 is intentional and θ_2 a nuisance ability (first panel) and the interest is in the linear combination $0.5\theta_1 + 0.5\theta_2$ (second panel)

both θ_1 and θ_2 intentional could also be considered as one with equal weights for the two abilities. But in this case $\mathbf{a}^T = (1.5, 1.5)$ was the least informative combination (see Figure 4.7). Apparently, both information measures I_M and K^P have opposite preferences for the items when the interest is in estimating them separately than as part of an (equally weighted) composite. This result illustrates the necessity of being explicit about the status of each of the ability parameters before choosing an item-selection criterion in multidimensional adaptive testing. The surface for K^B is at a constant height when $a_1 + a_2$ is a constant (see Figure 4.6). Although this feature may seem desirable for a test supposed to measure a linear composite of θ_1 and θ_2, the height is the same no matter the covariance between the ability parameters. Thus, in the case of the estimation of a composite ability, K^B also has less favorable properties than K^P and I_M.

Finally, as demonstrated by the limiting horizontal curves in Figures 4.7 and 4.8, for a higher covariance, the information measures become exclusively dependent on the sum of the discrimination parameters. This feature is another consequence of the Bayesian principles of borrowing on one parameter from the information in the data about the other as their correlation goes up.

4.8 Conclusion

The item-selection criteria studied in this chapter were all based on the Kullback-Leibler definition of the "distance" between two probability distributions. The first measure, K^B, was the distance between the response distributions at a current estimate of the ability vector, $\widehat{\boldsymbol{\theta}}$, and the true vector $\boldsymbol{\theta}$ integrated over the current posterior distribution of the latter. The second, K^P, was the expected distance between the prior distribution of $\boldsymbol{\theta}$ and its posterior distribution upon administration of the candidate item. The third criterion was a symmetric version of the KL distance known as the mutual information measure, which was defined between the current posterior distribution of $\boldsymbol{\theta}$ and the response distribution on the candidate item. We also proposed modifications of these criteria that are appropriate for the case in which some of the ability components are only intentional and the others should be treated as nuisance parameters as well as the case of the interest being only in a linear composite of the ability components.

An attractive feature of all these KL-based criteria is their immediate generalization from unidimensional to multidimensional adaptive testing. Except for the dimensionality of the ability parameter in the response model, no other changes are involved. No matter the dimensionality of $\boldsymbol{\theta}$, each of these criteria is thus always as scalar. Unlike item selection based on Fisher's information, no additional reduction of an information matrix to a unidimensional criterion is therefore necessary.

Although all three criteria appear to be meaningful for multidimensional adaptive testing when all ability components are intentional, this conclusion has to be modified somewhat when some of the ability components have the status of a nuisance parameter. In this case, the behavior of the posterior expected KL measure, K^B, appears to be independent of the discrimination parameters of the nuisance abilities when the ability estimate of the nuisance abilities is approximately zero, whereas it should have shown a decrease with them. Due to its independence of the posterior covariance for a constant sum of the discrimination parameters, K^B also has the least favorable behavior when the interest is in an equally weighted linear composite of the ability parameters. For the case of not all ability parameters intentional, we therefore only recommend the KL distance between the subsequent posterior distributions of the ability parameters, K^P, and the mutual information in the responses distribution on the candidate item about the test taker's posterior distribution, I_M, as item-selection criteria.

These conclusions were derived analytically assuming an item pool with any possible combination of discrimination parameters over a reasonable range of values. In real-world applications with less than ideal distributions of the item parameters, we may be forced to restrict the selection to special combinations of parameters, and the analysis in this chapter may have to be completed with computer simulations of adaptive testing from the actual item pool.

References

Bloxom, B. & Vale, C. D. (1987, June). *Multidimensional adaptive testing: An approximate procedure for updating*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Boughton, K. A., Yoa, L. & Lewis, D. M. (2006, April). *Reporting diagnostic subscale scores for tests composed of complex structure*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Chang, H.-H. (2004). Understanding computerized adaptive testing: From Robbins-Monro to Lord and beyond. In D. Kaplan (Ed.), *Handbook of quantitative methods for the social sciences* (pp. 117–133). Thousand Oaks, CA: Sage.

Chang, H.-H. & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58*, 37–52.

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.

Cover, T. M. & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Fan, M. & Hsu, Y. (1996, April). *Multidimensional computer adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1). Norwood, NJ: Ablex.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhof Publishing.

Lehmann, E. L. & Casella, G. (1998). *Theory of point estimation* (2nd ed.). New York: Springer-Verlag.

Luecht, R. M. (1996). Multidimensional computer adaptive testing. *Applied Psychological Measurement, 20*, 389–404.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs No. 15*.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 258–270). New York: Springer-Verlag.

Mulder, J. & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74*, 273–296.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69–92). Princeton, NJ: Educational Testing Service.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401–412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous items response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Samejima, F. (1974). normal-ogive model for the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111–121.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Silvey, S. D. (1980). *Optimal design*. London: Chapman & Hall.

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20*, 373–388.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24*, 398–412.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W. J. & Glas, C. A. W. (2007). Statistical aspects of adaptive testing. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 27: Psychometrics) (pp. 801–838). Amsterdam: North-Holland.

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575–588.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement*, *67*, 41–58

Yao, L. & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83–105.

Chapter 5
Sequencing an Adaptive Test Battery

Wim J. van der Linden

5.1 Introduction

Switching a testing program from a linear to an adaptive format increases its efficiency considerably. The gain in efficiency can be used to shorten the length of the test or increase the accuracy of the scores. The gain is especially relevant to testing programs in which a battery of tests has to be administered in a single session but the testing time has to remain feasible. Examples of such programs are diagnostic testing for instructional purposes (e.g., Boughton, Yao, & Lewis 2006; Yao & Boughton, 2007) and large-scale assessments of education. These programs generally involve the reporting of profiles of scores of students, schools, or districts. In order to use such profiles for decision making, each of their individual scores should have satisfactory accuracy. The more advantageous combination of testing time and score accuracy made possible by the use of a battery of adaptive instead of linear tests has been highlighted earlier, for instance, in Brown and Weiss (1977) and Giallucca and Weiss (1979).

In an adaptive test battery, each individual test is assembled from a different item pool. For the first test, an initial ability estimate is chosen and the first item is selected from the first pool to be optimal at this estimate. The response is then used to update the initial estimate, and the second item is selected to be optimal at the update. The process is repeated until a predetermined number of items or accuracy level is reached, whereupon a new test from a new pool is started. Using this format, a typical saving of the length of the individual tests by some 40–60% percent relative to a linear version of them is possible.

It is easy to confuse the case of a battery of unidimensional adaptive tests addressed in this research with a multidimensional adaptive test, particularly if the abilities correlate. Multidimensional adaptive testing has its own procedures of optimal item selection (Mulder & van der Linden, 2009; Segall, 1996, this volume,

W.J. van der Linden (✉)
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_5,

chap. 3; van der Linden, 2005, sect. 9.7.; Veldkamp & van der Linden, 2002). For these procedures, it is not necessary (and, in fact, even disadvantageous) to constrain the selection of items to one and the same subpool until a predetermined number of items is administered, as is done in testing with a battery of unidimensional tests.

In the current research, the focus was not on the question of how to select the individual items. Readers with an interest in this question should refer to Thissen and Mislevy (2000) or van der Linden and Pashley (this volume, chap. 1). Instead, the interest was in the optimal sequencing of the test battery, the idea being that instead of always administering the tests in the battery in the same predetermined order for each test taker, its efficiency could be increased further by optimizing the order for the individual test takers. In fact, the best approach seems to repeat the principle of adaptation at the level of the selection of the tests. We would then pick the first test to be optimal over the initial estimates of each of the test taker's abilities measured by the battery. The second test would be chosen to be optimal given the test taker's responses to the first test. And so on.

A statistical notion relevant to such an approach is that of collateral information, i.e., the information about the test taker's ability measured by one test available in the other tests. We expect the amount of collateral information in test batteries to be substantial because they are typically designed to measure a set of strongly related but distinct abilities, for instance, abilities in early mathematics or language acquisition in elementary education. It would be a waste to ignore such information when choosing a next test from the battery. The problem of how to improve subtest scores by borrowing information from other subtests has received considerable attention recently (e.g., Wainer et al., 2001). The solution to the problem of sequencing the tests in a battery addressed in this research can be viewed as an adaptive solution to this more general problem.

An appropriate framework for implementing the adaptive approach is multilevel item-response theory (IRT) in combination with an empirical Bayes procedure for the selection of the tests. The two-level model used in this research consists of distinct response models for the unidimensional item pools as first-level models and a specification of the joint distribution of their ability parameters for the population of test takers as a second-level model. The second-level model allows the borrowing of information from earlier response vectors about the abilities measured by later tests. An empirical Bayes approach represents this information in the form of prior and posterior distributions: each time a test is completed, the test taker's response vector is used to update the posterior predictive distributions of the abilities measured by the remaining tests, whereupon the updates are used as prior distributions for the selection of the next test. The approach is empirical in that the second-level model for the distribution of the abilities is estimated from test data collected during pretesting of the items.

For the results to be realistic, we have to allow for content constraints on the item-selection process. These constraints are necessary to implement the test agency's specifications for each individual test, such as its required content distribution, possible logical relations between the items in the pool (e.g., items that should not be combined in the same test administration), and the total amount of time available for the test. In addition, if item security is a problem (which we do not generally expect to be the case for a diagnostic battery), it would be necessary to control the exposure rates of the items in the pools through the imposition of random ineligibility constraints on the items for the test takers.

In the next sections, we define this multilevel structure more precisely and present an algorithm for sequencing the test battery that satisfies any real-world set of content constraints. In the final section, results from an extensive simulation study are presented. Because the correlations between the abilities for the individual tests are critical, we varied the correlation structure in this study and examined its impact on the sequence of tests as well as the accuracy of the ability estimates. The same was done for different test lengths as well as testing with and without content constraints. As a baseline, we used the statistical accuracy of the ability estimators for the traditional case of a fixed but arbitrary order of the tests that ignored the responses to the earlier tests in the battery.

5.2 Multilevel Model

Suppose the battery consists of individual tests from item pools $h = 1, 2, \ldots, H$ of size N_h. The items in the pools are denoted as $i_h = 1, 2, \ldots, N_h$. The length of the test from pool h is equal to n_h. The ability measured by pool h is represented by a (scalar) parameter θ_h. The parameter is assumed to be defined by the three-parameter logistic (3PL) response model, which describes the probability of a successful response ($U_i = 1$) by a test taker with ability $\theta_h \in (-\infty, \infty)$ on item i as

$$p_i(\theta_h) \equiv \Pr\{U_{ij} = 1\} \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta_h - b_i)]}{1 + \exp[a_i(\theta_h - b_i)]}, \tag{5.1}$$

where $b_i \in (-\infty, \infty)$, $a_i > 0$, and $c_i \in [0, 1]$ are the difficulty, discriminating power, and guessing parameters of item $i = 1, \ldots, N$ in the pool, respectively (Birnbaum, 1968).

The ability structure of the population of test takers is represented by a multivariate normal density,

$$f(\theta_1, \theta_2, \ldots, \theta_H) = \mathrm{MVN}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}). \tag{5.2}$$

Finally, it is assumed that all item and population parameters have been estimated with enough precision during the pretest of the battery to consider them as known during the test administration. When the battery begins, the only unknown parameters are thus the abilities $\theta_1, \theta_2, \ldots, \theta_H$.

5.3 Empirical Bayes Approach

5.3.1 Selection of Initial Pool

Items in an adaptive test are typically selected using a maximum-information criterion or a Bayesian criterion derived from the posterior distribution of θ (van der Linden & Pashley, this volume, chap. 1). We will use the Bayesian criterion of maximum posterior expected information in the responses to the candidate items. The criterion is defined more precisely below.

It may seem attractive to use the same item-selection criterion to pick the initial test in a battery, that is, to begin the battery with the item pool that contains the item expected to be most informative at the test taker's initial ability estimate across all item pools. But this strategy would be less prudent for two reasons. First, this most informative item may have neighboring items in its pool that are less informative than some of the neighbors of the best items in the other pools. Selection of the pool will then be penalized immediately by less than optimal item selection after the test has begun. Second, as already indicated, real-world adaptive tests typically have to meet constraints on their composition that represent their content specifications. These constraints cannot be satisfied if we focus on one item at a time. The first problem can be avoided by evaluating the best set of items of the size of the intended test lengths, n_h, from each pool. The second problem is resolved if we require each of these sets to meet the content constraints on the individual tests.

The problem of selecting sets of items of length n_h from item pools $h = 1, \ldots, H$ can be formalized as follows. For the response model in (5.1), Fisher's information about θ in the response to a test item i is defined as

$$\begin{aligned} I_i(\theta) &= -\mathcal{E}\left[\frac{\partial^2}{\partial\theta^2}\ln l(\theta; U_i)\right] \\ &= \frac{[p_i'(\theta)]^2}{p_i(\theta)[1-p_i(\theta)]}, \end{aligned} \tag{5.3}$$

where U_i is a random response to item i by a test taker with ability θ, $l(\theta; U_i)$ is the likelihood statistic associated with its distribution, and $p_i'(\theta)$ is the derivative of the response probability in (5.1) with respect to θ (Hambleton & Swaminathan, 1985, sect. 6.3).

The criterion we propose is the expected value of Fisher's information over the posterior distribution of θ. (The same procedure can be followed for criteria based on any other posterior expected quantity; for such criteria, see van der Linden, 1998.) If no items have been administered yet, the appropriate marginal distribution of (5.2) should be used. For item i_h, the criterion takes the value

$$\int I_{i_h}(\theta_h) f(\theta_h) d\theta_h, \tag{5.4}$$

where $f(\theta_h)$ is the marginal distribution of θ_h from (5.2).

Let x_{i_h} denote a binary decision variable for the selection of item i_h, $i = 1,\ldots,N_h$, $h = 1,\ldots,H$. That is, $x_{i_h} = 1$ if the item is selected and $x_{i_h} = 0$ otherwise. The set of items selected from pool h can be found by solving

$$\text{maximize} \sum_{i_h=1}^{N_h} \left[\int I_{i_h}(\theta_h) f(\theta_h) d\theta_h \right] x_{i_h}, \tag{5.5}$$

subject to

$$\sum_{i_h=1}^{N_h} x_{i_h} = n_h, \tag{5.6}$$

$$x_{i_h} \in \{0,1\},\ i_h = 1,\ldots,N_h. \tag{5.7}$$

The objective function maximizes the sum of the prior expected information in the items. The constraint in (5.6) requires the selection of n_h items. Observe that both the objective function and the constraint are linear in the variables x_{i_h}. The initial pool in the battery selected for the test taker is that with the largest value for the objective function for the solution of the optimization problem in (5.5)–(5.7).

It may seem somewhat overdone to formalize the selection of the sets of items as a 0–1 linear optimization program because we could directly pick the n_h items with the largest values for (5.4) from the pools. But it immediately becomes advantageous to do so when test specifications are to be imposed on the selection of the items.

Such specifications are conveniently imposed using the shadow-test approach to adaptive testing (van der Linden, 2005, chap. 9; this volume, chap. 1). In order to illustrate the approach, we impose a set of constraints on some of the categorical attributes of the items (e.g., a content category or item format), a quantitative attribute (e.g., word count or expected time to respond to the items), and the exposure rates of the items. Let V_{c_h} be the subsets of items in pool h for content categories $c_h = 1,2,\ldots,C_h$ and q_{i_h} the value of item $i_h = 1,2,\ldots,N_h$ for quantitative attribute q. A general representation of the constrained version of the optimization problem for pool h is

$$\text{maximize} \sum_{i_h=1}^{N_h} \left[\int I_{i_h}(\theta_h) f(\theta_h) d\theta_h \right] x_{i_h}, \tag{5.8}$$

subject to

$$\sum_{i_h=1}^{N_h} x_{i_h} = n_h, \tag{5.9}$$

$$\sum_{i_h \in V_{c_h}} x_{i_h} \gtreqless n_{c_h}, \quad c_h = 1,2,\ldots,C_h, \tag{5.10}$$

$$\sum_{i_h=1}^{N_h} q_{i_h} x_{i_h} \gtreqless b_q, \tag{5.11}$$

$$x_{i_h} = 0, \quad i_h \in R_{j},. \tag{5.12}$$

$$x_{i_h} \in \{0, 1\}, \quad i_h = 1, 2, \ldots, N_h, \tag{5.13}$$

where $\gtreqless$ denotes the appropriate choice of an equality or inequality.

The new constraints are those in (5.10)–(5.12). In (5.10), the numbers of items in category $c_h = 1, \ldots, C_h$ from pool h are constrained by bounds n_{c_h}. Likewise, (5.11) constrains the sum of the quantitative item attributes q_i (e.g., the total word count for the test) by a bound b_q. The constraints in (5.12) are ineligibility constraints for the items in subset R_j for test taker j in the pool. They are added to control the exposure rates of the items in the pool. Subset R_j is chosen randomly for each test taker; that is, before the test is administered to each new test taker j, a probability experiment is conducted to determine which items belong to R_j with probabilities of selection that depend on both a chosen upper limit $r^{\max}$ on the exposure rates of the items and the counts of certain events during the history of the tests. For a derivation of these probabilities of eligibility and implementation details, see van der Linden and Veldkamp (2004, 2007). A larger collection of examples of possible content constraints for adaptive testing is presented in van der Linden (2005). In real-world applications, multiple versions of the same type of constraint may be required to impose existing test specifications.

A solution to the optimization problem is a vector of zeros and ones for the decision variables that identifies the set of items that meets the constraints and has a maximum value for the objective function. These solutions are known as shadow tests. They are easily found by a call to an integer solver in a standard linear-programming computer program prior to the item selection; for a description of such solvers, see van der Linden (2005, sect. 4.2.5). Using a well-initialized solver, shadow tests from a pool of several hundreds of items are found within a second.

Shadow tests are *not* administered; their sole purpose is to identify the set of items in each of the pools from which the CAT algorithm should pick the best item for administration.

5.3.2 Selection of First Test

For each item pool, a shadow test is calculated. The first test is the test from the pool for which the shadow test has the largest value for the objective function in (5.8). Without loss of generality, we will denote this first pool by $h = H$.

5.3.3 Administration of First Test

Once the first pool is identified, the first item that is administered is that in the shadow test with the largest value for the expected information in (5.4).

Each next item is the best item among the remaining free items in an update of the shadow test. The update involves two operations: (i) the addition of the constraint $x_{i_H} = 1$ for the last item that was administered to the model for the shadow test; and (ii) the update of the posterior distribution of θ_H in the item-selection criterion in the objective function of the model. The addition of the constraint guarantees that earlier items are automatically included when the shadow test is reassembled. The update of the posterior distribution guarantees that the shadow test remains optimal when it is reassembled. Because of these two features, the adaptive test meets all test specifications and still adapts the selection of the items optimally to the test taker's responses.

The more complicated operation is the update of the posterior distribution of θ_H. We give the update for the selection of the kth item. Let $\mathbf{u}_{Hj}^{(k-1)}$ be the response vector of test taker j for the first $k-1$ items in the first test. The posterior distribution of θ_H after $k-1$ items is defined as

$$f\left(\theta_H|\mathbf{u}_{Hj}^{(k-1)}\right) = \frac{f\left(\mathbf{u}_{Hj}^{(k-1)}|\theta_H\right)f(\theta_H)}{\int f\left(\mathbf{u}_{Hj}^{(k-1)}|\theta_H\right)f(\theta_H)d\theta}, \tag{5.14}$$

where $f\left(\mathbf{u}_{Hj}^{(k-1)}|\theta_H\right)$ is the model probability associated with the current response vector,

$$f\left(\mathbf{u}_{Hj}^{(k-1)}|\theta_H\right) = \prod_{i=1}^{k-1} P(U_{i_H j} = 1|\theta_H)^{u_{i_H j}}\left[1 - P(U_{i_H j} = 1|\theta_H)\right]^{1-u_{i_H j}}. \tag{5.15}$$

Hence, the update of (5.4) for the selection of the kth item in the test is

$$\int I_{i_H}(\theta_H) f\left(\theta_H|\mathbf{u}_{Hj}^{(k-1)}\right) d\theta_H. \tag{5.16}$$

As the final estimate of θ_{Hj}, we suggest the mean of the last posterior distribution (expected a posteriori or EAP estimate),

$$\widehat{\theta}_{Hj}^{(n_H)} = \int \theta_H f\left(\theta_H|\mathbf{u}_{Hj}^{(n_H)}\right) d\theta_H, \tag{5.17}$$

or any other measure of its location. Alternatively, we could use the maximum-likelihood estimate (MLE) of θ_H, that is, the maximizer of (5.15) for the response vector $\mathbf{u}_{Hj}^{(n_H)}$. We then still profit from the extra information from earlier tests when picking a new test and selecting its items but report scores that are inferred only from the responses to the current test. This option should be used when it is deemed undesirable to report test scores based on any other statistical information than the test taker's responses to the proper test.

5.3.4 Selection of Subsequent Tests

The second pool has to be chosen from $h = 1, \ldots, H - 1$ using (5.8)–(5.13) with an update of the marginal density $f(\theta_h)$ from (5.2) to $f(\theta_h|\mathbf{u}_{Hj})$. Observe that this is the posterior distribution for the second ability given the responses to the items in the first test. This new posterior density can be written as

$$\begin{aligned} f(\theta_h|\mathbf{u}_{Hj}) &= \int f(\theta_h, \theta_H|\mathbf{u}Hj)d\theta_H \\ &= \int f(\theta_h|\theta_H)f(\theta_H|\mathbf{u}_{Hj})d\theta_H \\ &\propto \int f(\theta_h|\theta_H)f(\theta_H)f(\mathbf{u}_{Hj}|\theta_H)d\theta_H \\ &= \int f(\theta_h, \theta_H)f(\mathbf{u}_{Hj}|\theta_H)d\theta_H, \end{aligned} \tag{5.18}$$

where the second step follows upon the usual assumption of conditional independence of θ_h and $\mathbf{u}_{Hj}$ given θ_H ("local independence"). Observe that the second step also reveals the nature of the posterior distribution: it actually is a predictive posterior distribution with the second-level model probability $f(\theta_h|\theta_H)$ integrated over the posterior distribution of θ_H given $\mathbf{u}_{Hj}$.

The first factor in the last integrand of (5.18) is a known normal density that follows from (5.2) as

$$f(\theta_h, \theta_H) = \int \ldots \int f(\theta_1, \ldots, \theta_H)d\theta_1 \ldots d\theta_{h-1}d\theta_{h+1} \ldots d\theta_{H-1}. \tag{5.19}$$

The second factor of the same integrand is just the probability of the responses to the items in the first test given θ_H. The integral in the last step of (5.18) is therefore easily calculated from known expressions; see the Appendix.

Since the norming constant for the posterior density $f(\theta_h|\mathbf{u}_{Hj})$ is independent of θ_h, we can directly use the unnormed posterior density in (5.18) for comparison between the remaining item pools. The second pool is therefore found as the solution of the updated model for the shadow tests that has the maximum value over $h = 1, \ldots, H - 1$ for the new objective function in (5.8).

For the selection of the third pool, it is straightforward to show that (5.18) generalizes to

$$f\left(\theta_h|\mathbf{u}_{Hj},\mathbf{u}_{(H-1)j}\right) \propto \int\int f\left(\theta_h,\theta_{H-1},\theta_H\right) f\left(\mathbf{u}_{(H-1)j}|\theta_{H-1}\right) \times f\left(\mathbf{u}_{Hj}|\theta_H\right) d\theta_{H-1} d\theta_H. \tag{5.20}$$

For batteries with more than three tests, the expressions for the selection of the subsequent tests are analogous.

5.3.5 *Administration of Subsequent Tests*

In order to select the items in the second test, we need to update the posterior distribution $f\left(\theta_h|\mathbf{u}_{hj}^{(k-1)}\right)$ in (5.16) to

$$f\left(\theta_h|\mathbf{u}_{hj}^{(k-1)},\mathbf{u}_{Hj}\right) = \frac{f\left(\mathbf{u}_{hj}^{(k-1)}|\theta_h\right) f(\theta_h|\mathbf{u}_{Hj})}{\int f\left(\mathbf{u}_{hj}^{(k-1)}|\theta_h\right) f(\theta_h|\mathbf{u}_{Hj}) d\theta}. \tag{5.21}$$

The kth item is then selected to maximize

$$\int I_{i_h}(\theta_h) f\left(\theta_h|\mathbf{u}_{hj}^{(k-1)},\mathbf{u}_{Hj}\right) d\theta_h. \tag{5.22}$$

The second test should be scored using the version of (5.17) with $f\left(\theta_h|\mathbf{u}_{Hj}^{(n_H)}\right)$ replaced by $f\left(\theta_h|\mathbf{u}_{hj}^{(n_h)},\mathbf{u}_{Hj}^{(n_H)}\right)$, or, when ML estimation is more appropriate, by the maximizer of (5.15) for the new test. For the third test, we should replace $f\left(\theta_h|\mathbf{u}_{hj}^{(k-1)},\mathbf{u}_{Hj}\right)$ in (5.21) by $f\left(\theta_h|\mathbf{u}_{hj}^{(k-1)},\mathbf{u}_{(H-1)j},\mathbf{u}_{Hj}d\right)$. And so on.

5.4 Simulation Study

Adaptive testing from a real-world item pool was simulated to get a first impression of the empirical behavior of the method for sequencing a test battery presented in this chapter. The battery consisted of short tests from the three sections of the Law School Admission Test (LSAT), which measure analytic reasoning (AR), reading comprehension (RC), and logical reasoning (LR). (The current LSAT is a paper-and-pencil test with two subtests in its LR section.) A previous pool of operational items from the LSAT was used to run the adaptive tests. Although the items in the pool had been calibrated jointly under the 3PL model in (5.1), we treated the subpools for each of the three sections as a separate unidimensional pool. The sizes of the three subpools were 208 (AR), 240 (RC), and 304 items (LR).

To set a baseline, a simulation of the traditional version of an adaptive test battery with independent sequencing of the three tests was conducted.

5.4.1 Design of Study

The impact of the following factors was studied:

1. Adaptive versus independent sequencing of the tests.
2. Adaptive testing with and without content constraints.
3. Test lengths equal to $n = 5$, 10, 15, and 20 items.

The main comparison in this study was between adaptive and independent sequencing of the tests in the battery. The baseline procedure of independent sequencing consisted of 500 adaptive administrations for the true abilities $\theta_h = -2.0, -1.5, \ldots, 2.0$ for each of the three tests. Each test started with $\widehat{\theta}_h = 0$ as the initial ability estimate. All subsequent items were selected using the maximum posterior expected information criterion in (5.4). The tests were simulated entirely independently; no information from one of the tests was used in any of the others.

The procedure with the adaptive sequencing of the tests was as described above. AR was always chosen to be the first test. [For a test battery starting from the prior distribution in (5.2), the first test is automatically the same for all test takers; for a suggestion of adaptive selection of the first test as well, see the discussion at the end of this chapter.] For each of the true abilities $\theta_1 = -2.0, -1.5, \ldots, 2.0$, 500 administrations of the battery were simulated. The true abilities in the simulations of the second and third tests were sampled conditionally on θ_1 from the population in (5.2) to realize the correlational structure in (5.23) below. Again, the first test started with $\widehat{\theta}_h = 0$ as the initial ability estimate. But the subsequent items and tests were selected using the empirical Bayes approach in (5.14)–(5.22). The integrals in the procedure were calculated using the Monte Carlo approach described in the Appendix.

The content constraints adopted in the model for the shadow test were actual constraints for the LSAT. They were for specifications related to the content, item type, and answer-key distributions of the three sections, possible gender or minority orientation of their items, and word counts. An important difference was the omission of all constraints associated with the item-set structure for two of the tests. They had to be omitted to be able to study the impact of the different lengths of $n = 5$, 10, 15, and 20 items for these tests. The total numbers of constraints chosen for the tests were 12 (AR), 23 (RC), and 16 (LR).

Clearly, the pattern of covariances between the abilities in the covariance matrix in (5.2) is critical to the success of the method. The covariance matrix used was

$$\sum = \begin{pmatrix} 1.0 & 0.8 & 0.8 \\ & 1.0 & 0.3 \\ & & 1.0 \end{pmatrix}. \tag{5.23}$$

As already indicated, real-world test batteries usually consist of tests of highly related domains that need to be distinguished for practical reasons (e.g., diagnosis). The covariances between the first test (AR) and the two alternative second tests (RC and LR) in (5.23) are believed to represent a typical case. The covariance between the two alternative tests did not actually play a role in the sequencing of these two tests; when one of them was identified as the best second test for a test taker, the choice of the other as the third test was automatically fixed.

For each simulated administration of the battery, we recorded the error in the estimate of θ_h at the end of each test and counted the number of times each of the two possible paths through the battery (AR-RC-LR vs. AR-LR-RC) was taken. In the next section, we compare the estimated mean-square error (MSE) functions of the ability estimators between the different conditions in more detail. We made the same comparison for the estimated bias functions of the estimators but omit a discussion of the results because they matched those for the MSE functions.

5.4.2 Results

The estimated MSE functions for the baseline case of independent test selection are given in Figure 5.1. Because the impact of adaptive selection of the tests sets in after the first test, the functions are shown only for RC and LR. For the shorter tests, their curves tend to be convex because of the effect of the initial estimates $\widehat{\theta}_h = 0$. But for the longer tests, the benefits of adaptive testing become quickly visible in the form of flatter curves at a lower height. Generally, the constraints do not seem to have much impact on the MSE functions, a result typical of the efficient way the constraints are implemented by the shadow-test approach (van der Linden, 2005, chap. 9). Also, the MSE functions for the RC and LR tests do not differ much. In fact, the only noticeable difference was a tendency of the curves for RC to go up at the upper end of the θ scale in the condition with content constraints. This happened even for the longer tests. Because the curves run flat for the condition without the constraints, the tendency is no doubt the result of the constraints forcing the algorithm to select items for the highest ability level that are actually optimal at lower levels.

For every test length, the selection of RC as the second test yielded substantially lower MSEs than for the baseline case in Figure 5.1. The only exceptions were for the combinations of the two most extreme values of θ, $n = 5$, and no content constraints. For the selection of RC as the third test, the MSEs are substantially lower again and the curves run much flatter, even for test lengths as short as $n = 5$.

Table 5.1 shows the counts of the simulated test takers who took the two possible paths through the battery (AR-RC-LR vs. AR-LR-RC) for the conditions with and without content constraints in this study. Observe that the total number of simulated test takers was 18,000 (=i.e., 500 test takers at nine ability levels for each of the four test lengths). The counts reveal a strong preference for RC as the second test over LR. The preference must be due to the composition of their two item pools;

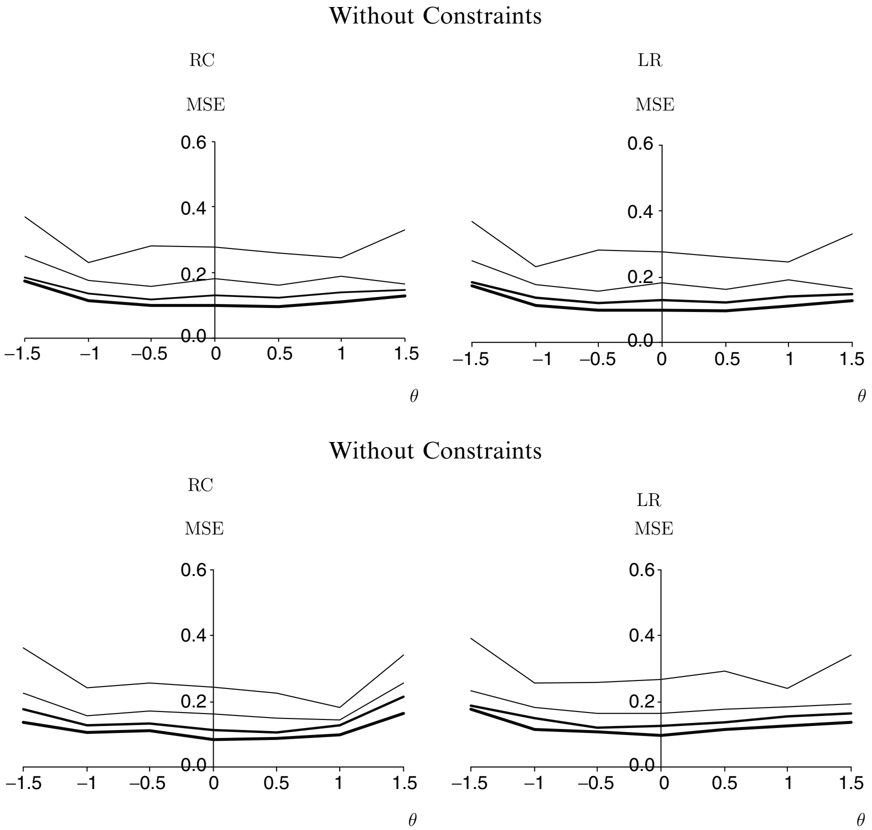


Fig. 5.1 Estimated mean-square error (MSE) functions of the ability estimators for independent administration of the reading comprehension (RC) and logical reasoning (LR) tests without and with content constraints on the item selection (baseline case). Note: The darker the line, the longer the test ($n = 5, 10, 15, 20$)

Table 5.1 Number of test takers with the paths RC-LR and LR-RC for the conditions with and without constraints

Path	RC-LR	LR-RC
Without constraints	11,295	6,705
With constraints	16,605	1,395

apparently, for the criteria in (5.5) and (5.8), the RC pool contained initial sets of items that were more informative than for the LR pool. A practical consequence of these effects was not enough data to estimate the MSE functions accurately for some of the conditions in this study. This somewhat unfortunate development was the price paid for the choice of a simulation study with a real-world item pool.

The estimated MSE functions for the conditions with adaptive sequencing of the tests are given in Figures 5.2 and 5.3. Observe that the curves in the upper right (LR

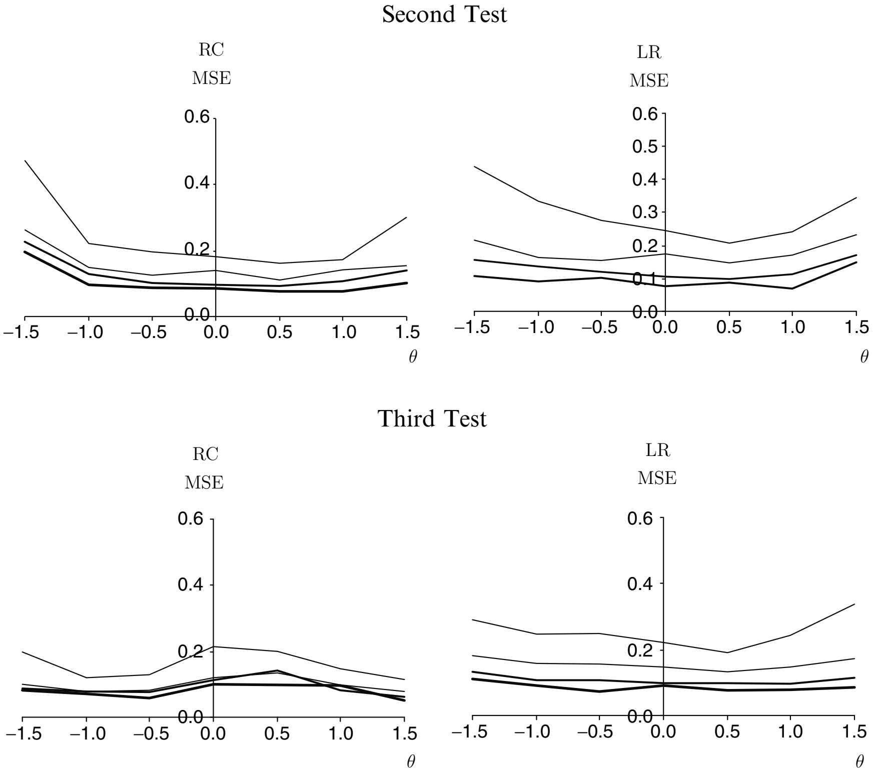


Fig. 5.2 Estimated mean-square error functions (MSE) of the ability estimators for administration of the reading comprehension (RC) and logical reasoning (LR) tests as the second and third tests in the battery (without content constraints). Note: The darker the line, the longer the test ($n = 5$, 10, 15, 20)

as second test) and lower left plots (RC as third test) of each of these figures show more variability because, as just indicated, they had to be based on much less data. For the same reason, for RC as the third test in Figure 5.3, only the estimates of the lower parts of the MSE functions could be estimated. Nearly all points in these omitted parts of these functions had less than 10 observations; a substantial portion of them even had no observations at all.

The best demonstration of the impact of the use of the collateral information in the responses to the earlier tests in the empirical study was the set of MSE functions estimated for RC as the third test in Figure 5.2. Even for a test length as small as $n = 5$, these functions ran already flat along the entire range of θ values. This feature shows that the impact of the use of collateral information from earlier tests is not only a general increase of the efficiency of the later tests but also enables them to start at initial ability estimates away from $\widehat{\theta}_h = 0$ and therefore move faster to true abilities that are at the upper or lower end of the scale—hence, these entirely flat curves.

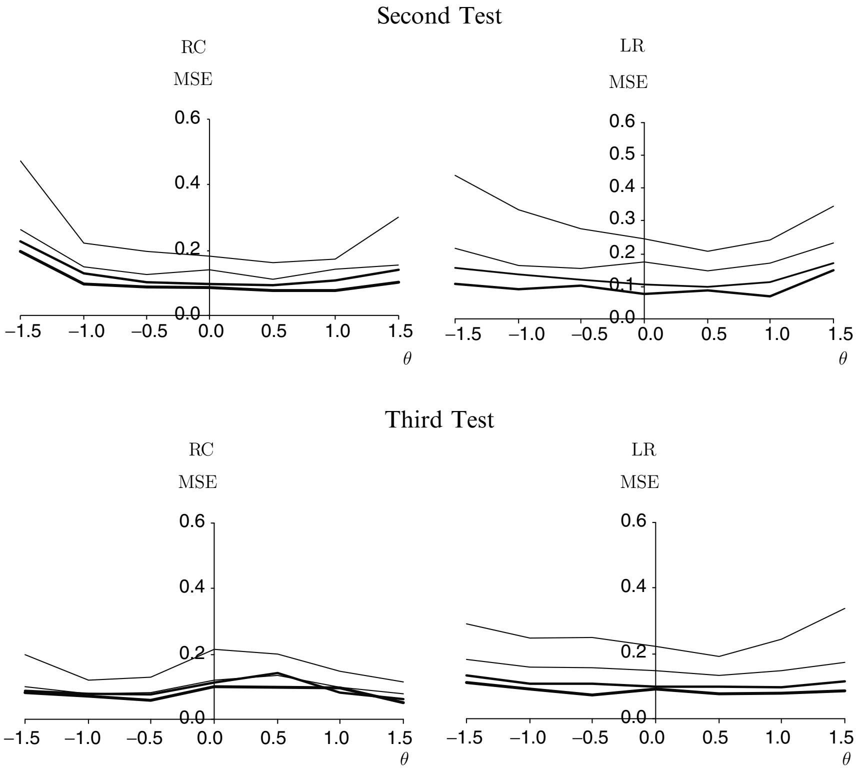


Fig. 5.3 Estimated mean-square error functions (MSE) of the ability estimators for administration of the reading comprehension (RC) and logical reasoning (LR) tests as the second and third tests in the battery (with content constraints). Note: The darker the line, the longer the test ($n = 5$, 10, 15, 20)

5.5 Concluding Remarks

Test batteries usually have severe time constraints on their tests, but the abilities they measure are highly correlated. Hence, as demonstrated in studies by Brown and Weiss (1977) and Gialluca and Weiss (1979), such batteries stand to profit substantially from adaptive testing. In this research, the idea of adaptation was extended to include the sequencing of the individual tests during the administration of the battery. While their sequence traditionally is arbitrarily fixed, this research suggests to adapting the sequence to the performances of the individual test takers. An appropriate framework for doing so is the combination of hierarchical item response modeling with an empirical Bayes approach presented above. The framework allows us to model the multivariate structure of the abilities measured by the tests and use this structure to translate empirical information from earlier tests directly into more efficient updates of the prior ability distributions for the later tests.

The empirical study was only the first of its kind; others have to be conducted to get a better insight into the precise quantitative effects of adaptive test sequencing, especially for batteries with larger numbers of tests and more complicated patterns of correlation between them. Such studies should also explore the additional effects of including known covariates for the abilities measured by the tests. Examples of useful covariates are observed scores on tests taken in the past or response times during the tests in the battery. The former is expected to adapt the selection of the first test to the individual test takers, which now was the same for all of them, and therefore to be automatically beneficial to all later tests as well. Response times have already proven to be a powerful source of collateral information for the case of a single adaptive test (van der Linden, 2008) and are expected to lead to even stronger advantages for a full battery of them.

5.6 Appendix: Computational Approach

Because of the population distribution in (5.2), the prior density $f(\theta_h)$ in (5.4) is that of $\mathcal{N}(\mu_{\theta_h}, \sigma_{hh})$ with σ_{hh} the hth diagonal element of $\Sigma_{\boldsymbol{\theta}}$. Although (5.4) could easily be calculated using Gauss–Hermite quadrature, in the empirical study we preferred simple Monte Carlo integration and approximated this criterion as

$$\int I_{i_h}(\theta_h) f(\theta_h) d\theta_h \approx R^{-1} \sum_{r=1}^{R} I\left(\theta_h^{(r)}\right), \tag{A1}$$

where $\theta_h^{(r)}$ is the rth draw from $f(\theta_h)$.

From (5.18), it follows that the criterion for the selection of the second test can be approximated as

$$\int I_i(\theta_h) f(\theta_h | \mathbf{u}_{Hj}) d\theta_h \approx \frac{\sum\limits_{r=1}^{R} I\left(\theta_h^{(r)}\right) f\left(\mathbf{u}_{Hj} | \theta_H^{(r)}\right)}{\sum\limits_{r=1}^{R} f\left(\mathbf{u}_{Hj} | \theta_H^{(r)}\right)}, \tag{A2}$$

where $\left(\theta_h^{(r)}, \theta_H^{(r)}\right)$ is the rth draw from $f(\theta_h, \theta_H)$. A composition method was used to draw $\left(\theta_h^{(r)}, \theta_H^{(r)}\right)$; that is, $\theta_H^{(r)}$ was drawn from $f(\theta_H)$ and $\theta_h^{(r)}$ from $f\left(\theta_h | \theta_H^{(r)}\right)$. Because of the multivariate normality in (5.2), either step involves a draw from a known normal density. [The ease of these steps explains our current preference for (A1) over numerical quadrature.]

In the study, the third test was automatically fixed when the first two were selected. But it may be interesting to note that the use of the generalization in (5.20) would have involved two simple operations: (i) the multiplication of the numerator and denominator of (A3) by the likelihood associated with the last test selected and

(ii) one more step of the composition method. That is, we could have approximated the criterion using

$$\int I_i(\theta_h) f\left(\theta_h|\mathbf{u}_{Hj},\mathbf{u}_{(H-1)j}\right) d\theta_h \approx \frac{\sum_{r=1}^{R} I\left(\theta_h^{(r)}\right) f\left(\mathbf{u}_{Hj}|\theta_H^{(r)}\right) f\left(\mathbf{u}_{(H-1)j}|\theta_{H-1}^{(r)}\right)}{\sum_{r=1}^{R} f\left(\mathbf{u}_{Hj}|\theta_H^{(r)}\right) f\left(\mathbf{u}_{(H-1)j}|\theta_{H-1}^{(r)}\right)}, \tag{A3}$$

with an extra step to draw $\theta_h^{(r)}$ from $f\left(\theta_h|\theta_H^{(r)},\theta_{H-1}^{(r)}\right)$.

The EAP estimates at the end of each test were calculated similarly with $I\left(\theta_h^{(r)}\right)$ in (A1)–(A3) replaced by $\theta_h^{(r)}$.

Generally, care should be exercised when using Monte Carlo integration for the current type of problem because the draws are from prior distributions that can be expected to be wider than the likelihoods. Consequently, the effective sample size tends to be smaller than its nominal value. When preparing the simulation study, several trial values for the sample size were used and the results were found to be stable for sizes larger than 3,000. This number should not be taken blindly as a recommendation of this computational approach, certainly not for applications in operational testing with larger test batteries than in this study.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Boughton, K. A., Yao, L. & Lewis, D. M. (2006, April). *Reporting diagnostic subscale scores for tests composed of complex structure.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Brown, J. M. & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries* (Research Report 77-6). Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Gialluca, K. A. & Weiss, D. J. (1979). *Efficiency of an adaptive inter-subtest branching strategy in the measurement of classroom achievement* (Research Report 79-6). Minneapolis, MN: University of Minnesota, Psychometric Methods Program.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhof Publishing.

Mulder, J. & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria for item selection. *Psychometrika, 74.* In press.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331–354.

Thissen, D. & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevey, L. Steinberg & D. Thissen. *Computerized adaptive testing: A primer* (pp. 103–135). Mahwah, NJ: Erlbaum.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics, 33*, 5–20.

van der Linden, W. J. & Veldkamp, B. P. (2004). Constraining item exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.

van der Linden, W. J. & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418.

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67*, 575–588.

Wainer, H., Vevea, J. L., Canachi, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert, K. A. & Thissen, D. (2001). Augmented scores–"Borrowing strength" to compute scores based upon small numbers of items. In H. Wainer & D. Thissen (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum.

Yao, L. & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83–105.

Part II
Applications in Large-Scale Testing Programs

Chapter 6
Adaptive Tests for Measuring Anxiety and Depression

Otto B. Walter

6.1 Introduction

Psychological constructs such as depression or anxiety, and health-related measures such as pain or physical functioning, can be reliably assessed today by means of standardized tests. In fact, such tests are now well established as being an important part of clinical practice. Over the last few years, the number of bio-medical publications citing the word *questionnaire* has risen exponentially (Figure 6.1).

As a result, a considerable item burden is often placed on patients. In this context, the application of computer adaptive tests (CATs) seems promising. However, most of the theoretical and practical contributions to the application of CATs are still in the area of ability and achievement testing. Although efforts have been made to develop CATs for health-related measures, there have been very few reports on using CATs as a means of psychometric assessment in a medical setting. This situation is about to change. A prominent example is a joint initiative working on building a "Patient-Reported Outcomes Measurement Information System" (PROMIS) sponsored by the U.S. National Institutes of Health (NIH). The aim of this network is to develop a large bank of items that measures patient-reported outcomes and to create a computerized adaptive testing system that allows for efficient assessment in clinical research of a wide range of chronic diseases. These tools are expected to be available to the general medical community in 2008 (Fries, Bruce & Cella, 2005; Cella et al., 2007).

Given the advantages that the application of CATs promises, this large-scale effort on the part of the NIH to advance the development of CATs in patient-reported outcomes measurement is not surprising. Many of the advantages of CATs seem to be well suited to assessments in clinical psychology (Embretson, 1996) or medicine.

A particularly attractive property of CATs is the possibility of determining the measurement precision conditional upon the level of the underlying latent trait θ. A low measurement precision often occurs for extreme (high or low) θ-values. In CATs constructed within the framework of item response theory (IRT), situations

O.B. Walter (✉)
Institut für Psychologie, RWTH Aachen University, Jägerstrasse 17/19, 52066 Aachen, Germany

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_6,

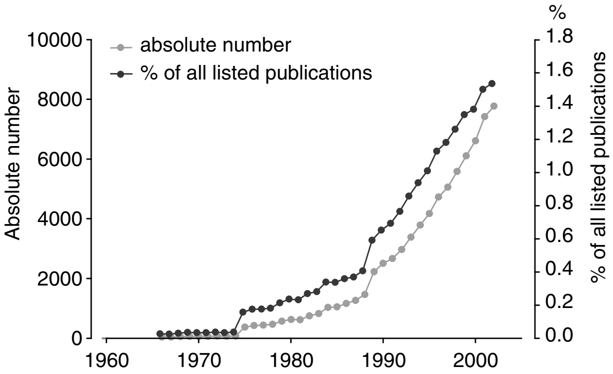


Fig. 6.1 Number of publications listed in Medline^{TM} that contain the word *questionnaire*

with undesirably low measurement precision can be identified and, if necessary, corrected by the administration of additional items. These additional items are not required when the measurement precision is already high. Substantial item savings may result, which would help to reduce the item burden placed on patients while still ensuring an efficient and precise measurement.

Response formats used in clinical psychology and medicine differ from those employed in ability and achievement tests. Typically, such clinical tests involve polytomous items, whereas a dichotomous coding of responses (correct vs. incorrect) is often a more natural choice for ability and achievement tests. These differences in response formats might have contributed to the delay in the use of CATs in medicine (compared to its use in educational contexts) even though many results holding for the dichotomous case have been generalized to polytomous items (e.g. Samejima, 1993).

As far as CATs are concerned, polytomous response formats open up the opportunity for substantial item savings. With the exception of highly discriminating dichotomous items, information in responses to polytomous items is usually considerably higher than in the dichotomous case. As the asymptotic standard error (SE) of a response pattern is the reciprocal of the square root of the sum of item information of this response pattern, a CAT algorithm that terminates as soon as the current error falls below a given error bond will terminate sooner if item information is high.

For instance, the Anxiety-CAT presented by Walter et al. (2005, 2007) requires, on average, about seven items to reach a predefined measurement precision of SE ≤ 0.32 (stopping criterion, corresponding to a reliability $\rho \geq 0.9$) for latent trait values within two standard deviations around the population mean. Similar reports on substantial item savings have been published by Ware, Bjorner, and Kosinski (2000), Ware et al. (2003), Fliege et al. (2005, 2009), or Haley et al. (2006).

If IRT, as a general framework, and computer adaptive testing, as a specific application of this framework, are recognized as being beneficial for overcoming the shortcomings of classical test theory (e.g., Hambleton, 2000; Hornke, 1999), the question remains as to why there are relatively few working applications of CATs in a clinical setting. In contrast, a number of authors have already employed an IRT-based approach to the analysis of health-related items (e.g., Bjorner, Kosinski & Ware, 2003c; Childs et al., 2000; Cooke et al., 1999; King et al., 1993; Krueger & Finger, 2001; Santor & Ramsay, 1998).

One possible explanation for this small number of applications of CATs in a medical context is the fact that there is, to date, no golden standard for constructing CATs. Crucial methodological issues in developing CATs include topics such as the assessment of item fit, the ensuring of the unidimensionality of the underlying construct, and the selection of an IRT model (e.g., for polytomous responses, the choice among rating scale, partial credit, generalized partial credit, or graded response model). Moreover, the construction of item banks requires substantial resources. For example, even though the exact number of respondents necessary for item calibration is still under debate, there is general agreement that this number should be rather large.

The majority of theoretical and practical contributions concerning the applications of CATs can be traced back to the seminal and ground-breaking work by Lord, Novick, and Birnbaum, which was focused on the ability and achievement testing context. In comparison, CATs in clinical contexts are still kittens. The development of clinical CATs is, however, moving forward and large-scale programs, such as by the PROMIS network, will play an important role in advancing them.

6.2 Development of CAT Systems

Using the IRT framework, we have developed and evaluated two item banks to measure anxiety (Anxiety-CAT) and depression (Depression-CAT) by means of a computerized adaptive method. The development of these CATs aimed at providing instruments that can be used in a real clinical setting for the assessments of these constructs in healthy persons, patients with somatic chronic diseases, and psychosomatic in- and outpatients. The development was motivated by the discrepancy between the general excitement about the theoretical advantages of CATs and the scarcity of reports of working CAT applications in clinical settings to date. A major aim of our studies was to evaluate the extent to which the theoretical advantages of CATs, namely precise and efficient measurement, would materialize in clinical practice.

The steps involved in the item bank construction process are summarized in Table 6.1. The following description focuses on the methodological decisions; a detailed description of the steps in development can be found in Walter et al. (2005, 2007) (Anxiety-CAT) and Fliege et al. (2005, 2009) (Depression-CAT).

6.2.1 Patient Samples for Empirical Item Analyses

The development of the item banks was based on data from the application of various standardized and well-established questionnaires used for routine psychometric

Table 6.1 Development steps of the Anxiety- and Depression-CATs

Definition of target construct
Unidimensionality checks
Inspection of item response functions
DIF analysis
Item calibration and item linking
Evaluation of model fit

assessment of in- and outpatients at the Department of Psychosomatic Medicine, Charité, University Medicine Berlin, Germany, obtained between 1995 and 2002. The item banks developed were comprised of subsets of items that were drawn from those standardized instruments considered pertinent to the constructs of anxiety and depression and met the statistical requirements of the IRT framework. The overall sample used for the construction of the Anxiety- and Depression-CAT consisted of $N = 2,348$ and $N = 3,270$ respondents respectively. These samples were used to conduct the empirical analyses and the item calibration described below.

6.2.2 Definition of Target Construct

In the first step of the construction process, the target constructs were defined conceptually. In the case of the Anxiety-CAT, the authors assented to Spielberger's (1972) definition of (state) anxiety as an "emotional state, characterized by strain, solitude, nervousness, inner discomposure and fear of future occasions" (Häcker & Stapf, 1998). This definition conforms to criteria for generalized anxiety disorders (F41.1) reported in the Tenth Revision of the International Classification of Diseases (ICD-10), where "fear, vegetative arousal and tenseness" are considered to be the main properties of anxiety disorders (Dilling, Mombour & Schmidt, 1999). Specific situations, activities, or objects pertaining to phobic disorders were not included. Depression was defined according to the criteria outlined in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition (DSM-IV), which include depressed mood, loss of interest and pleasure, activity disturbance, appetite or weight disturbance, sleep disturbance, fatigue or loss of energy, and thought of death/suicide.

6.2.3 Initial Item Pool

The initial item pool consisted of more than 300 items drawn from 13 standardized questionnaires used routinely in psychometric assessment at the Department of Psychosomatic Medicine, Charité, University Medicine Berlin, Germany. Administration of these questionnaires was conducted in a computer-assisted mode using personal digital assistants (PDAs). Each item was presented separately on the PDA screen (Rose et al., 2002). Among the instruments were German versions of questionnaires widely used internationally, for instance, the Center for Epidemiological Studies Depression Scale (CES-D), Beck Depression Inventory (BDI), Hospital Anxiety Depression Scale (HADS), SF-36 Health Survey, State Trait Anxiety Inventory (STAI), and Selfefficacy Scale and Life Orientation Test (SES/LOS). As the items were drawn from different instruments, only a subset of them was expected to be pertinent to the target constructs. All items were rated separately by members of the research group as to their relevance to anxiety or depression. Only those items upon which the raters agreed remained in the item pool (anxiety: 81 items; depression: 144 items).

6.2.4 Test Dimensionality

The question as to whether the items are measuring one underlying dimension or separate dimensions (Bjorner, Kosinski, & Ware, 2003b) is a crucial one in IRT (Embretson & Reise, 2000). Exploratory and confirmatory factor analysis can be employed to determine the extent to which items are unidimensional (Hays, Morales & Reise, 2000). To ensure unidimensionality, we conducted a one-factorial confirmatory factor analysis for categorical variables using MPlus (Muthén & Muthén, 2004) and excluded one item of each pair of items exhibiting residual correlations larger than 0.25. This particular choice of cut-off level was motivated by reports that item calibration is to some extent robust to slight violations of unidimensionality (Drasgow & Parsons, 1983; Reckase, 1979) and by the approach employed by Bjorner, Kosinski, & Ware (2003b), in which a similar, albeit slightly more conservative, cut-off of 0.20 was used.

6.2.5 Nonparametric Analyses

In our approach to item bank construction, visual inspection of item response functions (IRFs) computed nonparametrically (Gaussian kernel smoothing; see, Ramsay, 1995) proved to be a useful step during the analysis. The aim of this step was to compare the shapes of the observed response functions with those of parametrically modeled functions. An ideal category function exhibits steep trace lines with one sharp maximum and exceeds all other response functions in exactly one interval of the latent trait. Sorted in ascending order, the θ values for which a response function is maximal should match the order in which the response choices of an item are presented (Figures 6.2a and 6.2b). In some cases, the observed pattern could be brought into line with the ideal pattern by collapsing two or more response options (Figures 6.2c and 6.2d). In other cases, such amendments were not possible. Items with unsatisfactory response functions were excluded from further analysis.

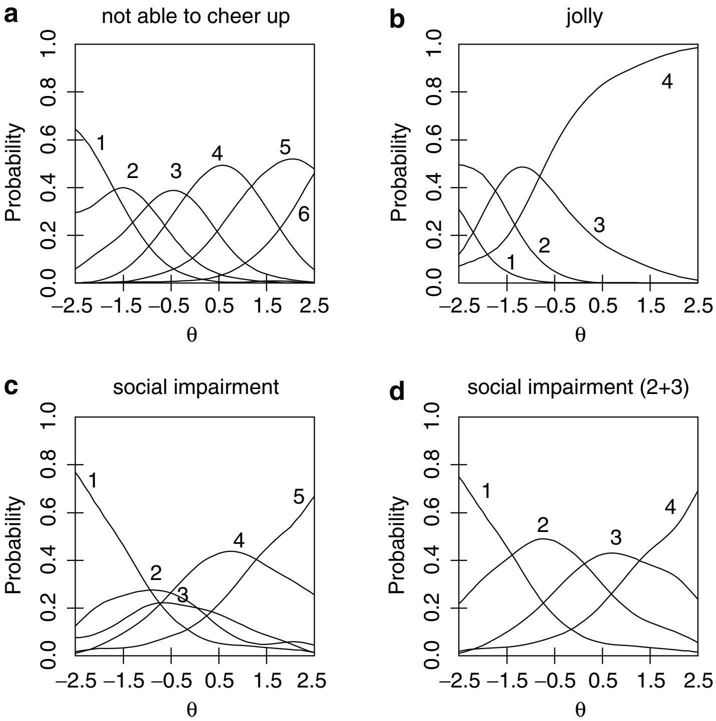


Fig. 6.2 Examples of nonparametric item response functions (IRFs) for standard normal θ scores of the Depression-CAT: (**a**) Item with ideal trace line pattern (steep curve and peaked maximum). (**b**) Easy item (i.e., low location parameter) with appropriate ICCs. (**c**) Item with unsatisfactory IRFs. (**d**) Amended item after collapsing response options 2 and 3

6.2.6 DIF Analysis

Systematic differences due to group biases can be determined by tests of differential item functioning (DIF; Holland & Wainer, 1993). To ensure that all items can be applied to individuals of any age or sex group, DIF analyses were conducted using a polytomous logistic regression model (Swaminathan & Rogers, 1990; Zumbo, 1999) with item response as the dependent variable and scale score and group membership (gender, age group, sample group) as independent variables. In this approach, uniform DIF is revealed by a direct effect of group membership on the item score when controlled for the scale score; nonuniform DIF is indicated by an interaction effect between item scale score and group membership (Zumbo & Hubley, 2003). Statistical significance and Nagelkerke's R^2 (Nagelkerke, 1991) can be used as criteria for determining whether or not an item exhibits DIF. The latter criterion quantifies the magnitude of DIF as the difference between variance before and after including the variable for group membership (Bjorner, Kosinski, & Ware, 2003b). This criterion is particularly useful when dealing with large sample sizes where even small DIF effects may become statistically significant. We saw an item as exhibiting DIF if the R^2-difference (ΔR^2) exceeded a cut-off value of 0.03. The method proposed by Zumbo (1999) was found to be less restrictive (i.e., to exclude fewer items), so we used the more conservative $\Delta R^2 > 0.03$ as the indicator of DIF. Recommendations on how to choose such cut-off values are rare, with Bjorner, Kosinski, & Ware (2003b), who used a similar, slightly more restrictive cut-off value of 0.02, being an exception.

6.2.7 Item Calibration

The items of the Depression- and Anxiety-CAT had polytomous response categories ranked appropriately. For instance, the item *I am worried* (Anxiety-CAT) had the following four response categories: *not at all*; *somewhat; rather; very*. Visual analysis of nonparametric item response curves (see above) suggested that the slopes of their trace lines ranked the response categories differently. Therefore, these items were calibrated using the generalized partial credit model (GPCM; Muraki, 1992), a two-parameter model for polytomous items. In contrast to the partial credit model (PCM; Masters & Wright, 1997), the item parameters in the GPCM are allowed to vary in slope. Item parameter estimation was conducted by the marginal maximum likelihood estimation procedure implemented in the Parscale software (Muraki & Bock, 1999). As in the two-parameter model for dichotomous items, the slope parameter in the GPCM largely determines the level of information in responses to the item. As in our CAT algorithm item selection was based on this information (the higher the information in an item at the current θ estimate, the more likely the selection of the item during the test), we decided to exclude items with slope parameters smaller than 0.80 (Anxiety-CAT) and 0.70 (Depression-CAT).

The sample of respondents consisted of several subsamples to which different sets of items were administered. The items in the sets had to be linked to a common scale. A prerequisite of such a linking is a set of common items used in all subsamples (Kim & Lee, 2006). These anchor items were used to estimate a linear transformation that linked the different item sets. More specifically, we used the mean/sigma method for the b parameter (Kolen & Brennan, 2004, p. 167) to adjust the item parameters of the anchor items in one sample to those of the other sample. These adjusted parameters then remained fixed while the other items in the second sample were recalibrated.

6.2.8 Investigation of Model Fit

At present, there are no widely accepted procedures for assessing the model fit of polytomous IRT models such as the GPCM. This is particularly true when dealing with large sample sizes where even tiny discrepancies may become statistically significant. In the construction of the item banks for the Anxiety- and Depression-CAT, we pursued two approaches to the investigation of model fit. To examine the relationship between the ratio of well-predicted and mispredicted scores, we conducted the likelihood-ratio χ^2 statistics provided by the Parscale software (Muraki, 1997; Muraki & Bock, 1999). For a more detailed analysis of the model fit, we computed a test statistic for each response category using an approach described by Bjorner and colleagues (Bjorner, Kosinski & Ware, 2003a). This approach was an extension of likelihood-based fit indices for dichotomous IRT models devised by Orlando and Thissen (2000). The IRT model was employed to predict the distribution of item responses for each response category and each level of the sum score of all items in the scale. From the observed versus expected frequencies, a χ^2 fit statistic was computed for each response category of each item. Additionally, the observed proportions of responses in each category were plotted as a function of the sum score. Predicted proportions with 95% confidence intervals were displayed in the same graph and allowed for graphical investigation of the model fit of each response category for each item.

6.2.9 Item Banks

The final item pool of the Anxiety- and Depression-CAT was comprised of 50 and 64 items, respectively. On the basis of the criteria described above, 31 (Anxiety-CAT) and 80 items (Depression-CAT) were excluded in total. The items excluded during the construction of the Anxiety-CAT pertained to specific physical aspects of anxiety, attention deficiencies, hypochondriac or social fears, and concerns regarding health or other people. Items excluded from the Depression-CAT mainly focused on side effects of depression, on social contacts, sexual function, work, and obligations.

6.2.10 CAT Algorithm

The CAT algorithm consists of several steps: (1) Initially, the person parameter estimate is set to $\hat{\theta}_0 = 0.0$, which is the assumed population mean. (2) For the current θ estimate, the item with the highest item information is selected and presented to the respondent. (3) After the respondent has answered, the response is used to compute both a new estimate $\hat{\theta}_1$ and standard error (SE) using the expected a posteriori (EAP) method (Bock & Mislevy, 1982). Steps (2) and (3) are repeated until either the current SE falls below 0.32 (stopping rule) or all items in the item bank have been presented to the respondent. The criterion SE ≤ 0.32 corresponds to a reliability of $\rho \geq 0.9$ ($\rho = 1 - \text{SE}^2$). When the algorithm has terminated, the θ estimate and standard error from the last step are reported.

In addition to person parameter estimation by the EAP method, the current implementation of the CAT algorithm allows for an extension of Warm's (1989) weighted likelihood estimation (WLE) to polytomous items. It has been noted that person parameter estimates using the EAP approach may be severely biased toward the prior mean (Chen, Hou & Dodd, 1998; Meijer & Nering, 1999). Results obtained from our own simulation studies supported these reports and also indicated that, for a given standard error, the EAP approach requires slightly fewer items than Warm's WLE. However, as for bias and root mean-square error, Warm's WLE is superior to the EAP method (Wang & Wang, 2001). The superiority is particularly noticeable for extreme values of θ, say, $|\theta| > 2$. However, for $|\theta| \leq 2$ (i.e., where some 95% of the standard normal population is expected to score), the difference between both methods becomes negligible, as the bias of EAP estimates tends to be small for this interval (Bock & Mislevy, 1982).

Warm's WLE is computationally more intensive. Its use is therefore recommended only in cases where unbiased test scores are needed, for instance, when cut-off scores are set on the ability scale or a comparison between scores on CAT and paper-and-pencil versions of a test are planned. Because of its ease of implementation, when only a ranking of test takers is required, the EAP method is an attractive option.

6.2.11 Delivery System

The CAT algorithm was implemented as a computer program written by the author. The core of this program (CAT engine) was written in standard C++ and can be run on several platforms. Various parametric IRT models for dichotomous and polytomous items are supported, such as the 1PL, 2PL, 3PL, partial credit, and generalized partial credit models. The CAT engine can be used for both simulation studies and psychometric assessment. In assessment mode, the engine can be attached to several graphical clients and CATs can be presented to respondents on PCs, laptops, or handheld devices. The most recent version of it runs on a web server and allows the delivery of CATs through a web browser interface.

6.3 Evaluation Studies

This CAT engine was used in various simulation studies in which the properties of the item banks of the Anxiety- and Depression-CAT were evaluated using the PC version of the CAT engine. In the first study, we used the method described by Wang (1999) and generated response patterns for simulated respondents. For each of the values of $\theta = -3.5(.25)3.5$, we generated the responses for 100 respondents. The stopping rule was set to $SE \leq 0.32$, and θ was estimated using EAP estimation. Figures 6.3a and 6.3b show the average number of items needed to measure the various levels of θ with the predefined precision.

For $|\theta| \leq 2$, the average number of items needed was 6.9 ± 2.6 and 7.15 ± 1.4 items (mean ±SD) for the Anxiety- and Depression-CAT, respectively. For extreme values of the latent trait ($|\theta| > 2$), substantially more items were required.

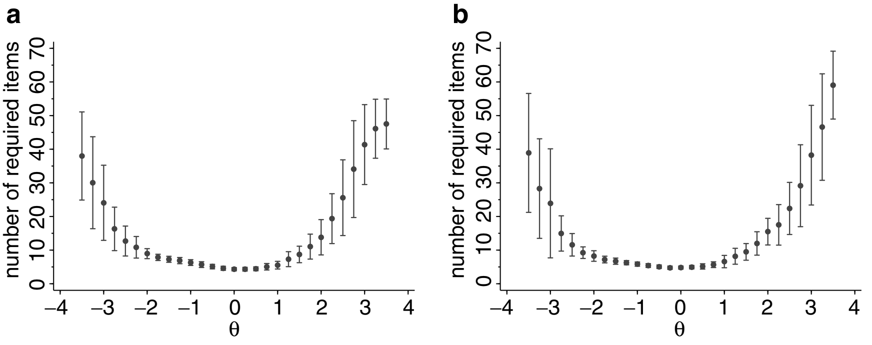


Fig. 6.3 Number of items required by the CAT algorithm as a function of θ. (**a**) Anxiety-CAT; (**b**) Depression-CAT

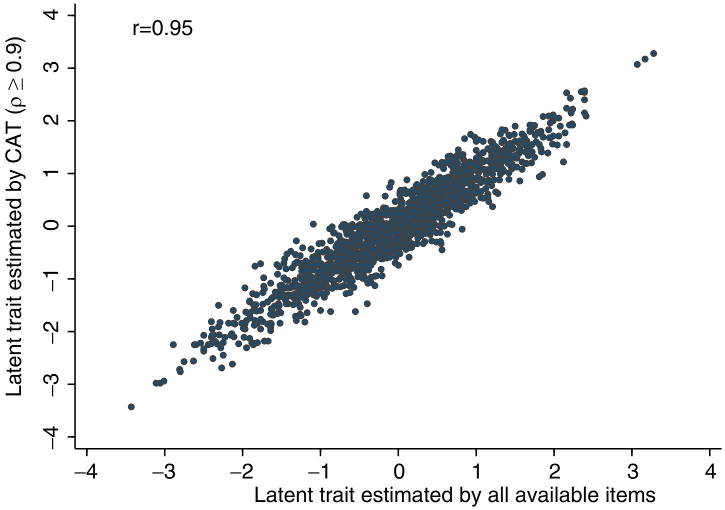


Fig. 6.4 Plot of the simulated scores on the Depression-CAT (EAP estimation; stopping rule SE $\leq$ 0.32) against the θ estimates computed from all available items in the pool. Note: the responses were from the patients in the sample used to calibrate the item bank for the Depression-CAT

In a second simulation, the ability levels were estimated from simulated CATs based on the real responses by the patients in the sample used to calibrate the item banks. Two runs were made. In the first run, the stopping rule was set to SE ≤ 0.32. In the second run, each item in the bank was used to estimate θ (provided a response to the item existed in the data set). As the respondents were administered different sets of items, not every item in the item bank was answered by all respondents. Nevertheless, the correlation between the θ estimates for the two runs was very high: $r = 0.97$ (Anxiety-CAT); $r = 0.95$ (Depression-CAT); see Figure 6.4. These findings indicate that, in spite of substantial savings of testing time for the CAT algorithm, not much information was lost and θ can still be estimated with as much precision as from the total item bank.

6.4 Discussion

The two CATs presented here were designed to measure the severity of symptoms of anxiety and depression. To this end, we selected items from those existing standardized questionnaires that met the statistical requirements of the IRT model and exhibited high discriminative power between individuals with different levels of anxiety or depression.

The IRT model we used assumes that an individual's responses to the items can be accounted for by just one latent variable. An evaluation of this assumption of unidimensionality showed that items assessing somatic symptoms of anxiety could not be located on the same scale as those focusing on emotional or cognitive symptoms. From a clinical perspective, the exclusion of somatic symptoms from the test is not ideal. However, some encouraging results concerning the validity of the two CATs have been reported recently (Becker et al., 2004; Rose et al., 2004). In one of our real-world applications of the Anxiety-CAT, a comparison between the Anxiety-CAT and two standardized questionnaires for the assessment of anxiety (STAI, HADS) yielded lower correlations than in simulation studies. Nevertheless, the correlations between the CAT scores and STAI and HADS were entirely comparable to those between STAI and HADS (Walter et al., 2007). These results suggest that the Anxiety- and Depression-CATs cover the underlying constructs in a manner similar to well-established standardized questionnaires.

It should be noted that differences in focus exist even between conventional, fixed-length questionnaires. STAI, HADS, and our Anxiety-CAT focus more on general distress and negative affects, whereas, for example, the Beck Anxiety Inventory (BAI) emphasizes somatic symptoms. These different foci correspond with the findings of several empirical studies favoring a three-factor model for anxiety, in which the first two factors reflect nonspecific aspects and somatic manifestations of anxiety (Joiner, Catanzaro & Laurent, 1996; Zinbarg & Barlow, 1996; Mineka, Watson & Clark, 1998). As long as the use of multidimensional IRT models has not yet arrived in the field of measurement of health-related constructs (e.g., Gardner et al., 2002), it seems recommendable to use a separate item bank to capture the somatic aspects of anxiety or depression.

In the Anxiety- and Depression-CATs, item selection is determined solely by the statistical criterion of maximum information. A possible refinement of the item selection could include the balancing of test content across test takers.

However, even in its present form, both the Anxiety- and Depression-CAs provide efficient measurement of clinically relevant constructs. Given the well-known fact that patients with depressive disorders often experience each extra questionnaire item as a burden, the substantial reduction of the test length not only reduces testing time but also alleviates the stress involved in the testing procedure. Even though the era of CATs in health-related measurement has just begun, and many problems still need to be solved, the results obtained so far indicate that medical applications of computer-adaptive testing have the potential to further improve psychometric assessment in the near future.

References

Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., Kocalevent, R., Schmid, G., Klapp, B. F. & Rose, M. (2004). Validating the German computerized adaptive test for anxiety on healthy sample (A-CAT). *Quality of Life Research*, 13, 1515.

Bjorner, J.B., Kosinski, M. & Ware, J. E. (2003a). The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. *Quality of Life Research*, 12 , 887–902.

Bjorner, J., Kosinski, M. & Ware, J. (2003b). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT-super^{TM}). *Quality of Life Research*, 12, 913–933.

Bjorner, J., Kosinski, M. & Ware, J. (2003c). Using item response theory to calibrate the Headache Impact Test (HIT-6-super^{TM}) to the metric of traditional headache scales. *Quality of Life Research*, 12, 981–1002.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431–444.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce B. & Rose, M. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH Roadmap Cooperative Group during its first two years. *Medical Care,* 45, I3–I11.

Chen, S., Hou, L. & Dodd, B. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement*, 58, 569–595.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M. & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 Depression scale. *Assessment*, 7, 37–54.

Cooke, D. J., Michie, C., Hart, S. D. & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist—Revised (PCL:SV): An item response theory analysis. *Psychological Assessment*, 11, 3–13.

Dilling, H., thinspace Mombour, W. & Schmidt, M. H. (1999). *Internationale Klassifikation psychischer Störungen. ICD-10 Kapitel V(F). Klinisch-diagnostische Leitlinien* (3. Aufl.). Bern: Huber.

Drasgow, F. & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189–199.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8, 341–349.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F. & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research*, 14, 2277–2291.

Fliege, H., Becker, J., Walter, O. B., Rose, M., Bjorner, J. & Klapp, B. F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. *International Journal of Methods in Psychiatric Research,* 18, 23–36.

Fries, J. F., Bruce, B. & Cella, D. (2005). The promise of PROMIS: Using item response theory to improve assessment of patient-reported outcomes. *Journal of Clinical and Experimental Rheumatology*, 23, 33–37.

Gardner, W., Kelleher, K. J. & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. *Medical Care*, 40, 812–823.

Häcker, H. & Stapf, K.-H. (1998). *Dorsch Psychologisches Wörterbuch*. Bern: Huber.

Haley, S. M., Pengsheng N., Hambleton R. K., Slavin M. D. & Jette A. M. (2006). Computer adaptive testing improved accuracy and precision of scores over random item selection in a physical functioning item bank. *Journal of Clinical Epidemiology*, 59, 1174–1182.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38, II60–II65.

Hays, R. D., Morales, L. S. & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care*, 38, II28–II42.

Holland, P. W. & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Hornke, L. (1999). Benefits from computerized adaptive testing as seen in simulation studies. *European Journal of Psychological Assessment*, 15, 91–98.

Joiner, T., Catanzaro, S. & Laurent, J. (1996). Tripartite structure of positive and negative affect, depression, and anxiety in child and adolescent psychiatric inpatients. *Journal of Abnormal Psychology*, 105, 401–409.

Kim, S. & Lee, W.-C. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement*, 43, 53–76.

King, D. W., King, L. A., Fairbank, J. A. & Schlenger, W. E. (1993). Enhancing the precision of the Mississippi Scale for Combat-Related Posttraumatic Stress Disorder: An application of item response theory. *Psychological Assessment*, 5, 457–471.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling and linking* (2nd ed.). New York: Springer-Verlag.

Krueger, R. F. & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. *Psychological Assessment*, 13, 140–151.

Masters, G. N. & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer-Verlag.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, 23, 187–194.

Mineka, S., Watson, D. & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. *Annual Review of Psychology*, 49, 377–412.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164) New York: Springer-Verlag.

Muraki, E. & Bock, R. D. (1999). *PARSCALE: Analysis of graded responses and ratings*. Chicago: Scientific Software International, Inc.

Muthén, L. K. & Muthén, B. O. (2004). *Mplus. The Comprehensive Modeling Program for Applied Researchers. User's Guide*. Los Angeles: Muthén & Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78, 691–692.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24, 50–64.

Ramsay, J. O. (1995). *TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data*. Montreal: McGill University. (http://www.psych.mcgill.ca/faculty/ramsay/TestGraf.html)

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207–230.

Rose, M., Walter, O. B., Becker, J., Kocalevent, R., Fliege, H., Schmid, G., Grimm, A. & Klapp, B. F. (2004). Evaluating a computer adaptive test for depression (D-CAT) in a healthy sample. *Quality of Life Research*, 13, 1515.

Rose, M., Walter, O. B., Fliege, H., Becker, J., Hess, V. & Klapp, B. F. (2002). Seven years of experience using personal digital assistants (PDA) for psychometric diagnostics in 6000 inpatients and polyclinic patients. In H.-B. Bludau & A. Koop (Eds.), *Mobile computing in medicine (Lecture notes in informatics)* (pp. 35–44). Bonn: Köllen.

Samejima, F. (1993). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika*, 58, 195–209.

Santor, D. A. & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. *Psychological Assessment*, 10, 345–359.

Spielberger, C. D. (1972). *Anxiety: Current trends in theory and research*. Oxford: Academic Press.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361–370.

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F. & Rose M. (2007). Development and evaluation of computer adaptive test for anxiety (Anxiety-CAT). *Quality of Life Research*, 16, 143–155.

Walter, O. B., Becker, J., Fliege, H., Bjorner, J. B., Kosinski, M., Walter, M., Klapp, B. F. & Rose, M. (2005). Entwicklungsschritte für einen computeradaptiven Test zur Erfassung von Angst (A-CAT). *Diagnostica*, 51, 88–100.

Wang, S. (1999). *The accuracy of ability estimation methods for computerized adaptive testing using the generalized partial credit model*. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburg, PA.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, 25, 317–331.

Ware, J. E., Bjorner, J. B. & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care*, 38, II73–II82.

Ware, J. E., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlöf, C. G. H., Tepper, S. & Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research*, 12, 935–952.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450.

Zinbarg, R. & Barlow, D. (1996). Structure of anxiety and the anxiety disorders: A hierarchical model. *Journal of Abnormal Psychology*, 105, 181–193.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D. & Hubley, A. M. (2003). Item bias. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment* (pp. 505–509). Thousand Oaks, CA: Sage.

Chapter 7
MATHCAT: A Flexible Testing System in Mathematics Education for Adults

Angela J. Verschoor and Gerard J.J.M. Straetmans

7.1 Introduction

One of the mathematics courses in adult basic education in the Netherlands is offered at three different levels. The majority of the students are foreign and, due to a large variation in background, most of their educational histories are unknown or can be determined only unreliably. In the program's intake procedure, a placement test is used to assign students to a course level. As the students' abilities vary widely, the paper-and-pencil placement test currently used has the two-stage format described in Lord (1971). In the first stage, all examinees take a routing test of 15 items with an average difficulty matching the average proficiency in the population of students. Depending on their scores on the routing test, the examinees then take one of the three follow-up tests. Each follow-up test consists of 10 items.

There are several drawbacks to this current testing procedure:

1. Test administration is laborious because of the scoring that has to take place after the routing test.
2. Preventing disclosure of the test items is difficult due to the flexible intake procedure inherent in adult basic education. Disclosed items can easily lead to misclassifications (assignment of prospective students to a course level for which they lack proficiency).
3. Because only one branching decision is made, possible misroutings cannot be corrected (Weiss, 1974) and measurement precision may be low.

A computerized adaptive placement test has offered a solution to these problems. First, such tests have as many branching decisions as items in the test. Erroneously branching on the items, because of incorrect responses to items that are too easy or correct responses to items too difficult, is corrected later in the test. Second, computerized test administration offers the advantage of immediate test scoring and

A.J. Verschoor (✉)
Cito Institute for Educational Measurement, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

G.J.J.M. Straetmans
Cito Institute for Educational Measurement, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_7,

feedback. As a result, remedial measures can be taken right after the test. Third, preventing disclosure of testing material is less of a problem because, in principle, each examinee takes a different test.

These features of computerized adaptive testing are very interesting, particularly because all colleges offering adult basic education in the Netherlands already have well-equipped computer rooms or are in the process of installing them. Besides, the technology of computerized adaptive testing is flexible enough to deliver tests from the same pool for other purposes than placement decisions, such as monitoring student achievements during the course or grading the students at the end of it. A testing system with these additional features has been thought to be very helpful in supporting the current movement toward a more flexible adult education system in the Netherlands.

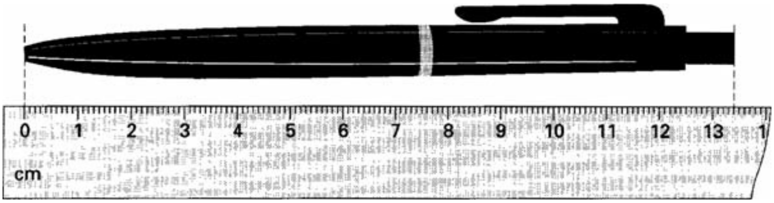
In this chapter, a description of MATHCAT, the current adaptive testing system that has replaced the old paper-and-pencil two-stage test, is given. MATHCAT delivers tests serving two different educational purposes. One purpose is placing examinees into courses in arithmetic/mathematics at three available levels. The other is achievement testing during these courses to monitor the students' achievements. We first describe the item bank because its quality has strong consequences for the utility of the test scores. Then, the testing algorithms for the placement and achievement tests are discussed and results from an evaluation of these algorithms are presented. Finally, we will discuss some features of the student and teacher modules in MATHCAT.

7.2 The Item Bank for Numerical and Mathematical Skills

Adaptive testing requires an item bank calibrated using an appropriate IRT model. The item bank currently used by MATHCAT contains 578 items, of which 476 were calibrated using the following model (Verhelst & Glas, 1995):

$$p_i(\theta) \equiv P(X_i = 1|\theta) \equiv \frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}. \tag{7.1}$$

The response to item i is either correct ($X_i = 1$) or incorrect ($X_i = 0$). The probability of answering an item correctly in (7.1) is an increasing function of the latent proficiency, θ, and depends on two item characteristics: difficulty b_i and discriminatory power a_i. All parameters of the items in the bank were estimated using the OPLM software (Verhelst & Glas, 1995). The software iteratively chooses integer values for the item parameters a_i, computes conditional maximum likelihood estimates (CML) of the item parameters b_i, and tests for model fit, until acceptable estimates of the values of the item parameters a_i and b_i are obtained. The distribution of θ was estimated using a marginal maximum likelihood (MML) method. The estimated mean and standard deviation were $\hat{\mu} = 0.074$ and $\hat{\sigma} = 0.519$. In addition, the item pool has a changing subset of items that are not yet operationally used but seeded into the tests to collect responses for their future calibration.



The length of the pen is ...... centimeters and ...... millimeters.

Fig. 7.1 Sample item #1 (domain: basic concepts and skills; level: 1; format: short answer)

Cutoff scores on the proficiency scale are used to define the three course levels. The cutoff scores were derived through the following procedure: First, content specialists defined subsets of items by labeling them as Level 1, Level 2, or Level 3 items. Second, the mean difficulty of each subset of items was computed. Third, using the basic equation of the OPLM model in (7.1), the cutoff scores were defined as the abilities that had a minimum probability of success equal to 0.7 for all items labeled as Levels 1 and 2, respectively. This procedure resulted in $\theta_{12} = -0.544$ (cutoff score between Levels 1 and 2) and $\theta_{23} = -0.021$ (cutoff score between Levels 2 and 3).

The item bank covers end-of-course objectives in the following four cognitive domains:

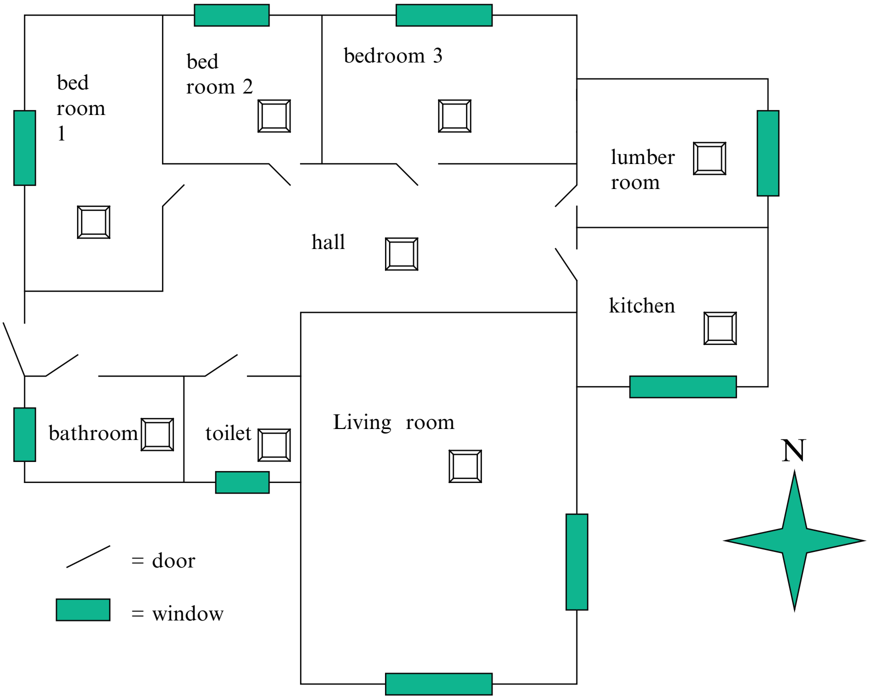
1. Basic concepts and skills (e.g., number operations, mental arithmetic, use of electronic calculator, measurement, fractions, percentages, and proportions);
2. Geometry (e.g., orientation in space, reading maps, identification of geometrical figures);
3. Statistics (e.g., interpreting tables and graphs, measures of central tendency, probability);
4. Algebra (e.g., relations between variables, equations).

Most items are of the short-answer type. Other item formats frequently used are the multiple-choice and multiple-response formats.

In Figures 7.1–7.4, four typical items from the MATHCAT pool are shown. The items were selected to represent the three different course levels, the four domains, and the dominant item formats.

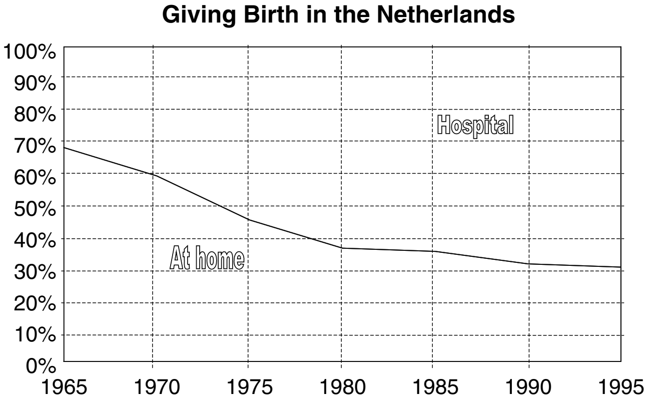
7.3 Item-Selection Algorithm

The item-selection algorithm drives the adaptation of the test; it determines how the test starts, continues, and stops. Different purposes of the test should be supported by different algorithms. The algorithms used for placement and achievement testing are discussed in the next two sections.



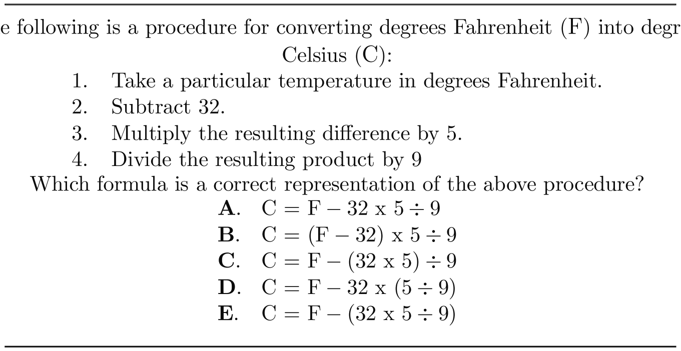
Eric and Fiona have bought a new house. This is the floor plan.
Which rooms have south-facing windows?

Fig. 7.2 Sample item #2 (domain: geometry; level: 2; format: select each alternative that applies)



The graph above shows the percentage of women giving birth to a child at home or in the hospital. What is the percentage of women giving birth in the hospital in 1985? .... percent.

Fig. 7.3 Sample item #3 (domain: statistics; level: 3; format: short answer)



The following is a procedure for converting degrees Fahrenheit (F) into degrees Celsius (C):

1. Take a particular temperature in degrees Fahrenheit.
2. Subtract 32.
3. Multiply the resulting difference by 5.
4. Divide the resulting product by 9

Which formula is a correct representation of the above procedure?

A. C = F – 32 x 5 ÷ 9
B. C = (F – 32) x 5 ÷ 9
C. C = F – (32 x 5) ÷ 9
D. C = F – 32 x (5 ÷ 9)
E. C = F – (32 x 5 ÷ 9)

Fig. 7.4 Sample item #4 (domain: algebra; level: 3; format: multiple choice)

7.3.1 Placement Testing

Purpose. The purpose of the placement test is to assign prospective students of adult basic education to three different course levels. An important practical requirement is that tests for this purpose should be as short as possible, with a maximum length of 25 items.

Administration Procedure. In adaptive testing, the choice of the next item is dependent on the current estimate of the examinee's proficiency. When testing begins, however, no previous information about the proficiency level of the examinee is available. This holds particularly for placement testing of new students to decide on their optimal level of instruction. In many CAT programs, this problem is resolved by selecting an item optimal at the average proficiency of the examinees in the calibration study.

In MATHCAT, a different strategy is used, the reason being that its examinees are often poorly educated and have bad recollections of attending school. In addition, many of them suffer from test anxiety. To make examinees feel more comfortable, the first two items in the placement test are selected at random from a subset of relatively easy (Level 1) items.

Mental arithmetic is an important topic in adult mathematics education. Its importance is reflected by the relatively large percentage of mental arithmetic items in the item bank (89 of 476 items). These items should be answered by performing mental calculations without the use of any paper or pencil. In order to meet this condition, the testing algorithm selects the first four items from a subset of mental arithmetic items. As soon as an examinee has responded to the fourth item in the test, he or she receives the following message from the software: "From now on you are free to use paper and pencil."

Wainer (1992) suggests that the use of adaptation in tests with a cutoff score is not worth the trouble. According to him: "The most practical way to make the test adaptive is to choose an adaptive stopping rule. Thus after each item we make the decision “pass”, “fail”, or “keep on testing”. If testing is continued, an item is selected whose difficulty matches that of the cut-score as closely as possible” (p. 4). The “trouble” Wainer refers to is the update of the examinee’s proficiency estimate each time an item has been responded to. However, modern computers have enough power to perform the required calculations very quickly.

Another reason why Wainer’s suggestion was difficult to follow was the very large numbers of items it would require at the cutoff scores. Also, the present item bank had to be designed to provide sufficient numbers of items along the full achievement continuum. This was necessary because of the two different purposes of MATHCAT (placement and achievement testing).

For the placement test, the items are selected using the maximum-information criterion (van der Linden & Pashley, this volume, chap. 1). This criterion selects items with maximum information at the current proficiency estimate for the examinee. The test stops as soon as the examinee can be assigned to a course level with 90% certainty, that is, when the 90% confidence interval for the examinee’s current proficiency estimate no longer covers either of the cutoff scores. This rule is used in combination with the requirement that the test length be between 12 and 25 items.

Figure 7.5 depicts an example of the process of administering the placement test to a high-proficiency examinee (Straetmans & Eggen, 1998). In the graph, the horizontal axis represents the successive items in the test. On the vertical axis both the difficulties of the selected items (denoted by crosses) and the proficiency estimates of the examinee (denoted by circles) are projected. The two horizontal lines represent the cutoff scores between Levels 1 (easy) and 2 (moderate) and between Levels 2 and 3 (hard). To put the examinee at ease, the first two items were selected at random from a subset of relatively easy items. After the examinee responded to the second item, the proficiency was estimated for the first time. Of course, this estimate cannot be very precise. The bar about each estimate represents the 90% confidence interval for the examinee’s proficiency. As both cutoff scores fell in the first confidence interval, it was not yet possible to determine which course level this examinee had to be assigned to; therefore, testing was continued. From this point

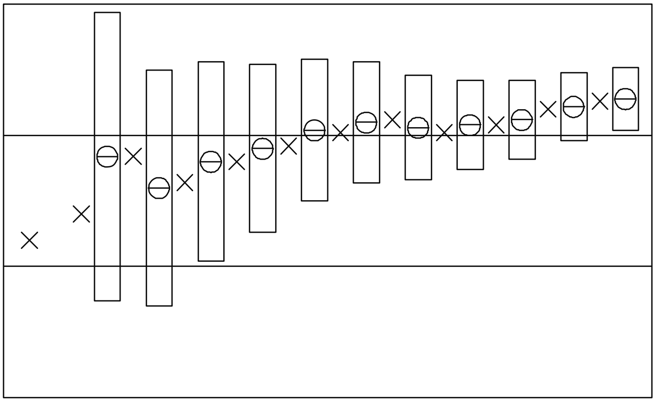


Fig. 7.5 Sample placement test taken by a high-proficiency student

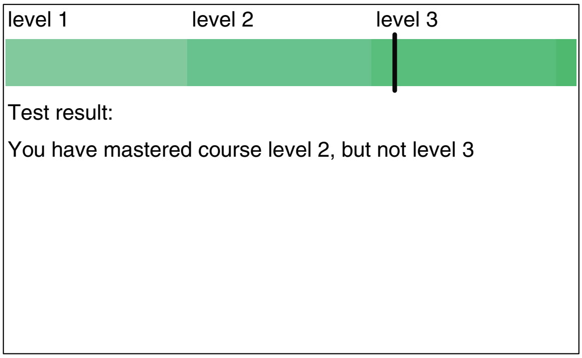
on, the adaptive test did what it is supposed to do and selected items with difficulties close to the proficiency estimates. If the examinee gave an incorrect answer, the estimate went down; for a correct answer, it went up. As demonstrated by the size of the 90% confidence interval, in either case the uncertainty about the examinee's proficiency level was decreased. After 12 items, the test was stopped because the confidence interval no longer covered either of the cutoff scores. Because the lower bound of the confidence interval was above the higher cutoff score, we could be fairly sure that the examinee had to be assigned to the Level-3 course.

Reporting of Results. Immediately after the test, the student gets information about his or her performance. It has been difficult to find a straightforward, yet sufficiently informative way of doing so. In adaptive testing, the number-correct score is not a good indicator of the student's performance. But reporting an estimate of θ only is not very informative either. Therefore, it was decided to report the examinees' performances on a graphical representation of the proficiency scale along with a short explanatory text. On the same scale, the three course levels are marked in various shades of gray. See Figure 7.6 for a sample report.

Evaluation of Placement Test. An important criterion for the evaluation of a placement test was the accuracy of the decisions based on it. In order to determine the accuracy of the MATHCAT system, the following simulation study was performed: Values for the proficiency parameter were drawn from the population distribution. For each course level, the proficiency range was divided into 10n equal-size intervals; 100 draws were made from each interval. For each value, an adaptive test was simulated using the algorithm above. The test was stopped using the rule above.

Table 7.1 shows how many simulees were placed at each of the levels by their test scores.

The accuracy was largest for proficiency values from Level 3 and smallest for values from Level 2. This result follows from the fact that Level 2 has two adjacent levels, while the two others have only one. It never occurred, for instance, that a Level 1 examinee was placed in Level 3, or vice versa. A simulation study for



Test result:

You have mastered course level 2, but not level 3

Fig. 7.6 Sample results for the placement test

Table 7.1 Accuracy of placement tests

True Course	Observed	Course	Level	% of Correct
Level	1	2	3	Decisions
1	920	80	0	92.0
2	93	831	76	83.1
3	0	53	947	94.7

Table 7.2 Length of placement tests

Course Level	Average Test Length (SD)	Percentage of Minimum Test Lengths	Percentage of Maximum Test Lengths
1	17 (5.9)	49	32
2	20 (5.6)	19	48
3	16 (5.4)	51	22

the previous two-stage, paper-and-pencil test resulted in percentages of correct decisions equal to 88.5% (Level 1), 81.6% (Level 2), and 91.1% (Level 3). Thus, the new adaptive test was more accurate.

Table 7.2 shows the average test lengths (with standard deviations) as well as the percentages of tests that had minimum (12 items) and maximum length (25 items). Compared to the current two-stage test (fixed length of 25 items), the average length of the adaptive tests was considerably shorter.

7.3.2 Achievement Testing

Purpose. The second purpose of MATHCAT was to monitor the achievements of the examinees after placement at a course level. The two questions that had to be addressed by the test results were (1) To which extent are the objectives of the course level met? (2) What are the strong and weak points in the achievements of the student?

To be able to identify the strong and weak points in the achievements, the relevant content domains had to be represented in the test by reasonable numbers of items. A test that is just generally most informative does not necessarily represent these domains well. Thus, the item-selection algorithm for the achievement test should deal with several content constraints; for a review of techniques to impose such constraints, see van der Linden (this volume, chap. 2).

Item-Selection Procedure. The procedure used to implement the content constraints on the item-selection process consists of three phases.

The goal of the first phase is to provide an initial estimate of the proficiency of the student. This information is then used to determine which test content specification should be used. The initial idea was to use the results from the placement test for this purpose, but the idea had to be abandoned for two reasons: First, students do

Table 7.3 Numbers of items from (sub)domains in achievement test

				Level		
	Domain		Subdomain	1	2	3
1.	Basic concepts and skills					$\geq$10
		1.1	Number operations, mental arithmetic	$\geq$10	$\geq$10	
		1.2	Electronic calculator			
		1.3	Fractions, proportions and percentages	$\geq$10		
		1.4	Measurement	$\geq$10	$\geq$10	
2.	Geometry					$\geq$10
3.	Statistics					$\geq$10
4.	Algebra					$\geq$10
	Total			30	35	40

not always take the placement test. Second, these test results might already have become obsolete, particularly if some time has elapsed between the administration of the placement test and the current achievement test.

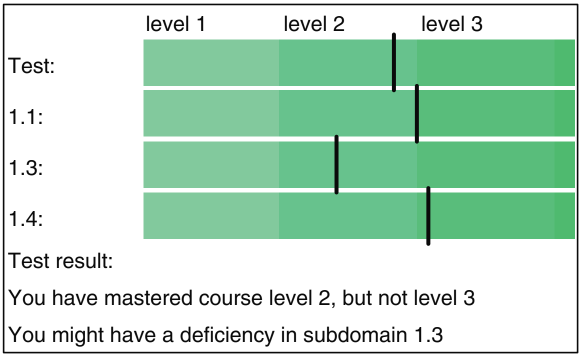
Therefore, in the first phase, 10 new items are administered; the first four are mental arithmetic items, the remaining six are drawn from Domain 1 (see the content specifications for the achievement test in Table 7.3). Depending on the proficiency estimate, the items administered in the second phase have to obey one of three different sets of content specifications. If $\widehat{\theta} \leq -0.544$, the specifications belonging to the objectives of Level 1 are chosen. If $\widehat{\theta} \geq -0.021$, the specifications for Level 3 are chosen. For the intermediate values, the content specifications for Level 2 are chosen.

In the second phase, 20 to 30 items are administered. The items for Level 1 are mainly from Domain 1. The other domains are covered only marginally at this level. Also, for Domain 1, most items are taken from Subdomains 1.1 and 1.4. As an additional constraint, for the first and second phases together, at least ten items should be taken from these two subdomains. The other 10 items are selected freely from all domains. The items for Level 2 are approximately from the same domain as those for Level 1, one exception being that 10 items should be selected from Subdomain 1.3. For Level 3, the items are to be selected predominantly from Domains 2, 3, and 4.

In the third phase of the test, five pretest items are administered. The responses to these items are not used for any proficiency estimation but for calibration purposes only. To prevent examinees from taking pretest items that are too easy or too difficult, the items are selected according to the following rule:

$\theta \leq 0.544$: random selection from the domains for Levels 1 and 2.
$-0.544 \leq \theta \leq -0.021$: random selection from the domains for all three levels.
$\theta \geq -0.021$: random selection from the domains for Levels 2 and 3.

Reporting of Results. Again, immediately upon completion of the test, the results are reported graphically along with a short explanatory text. The report not only depicts



Test result:

You have mastered course level 2, but not level 3

You might have a deficiency in subdomain 1.3

Fig. 7.7 Sample report for the achievement test

Table 7.4 Accuracy of phase 1 of achievement tests

	Observed	Course	Level	% of Correct
True Course Level	1	2	3	Decisions
1	848	152	0	84.8
2	122	760	118	76.0
3	0	122	878	87.8

the general proficiency level but also proficiency estimates for relevant domains or subdomains. If an estimate for a domain is significantly lower than the overall estimate, a warning is given that there might be a deficiency in the student's knowledge. An example of a report is shown in Figure 7.7.

Evaluation of Achievement Test. To assess the accuracy of the test, a simulation study was performed. The accuracy of the first phase in the test was evaluated by the percentage of correct branching decisions made for a typical population of examinees. The accuracy of the whole test was evaluated by the standard error of the proficiency estimate for the combination of the responses from the first and second phases.

Evaluation of First Phase. To this end, 1,000 proficiency values from each course level were drawn. That is, 1,000 values were drawn from the interval $[-1, -0.54]$; 1,000 values from the interval $[-0.54, 0.002]$; and 1,000 values from the interval $[-0.02, 0.46]$. The values were drawn from a uniform distribution over the intervals. Table 7.4 shows how many examinees were branched toward the three different levels.

The accuracy was largest for the true proficiencies at Level 3 and smallest for those at Level 2. The latter result can be explained by the fact that Level 2 has two adjacent levels, while the two others have only one. Examinees were never placed more than one level from their true proficiency.

Evaluation of Final Proficiency Estimate. To assess the accuracy of the final proficiency estimate, 5,000 proficiency values were drawn from the population distribution, which was estimated to be normal with mean 0.074 and standard deviation 0.519. For each value, an adaptive test was simulated and the test length and final

Table 7.5 Some statistics on achievement tests

	Mean	SD
Estimated Proficiency	0.071	0.530
SEM	0.088	0.022
Test Length	37.147	3.692

Table 7.6 Statistics on achievement tests per course level

Level	Mean Estimated Proficiency	SEM	Sample Size
1	−.761	.101	675
2	−.239	.082	1480
3	.430	.088	2842

Table 7.7 Statistics on achievement tests per (sub)domain

	Mean Estimated		Sample	# of Deficiency
	Proficiency	SEM	Size	Warnings
Subdomain 1.1	−.405	.170	2155	126 (5.8%)
Subdomain 1.3	−.252	.145	1480	52 (3.5%)
Subdomain 1.4	−.393	.164	2155	71 (3.3%)
Domain 1	.434	.181	2842	175 (6.2%)
Domain 2	.417	.189	2842	139 (4.9%)
Domain 3	.423	.230	2842	219 (7.7%)
Domain 4	.420	.186	2842	140 (4.9%)

proficiency estimate were recorded. Table 7.5 gives some statistics of the distribution of estimates and test lengths.

The same statistics were also calculated for each course level separately. The results are shown in Table 7.6.

Finally, a number of statistics specific to each of the (sub)domains were calculated. These results are given in Table 7.7. The deficiency warnings in the last column of Table 7.7 are erroneous; no true deficiencies were simulated.

7.4 MATHCAT Software

The MATHCAT software consists of both a student and a teacher module. The student module administers the test and reports the results to the students. The teacher module can be used to perform the following tasks:

1. Adding and removing students;
2. Planning a test for a student (a student is only allowed to take a test that the teacher has planned for him or her);
3. Viewing the most recent test results for all students (group report);
4. Viewing all available test results for a selected student (individual report).

Group Report. In Figure 7.8, an example of a group report is given:

Student ID	Student Name	Placement Test	Achievement Test
1	E. Long	79 (4/1/99)	86 (10/3/99)
2	R. Smith	91 (4/2/99)	
3	S. Baker	103 (4/2/99)	XXX

Fig. 7.8 Example of a group report

Name:	E. Long		
Date:	4/1/99	6/2/99	10/3/99
Placement Test:	79 - Level 1		
Achievement Test:		83 - Level 2	86 - Level 2
Subdomain 1.1:		87 - Level 2	88 - Level 2
Subdomain 1.3:		**75** - Level 1	85 - Level 2
Subdomain 1.4:		86 - Level 2	86 - Level 2
Domain 1:			
Domain 2:			
Domain 3:			
Domain 4:			

Fig. 7.9 Example of individual report

To report more realistic numbers, the estimated proficiencies in the reports are transformed by $\widehat{\theta}^* = 28.68\widehat{\theta} + 96.44$. Thus, the cutoff score between Levels 1 and 2 in the reports is at 82 and the one between Levels 2 and 3 is at 96. If no test result is shown, the student has not yet taken the test. If the test result is reported as "XXX", the student is currently taking the test.

Individual Report. An example of a review of all test results by one selected student is given in Figure 7.9. The review not only depicts the transformed proficiency estimates for this student but also the levels at which these estimates were classified. Besides the overall scores, the relevant profile scores are shown, together with their course levels, as well as the diagnostic warnings.

The report provides diagnostic warnings of two different types. The first type is detected deficiencies. In Figure 7.9, this type of warning is printed in boldface; for example, the score of **75** for Subdomain 1.3 taken on 6/2/99 is suspect. The second type is a warning of absence of progress issued when an estimated proficiency is not substantially higher than the previous estimate. In Figure 7.9, this type of warning is printed in italic; for example, the score of *88* for Subdomain 1.1 taken on 10/3/99 is not substantially higher than the previous result of *87* on 6/2/99.

7.5 Conclusions

Since January 1999, the MATHCAT testing system has been available to Dutch colleges for basic adult education. The use of MATHCAT has several advantages: greater accuracy, shorter test lengths, and easier usage. Decisions based on these tests are slightly more accurate than for the previous two-stage paper-and-pencil placement test (89.9% vs. 87.3% of correct placements). At the same time, however, the tests are considerably shorter. The software has been proven to be simple to use in practice. All test scoring, previously done by hand, is now done by the testing system. Unlike the previous two-stage test, no manual scoring after a first subtest is necessary. In sum, the main advantage of the system is less time-consuming test administration for both the students and the teachers. As a result, the teachers can now spend more time on their core activity: teaching.

References

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242.

Straetmans, G. J. J. M. & Eggen, T. J. H. M. (1998). Computerized adaptive testing: What it is and how it works. *Educational Technology, 38,* 45–52.

Verhelst, N. D. & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Their foundations, recent developments and applications* (pp. 215–237). New York: Springer-Verlag.

Wainer, H. (1992). *Some practical considerations when converting a linearly administered test to an adaptive format* (Research Report No. 92-13). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report No. 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Chapter 8
Implementing the Graduate Management Admission Test Computerized Adaptive Test

Lawrence M. Rudner

Wise and Kingsbury (2000) argue that the success of an adaptive testing program is a function of how well the various practical issues are addressed. Decisions must be made with regard to test specifications, item selection algorithms, pool design and rotation, ability estimation, pretesting, item analysis, database design, and data security. The test sponsor is ultimately responsible for each of these decisions and must work closely with the vendor to assure that the sponsor interests are met.

This chapter draws on the 12 years of experience of the Graduate Management Admission Council® in implementing a CAT-driven large-scale assessment. The chapter starts with an overview of the Graduate Management Admission Test® (GMAT), outlines the conversion to CAT in 1996, and then presents a range of practical issues. For each issue, we outline several options that are available and, to the extent possible, the approaches taken by GMAC.

8.1 Overview of the GMAT

The GMAT is a standardized assessment intended to help business schools assess the qualifications of applicants for advanced study in business and management and is comprised of three main components, the Analytical Writing Assessment (AWA), the Quantitative section, and the Verbal section. More than 200,000 examinees take the examination annually and GMAT scores are reported to more than 3,000 different programs. The test is continuously available, by appointment, through more than 400 testing centers worldwide.

An analysis of the results from 273 validity studies involving 41,338 students conducted during the calendar years 1997–2004 has shown the GMAT to be a good predictor of first-year grades (Talento-Miller & Rudner, 2005, 2008). The interquartile range of the predictive validity of the GMAT total score, AWA score, and undergraduate grade point average is 0.448 to 0.626, with a mean of 0.530. Of

L.M. Rudner (✉)
Graduate Management Admission Council, 1600 Tysons Boulevard, Ste. 1400, McLean, VA 22102, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_8,

special note is that the test is a much better predictor of performance in the first-year MBA program than prior grades, perhaps because of the wide diversity of students pursuing a degree in management.

The GMAT relies on the three-parameter logistic (3PL) response model. Items are calibrated and evaluated, in part, based on item parameters. Pools are formed to meet target conditional errors based on the model. The testing algorithm uses 3PL item parameters in selecting items to be adaptively administered.

8.1.1 Content

Table 8.1 provides and overview of GMAT content, allotted times, and scoring. Total examination time is 2.5 hours, not including a short questionnaire and optional breaks.

While the content titles may appear to be similar to those of a general-purpose admissions test, the GMAT test emulates business-like conceptualization through its emphasis on logical reasoning in both the verbal and quantitative sections and its use of business-related content.

With data sufficiency, an item type that is unique to the GMAT, the examinee is required to determine whether there is enough information to solve a problem; the examinee is not asked to solve the problem. These questions are designed to measure the examinee's ability to analyze a quantitative problem, to recognize which information is relevant, and to determine at what point there is sufficient information to solve the problem. An example is shown in Figure 8.1.

The correct answer is D. While data sufficiency and problem solving tap high-order skills, the content specifications call as well for a balance of items requiring basic arithmetic, algebra, and geometry skills. In addition, there are specified numbers of items that are applied mathematics problems and problems that are principally formula-driven. Within each of the three basic skills, there are upper bounds to the numbers of items that tap specific skills. For example, no more than a certain percentage of items can include triangles or percentages. There are also lower and

Table 8.1 Overview of GMAT content, allotted time, and scoring

	Number of Questions	Allotted Time	Scoring
Analytical Writing Assessment		60 minutes	0 – 6
Analysis of an issue	1	30 minutes	(half-point
Analysis of an argument	1	30 minutes	increments)
Quantitative	37	75 minutes	0 – 60
Problem solving			(1-point
Data sufficiency			increments)
Verbal	41	75 minutes	0 – 60
Verbal			(1-point
Sentence correction			increments)
Critical reasoning			

If a real estate agent received a commission of 6 percent of the selling price of a certain house, what was the selling price of the house?

(1) The selling price minus the real estate agent's commission was $84,600.

(2) The selling price was 250 percent of the original purchase price of $36,000.

(A) Statement (1) ALONE is sufficient, but statement (2) alone is not sufficient.

(B) Statement (2) ALONE is sufficient, but statement (1) alone is not sufficient.

(C) BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.

(D) EACH statement ALONE is sufficient.

(E) Statements (1) and (2) TOGETHER are NOT sufficient.

Fig. 8.1 Sample data sufficiency problem

upper bounds regarding gender content and a correct answer location. In total, the GMAT Quantitative exam has 27 constraints; the Verbal has many more. The problem in Figure 8.1 can be classified as a data sufficiency, algebra, percentage content, applied answer "D" problem. It does not count toward the gender limits.

8.2 Becoming a Computerized Adaptive Test

The GMAT first became an adaptive examination in October 1997, five years after the idea was first presented to GMAC management. The principal issue for the GMAC at the time was access. The paper-and-pencil GMAT examination was offered only four times each year. Test-taking volume was growing and prospective test takers were having an increasingly difficult time obtaining a seat, especially in locations outside the U.S. The second issue was that the more selective schools were having a harder time discriminating among the large number of test takers at the upper end of the score scale.

The first presentation to the GMAC Board of Directors was made in 1992 by Ernest Anastasio, then a vice president at Educational Testing Service. At the time, ETS provided comprehensive test development, administration, and scoring and reporting services for the GMAT, and ETS was interested in moving several of their clients (including GRE and TOEFL) to adaptive testing. Presumably a larger client base would mean more tests being administered and would make computer-based delivery economically feasible. Anastasio talked about increased access, opportunities for new item types, and the possibility of adding new assessments to the GMAC portfolio at some point in the future.

The first formal presentation to the GMAC Board in 1993 addressed the potential benefit of transitioning the GMAT to adaptive testing. CAT promised to address both of GMAC's principal issues—better access would be provided by more frequent testing opportunities, worldwide, and converting the test to adaptive format offered the promise of better discrimination at the upper end of the score scale. ETS told GMAC that the transition to an adaptive format would be principally a change in test delivery and that additional infrastructure costs would be incurred for changes in registration systems, item banking, score reporting, and the like. Because GMAC already had a fairly extensive item bank, ETS expected there would be no need for an appreciable increase in item production and the expected bill for conversion would be between $4 and $7 million.

The GMAC Board approved the move in 1995 and proceeded to communicate the plans to its membership and other GMAT score users. Because GMAC had no resident psychometric expertise (the entire staff was only 10 people) and the GMAC Board was comprised for the most part of deans and admissions directors, none of whom had measurement expertise, Barbara Plake of the Buros Institute was brought in as an independent third party to advise GMAC on the merits of the ETS plan and to review the migration of the test from paper-and-pencil (P&P) to adaptive format. One of Plake's major contributions was the insistence on a study to compare the results of CAT administration with P&P testing on the paper-and-pencil scales that GMAC knew so well.

In mid-1996, well after GMAC had told its clients of all the benefits and the need for the pending changes, ETS came to understand that it had substantially underestimated the need for additional, new item development, and communicated that to the GMAC Board. GMAC was already committed and reaffirmed its desire to implement GMAT CAT. The risk to GMAC was enormous. The final bill for the CAT transition, new item development and infrastructure changes came in at nearly $11.7 million—almost the entire cash reserves of GMAC. Improved access was needed and the CAT transition was viewed as essential to attaining that objective.

In October 1996, 12 months before launch, the comparability study was conducted. Details of the comparability study and a subsequent equating study are documented in Bridgeman, Wightman & Anderson (n.d.). The intent was a balanced design with examinees taking both P&P and CAT, with randomly assigned order. Invitations were issued to test registrants to participate in the first study. They were offered free examinations with only the highest score being reported. Of the 10,196 invitees, 4,300 examinees accepted, 3,606 satisfactorily completed the CAT version, and 2,545 took both versions. The members of the P&P-first group in the usable sample were notably different than the members of the CAT-first group on several important measurable variables, and the groups as a whole were different than all other people historically taking the P&P version.

The study concluded that P&P results were not comparable to CAT results and that sizable equating adjustments would be required. *"Between scores of 290 and 600, the equated scores (from the first equating study) were within plus or minus 10 points of the original scores. However, adjustments of 20 to 30 points were needed at the lower end of the scale and 20 to 40 points at the high end of the scale"* (Bridgeman, Wightman & Anderson, n.d.). In other words, the results were not comparable at the tails and differential adjustments were required.

Part of the issue was that the CAT test was unexpectedly speeded. Some 18% failed to answer the last two Quantitative items; many additional examinees clearly applied guessing strategies without reading the final questions. In an attempt to remedy this situation, ETS decided to add 5 minutes to the CAT Quantitative and to shorten the test by 2 operational items.

A second study to equate results was conducted in April 1997, a scant six months before launch. Because of the time constraint, a P&P-first-only design was used. Three thousand registrants were invited to participate, but only 773 who took the P&P version also took the CAT version. Apparently, many examinees were well satisfied with their P&P scores, and they did not return for the CAT administration.

Recognizing that the design and sample size of the April administration were not adequate for a defensible equating study, the final equating was based on a combination of data from the October and April data collections. While the details of how these data were combined are not clear, the resulting GMAT scaled scores would no longer be linearly related to theta.

There were numerous design and implementation issues. The comparability study was conducted in October 1996—a month with historically documented significantly higher mean GMAT scores. The second equating study was conducted in April 1997—a month with historically lower GMAT scores. Most important, the April administration used a P&P-first-only design. Participation rate was low, and it is highly unlikely that the samples were representative of the GMAT test-taking population. Most of these issues had been pointed out by Plake in her critique of the design document.

It is worth noting that the comparability study conducted in 2006, when GMAC transitioned the test contractor, used propensity score analysis (Rubin, 1997; Rudner & Peyton, 2006). Individuals taking the GMAT under the new contractor were matched to individuals having taken the test under the prior contractor. This rigorous methodology overcomes the issues encountered in the 1996 comparability study.

The impact of the 1997 equating study was as follows: (a) CAT-based scaled scores were not truly equivalent to the familiar P&P scores even though the scores were forced to that scale; (b) mean quantitative scores climbed dramatically once CAT was introduced; and (c) the new test failed to meet the goal of better differentiation in the upper end of the score scale.

Nevertheless, despite these outcomes, admissions officers and GMAC were quite pleased with the results. To them, there was no discernible difference in scores from P&P and CAT administrations and access was, in fact, greatly improved. Nine years later, focus groups were held to discuss the desirability of normalizing and extending the scale on the upper end. The overwhelming response was that this would be an unnecessarily disruptive refinement that would have very little practical advantage. Scores that are in the top 20th percentile are treated equally by almost all admissions representatives using the GMAT.

8.3 Implementation Issues

The following sections discuss several implementation issues that have arisen and the approach taken by GMAC to address those issues.

8.3.1 Meeting Content Specifications

Because the content specifications define the test and the construct being measured (Sireci, 1998), meeting the content specifications is of critical importance. The issue, then, is how to draw items from a larger pool and meet the specifications given a large number of desired specifications and the limited number of operational test item slots.

Kingsbury and Zara (1989) outline a constrained adaptive testing (C-CAT) procedure that provides content balancing by selecting the item within the content area that has the largest discrepancy and that provides the most information at the examinee's current achievement level estimate. A major disadvantage of this approach is that the item groups must be mutually exclusive. In this case, as the number of item features of interest increases, the resulting number of items per partition decreases.

Wainer and Kiely's (1987) testlet approach can provide excellent content balancing as each testlet can cover specific parts of the desired test specifications. Wainer, Kaplan, and Lewis (1992) have shown that when the size of the testlets is small, the gain to be realized in making the testlets themselves adaptive is modest.

Swanson and Stocking (1993) and Stocking and Swanson (1993) describe a weighted deviations model (WDM), which selects the subsequent item for which a weighted sum of deviations from the projected test attributes is minimized. WDM seeks to satisfy all the conditions by treating some as desired properties and moving them to the objective function (Stocking and Swanson, 1993, p. 280). However, WDM only assures that the test specifications will be meet on the margin. That is, on average, a given group of test takers will meet the specifications, however certain individual test takers in the group may not meet the test specifications. To GMAC, this is not acceptable. All test takers should receive the same content mix.

van der Linden (this volume, chap. 2) and van der Linden and Reese (1998) describe the shadow-test approach (STA) in which the items are not selected directly from the item pool but from a sequence of full tests (i.e., shadow tests) assembled in real time. With STA, large sets of content specifications can be met along with other desired constraints such as item cloning, item-exposure control, and control of speededness. The relative importance of each constraint can be specified and tradeoffs of objectives can be evaluated.

The approach taken for the GMAT is to separate the specifications for the individual and the specifications for the item pool. At the broadest level, GMAT Quantitative items can be classified using three categories: skill area (data sufficiency or problem solving), content base (algebra, arithmetic skills, or geometry), and application (applied or formula-based). GMAC specifies that each individual must receive a certain number of items in each of the seven categories just mentioned. We do not specify the interactions, e.g., the number of items on data sufficiency, algebra, or applied. The pools must contain the desired balance in terms of answer location, gender, and within-subject content. These specifications are implemented by having prespecified content for each item position, with varying order. This way we can assure that the critical content specifications are always met, and permit the less critical specifications, e.g., answer location, to vary slightly.

Test specifications for paper tests typically call for minimum reliabilities. For CAT we call for a minimum marginal reliability and specify a target for the conditional error curve. Rather than have the standard errors follow the U shape typical of observed data, the GMAT targets are completely flat across the center of the achievement scale. Rather than using conditional mean standard errors, we use conditional median standard errors. This way, once the target standard error for an examinee is met, the algorithm is free to select from all the items that maintain the target standard error rather than items that maximize information. The use of median rather than mean target values provides an opportunity to broaden the use of items within the pool.

8.3.2 Item Exposure, Item Use, and the CAT Algorithm

Test items are costly to develop, often in the range of US$1,500–2,500 per item. Given that expense, the test publisher is interested in assuring that all items are used and that no items are overused. An unconstrained greedy algorithm can cause a severe problem in that respect. Wainer (2000) describes an item pool consisting of 822 items. Upon repeat administrations of an exam utilizing this item pool with an information-greedy algorithm, 14% of the item pool, or 113 items, accounted for 50% of the items administered to examinees. If one considers a hypothetical situation in which the average ability of examinees is very high and the standard deviation of test scores is very low, an information-greedy algorithm would reduce the effective size of the item pool even further.

Figure 8.2 shows the observed item-exposure distribution for items in a past operational pool of GMAT items using a constrained algorithm based on maximum information. Approximately 28% of the items were never used, and 18% of the items in the pool were seen by more than 15% of the examinees. Because the sum of the exposure rates of the items across the pool is always equal to the test length (van der Linden & Veldkamp, 2007), for this pool and test length, the ideal exposure distribution, ignoring content constraints, would have been each of the 100% of the items being seen by slightly less than 3% of the examinees.

Without adequate exposure control, item selection based on maximum information will force some items to be underutilized and others to be overutilized. An example is shown in Figure 8.3 which presents item response functions and the corresponding information functions for three items. The response functions are nearly identical. Each of the three items would perform comparably if administered

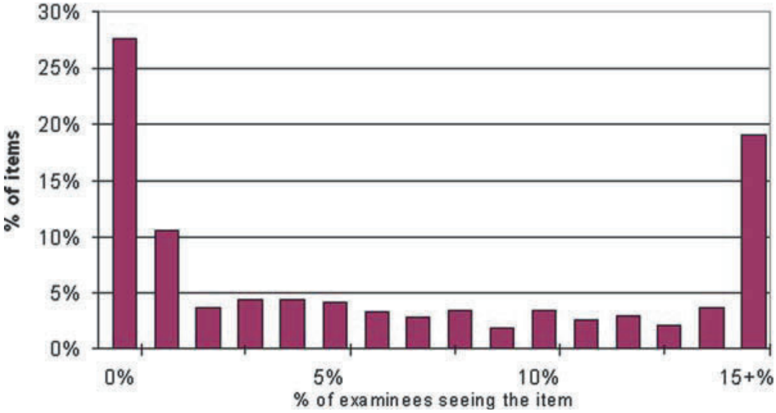


Fig. 8.2 Distribution of item exposure under a maximum information algorithm

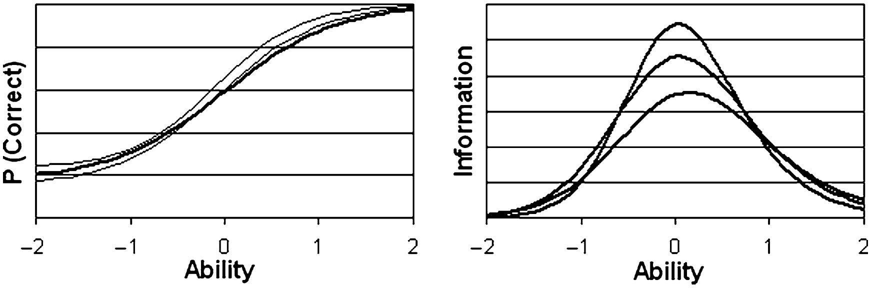


Fig. 8.3 Three item response functions and their associated information functions

to an examinee with near-average ability. However, if one is selecting items based just on maximum information, one item would supersede the others nearly every time. The end result is an exposure distribution similar to that shown in Figure 8.2.

An additional problem is that in the beginning of a testing session items are being selected that are targeted to the current, poor theta estimate. Because the theta estimate is poor, the difficulty of the selected items is often far from the examinee's true ability. An algorithm that selects the most informative, discriminating items at this stage is wasteful. The best items are being exposed while contributing little to the examinee's final ability estimate.

Overriding the item-selection process to limit exposure will better assure the availability of item level information and enhance test security. However, overriding also degrades the quality of the adaptive test. Thus, it is likely a longer test would be needed. However, if (a) a pool is made of sufficiently high-quality items, (b) the test is of sufficient length, and (c) the goal is to meet a target standard error rather than to minimize each examinee's standard error, then degradation is not an issue.

Sympson and Hetter (1985) developed an approach that controls item exposure using a probability model. The approach seeks to assure that the probability the item is administered, $P(A)$ is less than some value r—the expected, but not observed, maximum rate of item usage. If $P(S)$ denotes the probability an item is selected as optimal, and $P(A|S)$ denotes the probability the item is administered given that it was selected as optimal, then $P(A) = P(A|S) * P(S)$. The values for $P(A|S)$, the exposure control parameters for each item, can be determined though simulation studies. Sympson–Hetter control addresses the overexposure problem and only slightly helps the underexposure problem.

Another approach to control exposure is to randomly select the item to be administered from a small group of best-fitting items. Various randomization rules can be applied. For example, McBride and Martin (1983) suggest randomly selecting the first item from the five best-fitting items, the second item from the four best-fitting items, the third from a group of three, and the fourth from a group of two. The fifth and subsequent items would be selected optimally. After the initial items, the examinees would be sufficiently differentiated and would optimally receive different items. Kingsbury and Zara (1989, p. 369) report adding an option to Zara's CAT software to randomly select from two to 10 of the best items. The randomization rule developed by ACT and now used with the GMAT, which is a mixture of these two approaches, yields item exposures that are closely distributed around the ideal.

At GMAC, exposure risk is gauged by examining the probability that examinees with similar theta values will be administered items in common. Given a fixed number of examinees, as pool size gets larger, all items are exposed less and the conditional exposure rates will decrease. Another approach to reduce conditional exposure rates used for the GMAT has been to randomly select from multiple pools in the field at any one time. We have also staggered our pool rotation, have rotated pools frequently, and have used different pools in different regions. The closer the algorithm and pool are to achieving the ideal of administering a totally independent set of items, the less likely a given examinee can benefit from compromised items.

8.3.3 Item Pool Characteristics

In preparation for converting from paper and pencil to CAT in 1997, GMAC built up its item bank to include more than 9,000 quality items, and there has been a steady increase in the size of the available bank since that time. The challenge is to partition the item bank into pools that meet the specifications and to allow examinees to receive items that yield satisfactory standard errors.

The ideal item pool for an adaptive test would be one with a large number of highly discriminating items covering each content requirement at each ability level. The information functions for these items would appear as a series of peaked distributions across all values of theta. Another way to look at an item bank is to look at the sum of the item information functions. This Test Information Function shows the maximum amount of information the item bank can provide at each level of theta.

One approach to pool formation is to put all the available items from the item bank into the pool. Certainly this would yield a pool with the best available items. However, there may be dire consequences should that massive pool become compromised. As a test sponsor, we would like to see the smallest possible pools that permit the content specifications to be met. Weiss (1985) points out that satisfactory implementations of CAT have been obtained with an item pool of 100 high-quality, well-distributed items. He also notes that properly constructed item pools with 150–200 items are preferred. If one is going to incorporate a realistic set of constraints (e.g., random selection from among the most informative items to minimize item exposure, or selection from within subskills to provide content balance) or administer a very high-stakes examination, then a much larger pool would be needed. Given content constraints and standard error targets, pools of 600–1,000 items for tests such as the GMAT are not unrealistic.

Weiss was correct in that an item pool of 100 items can be used to produce a highly satisfactory CAT. In developing an online, 24-item, diagnostic adaptive version of the GMAT, we ran simulations to evaluate the needed pool size given our desired content balance, the quality of the item bank in terms of mean item discrimination parameter (a_i in the 3PL model), and our desire to permit examinees to use the same pool for up to three administrations with the constraint that an examinee would never see the same item twice. The criterion was the marginal reliabilities for examinees in the middle 90% of the distribution as a function of the number of times they took the test. The results of the simulation are shown in Figure 8.4 and Table 8.2. With an item bank having a mean a_i parameter of 1.0, one can develop a quality CAT, i.e., one having a marginal reliability of 0.90 or greater with as few as 48 items, provided one is only going to administer the test one time.

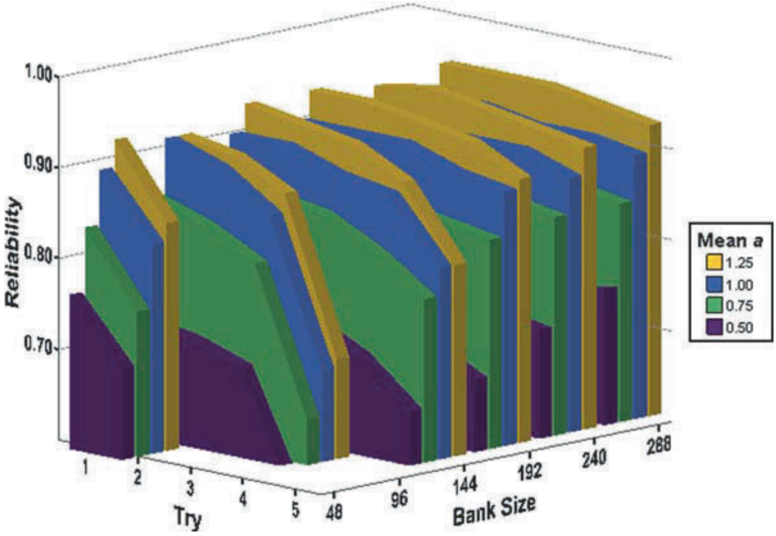


Fig. 8.4 Reliability as a function of testing attempt, item bank quality, and bank size

Table 8.2 Reliability as a function of test attempt, item bank quality, and bank size

		Number of Attempts				
Bank Size	Mean *a*	1	2	3	4	5
48	0.50	0.77	0.70			
48	0.75	0.84	0.76			
48	1.00	0.90	0.83			
48	1.25	0.93	0.85			
96	0.50	0.73	0.72	0.70	0.57	
96	0.75	0.86	0.84	0.81	0.65	
96	1.00	0.92	0.90	0.86	0.70	
96	1.25	0.92	0.91	0.88	0.71	
144	0.50	0.75	0.74	0.74	0.71	0.66
144	0.75	0.86	0.86	0.85	0.82	0.78
144	1.00	0.91	0.91	0.89	0.88	0.81
144	1.25	0.94	0.93	0.92	0.89	0.81
192	0.50	0.72	0.70	0.70	0.71	0.68
192	0.75	0.86	0.85	0.85	0.84	0.83
192	1.00	0.91	0.91	0.91	0.89	0.88
192	1.25	0.94	0.94	0.93	0.92	0.89
240	0.50	0.71	0.74	0.72	0.73	0.72
240	0.75	0.86	0.85	0.84	0.85	0.84
240	1.00	0.91	0.90	0.90	0.90	0.88
240	1.25	0.93	0.94	0.93	0.92	0.91
288	0.50	0.73	0.74	0.75	0.74	0.75
288	0.75	0.85	0.85	0.84	0.85	0.84
288	1.00	0.91	0.91	0.90	0.90	0.89
288	1.25	0.94	0.94	0.94	0.93	0.92

While these results are appropriate when one can hand-pick items and have few constraints, forming relatively small, effective pools for the actual GMAT administration is a more difficult task. The content specifications must be met, sufficient numbers of quality items at each score point are needed, and one wants to minimize the number of items that have been used extensively in the past. Because some items are clones of others or have similar content, pools are typically formed so that they do not contain any item enemies. In addition, because many examinees retake a test—Rudner (2005) reports that 61% of the GMAT examinees that retake the GMAT do so within three months—one would not want to use items that have appeared in recent pools. The GMAC pool formation rules require a certain rest period before items are considered for reuse.

Given the constraints, software tools have been developed to help form initial GMAT pools. Simulation studies taking into account the expected score distributions are used to evaluate the pools. These simulations provide a host of information including expected exposure rates and expected conditional standard errors. Gaps in the conditional errors are then corrected by manually replacing or adding items. The process is iterated until targets are met. The GMAT test development contractor, ACT, runs these simulations for each pool. GMAC reviews the results months in advance. Sporadically GMAC compared the simulations against actual data. To date, the simulations have mimicked reality exceptionally well.

Given the work needed to formulate a pool, it is tempting to reuse pools or large parts of previous pools. However, brain dumps, illicit test preparation sites, and other groups making operational items available make reusing portions of pools a risky proposition. GMAC pool formation rules also include a specification of the maximum percent of items that can overlap with any previous pool. GMAC now devotes more than two full-time-equivalent staff members to monitor the Internet, document infringements, and bring civil and criminal action against individuals, who infringe on GMAC's copyrighted material.

8.3.4 Item Bias

The common approach to investigating differential item functioning is to examine the item parameters resulting from group-specific calibrations. GMAC does this on a routine basis as part of the item pretest evaluation.

Pretest data, however, are often limited in terms of the number of examinees in subgroups. Accordingly, we also investigate bias using operational items. For example, we were interested in whether GMAT items show any bias when used in Europe (see Talento-Miller, 2008). Guo, Rudner, Owens, and Talento-Miller (2006) present a method suitable for operational adaptive tests that redefines item bias using an adverse impact perspective rather than a group difference perspective. Item bias was defined in that paper as the difference between the item response function (IRF) from a subgroup and the IRF defined by the operational item parameters. More specifically, an item is biased if examinees in a subgroup with the same ability do not have the same conditional probability of correct answers as the population (total group used in calibrating the operational item parameters).

8.3.5 Item Parameter Shift

The final practical consideration to be discussed in this chapter is shifting parameter estimates. Once an item is calibrated and found to be of sufficient quality, there is little reason, other than being compromised, to retire the item as long as it continues to function as it did when originally calibrated. Thus, there are the very real questions as to whether the individual item parameters have shifted and whether that shift makes a difference. Guo and Wang (2005) present a methodology used for the GMAT based on a set of commonly administered items. Other methodologies used by GMAC have included examining empirical IRFs from alternate administrations, calibrating operational items that have been placed in pretest slots, and calibrating adaptively administered items given examinee thetas and prior c parameter values. We have had the most success simultaneously recalibrating an entire pool's worth of item response data including the nonoperational items. Given the relatively large number of examinees seeing collections of nonoperational items, the resultant parameter estimates proved to be quite stable. Very few items had item parameters beyond the standard error of calibration. Parameters were updated for those that did.

8.4 Conclusion

A key component to any successful CAT program is the careful design and implementation of a system that provides quality information regarding examinee ability while minimizing item exposure and security risks. This chapter presented some of the practical issues considered by the Graduate Management Admission Council in the design and evaluation of the Graduate Management Admission Test.

Some of the key considerations are

1. Content specifications. Content specifications should assure similarity of content for every examinee, while balancing a wide range of considerations. GMAT Content Specifications identify the items to be received by every examinee, requirements for the pools, and specifications for the conditional errors.
2. Item exposure, item use, and the CAT algorithm. Most of the work on item exposure has addressed the issue of overexposure. We do not want the same items to be administered to large percentages of examinees. At the same time, the sponsor is interested in maximal use of the investment; that is, the test sponsor would like every quality item to be used. Quality test items are expensive to develop. Exposure and use issues associated with an algorithm based only on maximum information were identified.
3. Item-pool characteristics. While there are psychometric advantages to placing all available items in a test pool, there are practical issues to consider as well, not the least of which are the consequences of a security breach. Small pools are attractive from the test sponsor's perspective, but small pools raise issues with regard to conditional exposure rates. A practical balance and an approach to evaluating pools are discussed.
4. Item bias. While traditional approaches to investigating item bias are employed for pretest items, operational pools provide opportunities for investigations that are not possible for pretest items. This chapter presents an alternative: a practical viewpoint of differential item functioning focused on adverse impact for adaptively administered items.
5. Parameter shift. Items administered adaptively can have an extremely long shelf life. Approaches employed to investigate the consistency of GMAT item parameters over time are presented.

The papers offered by Georgiadou, Triantafillou, and Economides (2006) and Green, Bock, Humphreys, Linn, and Reckase (1984) provide excellent guidelines for evaluating adaptive tests. The issues presented here supplement these guidelines by examining practical concerns of test sponsors.

References

Bridgeman, B., Wightman, L. & Anderson D. (n.d.). *GMAT comparability study* [Internal Administrative Report]. McLean, VA: GMAC.

Georgiadou, E., Triantafillou, E. & Economides, A. A. (2006). Evaluation parameters for computer adaptive testing. *British Journal of Educational Technology*, 37, 261–278.

Green, B., Bock, R. D., Humphreys, L., Linn, R. & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347–360.

Guo, F., Rudner, L., Owens, K. & Talento-Miller, E. (2006, July). *Differential impact as an item bias indicator in CAT*. Paper presented at the International Testing Commission 5th International Conference on Psychological and Educational Test Adaptation across Language and Cultures, Brussels, Belgium.

Guo, F. & Wang, L. (2005). *Evaluating scale stability of a computer adaptive testing system* [Research Report RR 05-12]. McLean, VA: GMAC.

Kingsbury, G. & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2, 359–375.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223–236). New York: Academic Press.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical considerations in computer-based testing*. New York: Springer-Verlag.

Plake, B. P. (1996). *A review of the comparability study design* [Internal Administrative Report]. McLean, VA: GMAC.

Rosenbaum P. R. & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39, 33–38.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.

Rudner, L. M. (2005). *Examinees retaking the Graduate Management Admission Test* [Research Report RR-05-01]. McLean, VA: GMAC.

Rudner, L. M. & Peyton, J. (2006). Consider propensity scores to compare treatments. *Practical Assessment Research & Evaluation*, 11. (Available online: http://pareonline.net/getvn.asp?v=11&n=9)

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.

Sireci, S. & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement*, 66, 305–317.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17, 277–292.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17, 151–166.

Sympson, J. B. & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. In *Proceedings of the 27th Annual Meeting of the Military Testing Association*. San Diego: Navy Personnel Research and Development Center.

Talento-Miller, E. (2008). Generalizability of GMAT validity to programs outside the U.S. *International Journal of Testing,* 8, 127–142.

Talento-Miller, E. & Rudner, L. (2005). *GMAT validity study summary report for 1997 to 2004* [Research Report RR-05-06]. McLean, VA: GMAC.

Talento-Miller, E. & Rudner, L. (2008). The validity of Graduate Management Admission Test scores: A summary of studies conducted from 1997 to 2004. *Educational and Psychological Measurement*, 68, 129–138.

van der Linden, W. J. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42, 283–302.

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22, 259–270.

van der Linden, W. J. & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics*, 32, 398–418.

Wainer, H. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Erlbaum.

Wainer, H., Kaplan, B. & Lewis, C. (1992). A comparison of the performance of simulated hierarchical and linear testlets. *Journal of Educational Measurement*, 27, 1–14.

Wainer, H. & Kiely, G. (1987). Item clusters and computerized adaptive testing: the case for testlets. *Journal of Educational Measurement*, 24, 189–205.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789.

Wise, S. L. & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, 21, 135–155.

Chapter 9
Designing and Implementing a Multistage Adaptive Test: The Uniform CPA Exam

Gerald J. Melican, Krista Breithaupt, and Yanwei Zhang

9.1 Introduction

The Uniform CPA Exam (CPA Exam) was first administered in 1917 as a requirement in the licensing of certified public accountants. It included a series of accounting problems in auditing, accounting, and commercial law that were graded by members of the American Institute of Certified Public Accountants (AICPA). Since 1917, the CPA Exam evolved to include four independently scored sections that included not only accounting problems, but essays, multiple-choice questions, and extended multiple-choice formats. Until 2004, the CPA Exam was presented twice each year, in a paper-based format. At its peak in the early 1990s, over 300,000 individual CPA Exam papers were scored each year.

Beginning in the mid-1990s, consideration was given to transforming the CPA Exam from a paper-based test to one that could be administered by computer (Board of Examiners, 1995). The Board of Examiners explored computer-based formats to allow testing requisite CPA skills via high-fidelity technological tools (such as the ability to research accounting standards online), and other measures of skills that are not feasible in a paper-based examination. Other desirable advantages of computer-based testing were considered, such as improving exam security by scrambling content on multiple forms, imposing uniform secure proctoring in a greater number of test centers, test adaptation and reduced testing times, and allowing more flexible test scheduling for candidates.

After considerable study of the feasibility of implementing a computer-based CPA Exam (Professional Examination Service, 1999) and discussions of these findings with key stakeholders, the AICPA and the National Association of State

G.J. Melican (✉)
The College Board, 45 Columbus Avenue, New York, NY 10023–6992, USA

K. Breithaupt
American Institute of Certified Public Accountants, 1230 Corporate Parkway Avenue,
Ewing, NJ 08628–3018, USA

Y. Zhang
American Institute of Certified Public Accountants, 1230 Corporate Parkway Avenue,
Ewing, NJ 08628–3018, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_9,

Boards of Accountancy (NASBA) formed a joint Computerization Implementation Committee to plan and execute the transition to computer-based testing in accordance with state boards of accountancy, the AICPA, and CPA candidates (AICPA, 1997; AICPA & NASBA, 1998). After almost nine years of study, discussion, planning, and development, national administration of the computer-based CPA Exam was offered on April 5, 2004.

As of 2004, the Uniform CPA Exam has been a 14-hour test comprising four sections: Auditing and Attestation (AUD); Financial Accounting and Reporting (FAR); Regulation (REG); and Business Environment and Concepts (BEC). Candidates must pass all four sections within 18 months of passing the first section as part of satisfying the requirements for licensure as a CPA in each of the 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. Each section is administered separately and has a unique proficiency requirement for passing. If more than 18 months elapse, success on a previous section will be discounted and that exam must be retaken until all sections have been passed within an 18-month rolling window.

The purpose of the CPA Exam is to provide assurance to boards of accountancy that CPA candidates have mastered required knowledge and skills needed by entry-level CPAs to protect the public interest. While each board of accountancy has additional requirements for licensure, including specific education and experience requirements, the CPA Exam is one uniform standard that all 54 boards have adopted.

The Uniform CPA Exam is a high-stakes, highly visible examination with a long and valued history. Making changes of any magnitude, much less changes of the scale that were undertaken, requires a rigorous, comprehensive plan and the input of thousands of participants (CPAs, university faculty, psychometricians, and large numbers of students). The process of computerization was transparent to these stakeholders, and key decisions were vetted with representatives from the profession and the regulators. The decision to administer the CPA exam using a computerized adaptive multistage testing (MST) model for the multiple-choice portions of the tests is an example of this collaborative process.

The initial part of this chapter reviews the decision-making points that led to selection of the administration format for the MST. The rest of the chapter is organized around the development and implementation process and the major activities required to implement the MST exam program. Pretesting, calibrations of items, automatic test assembly, standard setting, score reporting, and quality control activities are described in later sections as highlights of the process that may have broad application in high-stakes testing programs considering an adaptive psychometric model.

9.2 Decision Making

9.2.1 Content

The accounting profession is constantly evolving in a modern business environment. The rise in multinational trade and the rapid changes in technologies require frequent reevaluation of the knowledge and skills required of entry-level CPAs. To ensure that the computer-based CPA Exam covers appropriate subject matter at the level needed by entry-level CPAs, a series of studies and evaluations was conducted. The keystone of this process was a 2001 national practice analysis study of accounting knowledge and skills involving thousands of CPAs (American Institutes for Research, 2001). The results of the study were then evaluated by the AICPA Board of Examiners' Content Committee for appropriateness to entry-level practice, impact on public protection, and feasibility for assessment on the computer-based exam.

This committee was comprised of nominees approved by the AICPA Board of Directors and the Board of Examiners. The membership on the committee was balanced by state representation, by years of experience, and by areas of expertise in regulation, accounting, auditing, and business. The Content Committee reviewed the results from the practice analysis and determined the number of test sections and lengths required to assess knowledge and skill for entry-level proficiency. Their recommendation for four test sections with independent passing requirements was based on the belief that strong knowledge in one content area could not compensate for lack of knowledge in another important content area. The resulting structure was a multiple-hurdle examination program, where candidates must demonstrate proficiency in all four content areas within the 18-month time frame.

Several of the new test sections included content areas that were similar to the previous four paper-based tests. However, one new test section was designed specifically to incorporate measurement of the market forces and technologies important to modern accounting practice. This section was titled Business Environment and Concepts (BEC). This expansion of required content reflected the very public and substantive changes in the accounting profession that had occurred over the preceding decade.

The Content Committee solicited comment on the content blueprint, length, and four-section structure proposed for the computer-based test through a series of invitations that were shared with all major stakeholders. State Boards, NASBA, focus groups of candidates, and review course providers were invited to review, discuss, and comment on the draft test blueprint. Their input was documented carefully and incorporated in the process of developing the test content and documentation. The Content Committee offered its recommendations, along with documentation describing the process, issues, and alternatives, to the Board of Examiners for approval in 2002.

This discovery and evaluation process for content specification resulted in an exposure draft describing the outline or blueprint for the computer-based CPA Exam. The Board of Examiners adopted this recommendation after key stakeholders were given the opportunity to comment. The content specifications for BEC, much of which represented subject matter never before tested on the CPA Exam, was specified in more detail to aid educators and candidates in preparing for this new material.

Information to candidates, review course providers, and state boards was designed to support and define each test outline well in advance of administration of the new exam. When test content or administration formats are revised, a long lead time is required to prepare all the educational materials and transitioning policies for candidates who will be or who are already in the process of certification. That effort is already underway for future revisions to the CPA Exam content and format, expected for implementation by 2010.

9.2.2 Administration Models

Identification of a test delivery model for the revised examination was an integral part of the development of a computerized Uniform CPA Exam. A test delivery model is a framework to organize test development activities, system design for item banking, test administration, scoring, and reporting. A firm foundation for these efforts was based on thorough psychometric research to ensure the validity of score interpretations and the adherence to best practices in high-stakes testing programs. To delineate important decision-making points, this section will highlight key aspects of that psychometric research program, including incorporation of adaptation for multiple-choice items and innovative item types for performance assessment, termed *simulations*.

In 2004, the CPA Exam was modified to include a new testing format made possible by the computerized delivery system. This new format incorporated complex performances to assess skills using simulations of tasks natural to CPA practice (simulations) and also real-time adaptation to individual candidate ability (adaptive testing). During the design and research phases a wide range of potential administration models was initially considered, including pre-constructed fixed-length, linear-on-the-fly, mastery, subtest-based adaptive, and item-level adaptive testing.

Initial research into the relative benefits of administration models favored three alternatives. These were identified as the most appropriate in consideration of use of test scores for the pass-fail decision required for credentialing, efficient use of testing time for score precision, capability to produce equivalent and valid test forms over time, and the characteristics of the multiple-choice questions in the CPA item bank. The three administration models that were considered best were linear-on-the-fly testing (LOFT), item-level computer adaptive testing (CAT), and multistage adaptive testing (MST). These three models are described below.

Linear-on-the-Fly Testing (LOFT)

A linear-on-the-fly test is assembled automatically from a subpool of test questions in real time during test administration or just prior to testing for individual candidates. The automated rules for selecting questions are usually derived from the content requirements (knowledge and skills) and some basic statistical requirements for item discrimination and difficulty. No information about candidate ability is used to optimize the precision of total scores, and there is no adaptation.

This method is sometimes preferred over other computerized delivery methods because test assembly and delivery software is available from existing vendors.

Also, scrambled selection of items provides less opportunity for collusion or cheating during the test. This model is similar to paper-based testing in some respects. For example, during the test, candidates may return to questions and change their answers. Unlike paper-based testing, it is not possible for experts to preview real-time assembled test forms for content validity.

Item-Level Adaptive Testing (CAT)

The item-level adaptive testing model, usually referred to as CAT, is the model most familiar to many. In fact, CAT is often the expected model when computerized administration is mentioned. With CAT, each test is automatically assembled during the testing session using an item-level adaptive routing system. In addition to the content and other item properties guiding test assembly, the choice of each successive question presented to the test takers is based upon their ability estimate calculated after each question is answered.

Item-level adaptive testing is attractive because it allows increased precision, given shorter testing times, when item difficulty is targeted to the ability level of candidates. Hence, a more precise total test score is possible using fewer test questions. However, unless more sophisticated forms of adaptive testing are chosen, such as the shadow-test approach (van der Linden, this volume, chap. 2), adaptive testing may raise concerns about the potential unfairness of the testing experience to candidates who might get a generally more difficult or less difficult test (Rotou et al., 2007): Candidates at different ability levels might use more or less time to answer the questions selected for them as a function of the difficulty of the test question and their ability level, leading to differential speededness for individual test takers. Consequently, they may have different overall experiences with the test.

CAT also depends on implementing in real time a large number of constraints on the selection of the items from the item bank. Also, a method of item-exposure control has to be used to prevent some questions from being overexposed because they have desirable statistical properties and also fulfill key content constraints. Finally, as with administrations based on multiple LOFT forms, it is usually not possible for content committees to review test forms in advance.

Multistage Adaptive Testing

Multistage adaptive test (MST) designs are structured adaptive tests that employ preassembled subtests as the basic units of test administration (Adema, 1990; Luecht, Nungester & Hadadi, 1996; Luecht & Nungester, 1998, 2000; Luecht, 2000; Glas & Vos, this volume, chap. 21). In contrast to item-level CAT designs, which result in different test forms for each test taker, MST designs use a modularized configuration of preconstructed subtests and embedded score-routing schemes to prepackage validated test forms.

Zenisky, Hambleton and Luecht (this volume, chap. 18) describe the MST concept in detail, and only a short description is offered here. Candidates sitting for the exam initially respond to a set of multiple-choice questions (MCQs) comprising a short routing test, or subtest. The routing subtest is assembled from items of medium difficulty from the item bank. The candidate will be presented new subtests following his or her responses to the first routing subtest. Review and revision of responses may be possible within any subtest, but may be restricted after completion of the subtest.

After completion of the routing test, the test taker is presented a second subtest, which is selected based on difficulty level, depending in his or her performance on the routing test. Also, the number of subsequent subtests that must be answered will depend on the desired test length, item formats, and the decision that is made based on test scores. For example, test questions that share a common exhibit might appear together in a subtest (set-based items), or the subtest may be composed of traditional MCQs selected to represent the content and skills to be measured. In any MST, the subtests are constructed and selected based on the statistical properties of items and some estimate of item difficulty.

The second subtest is chosen based on the performance of the test taker on the first subtest, and the choice of the third subtest is based on the performance of the test taker on the first two subtests. Scoring of the items for the purpose of selecting future subtests might be based on number-correct or more sophisticated scoring schemes, such as item response theory (IRT) ability estimation. In MST, the subtests are scored cumulatively to obtain provisional ability estimates that will be used to select the next subtests. These provisional scores are compared to routing scores, which determine whether the candidate is routed to a more or less difficult subsequent subtest.

MST provides some adaptation or tailoring to the candidate ability while allowing test takers to review and change answers within subtests during the test administration. An additional benefit of this mode of test administration is the possibility for review and validation of the test forms before these are administered to candidates.

The MST has some efficiency afforded by tailoring to the ability level of the examinee, and yields more precise scores, given fixed test lengths. However, this efficiency is not as great as CAT administration where adaptation occurs for each test item. However, Routou et al. (2007) compared a two-stage multistage test to a CAT of the same length, which both included only set-based items. The MST had slightly higher reliability under the one- and two-parameter models and equal reliability under the three-parameter model. Also, given a fixed item bank size, due to the possibility of building a variety of different forms that overlap at the module level, MST does allow the test developer to create a greater diversity of forms compared with LOFT.

The following section highlights some benefits and limitations of each model for the CPA licensing examination.

9.2.3 Evaluation Criteria

Given the general class of models considered, a selection decision for the CPA Exam was developed using a rationale anchored by the intended use of test scores. The intended purpose of the examination is to provide to State Boards a pass–fail decision indicating proficiency for entry into CPA practice. Passing each of the four sections of the Uniform CPA Exam within 18 months is just one of a set of requirements for licensure as a CPA in the U.S. Legal defensibility of the licensing decision rests on the validity of this pass–fail interpretation. Validity evidence for this decision must be built into the test design and administration model.

Considerations that assure test-score precision and efficient use of testing time were balanced against test development goals. Uniformly meeting content and other nonstatistical test specifications is important to assure the quality and equivalence of every test form. Operational considerations were also important in establishing test validity and acceptability to the profession, such as whether to allow candidate review of questions, how to code, maintain, and select test questions, and how to manage the exposure of test questions and to block memorable content for those candidates who retake the examination.

Each of the proposed computerized testing models involves automated test assembly (ATA) of items or questions into test forms. Because of this automated process, the formulation of exact rules to assemble the various test forms is required to ensure that scores from each computerized test are comparable and that each pass–fail decision is made accurately. The administration options, however, differ in the opportunity to modify and review the results of assembly. LOFT or CAT construction occurs during testing and depends on accurate selection of test questions from a large item bank in the test center.

The recommendation for an appropriate model for the CPA Exam was based on psychometric research and analysis of these considerations and offered to the AICPA Board of Examiners. The process toward decision-making included critical evaluation by outside psychometric consultants and a research consortium of three universities (University of Massachusetts–Amherst, University of North Carolina–Greensboro, and University of Illinois–Urbana/Champaign). These consulting teams implemented and published over 40 studies over a period of about four years on topics relevant to practical and theoretical issues for computerizing the CPA Exam. Guidance and participation in these efforts were available from a Technical Research Issues Oversight (TRIO) committee of three internationally respected psychometricians, who collaborated with the teams throughout the research and development cycle. In addition, there was a standing Psychometric Oversight Committee (POC) working under the auspices of the Board of Examiners.

The policy makers at the AICPA had been familiar with a twice-annual administration of the paper-based test, where each form was hand-crafted, underwent numerous reviews by test developers and their professional committees, and was administered on just one day. They would only be supportive of a mode of computerization that allowed expert review and quality control analogous to that used for paper administration at least until the efficacy of automated test assembly was firmly exhibited. Such review is usually not as convincing when items are administered using CAT or LOFT. Although sample forms can be generated under CAT and LOFT, as many as desired, it is possible that the sample forms may never be seen by a real candidate, whereas MST modules can be preconstructed and reviewed prior to packaging for administration.

Also, the adopted test model must allow passing scores to be set using a criterion-referenced standard-setting process, and established on a score scale easily understandable to candidates. All three models above were considered to meet this criterion equally well, although some scoring complexity is associated with the use of an IRT model (applicable to any of the administration formats).

Physical security of test items is essential to ensuring that test-score interpretations are valid. While this security can be established independently from the testing mode, there are some benefits to the MST format when security is considered. Since MST makes use of interchangeable modules that can be shuffled into various forms, it may be easier to manage lower item exposure for smaller banks of test questions. This also decreases the likelihood of adjacent test takers seeing the same test form.

Finally, stakeholders expressed a strong preference for giving candidates the capability to review and revise answers to test questions in the administration. The MST model allows review within modules, and the opportunity to restrict reviews between subtests.

9.2.4 The AICPA MST Model

The AICPA adopted a three-stage MST with one medium-difficulty subtest administered at the first stage and two subtests available for selection at the second and third stages of administration. One of the two subtests at the second stage is of medium difficulty while the other is slightly more difficult, and the same is true for the third stage. This format met all of the psychometric and practical criteria established by the BOE and the psychometricians. A full description of benefits is available in Luecht, Brumfield and Breithaupt (2006) and Melican et al. (2005).

The collection of all possible subtests for administration as an MST is referred to as a *panel*. Multiple panels can be simultaneously preconstructed, balancing quality control, security, and score precision across panels. This approach allows score precision to be built in exactly where it is needed for each panel. Test committees are able to review the content and quality of the test forms within each panel. Furthermore, trial runs could be made to ensure each panel is working properly, before activation in the live examination pool. To address security risks, panels can be randomly assigned to examinees, and the items can be randomly scrambled within subtests. These factors deter strategizing by candidates. Finally, the panels can be eliminated from selection if their content has been previously seen by a retesting candidate.

Any of the computerized delivery models evaluated offers significant gains in security when compared with paper-based linear forms. For the MST, the reuse of test questions and simulations can be controlled explicitly when designing the examinations. A pool of items forms the basis for test creation, and a very large number of different panels are created as a result. This is important when testing appointments are offered over a longer period of testing (e.g., ongoing for a two-month testing window). Assembly of tests and rotation of test items during the year are planned carefully so that only a small portion of candidates will see a given item or test form.

9.3 Implementation

9.3.1 Test Assembly

This section offers an overview of the assembly procedure that was used for the Uniform CPA Exam for the implementation of the MST model. Typical test requirements focus on content coverage described in a publicized test blueprint. The blueprint imposes a variety of other nonstatistical construction rules, such as content coverage, form lengths, and item formats. Another common rule imposes limitations on test questions that might be termed “item enemies”; these are test questions not allowed to appear on the same subtest because they cue other items or are close variations of other items on that subtest. There may also be rules for forced inclusion of items on key topics, or limits on the number of items of similar content.

It is important to carefully design the automated assembly system to ensure that high-quality subtests can be produced on a continual basis without overuse of pool content. The following paragraphs describe some design issues to consider in identification of the rules for subtest construction for the multistage testing model, and a solution used by the CPA Exam to solve the simultaneous assembly problem for MST.

Four important decisions were required to finalize the subtest assembly design. These were solved in the following order: (i) objective of test assembly; (ii) statistical targets for the subtests; (iii) number of subtests; and (iv) stringency of content constraints.

Objective of Test Assembly

The mandated goal of the CPA Exam is to protect the public by admitting only qualified CPAs to licensure. The required standard for competency is represented by the recommended passing score on the exam. Therefore, it is critical to obtain optimal measurement precision on the score scale in the area of the passing score. A secondary goal is to provide informative feedback to candidates who are not successful so they might prepare to retake the examination. Therefore, a lesser goal was defined to obtain good measurement precision and coverage of a broad range of content just below the area of the passing score. Finally, a pervasive goal for all areas of any high-stakes testing program is the security of test content to preserve the validity of scores and classification decisions. In the assembly process, this security requirement led us to engineer an assembly method that would yield reliable and valid test scores without depleting the best test content over the short term.

Statistical Targets for Subtests

Targets for subtest difficulty for an adaptive model of administration must be set with the properties of the item bank and the intended use of test scores in mind. Statistical criteria for selecting items for subtests must be in the range represented by the majority of MCQs in the item bank; otherwise, it will not be possible to assemble many (or any) subtests. In order to make good use of the existing pool of MCQs for the CPA Exam, the median difficulty of the item pool was used to adjust the preliminary difficulty targets for subtests. Since the MCQs in this licensing exam program are developed for entry-level knowledge and skill, most test questions have a difficulty level in a range close to the passing point on the examination. This means that the selection of statistical targets for the moderate (*M*) and difficult (*D*) subtest questions emphasize both difficulty around the passing score and the mean item bank difficulty.

Subtest reliability and cumulative reliability across all subtests on a route were also considered in designing assembly rules for the MST. Precision over each possible set of three MCQ subtests candidates would be administered from each panel was examined to ensure that marginal IRT reliability of total scores remained above 0.90.

Number of Subtests

Once the test length, the location and kind of targets for subtest information, and the content constraint ranges had been defined, it was possible to begin to determine the maximum number of possible subtests that should be created from the entire pool of items for an individual administration window. Several factors were considered in deciding the number of panels and subtests needed for continuous administration, including the size of the item pool, dependencies among items (e.g., item enemies), and the minimum number of items required for adequate content coverage on each content or skill per subtest.

In order to determine the feasibility of sustaining equally high-quality subtest creation across administrations, the maximum number of subtests possible was assembled from the master item pool. All of these subtests met the required content and other statistical constraints, and were of the required length. After this initial assembly, a small set of those subtests was used to actually create panels for the first administration, and the remaining subtests were returned to the item pool. In successive production cycles, new subtests were assembled using only a subset of the master item pool (including the unused subtest items and a selection of content used in prior administrations). In this way, it was possible to ensure that subtests of equivalent quality could be constructed during the short and longer terms. A full solution to the long-term inventory system is provided in Breithaupt, Ariel, and Hare (this volume, chap. 13).

Stringency of Content Constraints

The test blueprint defines what proportion of the examination will cover each content topic area. These are expressed as ranges in the assembly problem to allow some variation in coverage according to the available item pool. Competing goals exist in selecting the range of content to express in the assembly constraints. It is desirable in simultaneous assembly to make the ranges as small as possible so that subtests will be exactly parallel (e.g., each would have the same number of questions on a particular content area). However, tightly constrained solutions will have the effect of limiting the number of possible subtests. So, ranges were set with an understanding of the exact contents of the item pool and within the allowable variation in the published test blueprint.

Mixed Integer Programming (MIP) Solution

A general procedure for IRT assembly of parallel tests requires items to be selected from a calibrated item pool to fill a desired shape of the test information function (TIF). The reader is referred to Hambleton and Jones 1993) for an introduction of this topic as well as more classical approaches to test design. An extension of the general assembly method using TIFs has reformulated the problem in linear programming terms (Theunissen, 1987; van der Linden, 2005; Veldkamp, 2002) using discrete optimization to determine whether or not an item is selected. Typically, some real-valued decision variables are required as well. The resulting mixed integer programming (MIP) model may have a large number of constraints and variables. This is true of the CPA Exam, where large numbers of parallel subtests have to be assembled simultaneously to ensure equivalence in psychometric quality across forms and to reduce the risk of overuse of the highly discriminating items. For these multiple-forms assembly problems, each additional rule or requirement in the assembly model has a multiplicative effect on the total number of variables and often on the number of constraints that must be met as well. A solution that meets all constraints is often computationally demanding or even impossible. These kinds of problems used to be notoriously difficult to solve even when the models are small (most are NP-hard problems; for examples, see Nemhauser & Wolsey, 1999). However, modern solvers for MIP problems have overcome many historic problems and now easily solve real-world test assembly problems of substantial size.

A general model for the assembly process defined by van der Linden and Boekkooi-Timminga (1989) was adapted for use in the operational production of the MST subtests. The model maximizes the lower bounds of the subtest information at a specific ability point subject to the content constraints. The lower bounds represent a relative target for the TIF. Let x_{it} be the decision variable to indicate whether or not item $i = 1, \ldots, I$ in the pool is assigned to subtest $t = 1, \ldots, T$. For each subtest t, the desired shape of the TIF is specified as a series of weights for the TIF at ability values $\theta_k = k = 1, \ldots, K$. Let $w_{kt} > 0$ be the weight at θ_k for subtest t. The item pool has item types or content areas that are represented by subsets V_c, $c = 1, \ldots, C$. In addition, for each of the subtests critical content C_t has to be included. Finally, enemy sets of items in the pool are denoted as V_e, $e = 1, \ldots, E$, whereas bounds on sets are denoted by appropriate subscripts and superscripts.

The general form of the model is

$$\text{maximize } y \tag{9.1}$$

subject to

$$\sum_{i=1}^{I} I_i(\theta_k)x_{it} \geq w_{kt}y, t = 1, \ldots, T; k = 1, .., K; \tag{9.2}$$

(relative targets for subtests)

$$\sum_{i=1}^{I} x_{it} = n_t, t = 1, \ldots, T; \tag{9.3}$$

(length of subtests)

$$n_t^{(l)} \leq \sum_{i \in V_c} x_{it} \leq n_t^{(u)}, t = 1, \ldots, T; c = 1, \ldots, C; \tag{9.4}$$

(item types or content areas)

$$\sum_{i \in C_t} x_{it} \geq 1, t = 1, \ldots, T; \tag{9.5}$$

(critical content to be included in subtests)

$$\sum_{i \in V_e} x_{it} \leq n^{(e)}, t = 1, \ldots, T;\ e = 1, \ldots, E; \tag{9.6}$$

(enemy set)

$$\sum_{t=1}^{T} x_{it} \leq 1, i = 1, \ldots, I; \tag{9.7}$$

(no overlap between subtests)

$$x_{it} \in \{0, 1\}, i = 1, \ldots, I; t = 1, \ldots, T. \tag{9.8}$$

(definition of variables)

The expressions in (9.2) serve a double goal: they defines a common factor, y, in the lower bounds on the subtest information functions at the values θ_k but also introduces the weights w_{kt} for each of the subtests at these ability values that define the shape of their information function. The common factor is maximized in (9.1), and hence the functions are maximized subject to their shapes. For the CPA Exam, there were three values θ_k needed to represent the information functions for the medium-difficulty subtest and one value for the more difficult subtest. When a common value θ_k is used to express different targets, the subtest with a higher weight at this θ_k will be favored for the selection of the more discriminating items.

Perhaps the most distinctive element in this model should be emphasized, namely the fact that the MIP model maximizes the (weighted) information for all subtests in a single solution. The model therefore allows us to create the moderate and more difficult subtests for multiple panels simultaneously and ensures the best possible combination of items into subtests that can be found. Each constraint is satisfied completely for each individual subtest by the final optimal solution to the MIP. This computationally intensive process can be completed using available software in a reasonable amount of time. The subtest assembly using OPL Studio (ILOG, 2002) for any one section of the Uniform CPA Exam applied approximately 50,000 constraints and 60,000 variables. Using a common desktop computer (1.2 GHz), the solution time typically requires less than five minutes.

An example of the result from this assembly method is depicted in Figure 9.1. The curves represent overlapping TIFs for medium and difficult subtests. Each curve represents the information from a single subtest across the theta scale. The TIFs are very close in the amount of information provided at equivalent theta points. TIFs at the left (lower) ability level for the medium-difficulty subtests have a lower information target and a smaller weight, compared with the more difficult subtests on the right.

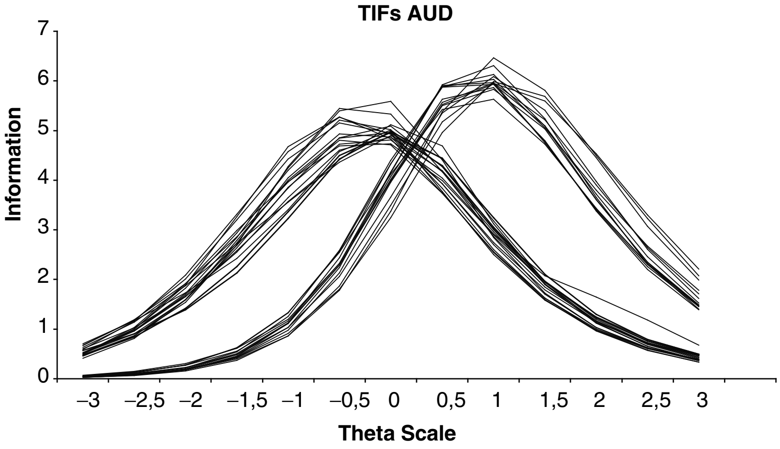


Fig. 9.1 Test information functions from simultaneous subtest assembly

When the optimization model is infeasible, no solution can be found. Detecting the reasons of infeasibility can be complex when the assembly problem has many constraints. An overview of literature on infeasibility analysis is given in Huitzing, Veldkamp and Verschoor, (2005).

Some researchers have offered techniques to detect infeasibility (e.g., Huitzing, 2004), whereas others offer an algorithm to flexibly deal with the constraints while choosing items for a test (e.g., Stocking & Swanson, 1993). When this flexibility or relaxation of constraints is not desirable (that is, the content and item-exposure constraints are necessary to ensure equality and test validity), careful examination of the constraint sets and data can be used to solve infeasibility (see Breithaupt, Ariel & Hare, this volume, chap. 13).

A simpler problem is the assembly and assignment of pretest questions to each panel. The script functionality of OPL is capable of solving both the operational subtests and the pretest assignment problems at the same time. Pretest blocks are designed to fill weak areas in the item pool and are assigned to each stage of the MST. Pretest questions are assigned to both the moderate and difficult subtests at stages 2 and 3 to reduce calibration problems related to constrained samples.

When the solution run has been completed, the psychometric and content teams review the results using a set of informative text files generated by the script. Output files that describe key properties of the subtests include information concerning content coverage and enemy conditions. In addition to these quality control files, a solution input file is generated by the script file to convey the composition of subtests to the next step in panel assembly.

Panel Construction

The panel construction step is accomplished via a custom software tool, but could easily be completed using logic implemented by linear programming rules similar to those used to assemble the subtests (Breithaupt & Hare, 2008). The panel builder tool reads the subtest build output and places together five subtests on a single panel for the MST design, according to specific rules. These rules include limits to enemy items across subtests, placement of items for pretesting, content balancing across routes, and a maximum reuse rule for any subtest on panels, among others. The panel builder also generates the routing score table that is used by the test driver in the field. The panel builder execution requires a few minutes to select rule conditions and is completed in a matter of seconds.

Creation of the routing rules is an important activity of the panel assembly software. The panel builder uses IRT properties of the subtests to identify a cutoff score on an estimated-number-correct scale that directs the test driver to administer the next subtest. Specifically, the point of intersection of the TIFs defined by each pair of medium and more difficult subtests is used as the cut-off score. There are separate cutoff scores for the stage 2 and 3 decisions. The routing score is expressed as the expected number of correct answers represented by the θ value at the point of intersection. This routing rule is enforced by the test administration software in the test center and is based on the number-correct score the candidate obtained. The routing table that is packaged for the test driver for each panel includes these number-correct cutoff scores.

Until this step in the process, all quality control work is completed using the features of items, subtests, and panels that are represented as data input to the automated test assembly process. The actual content of the test questions (stems, exhibits, and distractors for MCQs) is not needed during the assembly process. This procedure affords more protection for access to the actual test content, and therefore reduces security risks to the production process. The result from panel building is like a recipe for each administration of panels, including all required instructions but none of the live test content that candidates will see. All features of the test content important for valid panel and subtest assembly are extracted as codes in the item pool. Of course, the success of the build depends on accurate and consistent coding of each item, and the sufficiency of the item pool.

9.3.2 Calibration and Statistical Analyses of Multiple-Choice Items

In order to build and sustain an item bank and use IRT scoring for the MCQs, it is necessary to pretest and calibrate them. Therefore, these items must be administered first as pretest and then as operational items. Pretest items do not contribute to the scores of candidates. Instead, performances on them are evaluated to determine if their quality is adequate, and if so, estimate their IRT parameters. If the estimates meet minimum standards for quality, the pretest items may be promoted to operational status and used to assemble and score future examinations. This section offers a discussion of the methods used for the calibration of the item bank for the CPA Exams.

When the computerized version of the examination was launched, operational items were identified based on calibrations using candidate records from selected earlier paper-based administrations, where feasible. The items for the new section, BEC, were pretested in special studies at universities across the nation. Since 2004, the operational items have been selected using pretest calibrations from new items that were administered via computer.

Monitoring Item Quality

Monitoring the quality of operational and pretest items also depends on analyzing response data from operational administrations. For operational MCQs, the primary method of summarizing empirical quality for items is classical item analysis (proportion correct or p-value; biserial correlation between item and total scores). These summaries are created for each item, considering both the answer keys and distractors for MCQs. The purpose of this analysis is to flag any operational item that behaves unexpectedly as a result of a presentation error, or flawed or obsolete content. A set of guidelines is useful in identifying potentially flawed items. If an anomaly is found in an operational MCQ, that question is removed from scoring. Pretest items with problematic statistical properties are reviewed by content experts and generally removed from the bank and excluded from operational promotion.

Flagging criteria are based on a combination of statistical properties. However, the classical item difficulty and discrimination estimates used in initial evaluation are known to be conditional on the particular sample of test takers (Hambleton and Jones, 1993). Although easily computed and widely used, this classical analysis is thus insufficient when applied to MST administration models. Because of the routing, some operational items will be seen only by candidates within a defined range of ability. This includes the group of high-ability candidates who are administered the subtests of equal difficulty level at stages 2 and 3.

To alleviate the impact of the incomplete data on the estimation of classical estimates of item discrimination, individual item scores are not correlated with the number-correct scores of operational MCQs but with the ability estimates from the operational MCQs based on their IRT parameters. Because these ability estimates have been adjusted for the characteristics of the items including difficulty, the estimate of item discrimination is independent of the particular route or items the candidate saw during the administration.

Item Calibration

Responses to all objectively scored test components (MCQs and performance tasks in the simulations) were calibrated to fit the three-parameter logistic IRT model. This model for calibration established a common scale for all objectively scored items in the bank.

Initially, the location for this IRT scale was established using responses to the paper-based administration of the exam. New computer-administered item parameter estimates were linked to this scale using a joint calibration method where known item parameters were given fixed values. Once the common scale had been established, the scores on different panels drawn from this bank yielded consistent ability estimates for test takers of identical ability. In other words, scores for candidates who were administered different panels were comparable without any postadministration score adjustment (e.g., classical observed score equating).

A separate step was undertaken to validate and update the IRT item parameters. As described above, initial estimates for MCQs were based on large-sample response data from paper-based tests. This calibration established the initial scale for the IRT parameters and ability estimates. The new computerized simulations, obviously, were only administered to computer-based candidates. The items were calibrated and linked to the old scale through the MCQs. Also, MCQs for the parts of the test measuring new content, as well as all new pretest items, were also calibrated using computer-based response data and linked to the old scale.

Item Recalibration

Some time was needed for the computerized CPA Exam to allow candidate preparation and scheduling habits to stabilize. It was expected that the option to schedule tests at the convenience of the candidate, the availability of appointments on a nearly continuous basis, and growing familiarity of review course providers with the test blueprint and the format of the new examination would lead to a different ability scale. Therefore, a second calibration was conducted, taking advantage of larger samples to recenter the IRT scale using responses only from computerized administrations. It is important to note that the recalibration would not impact expected passing rates; in fact, the recalibration was validated based on replicating exactly the estimates of the scores for previous examinees.

A series of calibrations was required to link sparse matrices of accumulated exam data across multiple administrations. First, common MCQs between administrations were identified and calibrated together to establish a scale. Next, remaining MCQs were calibrated to this scale with common item parameter estimates fixed. Following this step, all simulations were calibrated with MCQ item parameter estimates fixed. Finally, all conversion tables used to compute expected number-correct scores were recreated using new item parameter estimates.

After the recalibration of computerized responses was completed in 2005, the new item parameter estimates and conversion tables were used to rescore past responses. The resulting scores matched the original scores well, within errors of measurement in the reported score scale. This was only one part of an extensive validation study, which also included evaluations of fit for the IRT model, data verifications, and staff peer reviews. These results were included in a review by the Psychometric Oversight Committee and an independent psychometrician to ensure no errors had been introduced during recalibration. After assurances and all quality checks were completed, the production version of the scoring system was updated with the new item parameter estimates and conversion tables to serve as the basis for operationally reported scores.

After the recalibration, new items pretested in subsequent administrations have been added to the item bank using the method of concurrent calibration with operational item parameters fixed. Periodic evaluation of item-parameter drift and trends in passing scores has been used to determine if an additional bank scale adjustment might be reasonable. Such a new recalibration may be needed if any significant changes in the preparation of candidates, in the administration mode, or in the content or skills occurred that would represent a change in the construct being measured.

9.3.3 Controls for Security and Score Accuracy

The CPA Exams are administered at the Prometric testing centers. Close cooperation with the NASBA and Prometric, our partner and vendor in the testing program, was essential to ensuring the test scores were accurate and valid. Physical and other security measures are important to preserve the veracity and validity of the test scores. These security measures include restricted access to software and data, and the creation of reasonable controls at each step of the registration, administration, test development, and scoring processes. This section offers a few examples of practices in the CPA testing program to illustrate those controls.

No scores are provided immediately following the administration of the examination. Instead, a secure continuous transmission of response data is sent from Prometric test centers to an AICPA-owned automated scoring system. This process affords at least two important benefits related to security. First, there is an opportunity to retain the final scores for a 100% replication on a separate scoring system at AICPA. The automated system scores immediately on receipt of responses from Prometric. Prior to approving release of those scores, all results are rescored on this separate scoring system at AICPA. Each result must match before it goes out to candidates. Second, the item calibrations required to construct scores are not distributed to test centers. As item calibrations are costly to develop and calculate, the scoring data itself are a valuable asset for the testing program and the ability to restrict access enhances both security and risks to accuracy.

Another issue concerns how and where test content is created. The test development occurs at AICPA and vendor-hosted authoring meetings in locations across the country. In each instance, only secure transmission of draft materials and completed test questions is allowed. The test content itself is housed in a secure remote location and has controlled access granted only by AICPA-approved administrators.

Test content and scoring information is backed up and stored in a separate secure location by AICPA and the vendors on a regular schedule. Background checks for all vendors and staff, in addition to strict confidentiality agreements and intellectual property rights, are all important factors in constructing business agreements among the vendors and partners. All these measures ensure that the best possible precautions against disaster or intentional disruption have been taken. The CPA testing program is also audited by as many as a dozen separate auditing bodies on an annual basis. Many of these audits focus on systems and security, and all are intended to exercise and test existing controls and quality checkpoints, and to identify new controls, as these are needed.

Another example of controls for score accuracy is provided in procedures used to review test content and scoring information for assembled panels. Committees of expert CPAs provide approval for each test item before it can be used on the examination. The AICPA staff validates the accuracy of the item bank coding that forms the basis for selection of questions into test forms with automated assembly. CPAs and software testing teams evaluate every panel prior to packing and release to Prometric test centers. Psychometric teams provide assurances and independent verification of data used for scoring each panel prior to the beginning of test administration. Also, there are psychometric and CPA reviews of any new content and scoring data analysis prior to score releases for every administration window.

In continuously testing programs, evidence of security risk usually appears too late in the process to protect the integrity of scores. Although we monitor changes in the difficulty of test questions (item drift) and track the number of administrations for any given item, most of the security focus is much earlier in the process. For example, overexposure of test content can lead to collusion and sharing of test content. Therefore, during the review of test panels, the CPA Exam staff check for unbalanced exposure of subtests among panels and the appropriateness of the routing rules for each panel. This helps to ensure expected item exposures are restricted to a minimum. In combination with a highly specific schedule for the authoring, use, and retirement of test questions, it is possible to set guidelines on item exposure. A more detailed description of the disciplined and specific inventory management system is offered by Breithaupt, Ariel, and Hare (this volume, chap. 13). It may be sufficient to note here that overlap of content between administrations is restricted, and that new subtests and panels are constructed on a continual basis.

Monitoring of item drift has become an operational procedure that is conducted every three months. Because an operational item may appear in subtests of different difficulty in the MST model across different administrations, an expected proportion of correct responses can be estimated for each item using IRT parameters. If the observed proportion of correct responses, conditioned on the ability of the candidate, is above projected estimates, the item is removed from the item bank and Prometric and State Boards are contacted to evaluate possible breaches of security.

It is also possible to use response times to evaluate changes in speededness of items (van der Linden, Breithaupt, Chuah, & Zhang, 2007) that might indicate possible cheating behavior. Because the keyboard and mouse movement of candidates during the administration are exactly recorded, their response time to items can be parsed, analyzed, and used to create a predictive model. Candidates spending too little or too long on items may indicate preknowledge or memorizing, for example. When the pattern of times does not match the expectation, based on previous time use for examinees, an investigation can be conducted into possible misuse of exam time.

Perhaps the most important responsibility for the overall security of the testing program lies with the testdelivery vendors. Impersonation of candidates is the most frequently identified source of security breach leading to invalid score decisions. Prometric test centers make use of photographs, fingerprinting, and controlled proctoring, and work with NASBA and AICPA to ensure that candidate data and scheduling are secure.

A unique identifier known only to NASBA is assigned once strict eligibility requirements are met and a test appointment may be scheduled. The registration at Prometric generates a fake identifier that will follow the result through the score calculation and transmission process to AICPA. Finally, if a candidate retakes the examination, these identifiers are the basis of restricting administration of any new panel that might contain the performance tasks the candidate has previously seen. AICPA does not have, nor have access to, any information that might identify candidates.

9.3.4 Standard Setting

In order to qualify for licensure, each candidate must pass the four separate examinations on Auditing and Attestation, Financial Accounting and Reporting, Regulation, and Business Environment and Constructs. Each of the tests is constructed somewhat differently (e.g., three have simulation tasks, including written communications, and one is MCQ only). All four tests have a unique performance standard that is represented as the required passing score. A series of research and operational studies was conducted over a period of four years to develop passing scores for each section. A detailed description is offered in the next section. Once the standard-setting methods were identified, it was necessary to collect recommendations from panelists of CPAs with knowledge of entry-level practice who were trained in the standard-setting procedure. The recommendations of the panel provide an important set of information for the final step, which was establishing the passing scores for the four exams by the Board of Examiners. The Board of Examiners used the panel's recommendations as well as other sources of relevant information.

Research on Standard Setting

Initial studies were performed to evaluate and contrast different standard-setting methods (Hambleton & Pitoniak, 2002; Mills, Hambleton, Biskin, Evans & Pfeffer, 2000; Pitoniak, Hambleton & Sireci, 2002). The methods reviewed included the Angoff, item-cluster, and direct consensus methods for the MCQs as well as the work classification and analytic method for the simulations.

The familiar Angoff method requires panelists to develop and accept a definition of borderline candidates and then to predict the proportion of such candidates who will answer an item correctly. In the item-cluster method, items are arranged into related sets and the responses of approximately 15–20 test takers to the items in each set are presented to the panelists. The panelists' task is to rate the test takers into one of four categories, from "hopeless" to "solid/exceptional." In the direct consensus method, the items are also clustered into related sets but the panelists' task is to predict the number of items in the set that a borderline candidate will answer correctly. The work classification method is similar to the item-cluster method in that the responses of 15–20 real test takers to the simulations are presented to the panelists who rate the responses from failing to solid. The analytic method is similar to the Angoff method in that panelists are required to predict the proportion of borderline test takers who will provide a correct response to a question.

The team of CPA Exam staff and University of Massachusetts at Amherst faculty who conducted the reviews recommended the use of the item-cluster method for the multiple-choice section and the work classification method for the simulations (Hambleton & Pitoniak, 2002; Pitoniak et al., 2002). The recommendation for the work classification method was based on the idea that sorting test takers' performances is a more meaningful task for panelists than predicting the performance of borderline candidates. The item-cluster method is similar to the work classification method. It was chosen given the preference for the latter and the secondary desire to have similar tasks for the multiple-choice and simulation-based tests. The implementation of that recommendation occurred in the standard-setting study one year before the examination was computerized.

Standard-Setting Study

Approximately 25 CPAs with requisite skills and experience were convened for four two-day standard-setting meetings, one for each of the four tests. With no overlap in panel membership, 100 CPAs participated in the standard-setting study. For the multiple-choice section a prototype form of each test, representing the content specifications and the target test characteristic curve, was generated using the underlying criteria discussed in the automated test assembly section. Using the prototype form of the multiple-choice section meant that it was not necessary for the panelists to evaluate each of the four possible pathways through the test.

The panelists were trained in the definition of borderline candidates and discussed what this meant and what behaviors could be expected on the job by such candidates. They were then trained in each method, made initial ratings, and discussed their ratings as a panel. Panelists could alter their ratings during a second round. Thereafter, a projected passing score was computed for each panelist and one last round of discussion concerning these passing scores ensued. Panelists could indicate whether they felt their recommendation was too easy, too hard, or accurate. Using the IRT parameters for the items in the prototype form, a test characteristic curve was generated and the raw passing scores were placed on the scales for the tests.

Three of the four tests include simulations. The panelists in these three standard-setting studies also reviewed two simulations using the work classification method. The recommended raw passing scores were placed on the scales for these tests using the same method as for the multiple-choice items. These two scores, the multiple-choice and the simulation, were weighted and combined to arrive at the panels' recommendations for total test-passing scores.

Establishing the Operational Standards

The AICPA Board of Examiners was convened for a final decision after the first window of administration of the new examination. The Board considered the panels' advisory recommendations and their comments about the entire standard-setting process they had followed. The Board also reviewed empirical data and issues related to level of preparation of the candidates in this first window, their facility with computerized administration, comparability to the paper-based examination, and many other topics. Geisinger (1991) presents an excellent review of the types of information that policy makers may want to address in making a final determination. Deliberations were documented and discussion was iterated with confidential voting between rounds. The Board members made their initial recommendations independently and then discussed them with regard to the above criteria to arrive at a consensus.

9.4 Conclusion

The CPA Exam was launched with the multiple-choice portions presented as a multistage adaptive test. The criteria for choosing this model were primarily the ability to capitalize on the benefits of adaptive testing without limiting the ability of the test takers to revisit and change answers to previous items within subtests. The format may also allow future changes to introduce better, more efficient diagnostic information for the test taker.

This chapter reviews the decision-making process from conceptualization of the new CPA Exam through the setting of the pass–fail standards to score reporting. The model has performed well since launch in 2004. Panels and subtests meet stringent validity and comparability standards, and the resultant scores are consistent across panels, routes, examinees, and administration windows. It is important to note that the computerized administration, assembly, and scoring model for MST has been entirely accepted by the test takers.

The process was intense, disciplined, informed by research and best practices, and completely transparent. In the future, test content will be updated based on a 2008 practice analysis, and even more flexible and secure content development policies will be implemented. In closing, the MST model is working well and expected benefits have been demonstrated in the computerized CPA Exam.

References

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27*, 241–253.

American Institute of Certified Public Accountants. (June 1997). *Conversion of the Uniform CPA Exam to a computer-based examination*. Jersey City, NJ.

American Institute of Certified Public Accountants & National Association of State Boards of Accountancy Computerization Implementation Committee. (1998, October). *Conversion of the Uniform CPA Exam to a computer-based examination* (Status Report Briefing Paper No. 1). Jersey City, NJ.

American Institutes for Research (2001). *Practice analysis of certified public accountants* [Technical Report]. Norris, Russell, Goodwin & Jessee.

Board of Examiners (1995). *Invitation to comment: Conversion of the Uniform CPA Exam to a computer-based examination*. Jersey City, NJ: American Institute of Certified Public Accountants.

Breithaupt, K. & Hare, D. R. (2008). *Automated assembly techniques for rapidly changing test designs*. Invited presentation at the annual meeting of the National Council on Measurement in Education, New York.

Geisinger, K. F. (1991). Using standard-setting data to establish cutoff scores. *Educational Measurement: Issues and Practice, 10*, 17–22.

Hambleton, R. K. & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*, 38–47.

Hambleton, R. K. & Pitoniak, M. J. (2002). Setting passing scores on the CBT version of the Uniform CPA Exam: Comparison of several promising methods [Unpublished report]. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Huitzing, H. A. (2004). Using set covering with item sampling to analyze infeasibility of linear programming test assembly models. *Applied Psychological Measurement, 26*, 355–375.

Huitzing, H. A., Veldkamp, B. P. & Verschoor, A. J. (2005). Infeasibility in automated test assembly models: A comparison study of different methods. *Journal of Educational Measurement, 42*, 223–243.

ILOG. (2002). *ILOG OPL Studio 3.6.1* [User Manual]. Mountain View, CA.

Luecht, R. M. (2000). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., Brumfield, T. & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage tests. *Applied Measurement in Education, 19*, 189–202.

Luecht, R. M. & Nungester, R. J. (1998). Some practical applications of computerized adaptive sequential testing. *Journal of Educational Measurement, 35*, 229–249.

Luecht, R. M. & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden & C. A. W. Glas (Eds). *Computer-adaptive testing: Theory and practice* (pp. 117–128). Boston: Kluwer-Nijhof Publishing.

Luecht, R. M., Nungester, R. J. & Hadadi, A. (1996, April). *Heuristics-based CAT: Balancing item information, content, and exposure*. Paper presented at annual meeting of National Council on Measurement in Education, New York.

Melican, G., Breithaupt, K., Mills, C., Hambleton, R. K. & Zhao, Y. (2005). *Multi-stage testing and case studies in a fully functioning licensing examination*. Invited symposium paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Mills, C. N., Hambleton, R. K., Biskin, B., Evans, J. & Pfeffer, M. (2000, January). *A comparison of two standard setting methods for Uniform CPA Exam* [Technical Report]. Jersey City, NJ: American Institute of Certified Public Accountants.

Nemhauser, G. L. & Wolsey, L. A. (1999). *Integer and combinatorial optimization*. New York: Wiley.

Pitoniak, M. J., Hambleton, R. K. & Sireci, S. G. (2002). *Advances in standard setting for professional licensure examinations*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 2002. [Also published as Laboratory of Psychometric and Evaluative Research Report No. 423. Amherst: University of Massachusetts, School of Education.]

Professional Examination Service (1999). *Final report of the conduct of a feasibility study for the computerization and implementation of a Uniform CPA Exam in fifty-four jurisdictions*. New York.

Rotou, O., Patsula, L., Steffen, M. & Rizavi, S. (2007). *Comparison of multistage tests with computerized adaptive and paper-and-pencil tests* [Research Report No. RR-07-04]. Princeton, NJ: Educational Testing Service.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.

Theunissen, T. J. J. M. (1987). Text banking and test design. *Language Testing, 4*, 1–8.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54*, 237–247.

van der Linden, W., Breithaupt, K., Chuah, S. C. & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement, 44*, 117–130.

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement, 26*, 133–146.

Chapter 10
A Japanese Adaptive Test of English as a Foreign Language: Developmental and Operational Aspects

Yasuko Nogami and Norio Hayashi

10.1 Introduction

With the advance and spread of computer technology, computerized adaptive testing (CAT) is gradually replacing traditional paper-and-pencil testing. One of the first computerized tests in Japan was the adaptive version of the Test of English as a Foreign Language (TOEFL). As a result of the interest in this test, some of the Japanese testing organizations are now in the process of computerizing their examinations. The Japan Institute for Educational Measurement (JIEM) developed an adaptive test of the proficiency of English as a foreign language—known as the Computerized Assessment System for English Communication. In this chapter, we introduce the test and describe its process of development. The information may be helpful to other testing agencies that also consider moving their tests to an adaptive format.

10.2 Overview of the Test

The current version of the Computerized Assessment System for English Communication (CASEC) has the following features:

1. Adaptation: The test is adaptive, meaning that the level of the difficulty of test items is automatically adjusted to the proficiency of the individual examinees during the test;
2. Online service: With Internet connectivity, examinees can take the test at any time and any place;
3. Immediate feedback: Examinees obtain feedback immediately upon completion of the test;
4. Application of item-response theory: In order to assess proficiency in English on a common scale, item response theory (IRT) is used. Consequently, the

Y. Nogami (✉) and N. Hayashi
Japan Institute for Educational Measurement, Inc., 162–8680 Tokyo,
55 Yokodera-cho Shinjuku-ku, Japan

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_10,

examinees are able to compare current with past results and keep track of the developments in their proficiency over time;

5. Short testing time: The average time to take the CASEC is 40 minutes, which is comparatively short for a test of this accuracy;
6. Broad-range measurement: The CASEC can be used to measure a wide range of proficiency levels—from basic to advanced.

The CASEC was developed by the Society for Testing English Proficiency (STEP). In 2000, JIEM became independent from STEP and took over the development and management of CASEC. Currently, the CASEC is one of the largest computerized adaptive tests in Japan. Since 2001, the number of examinees has been steadily increasing. In 2008, the test was taken over 120,000 times. The examinees range widely in background—from junior-high-school students to university graduates and adults in the workforce. The results of the test are used for different purposes and in different contexts, for instance, placement in schools and monitoring of educational achievements. Examinees also take the test to check their English proficiency for personal purposes.

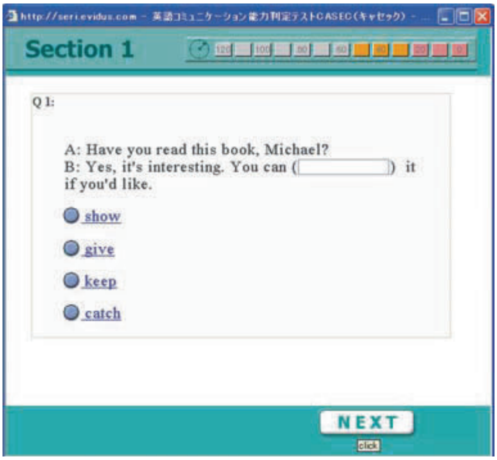
The CASEC consists of a total of 55 items divided into four different sections with an expected testing time of approximately 40 minutes. The four sections are described in Figure 10.2, which also shows sample test items and a sample score report that the examinees see on their screen.

Examinees answer each individual item when it is administered; they are not allowed to go back and review their responses to previous items. Each item has a time limit. The reason for setting this limit was to prevent the examinees from spending too much time on the entire test because of dwelling too long on an occasional item. The limit was set without giving the examinees a feeling of speededness. If an examinee is able to answer an item before the time limit passes, he or she can move on to the next item, so the total time on the test still varies somewhat among the examinees.

Although the CASEC is an adaptive test, it has a fixed number of test items. The reason for this is explained below. Scores on each of the four sections are reported on a scale from 0 to 250. The scores for the sections are also summed to report a total score.

The CASEC requires Internet connectivity, a computer to offer the environment described in Table 10.1, and, because it also has a listening component, either a headphone or speakers. It is possible to take the test only with a sound-enabled computer and a standard browser. Examinees who have taken the CASEC earlier can access its website from a log-in window by typing their ID and password. After log-in, the website performs a computer-environment check. If no problems are found, the examinee proceeds to the test; otherwise, they are required to adjust their computers or contact a support center for assistance. First-time examinees must register before logging in; they receive an ID and password immediately upon registration.

When the test is completed, the score report appears on the screen. The scores are stored for four years. The system allow examinees to check past records when they log in using the same ID and password. As mentioned earlier, one of the special features of the CASEC is that examinees can take the test anytime and can always keep track of how their English proficiency has developed over time.


http://seri.evidus.com - 英語コミュニケーション能力判定テストCASEC(キャセック) - ...
Section 1
120
100
80
60
40
20
0
NEXT
click

Q 1:

A: Have you read this book, Michael?
B: Yes, it's interesting. You can ( ) it if you'd like.

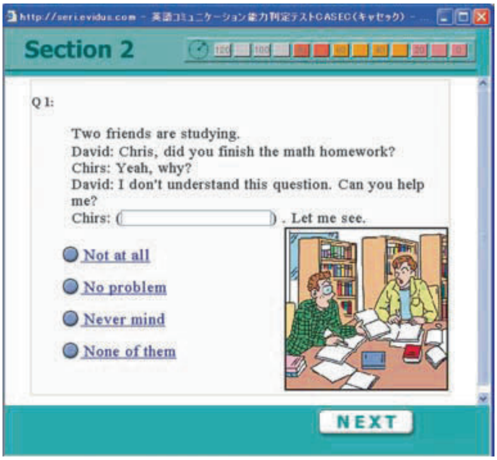
○ show

○ give

○ keep

○ catch

Section 1 assesses vocabulary knowledge. Examinees complete the presented sentence by selecting the most appropriate word from four alternatives to fill in the blank. The time limit for each item is 60 seconds. 15 items are administered in this section.


http://seri.evidus.com - 英語コミュニケーション能力判定テストCASEC(キャセック) - ...
Section 2
120
100
80
60
40
20
0
NEXT

Q 1:

Two friends are studying.
David: Chris, did you finish the math homework?
Chirs: Yeah, why?
David: I don't understand this question. Can you help me?
Chirs: ( ) . Let me see.

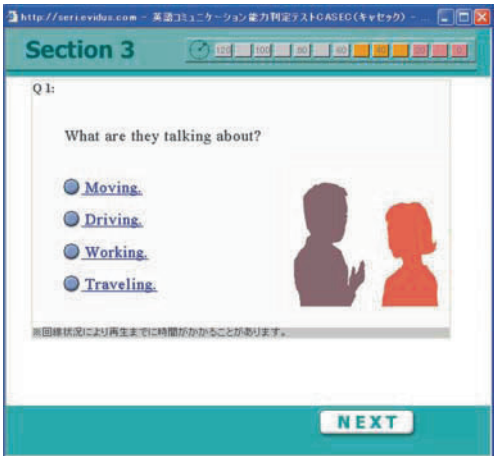
○ Not at all

○ No problem

○ Never mind

○ None of them

Section 2 assesses knowledge of phrasal expression and usage. Examinees read a given dialogue, look at a picture that describes the situation, and then select the most appropriate verbal response from four choices to fill in the blank. The time limit for each item is 90 seconds. 15 items are administered in this section.


http://seri.evidus.com - 英語コミュニケーション能力判定テストCASEC(キャセック) - ...
Section 3
120
100
80
60
40
20
0
NEXT

Q 1:

What are they talking about?

○ Moving.

○ Driving.

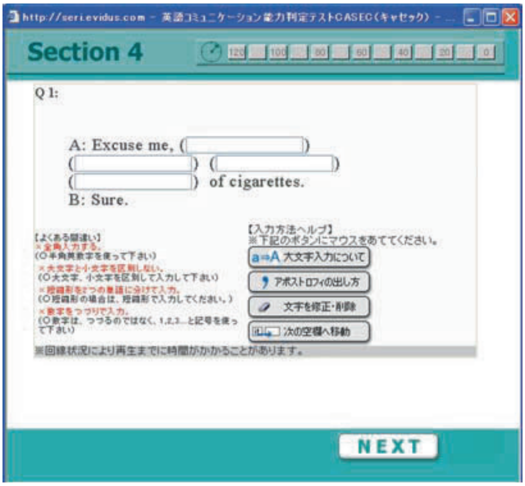
○ Working.

○ Traveling.

※回線状況により再生までに時間がかかることがあります。

Section 3 assesses listening proficiency, specifically the proficiency to understand the main idea. Examinees listen to sentences or dialogue through speakers or headphones and attempt to understand the content. Upon completion of a given passage, examinees are presented with a question and choose the most appropriate answer from four given alternatives. The time limit for each item is 60 seconds. 15 items are administered in this section.

Fig. 10.1 Sample test items and sample score report for the CASEC



Q 1:

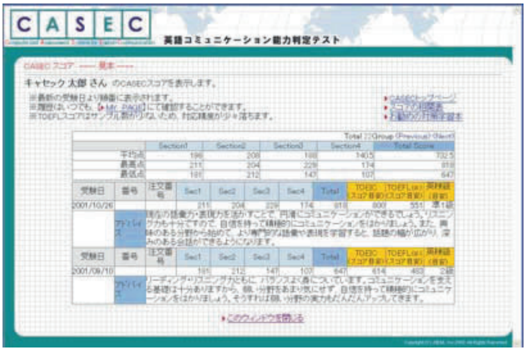
A: Excuse me, ( ) ( ) ( ) ( ) of cigarettes.
B: Sure.

【よくある間違い】
×全角入力する。
(○半角英数字を使って下さい)
×大文字と小文字を区別しない。
(○大文字、小文字を区別して入力して下さい)
×短縮形を2つの単語に分けて入力。
(○短縮形の場合は、短縮形で入力してください。)
×数字をつづりで入力。
(○数字は、つづるのではなく、1,2,3...と記号を使って下さい)

【入力方法ヘルプ】
※下記のボタンにマウスをあててください。

※回線状況により再生までに時間がかかることがあります。

Section 4 assesses listening proficiency, specifically the proficiency to understand specific information. This is a dictation task in which examinees listen to sentences being read while observing the written text on their computer screen. Examinees type in the missing words in the blanks within the text. The time limit for each item is 120 seconds. 10 items are administered in this section.



CASEC スコア ―― 見本 ――

キャセック太郎 さん のCASECスコアを表示します。

※最新の受験日より順番に表示されます。
※履歴はいつでも、【MY PAGE】にて確認することができます。
※TOEFLスコアはサンプル数が少ないため、対応精度が少々落ちます。

Total 2 Group (Previous) (Next)

	Section1	Section2	Section3	Section4	Total Score
平均点	196	208	188	140.5	732.5
最高点	211	204	229	174	818
最低点	181	212	147	107	647

受験日	番号	注文番号	Sec1	Sec2	Sec3	Sec4	Total	TOEIC(スコア目安)	TOEFL(pbt)(スコア目安)	英検級(目安)
2001/10/26			211	204	229	174	818	800	551	準1級
	アドバイス	現在の語彙力・表現力を活かすことで、円滑にコミュニケーションができるでしょう。リスニング力も十分ですので、自信を持って積極的にコミュニケーションをはかりましょう。また、興味のある分野から始めて、より専門的な語彙や表現を学習すると、話題の幅が広がり、深みのある会話ができるようになります。								
受験日	番号	注文番号	Sec1	Sec2	Sec3	Sec4	Total	TOEIC(スコア目安)	TOEFL(pbt)(スコア目安)	英検級(目安)
2001/09/10			181	212	147	107	647	614	483	2級
	アドバイス	リーディング・リスニング力ともに、バランスよく身についています。コミュニケーションを支える基礎は十分ありますから、弱い分野をあまり気にせず、自信を持って積極的にコミュニケーションをはかりましょう。そうすれば弱い分野の実力もだんだんアップしてきます。								

CASEC Score Report. Examinees receive a score report immediately after finishing their test. The report contains the total test score, a score for each section, and simple advice. The report also provides estimated equivalent *TOEIC*® and *TOEFL*® scores and an estimated grade level of the STEP Test as a reference. If examinees use the same ID and password to take the test multiple times, they can look back at past test scores and see how their English proficiency has varied over time.

Fig. 10.1 continued

Table 10.1 The minimum computer and Internet requirements for the CASEC

Component	Requirement
CPU	MMX 200 MHz or above
Memory	At least 128 MB of RAM
OS	Windows 98, Me, NT 4.0, 2000, XP, or Vista
Browser	Internet Explorer 5.0 or above
Sound	Windows Media Player 6.4 or above
Connection	At least 56 Kbps

10.3 CASEC Development Flow

Roughly, the developmental process of the CASEC can be divided into three stages: (i) test specification and item construction, (ii) choice of testing algorithms, and (iii) building the operational system.

During the first stage, the CASEC developers first determined which item formats were most appropriate for measuring English proficiency, and then constructed the items and calibrated them. During the second stage, the developers examined possible item selection rules and methods of proficiency estimation and made their choice. During the last stage, they investigated how to run the actual testing system efficiently and reliably. In the actual development process, these stages overlapped; their flow is shown in Figure 10.2. In the following sections, we discuss the first two stages in more detail.

10.3.1 Measurement Target

As already noted, the CASEC measures the English proficiency of a wide range of examinees, from a basic level (approximately first year of middle school) to university-graduate and working-adult levels. In fact, it was designed to match the

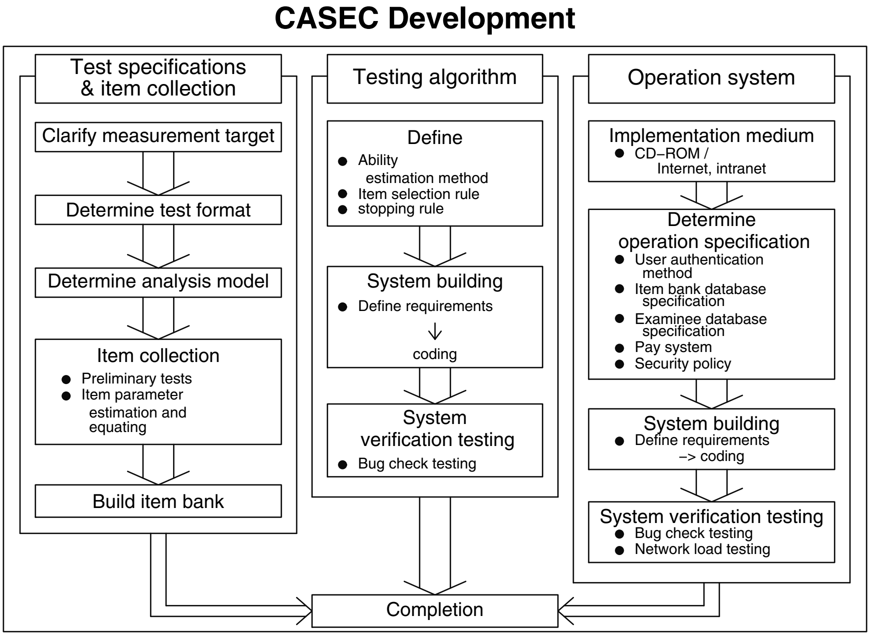

CASEC Development
Test specifications & item collection
Clarify measurement target
Determine test format
Determine analysis model
Item collection
● Preliminary tests
● Item parameter estimation and equating
Build item bank
Testing algorithm
Define
● Ability estimation method
● Item selection rule
● stopping rule
System building
● Define requirements
↓
coding
System verification testing
● Bug check testing
Operation system
Implementation medium
● CD-ROM / Internet, intranet
Determine operation specification
● User authentication method
● Item bank database specification
● Examinee database specification
● Pay system
● Security policy
System building
● Define requirements -> coding
System verification testing
● Bug check testing
● Network load testing
Completion

Fig. 10.2 Developmental flow of the CASEC

Table 10.2 Summary of the STEP levels

Grade Level	Vocabulary Acquisition	Performance
1	10,000 to 15,000 words	Understands English required for everyday life thoroughly and with sufficient expressiveness
Pre 1	7,500 words	Understands English required for everyday social life with an emphasis on oral expression
2	5,100 words	Understands English required for everyday life and at the workplace with an emphasis on oral expression
Pre 2	3,600 words	Understands simple English required for everyday life with an emphasis on oral expression
3	2,100 words	Understands basic English with an emphasis on oral expression
4	1,300 words	Understands elementary English with the ability to comprehend and speak simple English
5	600 words	Understands rudimentary English with the ability to comprehend and speak simple English

range from Grade 5 (the lowest) to Grade 1 (the highest) of the STEP test. This test, which is sponsored by the Society for Testing English Proficiency (STEP) and is designed according to the guideline provided by the Ministry of Education in Japan is the largest English proficiency test in Japan with nearly 2.5 million examinees per year. Over the past 40 years, more than 70 million examinees have taken it.

Examinees who take the STEP test select a level at which they want to test their English proficiency. Table 10.2 is a summary of each level of the STEP test. The score report for the test tells the examinee whether or not he or she has passed the chosen level. Although it was designed to cover the same range as the STEP test, the CASEC differs from it in that it measures the proficiency on a continuous scale.

10.3.2 Item Format

The test items were required to measure English communication proficiency. The developers examined dozens of types of item formats to find what would suit this objective best. They finally narrowed down their selection to 18 types and evaluated their appropriateness using trial items in a paper-and-pencil version. These items were administered along with a 20-minute interview, which was evaluated by two or three native English speakers. Furthermore, the developers thoroughly examined earlier empirical studies on the measurement of English proficiency. They also sought the opinions of teachers and researchers involved in the instruction of English.

Next, a short list of item formats with the highest correlations with the interview scores was put together. Based on their feasibility for computer administration and scoring, finally the four formats in Figure 10.1 were chosen.

10.3.3 Item Response Model

For use in adaptive testing, a variety of response models is available. Item calibration and proficiency estimation for the CASEC is based on the two- (2PL) or three-parameter logistic (3PL) models (van der Linden & Pashley, this volume, chap. 1). For each section, the choice of model was made to get compatibility with the response format of the items. The first three sections of the CASEC have a multiple-choice format with four alternatives for its items; for these sections, the 3PL model is used. Section 10.4 has a format in which the examinee listens to an English dialogue and types in the blanks on a keyboard; for this section, guessing is less likely and the 2PL model is used.

10.3.4 Item Construction and Pretesting

Once the item formats were chosen, the developers constructed and pretested the items for the CASEC. Over a three-year period, the items were pretested using past examinees of the STEP test. The total number of examinees used for these pretests was over 100,000.

Special test forms for these pretests were designed. Items were classified into three difficulty levels. Each form mainly consisted of items in one of the levels but also included those of one level up or down. For each level, we used multiple forms. Twenty percent of the items in each test form overlapped with other test forms. The items in the overlap thus served as anchor items, which were used to link the item parameters in different forms. After an examinee's level on the STEP test was identified, a test form appropriate at this level was given. At least 1,000 examinees were assigned to each form. To date, over 60 test forms have been assembled and administered. The process of pretesting the items still continues.

As a first step, the items were analyzed using classical test theory. Items with extreme p-values or a very low correlation with the total test score were eliminated. The IRT items parameter were then estimated using the computer program *PC-Bilog* (Mislevy and Bock, 1990). For each section, a separate proficiency parameter was fitted to the items. The final step was to use the anchor items to equate the estimates on common scales.

10.3.5 Description of the Item Bank

The advantage of CAT is measurement with a high degree of precision at any proficiency level. But this high degree of precision requires a well-stocked item bank to match any of these levels. So, the ideal shape of the difficulty distribution of an item bank for adaptive testing is a uniform distribution. As shown in Figure 10.3, the current shape of the item difficulty distribution of the nearly 4,000 items in

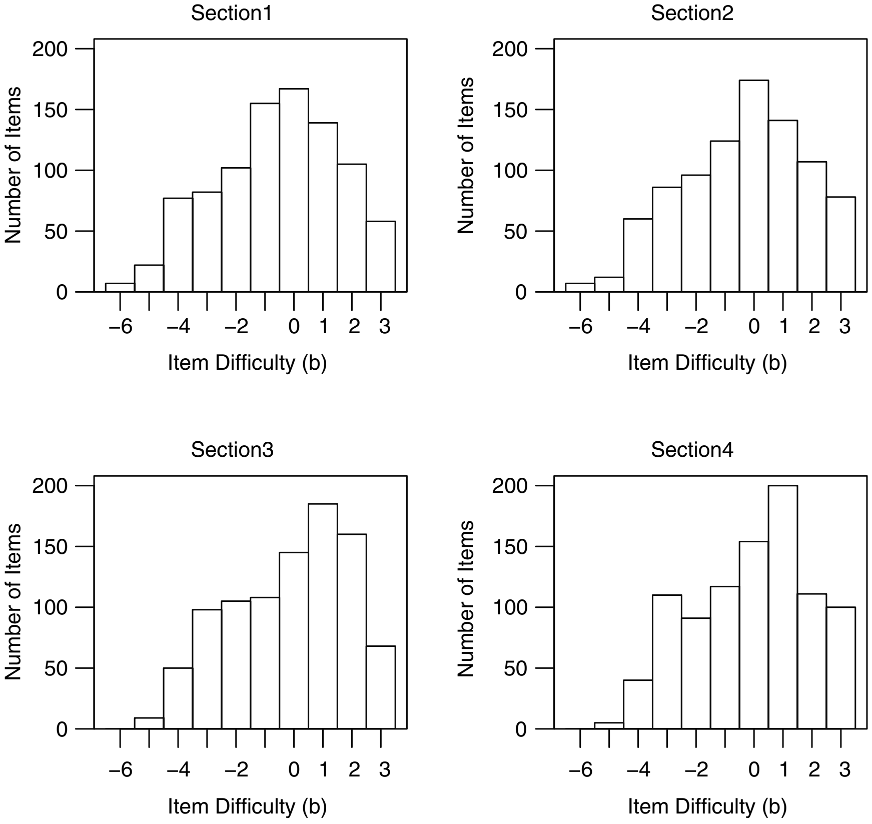


Fig. 10.3 Item difficulty distribution in each section of the item bank

the CASEC bank is unimodal, with sufficient numbers of items at the intermediate levels but fewer numbers at the lower and higher levels. The current distribution is not a major problem when taking an adaptive test for the first time. However, examinees at a high level who take the test multiple times have a probability of seeing past items that increases with each subsequent tests. But, of course, we will increase the number of easy and difficult items in the bank to accommodate lower-level and higher-level examinees who want to take the test as many times as they want.

Figure 10.4 shows that, even though the number of intermediate items is sufficient, items at this level are exposed relatively more frequently. As discussed in more detail below, the exposure frequency of intermediate items is mainly due to the CASEC item-selection algorithm. The algorithm presents all examinees with intermediate items at the beginning of the test; hence their higher exposure rate. From this perspective, it makes sense that the CASEC has more intermediate items than items at the other difficulty levels. As the CASEC becomes used more widely, the numbers of examinees at each level will increase. In order to prepare for this,

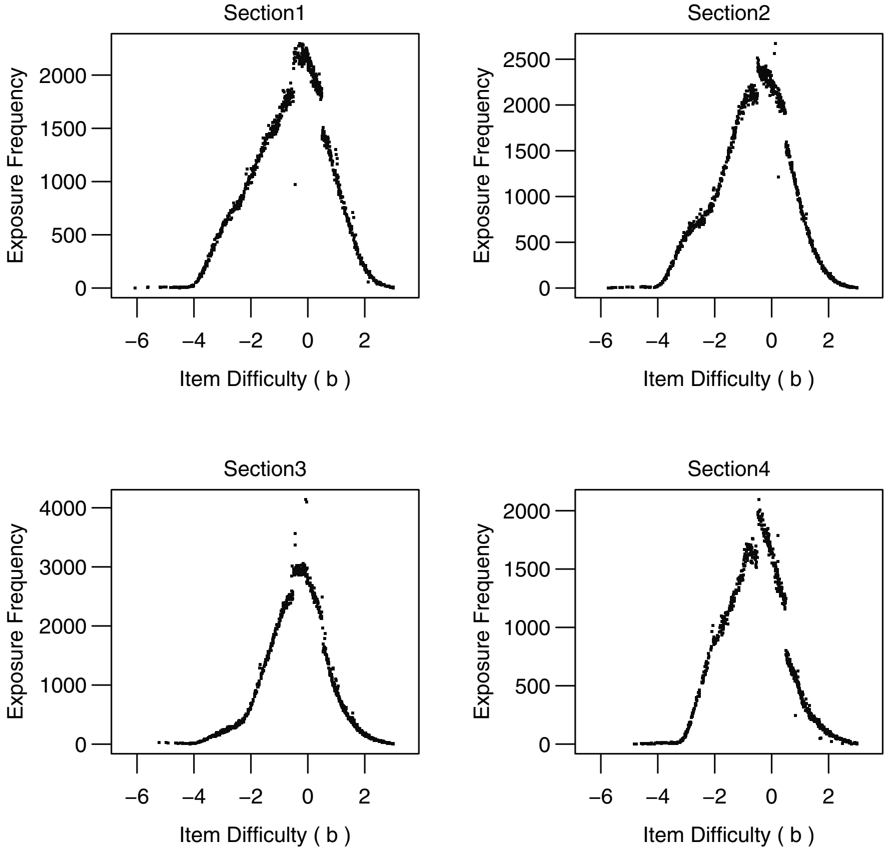


Fig. 10.4 Relationship between item difficulty and exposure frequency for each section in the item bank

the developers are continuously pretesting new items. It is expected that, before too long, examinees at any level who repeat the test will have a negligible probability of encountering an item from a previous test.

10.3.6 Testing Algorithms

In adaptive testing, the examinee's proficiency estimate is updated each time he or she answers a question and the next item is selected to be optimal at the estimate. Various algorithms for proficiency estimation and item selection have been proposed to achieve this adaptation. Before describing the sequential estimation method currently used for the CASEC, we will look at a tree method (Hayashi, 1998), which we experimented with during the development of the test.

Tree Method

The tree method was a very simple testing method. It was useful because it could be implemented in a short time and did not require complex calculations to estimate the examinee's proficiency.

As shown in Figure 10.5, the tree method involved a pyramidal arrangement. The vertical columns of items had approximately equal difficulties. After a correct answer, the next item was that on the lower right, whereas an incorrect answer invoked that on the lower left. Each time an examinee answered a question, points were given according to the difficulty level of the item block and whether the answer was correct or not.

We did a computer simulation to estimate the proficiency scores for the response patterns corresponding to each route. Then, a conversion table was prepared to convert the information about the examinee's route into a proficiency estimate. During the test, the algorithm followed the examinee's route and read the conversion table for an estimate of his or her proficiency score.

Although this method enabled us to run an adaptive test without complex calculations, it had some problems:

1. The proficiency estimates were fixed in advance and difficult to change.
2. The simulation to create a conversion table used the mean difficulty for each column in Figure 10.5 as the item parameter. The differences between the proficiency estimates based on the conversion table and the estimates based on the actual item parameters appeared not to be negligible though.
3. The proficiency estimator converged more slowly than necessary due to low measurement efficiency.

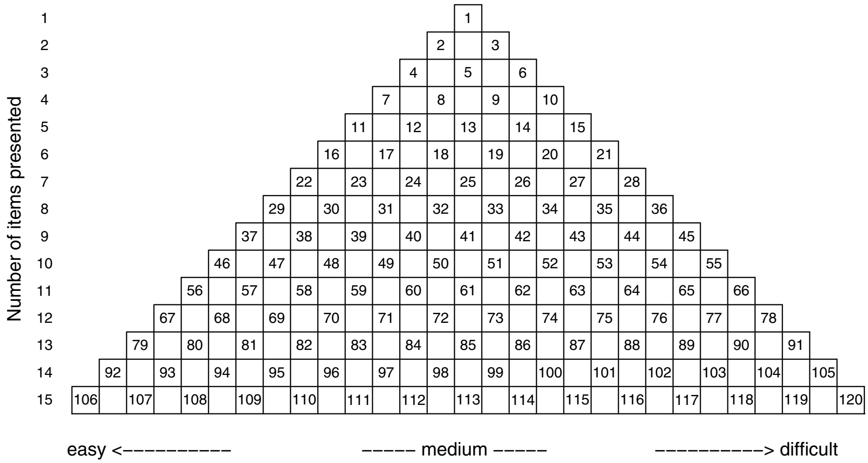


Fig. 10.5 Example of an item arrangement for adaptive testing using the tree method (15 items)

As soon as we had more advanced computational capacity at our disposal, the estimation method in the next section was developed for the CASEC to solve these problems.

Estimation Method

The current testing algorithm is based on Owen's method (Owen, 1975) (van der Linden & Pashley, this volume, chap. 2), though we use a logistic response model whereas Owen used a normal-ogive model.

As for the proficiency estimator, we use the Bayes expected a posteriori (EAP) estimator (Bock & Aitkin, 1981) just as Owen (1975) suggested. We also examined maximum-likelihood (ML) estimation. But the ML estimates appear to fluctuate widely in the beginning of the test when an examinee answers a difficult question correctly but misses the answer to an easy question.

Item-Selection Method

The item-selection rule is also based on Owen (1975). It selects the next item to have a difficulty parameter differing no more than 0.5 from the current proficiency estimate. That is, the kth item to be administered to an examinee i is chosen to satisfy

$$|b_{i_k} - E(\theta|u_{i_1} \cdots u_{i_{k-1}})| < \delta, \tag{10.1}$$

where b_{i_k} is the difficulty parameter of an item i in the pool administered as the kth item to the examinee, θ is the proficiency parameter, and the posterior mean of θ, that is, $E(\theta|u_{i_1} \cdots u_{i_{k-1}})$, is used as the proficiency estimate based on responses up to the previous $k-1$ items, $u_{i_1} \cdots u_{i_{k-1}}$. In our application, we chose $\delta = 0.5$. Notice that Owen (1975) requires δ to be a small positive number but does not specify any value.

The kth item is selected at random from the items that satisfy the condition given in (10.1).

When using the maximum-information criterion for item selection (van der Linden & Pashley, this volume, chap. 2), items with high discriminating power may appear to be selected too frequently early in the test (Chang and Ying, 1999). When selecting the items using the difficulty criterion in (10.1), this problem is reduced. From the point of view of measurement accuracy, a smaller range of item difficulty in (10.1) is better. Therefore, originally, δ was set at

$$\delta = 0.5\sigma_{k-1}, \tag{10.2}$$

where σ_{k-1} was the (asymptotic) standard error for the proficiency estimate after the first $k-1$ items. This choice implied that the more accurate the current proficiency estimate, the closer the difficulty of the next item to it.

However, the item bank does not always have an item that satisfies this criterion. We found this to happen more frequently when the proficiency of the examinees was very low or high. Therefore, the rule in (10.2) was changed to the value of $\delta = 0.5$ with further relaxation to $\delta = 0.75$ when no item is found.

The initial item for each section is selected to meet the item-selection rule at $\theta = 0$. That is, it is selected from those with difficulties $-0.5 < b_{i_1} < 0.5$. The value $\theta = 0$ is placed between the average proficiencies of the examinees who pass the Grade 2 and pre-Grade 2 levels of the STEP test in Table 10.2.

10.3.7 Stopping Rule

When the CASEC was developed, we experimented with different stopping rules. After examining both the results from these experiments and a survey of the potential users, we decided to choose a stopping rule with a fixed number of items instead of a fixed accuracy. The standard version of the test now has 15 items for Sections 10.1 through 10.3 and 10 items for Section 10.4.

The lengths of the sections were chosen balancing between the time available for the test and its measurement accuracy. Initially, the CASEC was expected to be used mainly by middle-school and senior-high-school students. Since a class in Japan usually takes 45 to 50 minutes, the maximum number of items feasible for this period was chosen.

Measurement accuracy for a test of 15 items is not very high. But as Figure 10.6 shows, even if the number of items increased to 20, measurement accuracy would not improve enough to offset the increased time required by the extra items. Therefore, the current number of items was judged appropriate.

10.3.8 Score Scale

Since proficiency estimates directly on the scale of θ are hard to interpret for the examinees, the estimates are converted to a scale that is easier to comprehend. In order to obtain a total score with a range from 0 to 1,000 points, the possible scores for each of the four sections run from 0 to 250 points. The scores are obtained from the $\hat{\theta}$s using a straightforward linear transformation. If all items are answered incorrectly the score is truncated at 0. If all items are correctly, the score is truncated at 250.

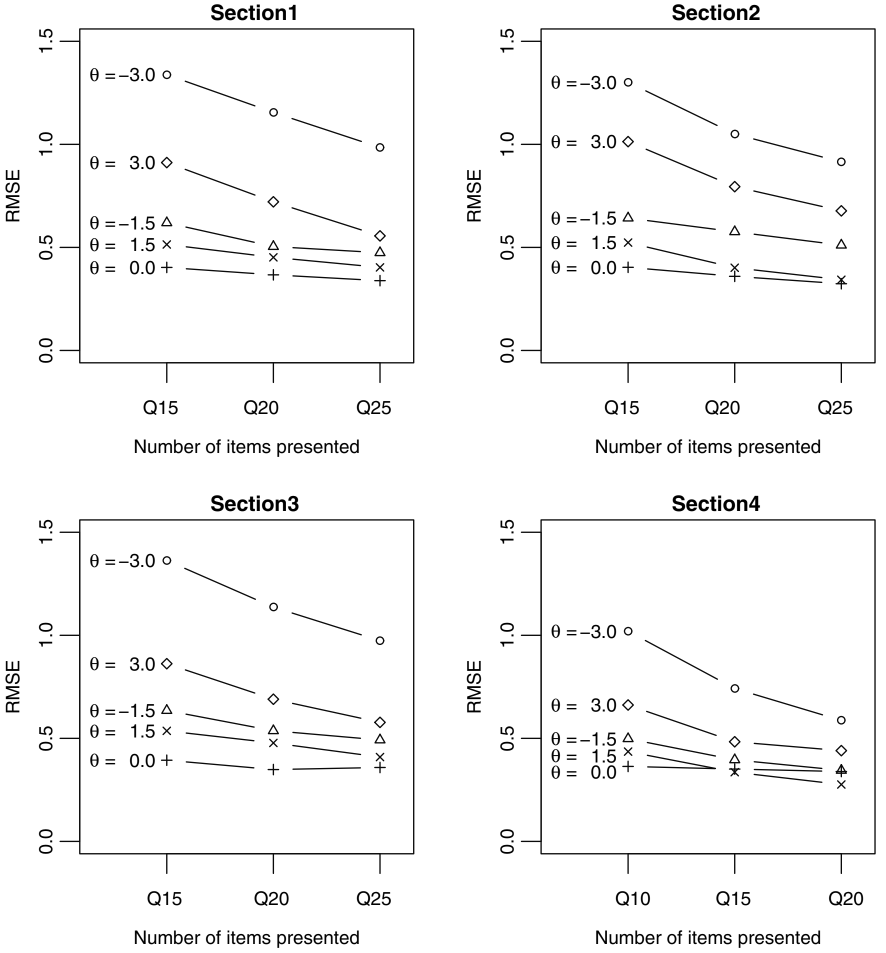


Fig. 10.6 The relationship between the number of items and the root mean-squared error (RMSE) for each section in the test

10.4 Validity Research

The development of the CASEC was supported by several empirical studies to assess its effectiveness and validity. Two of these studies are reported here. One study compares the measurement accuracies of an adaptive and a paper-and-pencil version of the test. The other study was to estimate the reliability and validity of the test.

10.4.1 Comparison of Measurement Accuracies Between CAT and P&P Versions of the Test

A total of 168 examinees took both a computerized adaptive (CAT) version and a paper-and-pencil (P&P) version of the test. For a description of both versions, see Table 10.3.

Since we experimented with an early version of the CASEC at the time, the stopping rule for the CAT version was different. It was stopped when either of the following criteria was satisfied (cf. Thissen & Mislevy, 1990):

1. The standard error of measurement was smaller than 0.5 and the difference with the preceding proficiency estimate was smaller than 0.001.
2. The number of items administered was up to 30.

The P&P version of the test was assembled from the same CASEC item bank as the CAT version. It had a test information function covering the same wide range of proficiency as the CAT version. Both versions had four sections. The P&P version consisted of 30 items per section. The order in which the sections were administered was the same for both versions. The total test time for the P&P version was two hours; one hour for Sections 10.1 and 10.2 and one hour for Sections 10.3 and 10.4.

In order to eliminate possible order effects, the administrations of the two tests were counterbalanced across the subjects. For both tests, the proficiencies were estimated using maximum-likelihood estimation. The ML estimates were converted to the regular score scale.

Table 10.4 shows the means, standard deviations, and correlations for the four sections of the test. The differences between the means yielded significant t-tests for all sections except Section 10.2. But no consistent trend for these differences was observed. All differences between the standard deviations yielded a significant t-test, where the standard deviations for the CAT version tended to be larger than those for the P&P version of the test. The correlations between the scores for the CAT and P&P versions were high, both for the individual sections (0.76 to 0.89) and for the total test (0.96).

Table 10.3 Description of CAT and P&P versions of the CASEC

	P&P	CAT
Number of items per section	30	≤ 30
Item selection	fixed	adaptive
Time limit		
Section 1	1)	60 sec/item
Section 2	1)	90 sec/item
Section 3	2)	60 sec/item
Section 4	2)	120 sec/item

1) Total of 1 hour for Sections 1–2,
2) Total of 1 hour for Sections 3–4.

Table 10.4 Mean, standard deviation, and correlation between identical sections of the CAT and P&P versions of the CASEC

		Section 1	Section 2	Section 3	Section 4
Mean	P&P	103.5	100.7	104.5	104.3
		(10.7)	(10.1)	(10.0)	(9.2)
	CAT	100.2	100.6	107.7	100.7
		(18.2)	(12.0)	(12.2)	(11.2)
Correlation		0.89	0.76	0.82	0.87

Note: Numbers in parentheses are standard deviations.

Table 10.5 Error of measurement for the P&P and CAT versions of the CASEC

	Mean Number of Items		Mean Standard Error	
Section	CAT	P&P	CAT	P&P
1	21.8	30.0	0.49	0.63
2	20.2	30.0	0.47	0.64
3	20.0	30.0	0.46	0.63
4	12.9	30.0	0.36	0.45

Table 10.5 shows the mean numbers of items and the mean standard errors of the proficiency scores for both versions of the test. The numbers of items for the P&P version were fixed at 30 items, but the numbers for the CAT version depended on the stopping rule discussed above. The CAT version had both smaller standard errors and smaller numbers of items than the P&P version. For example, the standard error for Section 10.1 of the P&P version was 0.63 for 30 items, but for the CAT version it was 0.49 for an average number of 21.8 items.

A much smaller error of measurement is one of the positive effects of adaptive testing. From Table 10.5, it follows that, on average, the CAT version required only 12 items to estimate the proficiency with the same degree of accuracy as the 30-item P&P version. This result confirms the rule of thumb that for an adaptive test to have the same accuracy as a paper-and-pencil test, only some 40% of the items is required.

10.4.2 Examining CASEC Reliability and Validity

In this experiment, 48 examinees took an early version of the CASEC three times in one day. This version had a total of up to 120 items, with at most 30 items in each section. For a fixed test, it is difficult to use the same setup because of retention effects with the examinees. For an adaptive test, such effects are much smaller.

Table 10.6 shows the correlations between the three test administrations and the means and standard errors of their proficiency scores. The correlations can be interpreted as test–retest estimates of the reliability of the test. It is clear from Table 10.6 that the three administrations yielded identical score distributions and

Table 10.6 Means, standard deviations (SDs), and correlations between test administrations (N = 48)

		Test Administration		
		1st	2nd	3rd
	1st	1.00		
Correlations	2nd	0.98	1.00	
	3rd	0.97	0.96	1.00
Mean		423.2	425.2	423.7
SD		42.6	42.5	42.9

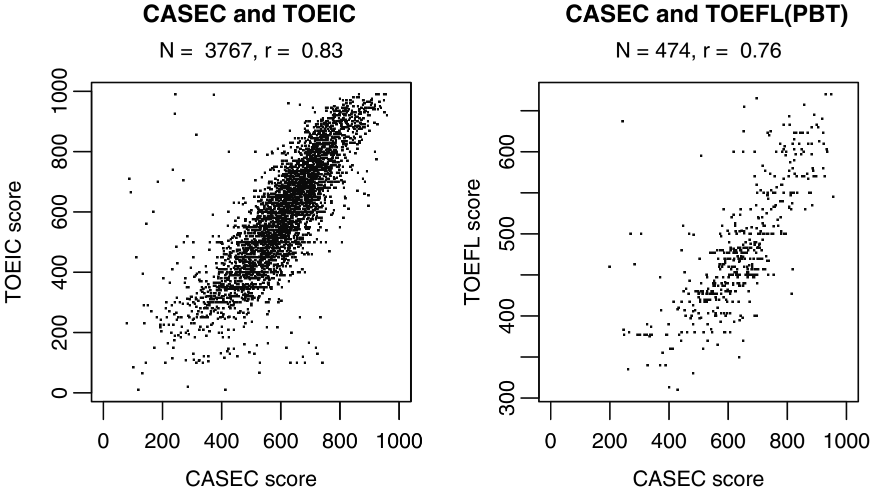


Fig. 10.7 Scatterplots of the CASEC and the TOEIC and TOEFL scores

their reliability estimates were high. The reliability estimates remained high when we corrected them for the difference in test length with the current version of the CASEC using the Spearman–Brown formula. After correction they ranged from 0.92 to 0.96.

To assess the validity of the current version of the CASEC, we correlated its scores with those on two well-established tests of English proficiency, namely the TOEIC and the TOEFL. Examinees who took the CASEC were asked if they had taken the TOEIC or the TOEFL within the last two years. They were then asked to submit their score on a voluntary basis. A total of 9,738 examinees submitted their score on the TOEIC and 88 their score on the TOEFL. Scatter plots of the scores on the CASEC and the TOEIC or TOEFL are shown in Figure 10.7 We found high validity coefficients. (Unfortunately, since we have no estimates of the reliability of the TOEIC and the TOEFL, we cannot correct them for attenuation of the criterion.)

Our conclusion from these two studies is that the CASEC has both high reliability and concurrent validity as a test of English proficiency.

10.5 Current Challenges

This section describes the challenges that the developers of the CASEC still face. The first category of challenges is related to psychometric issues, the second to the operational issues of running the testing program. For both categories, we outline the problems and discuss the potential solutions that are currently being researched.

10.5.1 Psychometric Issues

Item Bank Maintenance and Item Pretesting

Only when a large supply of items with a large variation in difficulty is available in the item bank can adaptive testing realize its potential of measuring examinees over a broad range of proficiency levels. As indicated earlier, items of low and high difficulty are currently underrepresented. Paradoxically, the same holds for items with a medium level of difficulty. The item bank does have more of them, but they are also needed more frequently because the test starts in the middle of the scale, exactly where the majority of the examinees are. The CASEC research lab is still in the process of administering pretest items to replenish the item bank.

Until recently, all items were pretested using fixed paper-and-pencil test forms. The advantage of this format was that large amounts of data were collected in a single test administration. Also, the quality of the data was high since the tests were administered in proctored sessions. A disadvantage was that the testing format used a different medium than the computer. Also, the expenses involved in finding test sites, remunerating the examinees, and travel were enormous. Such expenses must ultimately be incorporated in the examination fee.

By using an experimental testing system, developers are now experimenting with a method in which the examinees are unaware of which items are pretested and which are part of the real test. Currently, the number of pretest items is not larger than three per section. However, in order to collect enough data, the method requires a considerable number of examinees and is therefore not very efficient. On the other hand, we cannot increase the current testing time or reduce the length of the sections to create more space for pretest items. Also, we have to make sure that the examinees answer the items earnestly.

The problem of item pretesting is not unusual for a CAT program. It is not sufficient to build a program; once it is established, it must also be maintained. And a major portion of the costs of maintenance is for constructing and pretesting new items. Therefore, in order to make the test economically viable, we will continue our search for more efficient methods of item-bank design and item construction and pretesting. Ultimately, this would be beneficial not only for the testing agency but also for its examinees.

Initial Proficiency Estimate

Figure 10.8 shows for three actual examinees how their proficiency estimate and standard error of measurement varied with the number of items answered. In each panel, the black dots indicate the proficiency estimates and the bars the standard errors. The examples are for the same early version of the CASEC, with the maximum of 30 items per section, that was used in the empirical study in Section 10.4.1.

The first panel shows a typical case. The initial proficiency estimate was close to the final estimate value and the test stopped after 23 items. The second panel shows a case in which the proficiency estimate converged rapidly. This tendency is often observed for Section 10.4 of the test, which generally has items of high discrimination power. The third panel shows a case in which the initial estimate is far away from the final estimate value. Therefore, the process required more items to reach the stability at the final estimate. As a result, the test took 29 items.

A comparison among these three cases indicates that an important determinant of the length of the test and, hence, of its accuracy is how far the initial proficiency estimate is from the examinee's true proficiency. One possible improvement would be to predict the initial proficiency estimate from a previous test score. Also, depending on the correlations between the proficiencies for the sections of the test, the initial estimate for one section could be based on the scores for the previous sections. JIEM is currently examining a number of strategies and plans to incorporate the most effective of them into the CASEC.

Choice of Prior Distribution

The CASEC employs the Bayes EAP method as a proficiency estimation method. The method requires specifying a prior distribution of examinee's θ in advance. One possible choice of prior distribution is an estimate of the proficiency distribution of the examinees in the target population.

Selecting a prior distribution is very important because a test cannot distinguish examinees' proficiencies adequately when an inappropriate informative prior distribution is employed. Because EAP scores will regress to the mean of the prior (shrinkage), a short test may not be sufficiently able to detect differences between examinees whose proficiencies correspond to the extreme ends of the scale. For instance, proficiency estimates of two examinees may be almost the same even though the true difference between their proficiencies was twice as much as the standard deviation of the prior distribution. So when examinees whose proficiencies are, say, $\theta = -8$ or $\theta = 8$ are assumed to take a test, a distribution such as the standard normal distribution should not used as a prior distribution for the test.

Some test developers or managers may feel obliged to enlarge the range of proficiency that the test can measure properly. However, even if they added many easier or more difficult items to the item bank, those items would hardly be administered to examinees as long as they used EAP estimators with informative priors for short test lengths. To solve this problem, they should individualize the prior distribution based on earlier information about the examinees.

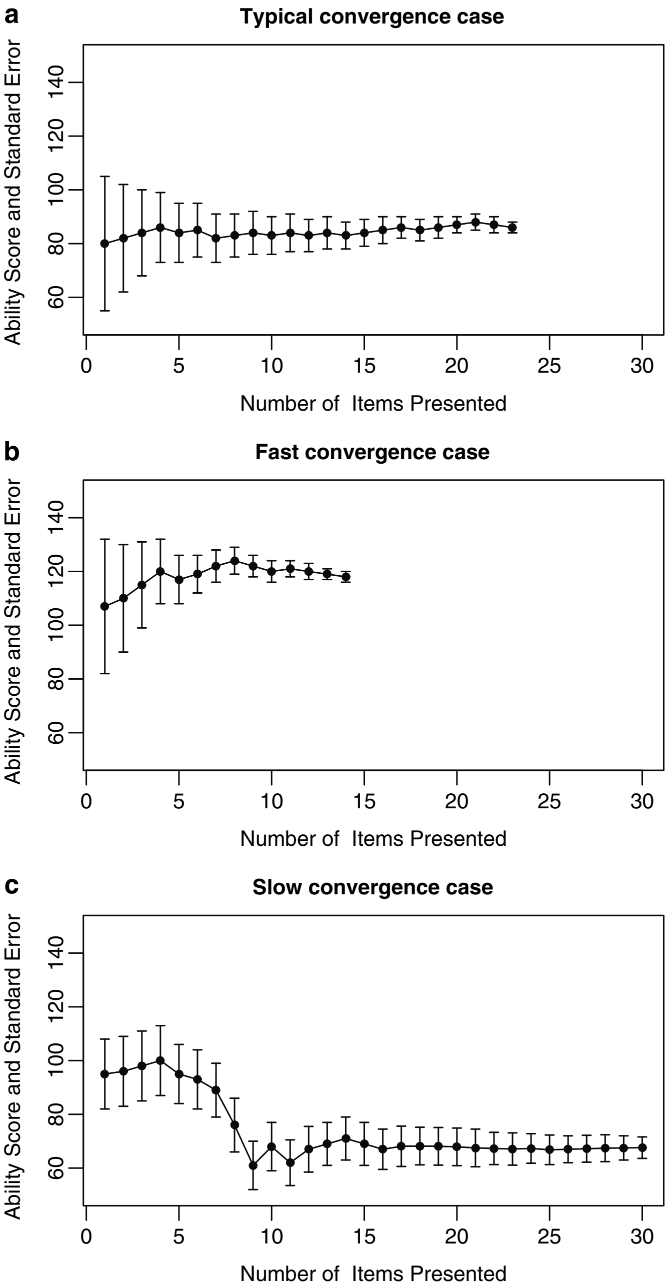


Fig. 10.8 Possible changes in the proficiency estimate and the standard error as a function of the number of items

10.5.2 Operational Issues

Cheating

Since the CASEC is delivered over the Internet, it is impossible to verify the identity of the examinees. This omission may stimulate cheating on the test or other unfair practices. The reason that the CASEC has nevertheless been used successfully under these conditions is largely due to the way it is used and implemented.

The test is taken both by individuals who use it to check their English proficiency and by organizations, such as schools and businesses, that use it for placement and educational purposes. When an individual takes the test at home, he or she usually wants to check his or her English proficiency or track it over time. In this case, the test is basically a low-stakes test and there is little merit in cheating.

When an organization wants to use the CASEC for its own purposes, the test may become high-stakes. It is then in the interest of the organization to check the identity of the examinees carefully and administer the test in a proctored environment.

Nevertheless, the possibility of cheating will be an important issue when the CASEC is used in new areas of application. Since the test is short, the risks of cheating will be extra high in high-stakes applications.

Item Bank Security

A test that uses the Internet basically exposes its items to many examinees. Compared to a group-based paper-and-pencil test, where the test booklets are collected at the end of the session, it is impossible to protect the CASEC 100% against every attack on its item bank. Since developing and managing a CAT system involves an enormous investment, the question of how to improve the protection of the item bank is very important. One possible strategy is not to use a single operational item bank but multiple banks that are periodically replaced. Also, it is important to determine when and how to refresh the items in these banks. This feasibility of this approach, which is described in more detail in Mills and Steffen (2000), is currently being examined.

System Stability

Computers and the Internet are not stable entities. For example, when all students at a school take the test at the same time, there is no guarantee that all computers are equally well maintained or that all students have the same computer skills. Also, computers may freeze without any warning or examinees may close the test window by mistake. The CASEC has been administered in computer rooms at numerous schools in the past, and interruptions due to system problems occurred 2% of the test administrations. Although not responsible for the hardware and Internet problems, it is important for JIEM to anticipate such problems and deal with them. If accidents occur during a test, all data up to the point of the accident are automatically saved. So when an examinee logs in again with the same ID and password, the test can resume from where it was interrupted. In fact, as a result of this, a CASEC examinee has never been unable to complete a test.

10.6 Conclusion

The CASEC is a CAT system developed in Japan to measure proficiency in English communication. The development of the system was supported by empirical research with trial versions of it. From the experiments it was found that the CASEC has (i) high measurement accuracy, even with fewer items than a traditional paper-and-pencil test format, (ii) very high reliability, and (iii) high concurrent validity in the form of correlations with other prominent tests of the same proficiency.

Many research issues have remained. In order to maintain the quality of the CASEC and extend its applicability, JIEM will continue its research agenda for the test.

References

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Chang, H.-H. & Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211–222.

Hayashi, N. (1998). *A research on a tree structure.* Unpublished, in Japanese.

Mills, C. N. & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* (pp.75–99). Boston: Kluwer-Nijhof Publishing.

Mislevy, R. J. & Bock, R. D. (1990). *PC-BILOG. Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351–356.

Thissen, D. & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer.* (pp.103–135). Hillsdale, NJ: Lawrence Erlbaum.

Woodcock, R. W. (1978). *Development and standardization of the Woodcook-Johnson Psycho-Educational Battery.* Hingham, MA: Teaching Resources Corporation.

Part III
Item Pool Development and Maintenance

Chapter 11
Innovative Items for Computerized Testing

Cynthia G. Parshall, J. Christine Harmes, Tim Davey, and Peter J. Pashley

11.1 Introduction

As computer-based testing (CBT) becomes a dominant, if not the dominant, medium for delivering assessments, interest in the potential of innovative items has grown. Innovative items are those that make use of features and functions of the computer to deliver assessments that do things not easily done in traditional paper-and-pencil assessments.

Innovative features that can be used by computer-administered items include sound, graphics, animation, and video. These can be incorporated into the item stem, response options, or both. Other innovations concern how items function. For example, examinees answering computerized items may highlight text, click on graphics, drag or move objects around the screen, or reorder a series of statements or pictures. The computer's ability to interact with examinees provides further possibilities. Items are not restricted to merely accepting a response. Instead, they can be designed to display content or provide new information, contingent on an examinee's actions. Finally, scoring algorithms have been developed that allow the computer to score items in which examinees generate, rather than simply select, their responses. This allows complex, performance-based tasks to be graded reliably and at minimal cost.

This chapter describes how innovative test items can make use of the computer's capabilities to improve measurement. Improvements can stem from innovations that enable tests either to measure more than they formerly did, or to measure it

C.G. Parshall (✉)
Measurement Consultant, 415 Dunedin Avenue, Temple Terrace, FL 33617, USA

J.C. Harmes
The Center for Assessment and Research Studies, James Madison University,
821 S. Main Street, MSC 6806, Harrisonburg, VA 22807, USA

T. Davey
Educational Testing Service, Rosedale Road, Princeton, NJ 08541, USA

P.J. Pashley
Law School Admission Council, P.O. Box 40, Newtown, PA 18940–0040, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_11,

better. The potential for improvement is real, but one caveat we offer is that if the innovation does not improve measurement of the construct in some way, it should not be used.

Our presentation is organized around a seven-faceted taxonomy for item innovation. Each facet of the taxonomy can be conceptualized as a continuum, ranging from less to more "innovative." However, this should be regarded as a description of innovations that *can* be developed rather than those that *should* be. For any given testing application, and any facet of the taxonomy, our recommendation is that the optimal level of innovation be targeted. In many cases, this optimal level will not be the "most innovative" level, but may be at another point along the continuum. A variety of other aspects of the exam program are likely to influence the determination of the appropriate target.

11.2 A Taxonomy for Innovative Items

Several ways of categorizing innovative items have been proposed (Koch, 1993; Harmes & Parshall, 2005; Luecht & Clauser, 2002; Parshall, Stewart & Ritter, 1996; Scalise & Gifford, 2006; Zenisky & Sireci, 2002). The first edition of this chapter (Parshall, Davey & Pashley, 2000) provided a comprehensive framework for innovative item types in terms of five dimensions: item format, response action, media inclusion, level of interactivity, and scoring method.

With the rapid advancement of technology has come an increasing number of options for innovation in CBTs. More elaborate structures for items and assessments have evolved, and greater sophistication has been introduced into CBT user interfaces and item functionality. To better capture assessments that are increasingly divergent from the traditional testing environment and that incorporate a broader array of innovations, a refinement to the taxonomy of testing innovations is necessary.

The revised taxonomy provided in this chapter is organized into seven dimensions: (1) assessment structure, (2) response action, (3) media inclusion, (4) interactivity, (5) complexity, (6) fidelity, and (7) scoring method. Assessment structure defines the structure of the item presentation and the kind of response collected from the examinee. These assessment structures range from selected response, through various forms of constructed response, and beyond. Response action refers to the means by which examinees provide their responses. Keyboard entry and mouse selection are common, while other input devices and user actions are less so. Media inclusion covers the use of elements such as graphics, sound, or video in an item. Interactivity describes the extent to which an item reacts or responds to examinee input. Complexity refers to the number and variety of elements examinees need to interpret and use in order to respond to an item. Fidelity considers the degree to which an item provides a realistic and accurate representation of the actual objects, situations, or tasks that are part of the construct being measured. Finally, scoring method addresses how examinee responses are translated into quantitative scores.

As will quickly become evident, these seven facets are far from independent in application. For example, including a video presentation in an item may also change the demands on examinees for responding to the item. The controls included for accessing and working with the video may require more innovative examinee response actions to correctly use them. Similarly, highly interactive items may require equally sophisticated scoring models. Although we attempt to differentiate the aspects of item innovation in the descriptions below, some of the items we present as examples of one sort of innovation complicate matters by being innovative in one or two other ways as well. However, each of the seven facets of the taxonomy relates to important decisions that test developers must make when designing and developing innovative items and their associated interfaces.

11.2.1 Assessment Structure

Conventional, paper-and-pencil tests generally make use of a limited number of item formats. Multiple-choice is clearly the most common, but the success of the format has spawned a number of variants. These include multiple-response, ordered response, and matching items. These items may be presented as discrete questions, or in sets in which all items relate to a common situation or stimulus. Formats not derived from multiple choice include fill-in-the-blank, short-answer, and essay. These require examinees to generate or construct rather than select their responses. Item types that require extensive construction may be less amenable to automated scoring; thus, these item types are less likely to be incorporated into a computerized adaptive test (CAT). Some of the most common item formats are detailed below.

Selected Response Items

The most familiar selected response item format is the multiple-choice; in this item type an examinee chooses an answer from a list of alternatives. Most often, the item consists of a question, or stem, and a set of anywhere from two to five possible responses. Computerized adaptations of this format can provide the potential to reduce guessing or to afford a more direct form of assessment. For example, items may ask examinees to click on and select the proper sentence from a reading passage, or to select one part of a complex graphic image. Because the number of available options can be much greater than the usual four or five, and can vary from item to item, the possibility of guessing correctly is substantially reduced.

Another selected response item type frequently adapted to CBTs is the multiple-response format. In this type of item, an examinee is asked to select more than one option; the examinee may be asked either to select a specified number of options, or to select “all that apply.” The ordered response item format is a further instance of selected response innovative item types. In this case, examinees are presented with a list of elements that they are then asked to place in the correct order or sequence. In a quantitative version of this item type, examinees may be asked to order numerical elements, perhaps from smallest to largest. In verbal versions of this item format, examinees may need to indicate the correct order of a series of events (O'Neill & Folk, 1996) or they may be asked to prioritize or rank a set of elements (Harmes & Parshall, 2000).

The "hot spot" or figural response item is an additional extension of the selected response item type. In figural response items, examinees respond by selecting a part of a figure or graphic. For example, an examinee may be asked to select a specific element or area within a spreadsheet, or a location on a diagram.

The primary goal in the design and use of each of these selected response item formats ought to be improved measurement. Innovative item types have the potential to improve measurement in various ways. Some of the formats may tap slightly different cognitive constructs than do traditional, text-based multiple-choice items. For example, the ordering and multiple-response types may change the examinee's cognitive task. Figural response items may improve measurement by eliminating a level of abstraction when they allow examinees to respond more directly with the material rather than referring to a lettered, indirect subset. Finally, these nontraditional formats may reduce the effect of guessing by expanding the range of possible responses.

Constructed Response Items

A wide range of constructed response items has also been considered, varying from fairly simple formats that are easy to score, to far more complex formats that require the use of elaborate scoring algorithms. The simplest examples are items that require the examinee to type a numerical answer to a quantitative question, or a very short response to a verbal question. Mathematical constructed responses may broaden the task by requiring examinees to enter formulas. The examinees' responses are scored by comparing each response to a list of acceptable answers that may include alternate mathematical formulations or acceptable misspellings.

Another constructed response item format extends the selected figural response type described above. The constructed version of this type allows examinees to mark on, assemble, or interact with various elements on the screen. In a typical example of this assessment structure, the examinees select from a set of reusable tools to draw an onscreen figure.

Requiring examinees to use verbal, mathematical, or figural elements to construct a response as opposed to merely selecting an answer represents an increase in cognitive challenge and can potentially result in acquiring a different type of information about the examinees' knowledge. Many of these innovative item types can also be incorporated within an adaptive computerized assessment. However, it should be noted that the development of these item types is likely to require additional effort in terms of item writing, programming, and possibly scoring.

Beyond Constructed Response

The descriptions of both selected response and constructed response items above referred to discrete items. However, an alternative approach is to present groups of items together within the structure of a single context. This type of structure may be created from sets of selected response items, various constructed response items, or a combination. In one approach, this might involve a situation or scenario in which examinees are asked to solve a problem, typically through a series of steps or items (Harmes & Parshall, 2005). Assessments structured in this way can provide a type of adaptivity within an item set or task (Zenisky & Sireci, 2002). Sets of items or tasks such as these can be designed to progress in either a structured or unstructured manner. A structured task will progress through the same steps in the same order for all examinees. An unstructured task will allow examinees to largely determine their own paths, based on any number of choices that may be made. This framework can be further extended to present representative items and tasks that use tools in an integrated context. Exam programs that are using these types of extended assessment structures include the AICPA (2004), the NCARB (Braun, 1994; Braun, Bejar & Williamson, 2006), and the USMLE (Melnick & Clauser, 2006). These more innovative assessment structures appear to have an exciting potential for expanding the construct validity of assessment, particularly through increasing construct representation (Huff & Sireci, 2001; Sireci & Zenisky, 2006). However, moving toward inclusion of these various types of integrated item sets or tasks may present significant challenges for the development of the assessment as well as the associated scoring method. Considerable effort is likely to be necessary to ensure that the extended assessments meet professional measurement standards.

11.2.2 Complexity

For an innovative assessment, complexity can be defined as comprising the number and variety of elements that an examinee needs to consider when responding to an item. This includes both conceptual and functional aspects, as an item may include both onscreen elements that need to be interpreted as well as item components that an examinee may use. For example, a complex innovative item might include informative text or graphics in several different locations on the screen, as well as various functional elements such as buttons, tabs, media players, or more.

Innovative items span a wide range of complexity. A low level of complexity is evident in a multiple-choice item with a simple, noninteractive graphic in the stem. This type of item would require little interpretation or inference beyond that required by a traditional item type. The difficulty of an item such as this is likely to be based almost entirely on the problem posed in the stem. Complexity is increased in an item type as additional visual elements are included on the screen. For example, text might appear in headers, labels, tabs, or item-specific instructions. Other types of visual elements include graphics, which may be either static or dynamic images. All of these forms of visual information need to be processed by the examinee, and thus affect the complexity of the item. The examinee's task is also made more complex as active or functional elements are added. The inclusion of a single functional component, such as a media player, might increase the complexity only slightly, but when numerous active elements are included, the task can become substantially more complex. The highest levels of complexity are perhaps evident in extended response assessments, as these tend to include numerous, varied screen elements that an examinee might need to interpret or use. An example of an innovative item with high complexity might be the AICPA's, 2004) forms completion task.

Complexity may tend to increase as other innovative dimensions of the item increase, particularly interactivity and fidelity (discussed below). Furthermore, increased item complexity may be associated with most contextualized, integrated assessments. It is important to note that in many cases increased complexity is likely to be associated with an increase in the item's cognitive challenge. From a measurement perspective, it is thus critical that the complexity of an item be construct-relevant. Finally, a spurious aspect of item complexity can also arise from an inappropriate complexity of the software's user interface. Just as we do not want traditional items to be made more difficult by "tricky" wording, so we should avoid artificial challenges in CBT items due to poor interface design or inadequate software usability (Parshall, Spray, Kalohn, & Davey, 2002).

11.2.3 Fidelity

In the context of this taxonomy, fidelity can be defined as the degree to which the assessment provides a realistic and accurate reproduction of the actual objects, situations, tasks, or environments that are part of the construct being measured (Harmes & Parshall, 2005). While this definition of fidelity would typically lead to face validity, it certainly goes beyond it. Instead of focusing just on examinees' perceptions of an item's merit, fidelity also incorporates the physical and functional correspondence of item elements to those in the target environment. Fidelity may relate to the way in which items or tasks are presented to examinees; it may also relate to the required response actions. An example of fidelity in the item presentation might include a video of a conversation (as opposed to a written transcript) or a detailed photograph of a complex machine (instead of a simple line drawing). An example of fidelity in the response action would be the inclusion of a simulated device such as a mannequin for use in testing medical skills.

Increasing the fidelity of an assessment often requires a greater investment of time and money in development. Various elements of an item or task may be evaluated as to their level of fidelity, and decisions are required on the part of the test developers regarding the appropriate target for the fidelity of these elements. Decisions about which level of fidelity to target will depend upon the purpose of the assessment. While a test in a computer-based flight simulator may provide enough information to differentiate between candidates for selection, a full-flight simulator with far greater fidelity would be necessary for qualifying a pilot to command an airplane.

Targeting a higher level of fidelity is not always recommended. Instead, test developers should carefully match fidelity levels to desired score inferences.

Developing an innovative assessment with a high degree of fidelity clearly has certain challenges. One potential challenge to the validity of an assessment is the risk of targeting a high-fidelity match to one environment, while providing a poor match to another environment. In some cases, increasing the level of fidelity may result in a reduction in control over certain elements, or unnecessary interference from elements occurring in a realistic setting (van der Linden, 2002). For example, including a high-fidelity audio clip of a patient breathing may include other sounds such as the heartbeat. By including this extra sound, the audio clip is more realistic; however, the additional aspects of the higher-fidelity sound could interfere with an examinee's ability to discern the essential breathing sounds. If it is not important that the examinee be able to undertake a task in the more difficult, albeit higher-fidelity, context, then increasing the level of fidelity at the expense of control is not appropriate. In this instance, if the assessment purpose is to demonstrate the ability to identify a particular breath sound in isolation (i.e, the more controlled situation), then a lower fidelity-audio clip would be appropriate. However, if the desired inference is whether or not an examinee can distinguish a particular breath sound within a realistic context of a physical assessment, then the higher-fidelity clip would be a better match.

As with the dimensions of complexity and interactivity, assessments with greater fidelity are likely to be more expensive to program. In addition, a high-fidelity assessment may require that specific computer hardware or simulated devices be available for test administration. It will be important for an exam program to target a useful level of fidelity, and not to divert program resources by unnecessarily exceeding that level. Fidelity may be an important component of an assessment and it may contribute to increasing the validity of an item or task. However, the relationship is not absolute: increasing fidelity does not necessarily increase validity.

11.2.4 Interactivity

Interactivity, as a facet of this taxonomy, describes the extent to which an item reacts or responds to examinee input. It does not refer to the adaptive nature of a CAT. Although interactivity can be incorporated into some types of discrete items, its primary use is with more elaborate assessment structures, particularly multiple-item sets.

The majority of innovative items are still discrete, single-step items. The examinee takes an action (e.g., makes a selection), and the item is complete. For many of these discrete innovative items, the only form of interactivity provided by the computer is a highlighted or shaded display of the response option selected by the examinee. At the next level of interactivity, a few item types provide a limited increase in item–examinee interaction. With these item types, the examinee acts and the computer then responds with some sort of reaction or information. Examples of this modest interactivity include ordered response items and constructed figural response items. When these item types are administered on the computer, a kind of contextual or informative feedback, along with a more direct means of responding to the material, can change the examinee's cognitive task.

More sophisticated use of interactivity typically occurs in assessment tasks that are situated within a representative context. This might include clicking tabs to access reference materials or clicking on elements within an item to view supplemental information. An example of moderate interactivity is evident in an assessment of research skills (Harmes & Parshall, 2000). In this test examinees order a set of article titles resulting from a literature search. They are also able to click on each article title to see a complete citation and abstract, in a manner that is similar to a real literature search. Moderate interactivity can also be seen in a test of conflict resolution skills (Olson-Buchanan, Drasgow, Moberg, Mead, Keenan & Donovan, 1998). In this assessment, after an examinee views a video displaying some type of workplace conflict, the response selected by the examinee branches to a specific additional video.

Higher levels of interactivity are characterized by an increase in the series of examinee actions and computer reactions. In the NCARB exam (Braun, 1994; Braun, Bejar & Williamson, 2006) examinees are presented with an architectural task and must use a palette of computerized drawing tools to design a solution to the problem within specified criteria. Another example of high interactivity comes from the NBME patient management tasks (Melnick & Clauser, 2006). Examinees are presented with a patient situation, and can order medical tests or procedures, receive and interpret the results of those procedures, diagnose the condition, and monitor changes in status over time and in response to actions taken.

There are numerous challenges inherent in developing interactive assessments. They require extensive design and development. A specific concern in the design of interactive assessments is determining the appropriate constraints on the examinee's possible actions. In particular, it is not beneficial for the interactive task to allow the examinee to proceed down a lengthy set of incorrect actions. Furthermore, highly interactive item designs can present significant scoring challenges.

11.2.5 Media Inclusion

Many innovative items are entirely text-based, providing innovation through assessment structure, interactivity, or automated scoring. However, a major advantage of administering tests via computer is the opportunity to include nontext media in the items. The appropriate use of these media can expand measurement of the construct, reduce unnecessary dependence on reading skills, and potentially enhance the validity of test scores.

Graphics

Graphics are the most common type of nontext media included in computerized tests. They are often used in innovative items such as the selected and constructed figural response types. While paper-and-pencil tests can also include graphics, they lack the computer's facility for interactivity. On computer, examinees may be able to rotate, resize, and zoom in or out of a scaled image, whether interacting with a graphical item stem or graphical response options. In the licensure exam for nurses, for example, candidates are presented with a graphic of a human torso, and respond by placing the stethoscope (mouse pointer) on the area appropriate for performing part of a cardiac assessment (NCSBN, 2005). The Teacher Technology Skills assessment (Harmes et al., 2004) includes items that provide graphics of computer applications, such as a web page viewed in a browser window. Examinees are asked to click on the area within the graphic that would be used to perform a specific action, such as returning to the previously visited page. In the graphical modeling items presented by Bennett, Morley & Quardt (1998), examinees respond by plotting points on a set of axes or grid, and then use either curve or line tools to connect the points. In a medical assessment, examinees are able to view high-resolution graphics, such as histopathology and other slides; the examinees are also able to pan across or zoom into these images to view them more closely (NBME, 2004). All of these examples contain visual elements that are highly content-relevant. There are broad-ranging content applications for the use of graphics, and this may be the easiest type of media to implement in a CBT. Many software programs will easily integrate and store graphics in a variety of file formats, their file sizes tend to be relatively small, and most examinees are comfortable with the inclusion of graphics.

Audio

Audio has been incorporated primarily into computerized tests of language skills and music, two content areas that have traditionally assessed listening skills. Audio may also find applications outside language and music (Parshall, 1999). As Vispoel, Wang, and Bleiler point out, "a substantial amount of general life experiences and academic activities involves the processing of information that comes to us through listening" (1997, p. 59). Tests of listening comprehension are important because the visual and audio channels of communication tap different cognitive processes. For example, there is evidence that multiple streams of information can be processed concurrently more easily and accurately when communicated aurally (Fitch & Kramer, 1994). There are clear advantages to administering audio in CBTs as compared to using cassette tapes with paper-and-pencil exams. In computer-based tests, the sound quality may be higher and examinees can typically control the volume, timing, and possibly even frequency at which the clips are played (Parshall & Balizet, 2001). While there are many potential applications of audio that could increase the validity of the assessment, there are also challenges to its use. One critical concern is that audio not be added in such a way as to create unnecessary or unfair disadvantages to hearing-impaired examinees. In addition, the logistical considerations of audio file type, storage requirements, and possible memorability need to be considered (Parshall & Balizet, 2001).

Video

Just as some conventional tests have long incorporated audio, so a few others have historically incorporated video. Video discs and video cassettes have been used in such areas as business and interpersonal interactions, medical diagnosis and treatment, and aircraft operations. Video incorporated within CBT has some definite technological advantages over these older media, including greater reliability and examinee control of timing and replaying.

Video can be incorporated into a CBT as item stimulus material for text-based responses, and may also be included in the response options or actions. A video-based test of conflict resolution skills was developed and validated by Olson-Buchanan, Drasgow, Moberg, Mead, Keenan & Donovan (1998). This test, which presents scenes of conflict in the workplace, also includes a level of interactivity in that an examinee's selection branches to the next video displayed. Additional research examples of video-based items are reported in Bennett et al. (1997).

Video appears to be a useful addition to an assessment when the construct relates to interpersonal communication or other aspects of human interaction. Furthermore, video has the capacity to display dynamic processes, such as moving pistons or a beating heart. While dynamic processes, or movement, may be displayed using either video or animation, video may be more appropriate when congruence with a "real-world" setting is important. Additionally, the proliferation of digital cameras and video editing software will mean that video is easier to obtain in many instances.

One rationale for the inclusion of full-motion video, as opposed to just audio, is that it adds the nonverbal component of communication. However, many of the logistical issues that apply to audio apply to video as well, perhaps to a greater degree. There are many possible file types, memory requirements are high, test security could be problematic, and production and editing costs may be quite high. In addition, as a stimulus type, video may have specific potential components that could contribute to construct-irrelevant variance. Examples of these elements include distracting features of the video setting, characteristics of the actors, and production elements such as camera angles and lighting.

Animation

Animation has the capacity to display dynamic processes, unlike the static medium of paper. Although minimal operational use of animation has yet been made, a few research examples can be found. Animated items were developed to assess middle and high school science standards. These items include animation of students conducting lab experiments (Chandler et al., 2006). Bennett et al. (1997) used a type of animation to display changes in national boundaries over time, by displaying a series of static maps in quick succession. Examinees responded by identifying the particular static map that answered a question. The researchers also developed a sample multimedia item that included an animated heart monitor trace, a static electrocardiogram strip, and an audio file of the related heart sound. Animation has a few potential advantages over video in certain applications. Animation uses far less computer memory than video to store or display, and in some cases may be less expensive to produce. On the other hand, with the relatively recent proliferation of video production and editing tools, there may be many instances in which video is actually more cost-effective. More substantively, animation is likely to be simpler; in some instances this could more specifically focus the examinee on essential aspects of the movement than a complex video might. For other applications, the realistic detail and contextual information inherent in video may be essential.

11.2.6 Response Action

While the assessment structure defines how the item is presented and what we ask the examinee to tell us, the response action defines how we ask them to do so. Thus, the term "response action" refers to both the physical action that an examinee makes to respond to an item and the input devices used.

The most common input device in traditional paper-and-pencil testing is the pencil, while the most common physical action required is bubbling in an oval. Computerized tests, on the other hand, most often use the keyboard and the mouse. Examinees respond through the keyboard by typing numbers, characters, and sometimes extended text. The mouse may be used for selecting onscreen elements such as a box associated with a chosen response. Examinees may also be asked to click on a graphic, on part of a graphic, or on part of a text passage. Potentially more challenging response actions would include using the mouse to drag icons to create or complete an image, or to drag text, numbers, or icons to indicate a correct sequence of events. Examinees may also need to use the mouse for purposes other than responding to questions. Examples of these uses include accessing computerized calculators or reference materials, playing audio or video files, or identifying a specific frame in an animation.

The specific response actions required of examinees can raise a number of issues. Most of these issues concern the characteristics of the CBT software's user interface. Vicino and Moreno point out that the user interface deserves serious attention when they state that, "reactions to computer tasks are largely dependent on the software interface" (1997, p. 158). Do examinees have the necessary computer skills or experience to read, interact with, and respond to items? Is the interface simple enough to be easily learned yet comprehensive enough to provide examinees the power to do all that is needed to efficiently process and respond to items? Are directions or tutorials both clear enough and detailed enough to impart the required skills? A superficial consideration of such issues might advocate computerized tests that use only the least common denominator of input devices and response actions. However, there may be good measurement reasons for requiring less common devices and actions. For particular applications, use of input devices such as touch screens, light pens, joysticks, or trackballs may benefit measurement. For example, young examinees and examinees with low literacy skills may be assessed with less error using touch screens or light pens. Particular skills, such as those that are highly movement-oriented, may be better measured using trackballs.

The input devices listed above are currently available, relatively prevalent, and fairly cheap. Furthermore, it is increasingly possible to utilize more advanced devices. For example, speech recognition software and microphones let us collect, and even score, spoken responses to oral questions.

The choice of input devices, along with the design of the software interface, shapes and defines the response actions available to an examinee. Bennett and Bejar (1998) have discussed the concept of "task constraints," which comprise the factors in the structure of a test that focus and limit examinee responses. Clearly, standardized paper-and-pencil testing is highly constrained. Constraints are also imposed on computerized tests, generally through the software interface and input devices. Task constraints affect not only the ways in which examinees can respond, but also the kinds of questions that can be asked and possibly even the ways in which examinees can think about these questions. The design and development of any assessment should include careful analysis of the task constraints that are included, to ensure that they are appropriate. For optimal measurement in a CBT, it will also be important to consider the implementation of the assessment within the software interface.

11.2.7 Scoring Methods

Many of the important practical benefits of computerized testing require that scoring of the assessments be automated. For example, tests can be adapted to examinees only when item responses are instantly scored by the computer. Furthermore, score reports can be issued immediately after testing only when test scores are determined by the computer. In part because these CBT features are so desired, considerable effort has been expended to develop automated scoring models for the great majority of online assessments (see Williamson, Mislevy, & Bejar, 2006).

Strategies for automated scoring of CBTs range from dichotomous through complex modeling. The dichotomous approach to scoring involves collapsing the information provided by the examinee's response into a score of correct or incorrect. Dichotomous scoring has been the primary method used in traditional testing, and is still used in most innovative assessments as well.

Some of the approaches to scoring innovative assessments make relatively modest attempts to extend beyond the dichotomous scoring model. These approaches are often incorporated into the design and development of the test itself. In determining the scoring approach to be used, test developers might consider questions such as how to define a correct response, whether to score on single or multiple outcomes, and whether the multiple outcomes should include such elements as aesthetics and efficiency. In one approach, various response options can be weighted for correctness so scores other than zero or one are possible. For a constructed response item, a weighting schema could be applied to the components of the response or to the steps within a task. For example, an IT certification exam might include simulated software that examinees use to complete a task. The simulated task could then be designed to allow for various types of responses, and for the collection of additional process information, such as time or number of steps taken. The score for the task might include both the correctness of the examinee's final response, as well as the efficiency of the process taken.

As assessments become more elaborate, more sophisticated scoring methods are often needed. Highly complicated assessments may include a larger set of acceptable variations in examinee responses than do constructed response items or simple situated tasks. A common approach to developing an automated scoring system involves identification and evaluation of the salient, measurable elements of the examinee's performance or product, followed by the development of a model for combining these elements into a score. Approaches to scoring using complex modeling include rule-based methods (e.g., the ARE exam; Bejar, 1991; Braun, Bejar & Williamson, 2006), regression-based methods (e.g., the USMLE exam; Margolis & Clauser, 2006), and the use of psychological task modeling and Bayesian networks (e.g., the DISC project; Mislevy et al., 2002). Furthermore, a variety of additional approaches have been developed for essay scoring (Shermis & Burstein, 2003), in which writing skills, rather than correctness of response, is usually being measured.

Assessment tasks should be designed based on the inferences that test developer or exam program owners wish to be able to make (Mislevy et al., 2002). To this end, the scoring procedures should be developed in conjunction with development of the assessment. This may help ensure that the assessment is designed in a way that will capture the evidence required in order to make the desired inferences.

Potential advantages of automated scoring (as opposed to scoring by human experts) include objectivity, reliability, and efficiency (Williamson et al., 1999). Furthermore, complex modeling approaches to scoring are that they have the potential to incorporate much more information about the examinees' thought processes or to evaluate a deeper level of detail in evaluating examinee products. However, complex modeling scoring methods also potentially have far greater challenges than simpler methods, in terms of developing, calculating, programming, financing, and communicating to stakeholders.

11.3 Conclusion

In this chapter, we have delineated seven dimensions along which innovative items might vary: (1) assessment structure, (2) complexity, (3) fidelity, (4) interactivity, (5) media inclusion, (6) response action, and (7) scoring method. Throughout this chapter we have attempted to be realistic about the amount of effort and resources that may be necessary to design and develop innovative items. Furthermore, we have recommended that exam programs select a type of innovation, and target the level of an innovative dimension, with the goal of supporting construct representation.

In general, as assessments become more innovative, the test development effort needed also increases. More elaborate assessments will typically take longer to develop, be more expensive to program, and require a more extensive validation effort. While the potential exists for increasing validity by adding innovative items or tasks that address content or construct areas previously unmeasured, there are certain measurement risks as well. When an assessment concentrates on greater depth of measurement, this sometimes comes at the cost of reducing breadth. More innovative assessment structures may also create item-writing challenges. Finally, the way in which the assessments are represented on the screen may have implications such as a dependence upon higher levels of computer skills than some examinees may have. A spuriously complex user interface can also contribute to construct irrelevant variance.

Clearly, the main purpose of testing is to measure proficiencies in valid, reliable, and efficient ways. Computerized assessments can provide highly interactive, media-rich, complex environments. They can also offer greater control over the testing process and provide increased measurement efficiency, especially when incorporated into a CAT environment. Nevertheless, it is important not to undertake innovations simply because they appear to be glitzy or cutting-edge. Innovation in and of itself does not ensure better measurement, nor is it equivalent to increasing validity.

While innovations such as those discussed in this chapter may be enticing, the primary objective should remain that of good measurement. Test developers should ask themselves, "Does the test cover a construct or content area that needs to be represented, and that can be measured best through the addition of innovative elements?" Efforts must be taken to ensure the appropriate match of technology to construct, and to design innovative assessments that fulfill the promise of better measurement.

References

AICPA. (2004, February). *AICPA, NASBA, and Prometric successfully pilot computer-based exam for CPAs.* Retrieved April 9, 2005, from http://www.aicpa.org/download/news/2004_02_02.pdf.

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76,* 522–532.

Bennett, R. E. & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues & Practice, 17,* 9–17.

Bennett, R. E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H. & Zack, J. (1997). *Using multimedia in large-scale computer-based testing programs* (Research Report No. RR-97-3). Princeton, NJ: Educational Testing Service.

Bennett, R. E., Morley, M. & Quardt, D. (1998, April). *Three response types for broadening the conception of mathematical problem solving in computerized-adaptive tests.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego.

Braun, H. (1994). Assessing technology in assessment. In E. A. Baker & H. F. O'Neil (Eds.), *Technology assessment in education and training* (pp. 231–246). Hillsdale, NJ: Lawrence Erlbaum Associates.

Braun, H., Bejar, I. I. & Williamson, D. M. (2006). Rule-based methods for automated scoring: Application in a licensing context. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 83–122). Mahwah, NJ: Lawrence Erlbaum Associates.

Chandler, L., Zimmerman, L., Castro, J. & Way, W. D. (2006). *Using technology to create innovative state science assessments: Pilots and policy.* Presentation at the Council of Chief State School Officers Annual Conference on Large-Scale Assessment, San Francisco, CA.

Drasgow, F., Olson-Buchanan, J. B. & Moberg, P. J. (1999). Development of an interactive video assessment: Trials and tribulations. In F. Drasgow & J. B. Olson-Buchanan, (Eds.), *Innovations in computerized assessment.* (pp 177–196). Mahwah, NJ: Lawrence Erlbaum Associates.

Fitch, W. T. & Kramer, G. (1994). Sonifying the Body Electric: Superiority of an auditory over a visual display in a complex, multivariate system. In G. Kramer (Ed.), *Auditory Display*, (pp. 307–325). Reading, MA: Addison-Wesley.

Harmes, J. C. & Parshall, C. G. (2000). *An iterative process for computerized test development: Integrating usability methods.* Paper presented at the annual meeting of the Florida Educational Research Association, Tallahassee, FL.

Harmes, J. C. & Parshall, C. G. (2005). *Situated tasks and simulated environments: A look into the future for innovative computerized assessment.* Paper presented at the annual meeting of the Florida Educational Research Association. Miami.

Harmes, J. C., Parshall, C. G., Rendina-Gobioff, G., Jones, P. K., Githens, M. & Dennard, A. (2004, November). *Integrating usability methods into the CBT development process: Case study of a technology literacy assessment.* Paper presented at the annual meeting of the Florida Educational Research Association, Tampa, FL.

Huff, K. L & Sireci, S. G. (2001, Fall). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20,* 16–25.

Koch, D. A. (1993). Testing goes graphical. *Journal of Interactive Instruction Development, 5,* 14–21.

Luecht, R. M. & Clauser, B. E. (2002). Test models for complex computer-based testing. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89–102). Mahwah, NJ: Lawrence Erlbaum Associates.

Margolis, M. J. & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–168). Mahwah, NJ: Lawrence Erlbaum Associates.

Melnick, D. E. & Clauser, B. E (2006). Computer based testing for professional licensing and certification of health professionals. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances* (pp. 163–186). West Sussex, England: Wiley.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G. & Johnson, L. (2002). Making sense of data from complex assessments. *Applied Measurement in Education, 15,* 363–389.

National Board of Medical Examiners. (2004, Fall/Winter). *Continuing developments in computer-based testing. NBME Examiner.* Retrieved May 9, 2005, from http://www.nbme.org/Examiners/fallwinter2004/news2.asp.

National Council of Architectural Registration Boards. (2004). *ARE Guidelines 3.0.* Retrieved April 9, 2005, from http://www.ncarb.org/are/Areguide.html.

National Council of State Boards of Nursing. (2005). *Fast facts about alternate item formats and the NCLEX examination.* Retrieved April 20, 2005, from http://www.ncsbn.org/pdfs/01_08_04_Alt_Itm.pdf.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., P. A. Keenan & M. A. Donovan (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24.

O'Neill, K. & Folk, V. (1996, April). *Innovative CBT item formats in a teacher licensing program.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Parshall, C. G. (1999, February). *Audio CBTs: Measuring more through the use of speech and non-speech sound.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Parshall, C. G. & Balizet, S. (2001). Audio computer-based tests (CBTs): An initial framework for the use of sound in computerized tests. *Educational Measurement: Issues and Practice, 20,* 5–15.

Parshall, C. G., Davey, T. & Pashley, P. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* (pp. 129–148). Boston: Kluwer-Nijhof Publishing.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Parshall, C. G., Stewart, R & Ritter, J. (1996, April). *Innovations: Sound, graphics, and alternative response modes.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Scalise, K. & Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing ”Intermediate Constraint” questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment, 4.* Retrieved January 29, 2007, from http://www.jtla.org.

Shea, J. A., Norcini, J. J., Baranowski, R. A., Langdon, L. O. & Popp, R. L. (1992). A comparison of video and print formats in the assessment of skill in interpreting cardiovascular motion studies. *Evaluation and the Health Professions, 15,* 325–340.

Shermis, M. D. & Burstein, J. (Eds.), (2003). *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sireci, S. G. & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representations. In S. M. Downing & T. M. Haladyna, (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Earlbaum Associates.

van der Linden, W. J. (2002). On complexity in CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89-102). Mahwah, NJ: Lawrence Erlbaum Associates.

Vicino, F. L. & Moreno, K. E. (1997). Human factors in the CAT system: A pilot study. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 157–160). Washington, DC: APA.

Vispoel, W. P., Wang, T. & Bleiler, T. (1997). Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. *Journal of Educational Measurement, 34,* 43–63.

Williamson, D. M., Bejar, I. I. & Hone, A. S. (1999). “Mental model” comparison of automated and human scoring. *Journal of Educational Measurement, 36* 158–184.

Williamson, D. M., Mislevy, R. J. & Bejar, I. I. (Eds.), (2006). *Automated scoring of complex tasks in computer-based testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Zenisky, A. L. & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337–362.

Chapter 12
Designing Item Pools for Adaptive Testing

Bernard P. Veldkamp and Wim J. van der Linden

12.1 Introduction

In existing adaptive testing programs, each successive item in the test is chosen to optimize an objective. Examples of well-known objectives are maximizing the information in the test at the ability estimate for the test taker or minimizing the deviation of its information from a target value at the estimate. In addition, item selection is required to realize a set of content specifications for the test. For example, item content may be required to follow a certain taxonomy or the answer-key distribution for the test must not deviate too much from uniformity. Content specifications are generally defined in terms of combinations of attributes the items in the test should have. They are typically realized by imposing a set of constraints on the item-selection process. The presence of both an objective and a set of constraints in adaptive testing leads to the notion of adaptive testing as constrained (sequential) optimization problem; for a more formal introduction to this notion, see van der Linden (this volume, chap. 2).

In addition to content constraints, item selection in adaptive testing is often also constrained with respect to the exposure rates of the items in the pool. These constraints are necessary to maintain item-pool security. Sympson and Hetter (1985) developed a probabilistic method for item-exposure control. In their method, after an item is selected, a probability experiment is run to determine whether or not the item is administered. By manipulating the probabilities in this experiment, the exposure rates of the items are kept below their bounds. Several modifications of this method have been developed (Stocking & Lewis, 1998, 2000), whereas different implementations of it are described in van der Linden (2003). van der Linden and Veldkamp (2004; 2007) propose an item-eligibility method for exposure control. This method realizes the desired exposure rates by imposing random eligibility

B.P. Veldkamp
Department of Research Methodology, Measurement, and Data Analysis,
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden (✉)
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_12,

constraints on the items in the test for each test taker with probabilities that are a function of the current exposure rates of the items. This method does not need any time-consuming simulation studies to set values for these probabilities; it can just be implemented on the fly.

Although these methods of item-exposure control guarantee upper bounds on the exposure rates of the items, they do not impose any lower bounds on them. In fact, practical experience with adaptive testing shows that item pools often have surprisingly large subsets of items that are seldom administered, the reason being poor contributions by these items to the objective function optimized by the item-selection algorithm or combinations of item attributes that are overrepresented in the pool relative to what is required to meet the constraints on the test. Since item production usually involves a long and costly process of writing, reviewing, and pretesting the items, the presence of unused items in the pool is an undesired waste of resources.

Adaptive testing algorithms could be developed to guarantee a lower bound on the exposure rates for the items in the pool as well (Revuelta & Ponsoda, 1998) but a different approach to over- or underexposure of items is trying to prevent the problem at all and *design* the item pool to guarantee a more uniform usage of all items for the population of test takers. It is the purpose of this chapter to propose a method of item-pool design that addresses this goal. The main result from the method is an optimal blueprint for the item pool, that is, a document specifying for each possible combination of item attributes how many items are needed.

Such blueprints could be used as a starting point for the item-writing process. As will be explained below, by using the authors of the items as a set of item attributes in the design process, blueprints can also be used to find an optimal division of labor among these authors. However, since some quantitative item attributes, in particular those that depend on statistical parameters estimated from empirical data, are difficult to realize exactly, a more realistic approach is to use the method proposed in this chapter as a tool for continuous management of the item-writing process. Repeated applications of it can then help to optimally adjust the next stage of the item-writing process to the part of the blueprint that has already been realized.

12.2 Review of Item-Pool-Design Literature

The topic of item-pool design has been addressed earlier in the literature, for pools for use with both adaptive and linear-form testing. A general description of the process of developing item pools for adaptive testing is given in Flaugher (1990). This author outlines several steps in the development of an item pool and discusses current practices at each of these steps. A common feature of the process described in Flaugher and the method in the present chapter is the use of computer simulation. However, in Flaugher's outline, computer simulation is used to evaluate the performance of an item pool once the items have been written and field-tested, whereas in the current chapter computer simulation is used to design an optimal blueprint for the item pool.

Methods of item-pool design for the assembly of linear test forms are presented in Boekkooi-Timminga (1991) and van der Linden, Veldkamp and Reese (2000). These methods, which are based on the technique of integer programming, can be used to optimize the design of item pools that have to support the assembly of a future series of test forms. The method in Boekkooi-Timminga uses a sequential approach to calculating the numbers of items needed for these test forms, each time maximizing their information functions. The method assumes an item pool calibrated under the one-parameter logistic (1PL) or Rasch model. On the other hand, the method in van der Linden, Veldkamp, and Reese directly calculates a blueprint for the entire pool, minimizing an estimate of the costs involved in producing the items. All other test specifications, including those related to the information functions of the test forms, are represented by constraints in the integer programming model that produces the blueprint. This method can be used for item pools calibrated under any current IRT model. As will become clear below, the current proposal shares some of its logic with the latter method. However, integer programming is not used for direct calculation of the numbers of items needed in the pool—only to simulate constrained adaptive testing.

Both Stocking & Lewis (1998), and Way, Steffen and Anderson (1998; see also Way & Steffen, 1998) address the problem of designing a system of rotating item pools for adaptive testing. This system assumes the presence of a master pool from which operational item pools are generated. A basic quantity is the number of operational pools each item should be included in (i.e., degree of item-pool overlap). By manipulating the number of pools for each of the items, their exposure rates can be controlled. A heuristic based on Swanson and Stocking's (1993) weighted deviation model (WDM) is then used to assemble the operational pools from the master pool such that they both realize the desired degree of overlap between the operational pools and are as similar as possible.

A different approach to the assembly of rotating item pools is proposed by Ariel, van der Linden, and Veldkamp (2004). Motivated by Gulliksen's (1950) matched-random-subtests method, these authors propose a method that divides a master pool into (possibly overlapping) smaller operational item pools that are required to have similar distributions of content and statistical attributes.

The problem of assembling an operational pool for adaptive testing has been approached from an entirely different angle in van der Linden, Ariel, and Veldkamp (2006). Their method is motivated by the idea that an optimal item pool would consist of a maximum number of combinations of items that (1) meet all content specifications for the test and (2) are informative at a series of ability levels reflecting the shape of the ability distribution of the population of test takers. The first condition is met when the operational pool is assembled as a set of linear test forms each with the same content specifications as for the adaptive test. The item-selection algorithm is then able to mix and match between these forms pool and in doing so has access to a much larger number of combinations of items meeting the content specifications. In order to meet the second condition, the set of linear test forms can be forced to have maximum information at a distribution of ability levels that approximates the shape of the ability distribution in the population of test takers.

The major difference between each of these methods and the method described in this chapter is that the former are methods for the *assembly* of an item pool from a mastery pool, whereas the current method is for the design of an item pool.

Recently, Belov and Armstrong (2005) proposed the use of a Monte Carlo test-assembly method to improve an existing item pool. Because of the random nature of the Monte Carlo test assembly, the frequency by which the method selects the items indicates how well its combination of attributes is represented in the item pool. (The more frequently an item is selected, the scarcer its combination of attributes in the pool.) The information can be used to instruct item authors to write new items. This method could also be conceived of as a method of item-pool design.

12.3 Designing a Blueprint for the Item Pool

The process of designing an optimal blueprint for an item pool for adaptive testing presented in this chapter involves the following three stages: First, the set of specifications for the adaptive test is analyzed and all item attributes figuring in the specifications are identified. The result of this stage is a definition of the design space for the test-assembly program. Second, a Monte Carlo method is used to simulate adaptive test administrations over the design space. Third, the optimal blueprint is derived from the number of times an item was sampled from each of the design points during the simulation study.

12.3.1 Identifying the Design Space

A design space D is defined as the Cartesian product of all item attributes. These attributes can be of different types: (i) categorical, (ii) quantitative, or (iii) logical (van der Linden, 2005). Categorical item attributes, such as content, format, or item author, partition an item pool into a collection of subsets. If the items in a testing program are coded by multiple categorical attributes, their Cartesian product induces a partitioning of the pool.

Classifications based on quantitative attributes are less straightforward to deal with. Several of them, such as item-difficulty parameters or expected response times, have a continuous range of possible values. Since it does not make much sense to use their full range, we partition the range into intervals of adjacent values that are represented by single values. The midpoints of the intervals are an obvious choice for these values.

Combining all categorical and discretized quantitative attributes, D could be thought of as a large multivariate table, where each of its cells represents a design point d with a different possible combination of attributes. An example of a design point is

$$d = (\text{content, answer key, } a_d, b_d, c_d, \text{ type of stimulus, etc.}), \tag{12.1}$$

where (a_d, b_d, c_d) are the values of the item parameters in the response model at design point d. In the empirical example, we used the three-parameter logistic (3PL) response model in (12.23) below. For convenience, and without any restriction of generality, we assume the same model to hold in our presentation of the method of item-pool-design method in the next sections of this chapter.

Logical attributes deal with the relations between the items in the pool. Enemy sets, for example, consist of items that are to be excluded from the same test because they clue each other. The appropriate place to deal with such constraints is not during the design of the item pool but during test assembly from it. Because of this, some logical attributes can be ignored in the item-pool-design process.

On the other hand, attributes that control an item-set structure in the pool have to be addressed when designing the pool. Items in a set address a common stimulus, for instance, a common passage in a reading comprehension test. If item sets occur, a separate design space E for the stimuli will be used. Analogously to D, its points e represent all possible combinations of the attributes used in the specifications for the stimuli in a test. The number of stimulus attributes is usually much smaller than the number of item attributes; for example, statistical attributes at stimulus level are rare. Thus, E is typically much smaller than D.

12.3.2 Simulation of Adaptive Test Administrations

A blueprint for an item pool is a distribution of the numbers of items for the pool over its design space. To estimate the optimal distribution, a Monte Carlo method is used to simulate adaptive test administrations for test takers $j = 1, \ldots, J$ with ability levels θ_j randomly drawn from the ability distribution of the population for which the testing program is planned. The test administrations are simulated over the design space instead of a real item pool; that is, rather than selecting existing items, the algorithm selects design points. Each time a design point d is selected, an item with all of its attributes is assumed to be administered. The simulation is based on the shadow-test approach (STA) to adaptive testing, which allows us to impose each of the specifications for the test as an explicit constraint on the selection of the items during test administration (van der Linden, this volume, chap. 2).

Cost Function

A criterion of optimality is needed to calculate an optimal blueprint. An obvious candidate is minimization of the costs involved in the production of the items. Because the costs of field testing and calibration can be assumed to be equal for each item, basically the relevant costs are those of item writing. If direct estimates of the costs of writing items with the combinations of attributes at the design points are available, they should be used. But typically they are not and have to be approximated. A useful proxy recommended in van der Linden, Veldkamp and Reese (2000) is based on the assumption that items with combinations of attributes that occur frequently in previous item pools for the same program are relatively easy to produce and hence involve lower costs. In other words, the costs of writing the items were assumed to be inversely related to the frequency of the combinations of their attributes in a representative earlier version of the item pool. This cost function was used in the empirical example at the end of this chapter. Because the function is used in an optimization problem, it need not be specified beyond being monotonically decreasing in the frequencies of the items; any monotonic transformation of it will produce the same solution to the optimization problem.

When the design space is high-dimensional, the previous pool has to be large to obtain stable frequency estimates. In order to increase the stability of the estimates, Ariel, van der Linden, and Veldkamp (2006) recommend smoothing the cost estimates over the quantitative attributes of the design space. A useful smoothing method is k-nearest-neighbor regression, which replaces the estimate at point d by the average found in a small neighborhood of it.

Shadow-Test Model

The shadow-test approach (STA) for the general case of an item pool with item sets is addressed here. Therefore, the test-assembly model for simulating adaptive test administrations from this type of pool has separate variables at the stimulus and item level. At stimulus level, variable z_e is for the number of stimuli required at design point e, while x_{d_e} is for the number of items at design point d for stimulus e. Variables z_e and x_{d_e} are integer variables not restricted to 0-1 values as in regular adaptive testing because more than one item or stimulus with the same combinations of attributes can figure in the same test.

Each item in the adaptive test simulations is selected in two steps. First, a shadow test is assembled. Shadow tests are full-length tests optimal at the current ability estimate that meet all test specifications and contain all items already administered to the test taker. Second, the next item is selected to be the best among the active design points in the shadow test for the current stimulus. To be more precise, let $e^{(l)}$ be the point at which the current stimulus was chosen. This point remains active until a shadow test with a lower value for its decision variable $z_{e^{(l)}}$ occurs. As long as the stimulus point is active, the item with the smallest cost contribution to the objective function in the model among the active item points for the stimulus (i.e., for which $x_{d_{e^{(l)}}} \geq 1$) is selected.

The model allows for specifications for the adaptive test at test level, stimulus level, item-set level, and item level. Also, we now assume the variables z_e and x_{d_e} represent the free items in the shadow test only and use counters $\eta_e^{(j)}$ and $\eta_{d_e}^{(j)}$ for the numbers of items at design points d and e that have already been administered up to the current test taker, j. In addition, $\widehat{\theta}_j^{(k-1)}$ is the estimate of ability parameter θ after $k-1$ items have been administered to test taker j. Further, $I_d\left(\widehat{\theta}_j^{(k-1)}\right)$ is the (Fisher) information about $\theta = \widehat{\theta}^{(k-1)}$ in the candidate item at design point d.

The information is known because the design points are assumed to include the item parameters (a_d, b_d, c_d) in the 3PL response model used to calibrate the items; see (12.1). Categorical attributes partition the design spaces into collections of sets V_c^{item} and V_c^{stim} for the items and stimuli, respectively. Each of these sets contains the items or stimuli with a different categorical attribute (for instance, a different content category). The items and stimuli are assumed to have general quantitative attributes q_d and q_e (for instance, expected response times on the items or word counts for the stimuli). To allow for the exclusion or inclusion of items or stimuli with special combinations of attributes, we denote the design points with these combinations as V_0 and V_1, respectively, with an appropriate label to indicate the items or the stimuli. Finally, the notation used for the bounds to be imposed on each of the item or stimulus attributes is self-explanatory.

The standard form of the model for the selection of the kth item for the jth simulated test taker is

$$\min \sum_{e=1}^{E} \varphi_e z_e + \sum_{e=1}^{E} \sum_{d=1}^{D} \varphi_{d_e} x_{d_e}, \qquad \text{(minimize costs)} \tag{12.2}$$

subject to possible constraints at the following levels:

Test Level

$$\sum_{e=1}^{E} \sum_{d=1}^{D} I_d\left(\widehat{\theta}_j^{(k-1)}\right)(\eta_{d_e}^{(j,k-1)} + x_{d_e}) \geq T\left(\widehat{\theta}_j^{(k-1)}\right), \quad \text{(test information)} \tag{12.3}$$

$$\sum_{e=1}^{E} \sum_{d=1}^{D} \left(\eta_{d_e}^{(j,k-1)} + x_{d_e}\right) = n, \qquad \text{(test length)} \tag{12.4}$$

$$\sum_{e=1}^{E} \left(\eta_e^{(j,k-1)} + z_e\right) = m, \qquad \text{(number of stimuli)} \tag{12.5}$$

$$\sum_{e=1}^{E} \sum_{d \in V_c^{\text{item}}} \left(\eta_{d_e}^{(j,k-1)} + x_{d_e}\right) \gtreqless n_c^{\text{item}}, \text{ (categorical attributes)} \tag{12.6}$$

$$\sum_{e=1}^{E} \sum_{d=1}^{D} q_d \left(\eta_{d_e}^{(j,k-1)} + x_{d_e}\right) \gtreqless b_q^{\text{item}}, \text{ (quantitative attributes)} \tag{12.7}$$

$$\sum_{e \in V_c^{\text{stim}}}^{E} \left(\eta_e^{(j,k-1)} + z_e\right) \gtreqless n_c^{\text{stim}}, \text{ (categorical attributes)} \tag{12.8}$$

$$\sum_{e=1}^{E} q_e \left(\eta_e^{(j,k-1)} + z_e\right) \gtreqless b_q^{\text{stim}}, \text{ (quantitative attributes)} \tag{12.9}$$

Item-Set Level

$$\sum_{d=1}^{D}\left(\eta_{d_e}^{(j,k-1)}+x_{d_e}\right)\gtreqless n^{\text{set}}z_e,\ \text{for all } e, \tag{12.10}$$

(number of items per set)

$$\sum_{d\in V_c^{\text{item}}}\left(\eta_{d_e}^{(j,k-1)}+x_{d_e}\right)\gtreqless n_c^{\text{set}}z_e,\ \text{for all } e, \tag{12.11}$$

(categorical attributes)

$$\sum_{d=1}^{D}q_d\left(\eta_{d_e}^{(j,k-1)}+x_{d_e}\right)\gtreqless b_q^{\text{set}}z_e,\ \text{for all } e, \tag{12.12}$$

(quantitative attributes)

Stimulus Level

$$\sum_{e\in V_1^{\text{stim}}}^{E}\left(\eta_{e}^{(j,k-1)}+z_e\right)=n_1^{\text{stim}}, \tag{12.13}$$

(special combination of attributes)

$$\sum_{e\in V_0^{\text{stim}}}^{E}\left(\eta_{e}^{(j,k-1)}+z_e\right)=0, \tag{12.14}$$

(special combination of attributes)

Item Level

$$\sum_{d\in V_1^{\text{item}}}\left(\eta_{d_e}^{(j,k-1)}+x_{d_e}\right)\gtreqless n_1^{\text{item}}, \tag{12.15}$$

(special combination of attributes)

$$\sum_{d\in V_0^{\text{item}}}\left(\eta_{d_e}^{(j,k-1)}+x_{d_e}\right)\gtreqless n_0^{\text{item}}, \tag{12.16}$$

(special combination of attributes)

Definition of Variables

$$x_{d_e}\in\{0,1,\ldots\},\qquad \text{for all } d \text{ and } e,\qquad \text{(range of variables)} \tag{12.17}$$

$$z_e\in\{0,1,\ldots\},\qquad \text{for all } e.\qquad \text{(range of variables)} \tag{12.18}$$

This model only has a standard set of constraints for the adaptive test. For specific applications, several versions of the same types of constraints or entirely different constraints may be needed. A complete overview of all possible types of constraints is offered in van der Linden (2005). When the item pool does not have a set structure, a less complicated version of the model can be used. The first term of the objective function in (12.2) should then be deleted and the constraints in (12.8)–(12.14) are no longer required. The remaining model has only the item variables x_d.

The model in (12.2)–(12.18) is linear in the decision variables and can be solved using a standard integer solver, e.g., the one in CPLEX 9.0 (ILOG, 2003). A solution is a string of integer values for the decision variables z_e and x_{d_e}. As already indicated, the item that is selected is the one at the point in the design space for the items with $x_{d_e} \geq 1$ for the current stimulus that has the smallest value for the cost function φ_{d_e} in (12.2). After the item has been administered, both the ability estimate and the counters for the design points in the model are updated, and a new solution is calculated.

Blueprint for the Item Pool

For every simulated test taker j, the numbers $\eta_{d_e}^{(j,n)}$ and $\eta_e^{(j,n)}$ denote how many items at design points d_e and e were administered. These counts enable us to calculate the following numbers:

$$N_e = \sum_{j=1}^{J} \eta_e^{(j,n)}, \tag{12.19}$$

and

$$N_{d_e} = \sum_{j=1}^{J} \eta_{d_e}^{(j,n)}. \tag{12.20}$$

These numbers define the blueprint for the item pool as the combination of

$$(N_1, \ldots, N_E) \tag{12.21}$$

and

$$\begin{pmatrix} N_{1_1} & \cdots & N_{1_E} \\ \cdots & \cdots & \cdots \\ N_{D_1} & \cdots & N_{D_E} \end{pmatrix}. \tag{12.22}$$

For every design point d_e and e, these arrays describe how many items are needed. From the definitions of these points we know exactly what combination of attributes its items should have. This information is all we need to instruct the item writers.

12.4 Empirical Example

To illustrate the use of this method of item-bank design, an application to an adaptive version of one of the sections of the Law School Admission Test (LSAT) is presented. The section consists of items organized around common stimuli. Besides, the section has to meet several constraints both at item level and stimulus level. A previous item pool of 1,508 items for this section of the LSAT was available and could be used to define the cost function in the example.

12.4.1 Design Space

As the items in the pool had common stimuli, there were two design spaces, D and E.

At stimulus level, we had one categorical (content) and one quantitative attribute (word count). The categorical attribute had three possible values. The quantitative attribute had to be categorized. In the previous item pool, word count ranged from 58 to 182; this range was split into four intervals: (0–75], (75–100], (100–125], and (125–∞). As a result, design space E corresponded to a table with 12 cells.

At item level, two categorical attributes were relevant (content and answer key). Five different content classifications and five possible answer keys were distinguished.

All items were calibrated under the three-parameter logistic (3PL) model:

$$P_i(\theta_j) \equiv c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}, \tag{12.23}$$

where $P_i(\theta_j)$ is the probability of $j = 1, \ldots, J$ with an ability θ_j giving a correct response to an item $i = 1, \ldots, I$, a_i is the discrimination parameter, b_i the difficulty parameter, and c_i the guessing parameter of item i. The item parameters in the model were the quantitative item attributes. The range of values for the discrimination parameter is $[0, \infty)$. The range was split into four intervals, with the fourth interval extending to infinity. The difficulty parameters take values in $(-\infty, \infty)$. Likewise, this range was divided into six subintervals. Finally, in this previous item pool, all items had approximately the same value for the guessing parameter. Therefore, in the simulation c_i was fixed at the average value. The product of the categorical and quantitative attributes resulted in design space D with 600 cells.

Model for the Shadow Tests

The actual specifications for the LSAT section were used to formulate the version of the integer programming model in (12.2)–(12.18) for the shadow tests in this adaptive testing simulation. The model had 23 constraints dealing with the various attributes. As we had no absolute target for the test information function in (12.3), the objective function was chosen to be the following linear combination of test information and item-writing costs:

$$\max\left\{\lambda \sum_{d_e \in D_E} I_{d_e}\left(\widehat{\theta}_{k-1}\right) x_{d_e} - (1-\lambda)\left[\sum_{e \in E} k_e z_e + \sum_{d_e \in D_E} k_{d_e} x_{d_e}\right]\right\}, \tag{12.24}$$

where $I_{d_e}(\widehat{\theta}_{k-1})$ is the information in an item at design point d_e at $\theta = \widehat{\theta}_{k-1}$ and k_e and k_{d_e} were the reciprocals of the frequencies of the items in the previous item pool at design points e and d_e used as proxies of the item-writing costs. For points with zero frequencies, an arbitrary large number was chosen (see the earlier argument about the monotonicity of objective functions in optimization problems).

Simulation Study

The test takers were sampled from $N(0, 1)$. The simulations were executed using software for constrained adaptive testing with shadow tests developed at the University of Twente. The integer programming models for the shadow tests were solved using calls to the linear-programming software package CPLEX 9.0 (ILOG, 2003). The initial ability estimate of each new simulee was set at $\widehat{\theta} = 0$. The estimate was updated using expected a posteriori (EAP) estimation with a uniform prior distribution of θ.

Calculation of the Blueprint

The percentages of the numbers of times an item or stimulus attribute was hit in the simulation study are reported in Tables 12.1–12.4. Together, the percentages define the blueprint for the optimal item pool for the adaptive version of the LSAT section. For design space D, only 157 of the 7,200 possible cells were hit. The number of combinations of attributes actually needed for this test was thus much smaller than the Cartesian product of all attributes would suggest. This reduction shows the gain

Table 12.1 Counts of categorical stimulus attributes

Set type	I	II	III
Counts	8606	9022	8372

Table 12.2 Counts of quantitative stimulus attributes

Word count	0–75	75–100	100–125	>125
Counts	0	13690	11878	432

Table 12.3 Counts of categorical item attributes

Item type	1	2	3	4	5
Counts	4232	2728	4386	6919	7735

Table 12.4 Counts of quantitative item attributes

	Difficulty					
Discrimination	$(-\infty, -1)$	$(-1, -0.5)$	$(-0.5, 0)$	$(0, 0.5)$	$(0.5, 1)$	$(1, \infty)$
$(1, \infty)$	234	1820	10608	6162	1560	1170
$(0.75, 1)$	572	1170	910	936	208	390
$(0.5, 0.75)$	104	130	26	0	0	0
$(0, 0.5)$	0	0	0	0	0	0

in focus due to previous optimization of the blueprint for the item pool design relative to "blind item writing". Also, a small correlation ($\rho = 0.192$) between the difficulty and discrimination parameters in the blueprint was found. Not surprisingly, the most remarkable difference between the blueprint and the distribution of the items in the previous item pool was with respect to the discrimination parameters. In the previous item pool, only 8.1% of the items fell in the highest category, whereas 82.9% of the blueprint fell in this category.

12.5 Some Related Issues

12.5.1 Exposure Control

In the method described above, the numbers in the blueprint are based on observed frequencies in simulated test administrations. The expected exposure rates of the items in the blueprint are equal to the counts divided by the number of simulated test takers. In the preceding examples, no correction was made for a maximum exposure rate that might have been imposed on the items in the adaptive test if this had been a real-world test.

Such a correction would, however, simply consist of dividing the numbers in (12.20) by their maximum exposure rate $r^{\max}$. The blueprint for the item pool is then defined by

$$\widetilde{N}_{d_e} = \left\lceil \frac{N_{d_e}}{r^{\max}} \right\rceil, \tag{12.25}$$

with upward rounding of the resulting numbers to their nearest integer value. When the exposure rates of the stimuli also have to be restricted, a comparable adjustment should be made to the numbers in (12.19).

12.5.2 Rotating Item Pools (Calculating the Blueprint)

When the method is used to design a system of rotating item pools from a master pool, the numbers in (12.19) and (12.20) have to be adjusted slightly. A master pool can be viewed as a combination of individual item pools, whereby overlap between pools is allowed. Let G be the number of parallel item pools the master pool has to support, and n_g the number of overlapping pools in which an individual item can be present. The number of items required for the master pool can be calculated by multiplying the numbers in (12.19) and (12.20) by G/n_g. The blueprint for a master pool then consists of

$$\widetilde{N}_{d_e} = \left\lceil N_{d_e} \frac{G}{n_g} \right\rceil \tag{12.26}$$

and

$$\widetilde{N}_e = \left\lceil N_e \frac{G}{n_g} \right\rceil, \tag{12.27}$$

where, again, the resulting numbers have to be rounded upward to their nearest integer value.

12.5.3 Multidimensionality

When items in the pool are calibrated under a multidimensional IRT model, the constraint on Fisher's information in the design model in (12.2)–(12.18) needs to be modified. In Veldkamp and van der Linden (2002), the Kullback–Leibler information was used to replace the Fisher information in the case of multidimensionality. One of the advantages of Kullback–Leibler information is that it remains a scalar for a multidimensional ability parameter whereas the Fisher information becomes a matrix. Alternative information measures for item selection in multidimensional adaptive testing with the same feature are examined in Mulder and van der Linden (this volume, chap. 4).

12.6 Concluding Remarks

The method presented in this chapter produces a blueprint for an item pool that should be used as a guide for the item-writing process. The first type of guidance consists of preparing the instructions for the item writers. If the previous item pool was written by the same item writers and costs estimates are available for them, we should use their identity as one of the attributes for the design space. The blueprint then optimally assigns item blueprints to the item writers.

Typically, both the categorical item attributes as well as some of the quantitative attributes (e.g., word counts) can easily be realized during item writing. However, as already discussed, some of the other quantitative attributes, in particular those of a more statistical nature, are more difficult to realize. If an existing item pool is used to estimate the item-writing costs, the blueprint for the item pool automatically builds on the empirical correlations between such statistical attributes and all other attributes. For example, if the more difficult items tend to have other categorical attributes, the optimal blueprint automatically accounts for this fact.

Although this feature may improve the results of item writing, exact realization of statistical item attributes remains an optimistic goal. The best way to implement the blueprint is, therefore, not as a one-shot approach but in a sequential fashion, recalculating the blueprint after a certain portion of the items has been written and field-tested so that their actual attribute values are known. Repeated applications of the method help to adapt the item-writing efforts to the distribution of the items already present in the pool (van der Linden, 2005; van der Linden, Veldkamp and Reese, 2000).

References

Ariel, A., van der Linden, W. J. & Veldkamp, B. P. (2006). A strategy for optimizing item pool management. *Journal of Educational Measurement, 43*, 85–96.

Ariel, A., Veldkamp, B. P. & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41*, 345–359.

Belov, D. I. & Armstrong, R. D. (2005). Monte Carlo test assembly for item pool analysis and extension. *Applied Psychological Measurement, 29*, 239–261.

Boekkooi-Timminga, E. (1991). *A method for designing Rasch model based item banks*. Paper presented at the annual meeting of the Psychometric Society, Princeton NJ.

Flaugher, R. (1990). Item pools. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 41–64). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

ILOG, Inc. (2003). *CPLEX 9.0* [Computer program]. Incline Village, NV: Author.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311–327.

Stocking, M. L. & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *23*, 57–75.

Stocking, M. L. & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston: Kluwer-Nijhof Publishing.

Stocking, M. L. & Swanson, L. (1998). Optimal design of item banks for computerized adaptive testing. *Applied Psychological Measurement, 22*, 271–279.

Swanson, L. & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151–166.

Sympson, J. B. & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–77). San Diego: Navy Personnel Research and Development Center.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249–265.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer-Verlag.

van der Linden, W. J., Ariel, A. & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics*, *31*, 81–100.

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

van der Linden, W. J. & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273–291.

van der Linden, W. J. & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398–418.

van der Linden, W. J., Veldkamp, B. P. & Reese, L. M. (2000). An programming approach to item bank design. *Applied Psychological Measurement*, *24*, 139–150.

Veldkamp, B. P. & van der Linden, W. J. (2002). Multidimensional constrained adaptive testing. *Psychometrika, 67*, 575–588.

Way, W. D. & Steffen, M. (1998, April). *Strategies for managing item pools to maximize item security*. Paper presented at Annual Meeting of the National Council on Educational Measurement, San Diego.

Way, W. D., Steffen, M. & Anderson, G. S. (1998). Developing, maintaining, and renewing the item inventory to support computer-based testing. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89–102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chapter 13
Assembling an Inventory of Multistage Adaptive Testing Systems

Krista Breithaupt, Adelaide A. Ariel, and Donovan R. Hare

13.1 Introduction

There exists a natural tension between the goal of creating a large enough item bank to preserve the equivalency and security of test questions and that of cost reduction and efficiency for inventory creation.

The security of many high-stakes testing programs depends on ensuring a sufficiently large bank of test content and replenishing that inventory with new test content over time. In testing programs where the test content is used multiple times before being disclosed or retired, the validity of score interpretations is at risk when test content becomes overused. At the same time, the decisions made from test score use also depend on ensuring consistently high quality and equivalence of test forms throughout the test-administration timeline.

As test use continues to expand globally, and computerization allows for greater flexibility in scheduling and test-administration designs, our need to find a solution to these competing goals has become acute. Inventory assembly solutions that support the creation of equivalent and valid tests within administrations and over time have become strategic business planning tools, necessary for item-bank management in any large operational testing program.

Item-bank quality fluctuates naturally over time as a result of traditional item-writing practices, where experts are given broad instructions for writing test questions and a general description of topic areas. As a result of a fairly unstructured content development process, item replacements will range in quality and equivalency (Way, Steffen & Anderson, 1998). Most often, the equivalency of forms is

K. Breithaupt (✉)
American Institute of Certified Public Accountants, 1230 Corporate Parkway Avenue,
Ewing, NJ 08628–3018, USA

A.A. Ariel
Department of Research Methodology, Measurement, and Data Analysis,
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

D.R. Hare
Department of Mathematics & Statistics, University of British Columbia Okanagan,
3333 University Way, Kelowna, BC V1V 1V7, Canada

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics
for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_13,

managed in the assembly or scoring process, using statistical or other information as design rules. This approach fails to address the inventory problem at its source. For example, it is common to deal with item shortages in the operational bank by imposing item-exposure controls during the test-administration process (e.g., Chang & Ying, 1999; Stocking & Swanson, 1993; Sympson & Hetter, 1985; van der Linden, 2005). Although such item-exposure controls can be effective for limiting item exposure at administration time, they are a short-term solution. Ultimately, administration and assembly design rules do not improve the equivalency, quality, and uniformity of item use in the item bank.

An obvious resolution to our security and quality problems for multiple test forms is to create a very large item bank and replenish it often so that there is always an ample variety of appropriate high-quality test questions. This becomes costly as item authoring, reviewing, pretesting, and analysis require much time and effort. In order to make this expenditure as efficient as possible, careful item-bank design and deliberate inventory-development scheduling are prerequisite. Empirical research suggests that item bank maintenance based only on continually adding a number of items is not sufficient to maintain equivalent item-bank quality over time (Ariel, van der Linden & Veldkamp, 2006). In their simulation study, Ariel et al. emphasize the importance of putting in place an optimal planning strategy for inventory maintenance. Adequate strategies incorporate a comprehensive knowledge of properties for all required items in the bank, and consider expected attrition rates due to variances in item quality and retention.

A comprehensive set of alternative models for item-bank designs can be found in van der Linden (2005, chaps. 10–11). These models consider creation of forms, administration scheduling, and overlap, and make use of integer programming solutions to calculate item-bank blueprints that minimize costs and restrict item exposures while maintaining desirable test quality. Illustrations and examples are also provided to take into account the administration format of the test as linear or adaptive.

Veldkamp and van der Linden (this volume, chap. 12) illustrate the use of an item-bank blueprint to identify shortfalls in an available inventory, and how this blueprint might be used to guide item-development planning. Their analysis of the required inventory is based on the assumption that items in short supply are relatively more costly to produce (or more difficult to write), and they use a previous pool to calculate such cost estimates. As a result, the inventory-development schedule avoids overuse of rare items to preserve longer-term viability and equivalent quality of test forms over time.

In this chapter, we will integrate item-bank design and inventory-development schedules through an extension of mixed-integer programming and optimization techniques. In this inventory system, time periods are explicitly modeled for item development and administration, based on assumptions drawn from the administration format, composition, and scheduling for the test. This system illustrates how closely connecting test design, administration models, and item-bank development goals can guard test security and the validity of decisions made from test use.

13.2 Mixed-Integer Programming Concepts

A short review of some fundamental concepts and some applications of integer programming and optimization methods to schedule inventory and supply problems might provide a useful framework for our illustration.

There are many industrial problems whose solutions require a group of discrete choices to be made. In production industries, these choices might take the form of the number of widgets of a given type that should be made and a schedule for the group of machines that make them. The choices usually have natural dependences that constrain the ideal solution. Perhaps there is an order for some of the machines that build a type of widget or a time delay for a machine to paint widgets differing colors. These situations are analogous in many respects to our test construction problem.

In the process required to build a test form of traditional multiple-choice questions (MCQs), a viable solution will require us to choose a number of questions from a bank of potential questions. Selection onto test forms is ordinarily constrained by content specifications and other design or business rules, such as form length and item-exposure restrictions. In the case that there are many forms to create, our objective is to choose items for forms so that the total solution of all tests created is optimal with respect to the most important design goal. That is, the set of forms will be optimal, given the items available and all the design constraints defined in the problem.

This selection of items onto test forms has important implications for inventory planning, and a solution can be defined for the required item development in a very similar way. In operational planning for item banking, we may require that the supply schedule has the earliest end time to get new items into test forms (i.e., shortest makespan). Or, when building a group of subtests, it may be desirable to maximize some function of statistical properties of items to ensure score precision, or to allow for adaptive subtest designs, based on the difficulty of test questions. One example of the importance of statistical properties of items for test and inventory designs is the popularity of the use of item-response theory (IRT) for ensuring equivalence across test forms, or in building multistage tests (e.g., Luecht, 1998). The use of statistical properties of test questions, in addition to the discrete selection variables in the problem, introduces complexity in the overall problem. In the mathematical literature, these kinds of combinatorial problems are modeled as discrete optimization problems.

Discrete optimization problems range widely in their difficulty to be solved efficiently and in their solution techniques. Some of these problems can be modeled as shortest-path problems that are generally efficient to solve for most problem sizes. Others, like the “traveling salesperson” problem, require exponential time. These last problems define a frontier of knowledge in theoretical mathematics and computer science, and have a million-dollar bounty for their resolution. In this category, there are other problems, like finite-capacity scheduling, that obtain reasonable solution times using constraint programming on industrial-sized problems. The structure of the optimization problem solved in our illustration in this chapter makes use of mixed-integer programming (MIP), an area of study derived from linear programming. MIP has already been also used extensively in a variety of other test-design problems (van der Linden, 2005).

The choices in MIP are represented by decision variables, some of which are restricted to have values in a set of integers. In many MIPs, a value of 1 will represent that a selection is made and a value of 0 that the selection is not made. In scheduling and selection problems, time is usually discretized so that the schedule is divided into individual time blocks. For selection, a decision variable is used for every pairing of resources with time blocks to model when resources are needed along the time horizon. In our illustration for inventory systems, the blocks are subtests required for administration windows and the objects to be scheduled will be MCQs.

Once the decision variables have been defined, linear inequalities are required to model the dependences between the choices. For example, if two selections are not allowed simultaneously, then this can be modeled as the sum of their corresponding decision variables must be at most 1. Or, if exactly one item is allowed in a position in a subtest, then the sum of the decision variables of items allowed in that position of the subtest is required to be 1, another linear (in)equality. This will require that exactly one of the variables is 1 (selection is made) and hence the rest are 0 (no selections made). There are many problems that cannot be represented as MIP models because of nonlinear inequalities. However, linear inequalities can model a vast array of dependences and have many applications for problems related to inventory and supply scheduling.

Solving an MIP problem in a reasonable amount of time often depends on careful expression of the problem, and performance tuning may be needed using representative rules and data. The efficiency of the solution depends on the size of the input, the type of constraints, and the objective function. The majority of traditional solutions rely on the linear relaxation of the MIP problem. A linear relaxation of an MIP problem is a linear program that relaxes the restriction that its decision variables take integer values only. Linear programming has been used to solve optimization problems since the 1950s and although it was suspected to have an efficient solution strategy, this was not found until late 1980s.

Relaxation has important consequences for the usefulness of a solution. For example, when dealing with a discrete 0–1 choice, the solution obtained after relaxing its decision variable could be meaningless. Suppose a given relaxed 0–1 decision variable has a solution of 0.789. Do we make the selection decision, or not? At this point, there are two basic options: to branch or to cut (or both). A branch occurs when we introduce a pair of nonredundant linear inequalities that split the search space into two separate linear problems, one that adds the linear constraint that the decision variable be equal to 0 and another that constrains the variable to be equal to 1. The two new linear problems are then solved efficiently and their results can be used with other branches. The branches are managed in a search tree. Bounding inferences from the results of the branches are used so that searching through the entire tree occurs only in the worst case. But choices still need to be made regarding which fractional variable to branch on and which branch to consider next in traversing the search tree. Most commercial MIP packages, e.g., OPL Studio 3.6.1

(ILOG, 2002), allow the user to specify these choices or let them be made automatically. Different choices of solution strategies are usually tested in performance tuning. Once the proper structure for the problem is known, and an efficient solution strategy is identified, the model will be efficient and generally applicable for resolving scheduling and selection problems given a data set and objective function.

An inequality that is redundant in the original MIP is called a *cut*. When introduced in the relaxation stage (or at any stage of a branch-and-bound search), it may maintain the solution structure while cutting off an unacceptable solution to a decision variable (such as the 0.789 example). There are algorithms that produce different types of cuts (e.g., clique cuts, Gomory cuts, and disjunctive cuts). MIP practitioners need to choose their cuts wisely since poor choices can be relatively time-consuming with respect to the overall solution time and may end up contributing little to the quality of the solution. Many commercial software packages also allow the user to make their own cuts or to select which cuts should be generated automatically.

Most MIPs are solved using variations of these two strategies implemented in a commercial "solver," that is, the main processor that resolves constraints for the objective function to find the best solution for an input data set. Once the MIP is set up, the only two basic concerns are whether there is enough memory to process the MIP (an issue of size) and whether the solving process takes too long (an issue of time). With some combinatorics and perhaps some remodeling, the size of the MIP can readily be evaluated and modified if too large. As for the performance of the MIP solver, time can easily become the issue and small changes in search strategy can be shown to greatly reduce the time. Exploration of the branch-and-bound tree also gives useful performance-tuning information, such as how close the current feasible solution may be to the optimal solution (also referred to as an upper bound on optimality). Users can decide to stop the process if this solution is acceptable, or if a certain amount of time has passed. This is not an algorithmic or heuristic approach, and any solution returned will satisfy all the constraints. Thus, for the problem solved in our example of subtest creation, any feasible solution will meet our design and business rules and can be used to schedule the items. The upper bound or time limit improves efficiency but risks that the schedule generated is less than optimal. In our example of scheduling test-question development for the item bank, this may mean that another feasible schedule exists whose maximum exposure of any item is less than that in the interim accepted solution. These are the main considerations when solving MIP problems, and devoting some effort to resolving them can have great benefits in efficiency for operational use of optimization.

A system of inventory planning and scheduling used for the computerized Uniform Certified Public Accountancy licensing examination (CPA Exam) will be used to illustrate a practical solution for operational programs producing large numbers of forms. The methods described here use MPI as implemented by the Optimization Programming Language (OPL) Studio 3.6.1 software (ILOG, 2002) and illustrate the close connection among test design, administration models, and inventory scheduling.

The next section will provide a step-by-step description of the theoretical approach, and the analyses that were used to define an on-hand inventory required, given the test design, exposure rules, and administration schedules. This general methodology can be adopted to develop an initial specification for the resting state of the item bank, appropriate for a variety of psychometric models for testing. The method is useful for any kind of performance testing, including MCQ forms, linear test designs, and other complex performance tasks currently popular in competency testing.

The section thereafter uses the CPA Exam to illustrate how the ideal bank and the specific psychometric model used for a computer-based administration define a supply model to sustain the item bank over a five-year time horizon. In this illustration, the optimization formulation of the problem ensures a variety of constraints are always met for subtests of differing difficulties required for the computerized multistage testing design for the CPA Exam described by Melican, Breithaupt & Zhang (this volume, chap. 9).

The inventory production and management system described in the illustration is based on the goal of limiting the exposure of test content, while maintaining test equivalence and quality over time with the smallest practical bank size. There are two parts to the solution, described separately. The ideal bank is determined first, using minimal algebraic bounds required for the steady state (on-hand inventory). Second, a schedule is developed for MCQs that optimally restricts exposure of content and overlap between administrations while conforming to all the requirements necessary to ensure continuous high-quality testing.

13.3 The Ideal Bank (Steady-State Model)

In a growing number of testing programs, complex performances are being measured using innovative formats beyond those of the traditional MCQs. Examples of complex performance testing may be found in the certification of architects, physicians, accountants, and information technology personnel. These testing formats tend to be more costly to develop, administer and score (Drasgow, Olson-Buchanan, 1999; Tekian, McGuire, McGaghie & Associates, 1999). The two factors of continuous delivery of computerized assessments and complexity of performance tasks have underlined the importance of creating more efficient methods of inventory management and test assembly. For the CPA Exam, this adds to the existing problem of maintaining a large, high-quality bank of traditional test questions, such as MCQs, for secure delivery of a nondisclosed test. Program directors must determine the minimum amount of test content that must be on hand to maintain a testing program over time, and test developers have to work with administration and development schedules to continuously produce valid test forms while preserving the security of test content.

13.3.1 Assumptions Required

The useful lifespan of test content and policies for maximum item exposure must be clearly specified as fundamental rules to determine the minimum practical on-hand inventory. The latter are often expressed with respect to the total number of test takers in any administration, the number of discrete test forms or testing periods, or the overlap of operational banks of test questions.

The first consideration in determining the ideal minimum-size item bank is the content represented by each test question on a typical test form. Most certification or licensure tests are built to specifications derived from a task or practice analyses (Raymond & Neustel, 2006). Each test question (or performance task) is usually coded according to an outline that links the knowledge, tasks, or skills measured to a requirement in the test blueprint. The content outline is built from the findings of the job or practice analysis, and is approved by the policy board responsible for the testing program. In many programs, the length of the test forms and any range for the target proportion of test items on each content or skill area are also defined by policy. These are important guidelines for the content-related properties of items needed in the on-hand inventory. As is well explored in van der Linden, Ariel, and Veldkamp, (2006), an item bank might best be described as a large collection of valid test forms, and thus is representative of a valid test outline in the proportion of content, skills, or tasks measured by each test question. When the proportion of test questions required for each content or skill area and the test lengths are known, we can determine the proportions needed for test questions matching each attribute from the number of questions of that type on a valid test form.

Useful policies should also consider the schedule for retirement or disclosure of used test questions, which are dependent on the memorability of the test content and on the rate of change in the content itself in a discipline. Of course, security and nondisclosure policies are strictest when the test is used for selection decisions, such as licensing, certification, or admissions tests. For example, case studies used for licensing in medicine might both be more memorable and change more frequently with the evolution of medical science, as compared with traditional test questions included on a test used for classroom assessment in algebra. As a consequence, a shorter useful life-span policy for clinical case studies might be expected, compared with a relatively long lifespan for an item bank used for classroom testing. Other exposure policies might include a maximum of the number of times any given item is seen by any test taker, or the proportion of test questions that might be reused in adjacent test administrations.

When new test questions are included in assembly to allow for item analysis or calibration, a minimum exposure policy might also be needed. In order to apply some scoring models, such as item-response theory (IRT) models, response data must be gathered from a relatively large sample in advance of operational scoring (Hambleton & Jones, 1991). In this case, placement of new test questions on test forms would depend on the number of alternate forms and on the volume of candidates who take the test during a defined period. A minimum number of exposures may be an important consideration in the quantity of new content that can be pretested in any given period.

13.3.2 Considerations for Linear and Adaptive Testing Formats

The administration format is a key consideration in the analysis of the requirements for an ideal bank. The statistical properties of required test items will differ, depending on whether forms are parallel linear forms, adaptive multistage tests (MST), item-level computerized adaptive tests (CAT), or other format. Not only does the composition of the test depend on suitable statistical properties for all items (e.g., to ensure equivalence across forms and to garner desirable properties such as reliable test scores), each test design will have somewhat different implications for inventory needs and item-exposure projections. The topic of adaptive designs will be explored more fully after a simpler example of ideal bank-size calculation.

The selection of any item for a test form will be constrained within the rules governed by these kinds of design and other policies. Recall also that any valid test form represents a miniature item bank, one with a size to support just this one form. With this representation in mind, it is possible to use the properties of the actual items on a typical test form and think of each item assignment as conforming to the specific design rules, such as test length, statistical properties, skills, and content coding, that were used to create the form. The notion of each item assignment as a unique combination of these properties will be used to develop our ideas of the ideal bank and the supply schedule for the inventory system. So, the ideal bank inventory will determine the number of item assignments required, with each item assignment defined by the combination of key properties that define an item in that position on a test form.

With this set of basic assumptions in hand, we can solve an arithmetic expression that will yield the minimum on-hand inventory required for the ideal bank. One overarching decision is the time horizon we want to use for planning replenishment of the item bank, and this might be based on the useful life of the test questions. Below is a worked example for linear test forms, which could be modified easily to determine on-hand inventory required for other programs where assumptions may differ.

Suppose the inventory has to support

- a planning horizon of five years;
- two administrations a year;
- 25 forms for each administration;
- linear forms of 40 items; and
- a maximum reuse rate of five for any test question over five years.

As a result, there must be a minimum of $250/5 = 50$ suitable items in the ideal bank for each assignment to a test form, and the minimum number of items required to run the program is $50 \times 40 = 2{,}000$. Also, the item bank would need to include a minimum of 50 of each item assignment at any given time. This does not dictate the supply or retirement schedules for items, nor does it involve any control of the exposure rates of the items. We will return to scheduling after we have discussed the ideal item bank.

The case for MST has the additional factor of tailored selection when calculating exposure. It is possible to create a predictive model to determine the proportion of candidates who will receive each subtest in a multistage adaptive test. The probability of exposure depends on the number of forms as well as the number and position of subtests, based on their average difficulties. For example, if the alternate forms rotate randomly during the test administration, the probability that the first subtest (or routing test) will be seen is simply a function of the number of alternative forms for it (e.g., if there are 25 alternate routing tests, the probability for exposure for any of them is 1/25). At the end of the routing subtest, branching occurs. Suppose there are two choices of subtests at the second stage, their probabilities of exposure depend on the particular pairing of subtests at that stage. For example, if we assume one moderate (M) and one difficult (D) subtest at the second stage, the probability of any candidate being administered the M subtest of a pair at the next stage can be expressed as

$$\Pr\{M_{m,d}\} = \Pr\{X \leq \theta^*(m,d)\}. \tag{13.1}$$

Here, $\Pr(M_{m,d})$ is the probability of seeing the M subtest in a given M and D pair (indexed by m and d). Equation 13.1 represents this probability as a function of the candidate's number-correct score, X, and the cut-off score for the M and D pair, $\theta^*(m,d)$. Similarly, the probability of any candidate receiving the D subtest is simply $1 - \Pr(M_{m,d})$. In this way, a probability is computed for every subtest in the set of forms constructed for a given administration. The total predicted exposure rate of the item is the sum of the exposure rates across all forms where it appears.

Now we come to the question of exposures when adaptive tests are administered and the distribution of ability in the test-taker population is known. To determine empirical estimates of the exposure rates, it is only necessary to generate a representative distribution for candidate ability on the IRT θ scale. In our example, a normal $N(0,1)$ distribution of ability can be used to simulate the candidates who would be taking the examination. Lastly, it is only necessary to sum the empirical probabilities of exposure for all subtests in the administration to determine the predicted average and maximum exposure for a set of panels. All of this can be done without actual administration of the adaptive multistage subtests, as long as a calibrated bank is available. It makes sense to consider expected exposures when selecting design targets for adaptive subtests, given that the testing program has a calibrated item bank.

13.3.3 Inventory-Scheduling Problem

Several authors have proposed methods for limiting exposure in adaptive testing, including assembly-based control for content exposure (Luecht & Burgin, 2003). At the same time, the memorability of complex performance tasks poses a particular problem. These tasks require more time to complete and are also costly to create. Consequently, smaller banks of such performance tasks are often encountered and scheduling or inventory rotation becomes an important security concern. In such high-stakes testing programs as for licensure, there is often no way to anticipate and prevent the security breaches that may occur when cheating conspiracies are organized. The impact on score validity when security is breached can pose a very serious risk to the testing program (De Champlain et al., 2000).

The inventory system proposed in this section extends some basic assumptions we discussed for calculating the ideal minimum bank size, based on traditional test questions and linear forms, to refreshing the item bank over a longer time horizon. The schedule is designed to maintain the item bank for continuous administration of an examination of specific structure, preserving item exposure while ensuring forms of equivalent quality. Our example here will focus only on MCQs, although similar models can easily be developed to deal with combinations of complex performances that are selected to conform to test blueprints.

A prominent goal is to ensure the inventory-supply schedule will minimize the exposure rate of any given operational item while maintaining a reasonable item-bank size. Given practicalities of resources and continuous computerized administration, operational programs often struggle with limited item banks. Given that the minimum on-hand item bank is now known, the key questions in the supply planning problem can be summarized in the following way: when should we create, administer, and retire test content?

Our intent in designing an inventory system is to determine the minimal possible number of MCQs and simulations required for our item bank and an optimal schedule for the production and rotation of content to maintain our program over a specified period of time. Possibly the most important benefit of this solution is the schedule for content use. The schedule assigns each item or simulation in the item bank to subtests in such a way that exact statistical and content requirements are maintained. With this schedule, resource files for computerized delivery can be preassembled and authoring schedules can be aligned far in advance. Automated checks and controls for quality assurance are built into this inventory system.

13.3.4 Analysis Methods

Tools used to generate a solution for the item-bank model and the schedule relied on the mixed-integer programming (MIP) methods mentioned in the introduction. We made use of efficient search strategies as well as logically redundant constraints (cuts) to find integer (here, 0–1) solutions. In this case, the problem was modeled as an extension to a set-covering problem. The algorithmic solution applied using CPLEX in OPL Studio 3.6.1 (ILOG, 2002) was selected as an efficient and appropriate solution method because of its state-of-the-art capabilities with respect to solving difficult MIPs.

A general approach to this problem has been offered by van der Linden (2005, chap. 10) in which the integer program would choose the item-bank inventory whose items minimized authoring costs. Some reasonable modifications and simplifications were made to this more general formulation in order to make the problem tractable. First, the cost to write different items was assumed to be the same. Second, the input to our problem involved assumptions of volume predictions and the number of forms, subtests, and simulations. Cost was not an explicit component in the model but is represented by proxy by the maximum number of possible pretest MCQs and simulations our testing program can afford to produce. Cost could be modeled as a constraint, but this was not explicit in our application. The output from our model includes a schedule for use of items over the time horizon of five years.

In our formulation of the problem, a sample MCQ subtest from a previously optimized assembly was used as a template to define the ideal for the steady state of the item bank. In this formulation, it was necessary only to know the position and characteristics (e.g., content codes, skill codes, enemy codes, and IRT parameters) of each MCQ in the template subtest. The results from previous simultaneous optimal subtest assembly were used to represent all the constraints required for a valid panel. Visual Basic programming and common Excel spreadsheet functions were used to interpret the output from the optimization of schedules in such a way that production schedules were clearly articulated. The following section begins with a statement of the inventory-scheduling problem, followed by a detailed presentation of the model and solution.

13.4 Illustration from the Uniform CPA Licensing Exam

The multistage testing format for the CPA Exam is described by Melican, Breithaupt, and Zhang (this volume, chap. 9). It consists of five subtests of MCQs of equal length for each section of the exam, which, along with two complex performance simulations to measure accountancy skills administered after the MCQ portion of the examination, form a panel. Figure 13.1 depicts only the MCQ subtests

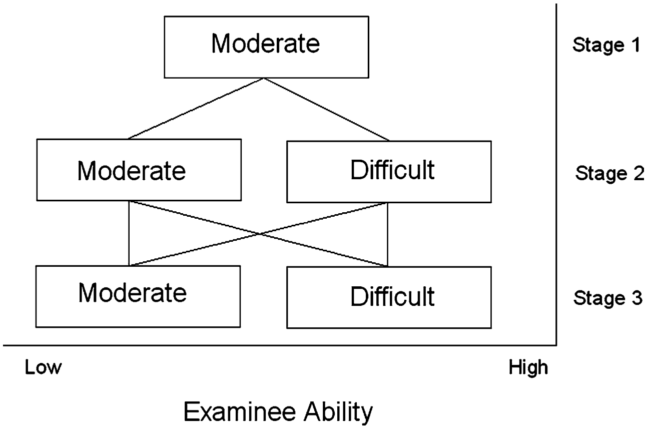


Fig. 13.1 A three-stage MST design

in one panel. Each MCQ subtest is targeted at a specific difficulty level. The first subtest the candidate receives is of medium (M) difficulty. Based on the performance of the candidate, the test driver will administer either a medium or a difficult (D) subtest in the second stage. At the completion of the second subtest, the candidate's performance on the first two subtests is evaluated to determine whether to administer an M or D subtest at the third stage. Figure 13.1 is simplified for illustrative purposes; the actual difficulty of the questions in M and D subtests overlaps to a greater extent than is shown in the figure.

Since our item banks are built near the passing score for the examination, and at a higher difficulty compared to the average ability of candidates, the likelihood of receiving the M subtest is greater than that of the D subtest. Also, each panel starts with an M subtest, so overall we expect M subtests to have greater exposures. Earlier empirical analyses have confirmed that the probabilities of getting the M subtests were always greater. Our assembly design allows for this, and the number of M subtests created was double that of the D subtests.

13.4.1 Ideal Bank and Inventory-Supply Schedule

The assumptions based on the CPA testing program considered as input to our model for the required steady state of the content item bank included the following:

- Four two-month administration cycles occur each year;
- The inventory model and schedule are for a five-year period;
- Given the expected volume of candidates each quarter and desired maximum exposure content, a fixed number of subtests, simulations, and panels can be specified as required for any administration window;
- Each MCQ has a defined lifecycle; after this time it is retired or rested.

A central notion in the model and the optimized schedule solution is the idea of an item class. For example, for the CPA Exam, the template for an M subtest has 25 MCQs with particular features that meet all the rules for subtest assembly. The combinations of these features, allowing for slight differences, define the item classes. For example, the MCQs selected for a given position on the subtests will have defined a similar range of difficulty, and will belong to the same content and skill categories as an item in the same position on an equivalent subtest. The bank will have a set of items that fit this description, and these can be understood as indexed items of a class. This notion is similar to that of the design point in an item-bank blueprint proposed by van der Linden, 2005, chap. 10). It follows that our subtest template is represented by 25 item classes with distinctive content and statistical characteristics. The master MCQ bank for the M and D subtests can be understood as a combination of subsets of the item banks for each of the 25 item classes. Some simple algebra can be used to decide what a necessary minimal bank size would be for both kinds of subtests, given the assumptions in our example.

In our example

- we have 4 builds over 5 years (i.e., 20 builds in total),
- and use 20 M and 10 D subtests for each build.

Then, for the set of M subtests,

- we have 20 builds with 20 subtests per build, which requires 400 item assignments in the schedule for each given item class,
- and, fixing the maximum exposure rate for any MCQ in an M subtest to 5, the subset of the item bank for each of these given item classes should have $400/5 = 80$ items.

It follows that the minimum number of items in the bank for the M subtests is 80×25, or 2,000, MCQs.

Likewise, for the set of D subtests,

- we have 20 builds with 10 subtests per build, which requires 200 item assignments in the schedule for each given item class,
- and, fixing the maximum exposure rate for any MCQ in a D subtest to 5, the subset of the item bank for each of these item classes should have $200/5 = 40$ items.

The minimum number of items in the bank for the M subtest is 40×25, or 1,000, MCQs.

From this analysis, the ideal bank size for our testing program requires a steady-state size of 3,000 MCQs in order to support the use of the M and D subtests for a typical administration window. The algebra, however, does not guarantee that such a bank is possible (only that there is no way to find a smaller one with the same input). We also know that if a minimal bank exists, then there must be 80 indexed items of each class for the M subtests, and 40 indexed items of each class for the D subtests.

The optimization solution produced by our models for MCQ banks creates a schedule to indicate when each item or simulation must be ready for use, how long it is operational, and when it will be retired. The optimization function minimizes exposure based on the fixed minimum bank size as input to the problem. Infeasibility arises if the bank cannot support the maximum exposure parameter.

Exposure is defined for our purposes as the number of times the MCQ is used over the five-year planning horizon. There are alternatives possible for defining exposure, including the proportion of candidates who see the item, or the actual count of candidates who received the item in any period of time. However, in continuous testing, it is often the length of exposure over time that is more important to security than the number of candidates who see the item in a particular build. Other aspects of exposure are accounted for in our model assumptions. For example, the minimum bank size was specified with assumptions about the number of candidates taking the exam each window, the number of subtests needed to meet reasonable empirical exposures for every item, and the number of new MCQs that can be produced each quarter.

In our example, the optimized schedule uses an item for only four of the five years. Recall that the time variable is discrete also, which means a five-year horizon includes four years of useful life for an item, and a possible selection for 16 test-administration cycles (four per year). After four years of operational use, the item could be retired, rested, or replaced by a new item. When an item type is scheduled to reappear, it could be as a pretested newly indexed item of that class. If all items were replaced on this schedule, the bank would be entirely new after a period of five years.

13.4.2 Supply-Scheduling Model

The supply-scheduling model is expressed as a separate set of expressions with an objective function and constraints. We only discuss the model for the M subtests for one given item class; the models for the other item classes and the D subtests are analogous.

Some notation is required for our problem. Let $b = 1, \ldots, 20$ represent the builds, $t = 1, \ldots, 20$ the subtests, and $i = 1, \ldots, 80$ the items in the given class for the subtests. The decision variables x_{bti} are equal to one if item i is assigned to subtest t of build b; otherwise, they are is equal to zero.

The integer program that schedules the given item type for the given years is

$$\text{minimize } z \tag{13.2}$$

subject to

$$\sum_{b=1}^{20}\sum_{t=1}^{20} x_{bti} \leq z, \qquad \text{for } i = 1, \ldots, 80; \tag{13.3}$$

(exposure equals sum of assignments per item)

$$\sum_{i=1}^{80} x_{bti} = 1, \qquad \text{for } b = 1, \ldots, 20;\ t = 1, \ldots, 20; \tag{13.4}$$

(one item in each position in subtest)

$$\sum_{t=1}^{20} x_{b-1,t,i} + x_{bti} \leq 1, \qquad \text{for } b = 2, \ldots, 20;\ i = 1, \ldots, 80; \tag{13.5}$$

(each item at most once in consecutive builds)

$$\sum_{b=4(y-1)+1}^{4(y-1)+4}\ \sum_{t=1}^{20} x_{bti} = 0, \qquad \text{for } y = 1, \ldots, 5;\ i = 16(y-1)+1, \ldots, 16y; \tag{13.6}$$

(each item operational for four years)

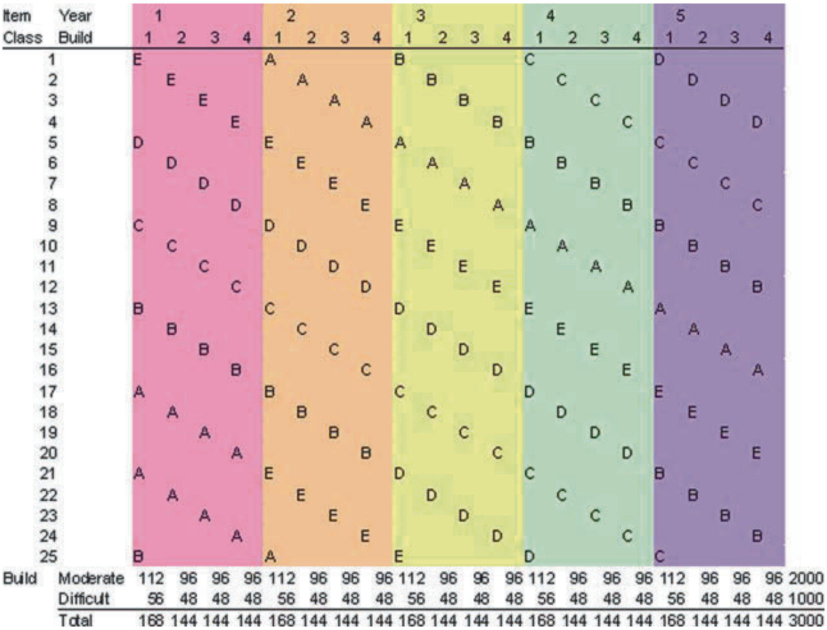
$$x_{bti} \in \{0, 1\}, \qquad b = 1, \ldots, 20;\ t = 1, \ldots, 20;\ i = 1, \ldots, 80. \tag{13.7}$$

(definition of variables)

In (13.6), we used constraints on exposure rates across the 20 builds (assuming five re-uses as a maximum for any item). Other constraints mirrored our actual assembly process, such as any item is used at most once in two consecutive builds in (13.5). However, the constraints in the optimization model can be varied to meet other practical considerations or generate a set of alternative solutions for decision-making.

Using OPL Studio, the optimal schedule was found for each of the 25 item classes in the MCQ subtests. The optimal solution provided the best schedule based on all the constraints in our model. Recall that in our example we are working with a steady-state item bank of 1,000 D items, and 2,000 M items. In Figure 13.2, each row represents one indexed item of a class on a 25-item subtest. The columns are numbered for a particular year and build quarter. All five years are depicted. The content-development needs, based on this example, would be a maximum of 112 M and 56 D items to be written and pretested (i.e., ready for operational use) every quarter.

The optimal schedule replaces all the items in the bank over a period of five years. This part of the scheduling is straightforward (20% of the items are replaced each year). We know that we needed 80 items of each type in our steady-state bank of M items (2,000 items total). The replacement schedule designates 16 of the 80 items in each build for retirement/resting for a period of one year. At the same time, 16 new items of that class become available for use for four years. Figure 13.2 offers the



Item	Year	1				2				3				4				5				
Class	Build	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	
1		E				A				B				C				D				
2			E				A				B				C				D			
3				E				A				B				C				D		
4					E				A				B				C				D	
5		D				E				A				B				C				
6			D				E				A				B				C			
7				D				E				A				B				C		
8					D				E				A				B				C	
9		C				D				E				A				B				
10			C				D				E				A				B			
11				C				D				E				A				B		
12					C				D				E				A				B	
13		B				C				D				E				A				
14			B				C				D				E				A			
15				B				C				D				E				A		
16					B				C				D				E				A	
17		A				B				C				D				E				
18			A				B				C				D				E			
19				A				B				C				D				E		
20					A				B				C				D				E	
21		A				E				D				C				B				
22			A				E				D				C				B			
23				A				E				D				C				B		
24					A				E				D				C				B	
25		B				A				E				D				C				
Build	Moderate	112	96	96	96	112	96	96	96	112	96	96	96	112	96	96	96	112	96	96	96	2000
	Difficult	56	48	48	48	56	48	48	48	56	48	48	48	56	48	48	48	56	48	48	48	1000
	Total	168	144	144	144	168	144	144	144	168	144	144	144	168	144	144	144	168	144	144	144	3000

Fig. 13.2 Eligible-for-use schedule. Indexed groups of M items: A: 1–16; B: 17–32; C: 33–48; D: 49–64; E: 65–80. Indexed groups of D items: A: 1–8; B: 9–16; C: 17–24; D: 25–32; E: 33–40

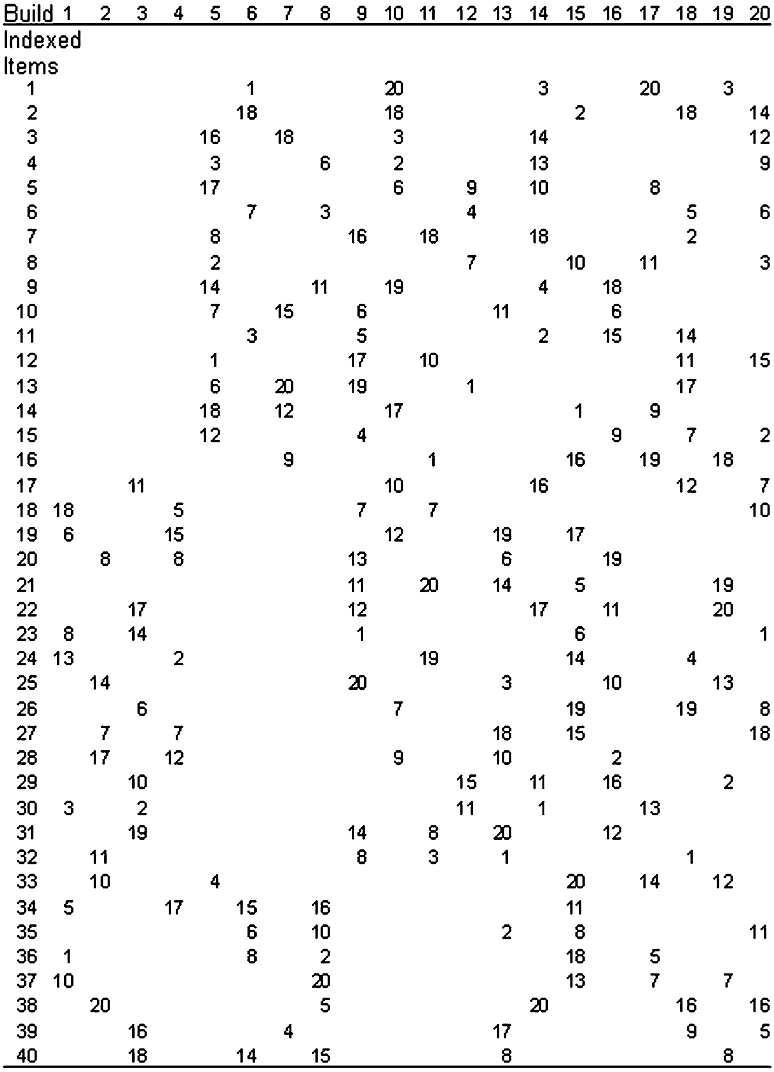
eligible-for-use view of the schedule. Within any item class, the items are indexed from 1 to 80. Items in the class are further divided into lettered groups. To use this schedule, the specific group of items of class 1 for the M subtests (those indexed from 1–16) will be required for operational use the first quarter of year two. This is apparent because the lettered group for the first instance when this group of items is used on subtests is A. Group A of these items are first selected for the first build of subtests in the second year. See the legend of Figure 13.2 for when indexed items from each class are required for the M and D subtests.

The eligible-for-use schedule can be modified easily to become a not-used schedule. We recreated the table so that classes of items of each type are scheduled not to be used for the four consecutive builds prior to their first date as an eligible-for-use item. During this period, final pretesting of the items and approval for operational use must be completed.

This aspect of scheduling is not complicated. However, it is based on the optimization solution we obtained using our 25-item subtest templates and an ideal bank size on hand to define the MIP model. The simplicity of the eligible-for-use schedule hides a fairly complex underlying schedule that designates when each actual item (indexed within an item class) is selected for a subtest. Specifically, when the class of items is available for use, only a specific indexed item from it is actually used to build subtests. The assignment of individual items over time to subtests cannot be replicated without a full representation of all the model elements, the objective function, and performance-tuning details. This allows the testing program to anticipate and plan exactly when and where every item in the bank will be exposed, which has obvious security benefits. Figure 13.3 illustrates an assignment schedule for each indexed item in one class across the 20 subtests for a four-year administration timeline with 20 builds. For simplicity, only 40 of the indexed items of this class are depicted. A similar assignment schedule is produced for every class of item on the M and D subtests, resulting in a complete recipe for item production over five years.

In the optimized solution in Figure 13.3, the exposure rate of any item was fixed at a maximum of five uses over the four years of availability. Where there are no subtest ID entries in the table for a given item index, the group of items is not available for use (e.g., items 1–16 in the first year; item 17–32 in the second year). This fact represents the time when new items are readied for operational use in the first quarter of the subsequent next year. It is also evident that no item is ever used in two adjacent builds or in more than one subtest for a given build. The actual resting period before an item is assigned to a new subtest is often longer than one year.

Because the template used was an optimized result from subtest assembly, it is possible to schedule assignments from the item bank in such a way that all subtests and panels retain equivalent and desirable psychometric properties and meet content constraints. The degree of a match in quality depends entirely on the interchangeability of the indexed items within each class. In this way, we can contemplate a time when individual optimized assemblies are no longer used for each individual build. Rather, the selection for subtests can be driven entirely from the optimized supply schedule, where items are selected from the bank based only on their item class and index numbers.



Build	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Indexed Items																				
1						1				20				3			20		3	
2						18				18					2			18		14
3					16		18			3				14						12
4					3			6		2				13						9
5					17					6		9		10			8			
6						7		3				4						5		6
7					8				16		18			18				2		
8					2							7			10		11			3
9					14			11		19				4		18				
10					7		15		6				11			6				
11						3			5					2		15		14		
12					1				17		10							11		15
13					6		20		19			1						17		
14					18		12			17					1		9			
15					12				4							9		7		2
16							9				1				16		19		18	
17			11							10				16				12		7
18	18			5					7		7									10
19	6			15						12			19		17					
20		8		8					13				6			19				
21									11		20		14		5				19	
22			17						12					17		11			20	
23	8		14						1						6					1
24	13			2							19				14			4		
25		14							20				3			10			13	
26			6							7					19			19		8
27		7		7									18		15					18
28		17		12						9			10			2				
29			10									15		11		16			2	
30	3		2									11		1			13			
31			19						14		8		20			12				
32		11							8		3		1					1		
33		10			4										20		14		12	
34	5			17		15		16							11					
35						6		10					2		8					11
36	1					8		2							18		5			
37	10							20							13		7		7	
38		20						5						20				16		16
39			16				4						17					9		5
40			18			14		15					8						8	

Fig. 13.3 Schedule of item assignments to subtests

13.5 Discussion

Many advantages for operational testing programs exist when their competing and often complex test-design rules can be automated. The design of the computer-based CPA Exam is one example where a tailored item bank and supply schedule can be used to control inventory. The solution takes into account the particular multistage testing format and business rules for the examination in creating an inventory system to match its test-assembly problem. The current illustration could easily be adapted to design inventory systems for a variety of test-administration formats. The same basic methodology has also been used at the CPA Exam program to schedule and select performance tasks for placement on examination forms. This flexible approach, as well as the availability of commercial software, are important advances for high-volume, high-stakes testing programs.

As we continue to research and modify these systems and procedures, the CPA Exam team hopes to extend our test-assembly approach to solve new production problems and to reduce our reliance on resources (in particular, the time required from subject-matter experts). The attention of our review committees is already shifting from the traditional evaluation of the completed examination forms to quality assurance for content development and rules for automated assembly and inventory planning. This change in process allows us to generate a large number of forms for computerized delivery. For example, our review of individual subtests was replaced by audits of reports completed using the summary information from the assembly system and not the actual test items. Our goal was to confirm that the assembly logic has been exactly followed for every subtest. During operational assembly, our internal content experts examine all of the completed panels but only a sample is reviewed by committees.

The inventory system provides an efficient method of determining the minimum bank to support a testing program, given cost and production capacity, and guides our authoring schedule to maintain the program for a specified time horizon. This kind of scheduling would be useful for long-term planning by test developers who need to project well in advance what new content must be authored. A necessary assumption of the model is that authors can write items to content specifications and to general difficulty levels. For example, item vendors would be instructed to write a fixed number of items for a content area having moderate difficulty. Some researchers have begun to explore the relationship between statistical and other properties of MCQs (Keller & Davey, 2002) and this knowledge can be used to train MCQ authors. In future content-development work, the relationship between content and the difficulty or discrimination of MCQs could be better exploited by item writers. For example, items within a content category that represent desired levels of difficulty could be incorporated into training materials.

Optimized eligible-for-use and not-used schedules also have the benefit of complexity that will deter organized cheating. It is only possible to replicate the optimized solution for administration of content when all features of the model are exactly known (including bank size). If one of these input parameters in the model is changed, the resulting schedule will no longer be representative. There are also many user-defined aspects to the optimization solution that could be changed without being apparent in the administration schedule for MCQs (e.g., items from the same content category rest for varying lengths of time or are administered on different schedules). The likelihood that the schedule could be determined by anyone outside the particular program is therefore miniscule.

The optimization solution for assembly and inventory systems presented here seeks to expand on the authoritative methodology offered by other practitioners and researchers (Veldkamp, 2001; Stocking & Swanson, 1993; van der Linden, 2005; and others). New applications of this kind of solution in testing programs with different assembly and administration methods, or different item types, would provide valuable insight into the potential usefulness of the general system.

Also, at the CPA Exam program improvements have been implemented extending this system, including gap analyses for new test specifications and item banks, simplified interfaces for subtest assembly, and generalized optimization models to produce subtests and panels for different administration formats and adjacent testing periods simultaneously. These developments are intended to reduce costs, improve our inventory controls, and increase the security of our test content. This illustration is one instance where advances in psychometric theory and practice can lead to benefits beyond elegant mathematics and offer immediate and practical gains for high-stakes testing programs.

References

Ariel, A., van der Linden, W. J. & Veldkamp, B. P. (2006). A strategy for optimizing item pool management. *Journal of Educational Measurement, 43,* 85–96.

Breithaupt, K. & Hare, D. (2004, February). *Automated simultaneous assembly of multi-stage subtests for the Uniform CPA Exam* [Technical Report]. Ewing, NJ: American Institute of Certified Public Accountants.

Breithaupt, K., Ariel, A. & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5*, 319–330.

De Champlain, A. F., MacMillan, M. K., Margolis, M. J., Klass, D. J., Lewis, E. & Ahern, S. (2000). Modelling the effects of a test security breach on a large-scale standardized patient examination with a sample of international medical graduates. *Academic Medicine, 75*, 109–111.

DeVore, R. (2002). *Considerations in the development of accounting simulations* [Technical Report]. Ewing, NJ: American Institute of Certified Public Accountants.

Downing, S. M. (2004). Item response theory: Applications of modern test theory in medical education. *Medical Education, 37,* 739-745.

Drasgow, F. & Olson-Buchanan, J. B. (1999). *Innovations in computerized assessment*. Mahwah, NJ*:* Lawrence Erlbaum Associates.

Hambleton, R. K. & Jones, R. W. (199l). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice* [Instructional Topics in Educational Measurement Series], 38–47.

Huitzing H. A. (2004). Using set covering with item sampling to analyze infeasibility of linear programming test assembly models. *Applied Psychological Measurement, 28,* 355–375.

Huitzing, H. A., Veldkamp, B. P. & Verschoor, A. J. (2003). Infeasibility in automatic test assembly models: A comparison study of different methods. *Journal of Educational Measurement, 42*, 223–243.

ILOG. (2002). *ILOG OPL Studio 3.6.1* [User Manual]. Mountain View, CA.

Keller, L. A. & Davey, T. (2002). *Using collateral information in IRT parameter estimation.* Paper presentation at the annual meeting of the International Test Commission, Winchester, UK.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22,* 224–236.

Luecht, R. M., Brumfield, T. & Breithaupt, K. (2002). *A subtest-assembly design for the uniform CPA Exam.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, Luecht, R. M. & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs.* Paper presentation at the annual meeting of the National Council on Measurement in Education, Chicago, IL.LA.

Luecht, R. M. & Nungester, R. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229–249.

Nemhauser, G. L. & Wolsey, L. A. (1999). *Integer and combinatorial optimization.* New York: Wiley & Sons.

Raymond, M. & Neustel, S. (2006). Determining the content of credentialing examinations. In S. Downing & T. Haldyna (Eds.), *Handbook of test development (*pp. 191–223). Mahwah, NJ: Erlbaum.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Sympson, J. B. & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.

Tekian, A., McGuire, McGaghie & Associates (1999). Innovative simulations for assessing professional competence: From paper-and-pencil to virtual reality. Chicago: Department of Medical Education, University of Illinois at Chicago.

Theunissen, T. J. J. M. (1985). Binary programming and test design. *Psychometrika, 50,* 411–420.

van der Linden, W. J. (2005) *Linear models for optimal test design.* New York: Springer-Verlag.

van der Linden, W. J. & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35,* 185–198.

van der Linden, W. J., Ariel, A. & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear test forms. *Journal of Educational and Behavioral Statistics, 31*, 81–100.

van der Linden, W. J. & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 54,* 237–247.

van der Linden, W. J., Veldkamp, B. P. & Reese, L. M. (2000). An integer programming approach to item bank design. *Applied Psychological Measurement, 22,* 259–270.

Veldkamp, B. P. (2001). *Principles and methods of constrained test assembly.* Doctoral dissertation, University of Twente, Enschede, The Netherlands.

Way, W. D., Steffen, M. & Anderson, G. S. (1998). Developing, maintaining, and renewing the item inventory to support computer-based testing. In C. N. Mills, M. Potenza, J. J. Fremer & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 89–102). Hillsdale, NJ: Lawrence Erlbaum Associates.

Williams, H.P. (1990). *Model building in mathematical programming (*3rd ed.). New York: Wiley & Sons.

Part IV
Item Calibration and Model Fit

Chapter 14
Item Parameter Estimation and Item Fit Analysis

Cees A.W. Glas

14.1 Introduction

Computer-based testing (CBT), as computerized adaptive testing (CAT), is based on the availability of a large pool of calibrated test items. Usually, the calibration process consists of two stages.

(1) A pretesting stage: In this stage, subsets of items are administered to subsets of respondents in a series of pretest sessions, and an item response theory (IRT) model is fit to the data to obtain item parameter estimates to support computerized test administration.

(2) An online stage: In this stage, data are gathered in a computerized assessment environment, proficiency parameters for examinees are estimated, and the incoming data may also be used for further item parameter estimation.

The topic of this chapter is the estimation of the item parameters and the evaluation of item fit, both in the pretest phase and in the online phase. Especially differences in item parameter values in the pretest and online stages are of interest. Such differences are often named *parameter drift*. Evaluation of parameter drift boils down to checking whether the pretest and online data comply with the same IRT model. Parameter drift may have different sources. Security is one major problem in adaptive testing. If adaptive testing items are administered to examinees on an almost daily basis, after a while some items may become known to new examinees. In an attempt to reduce the risk of overexposure, several exposure control methods have been developed. All these procedures prevent items from being administered more often than desired. Typically, this goal is reached by modifying the item selection criterion so that "psychometrically optimal" items are not always selected. Examples of methods of exposure control are the random-from-best-*n* method (see, e.g., Kingsbury & Zara, 1989, pp. 369–370), the count-down random method (see, e.g., Stocking & Swanson, 1993, pp. 285–286), and the method of Sympson and Hetter (1985; see also Stocking, 1993). With relatively low exposure

C.A.W. Glas (✉)
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_14,

rates, items will probably become known later than with high exposure rates. Still, sooner or later some items may become known to some future examinees.

Differences between the pretest and the online stages may also result in other forms of parameter drift. One might, for instance, think of differences in item difficulty resulting from the different modes of presentation (computerized or paper-and-pencil administration) or resulting from a changing curriculum. Also, differences in motivation of the examinees between the pretest and online stages might result in subtle shifts of the proficiency that is measured by the test. Response behavior in these stages might not be properly modeled by the same set of IRT parameters when examinees in the pretest stage are significantly less motivated than those in the high-stakes online stage.

In this chapter, two methods for the evaluation of parameter drift are proposed. The first method is based on a global item-oriented test for parameter drift using a Lagrange multiplier statistic. The method can be viewed as a generalization to adaptive testing of the modification indices for the 2PL model and the nominal response model introduced by Glas (1998, 1999; also see, Glas & Suarez-Falcon, 2003). The second method is targeted at parameter drift due to item disclosure. It addresses the one-sided hypothesis that the item is becoming easier and is losing its discriminative power. The test for this hypothesis is based on a so-called cumulative sum (CUSUM) statistic. Adoption of this approach in the framework of IRT-based adaptive testing was first suggested by Veerkamp (1996) for use with the Rasch model. The present method is a straightforward generalization of this work.

This chapter is organized as follows. First, the most common method of item calibration, marginal maximum likelihood, will be explained. Then the Lagrange multiplier test and the CUSUM test for parameter drift will be explained. Finally, the power of the two classes of tests will be examined in a number of simulation studies.

14.2 Item Parameter Estimation

14.2.1 MML Estimation

Marginal maximum likelihood (MML) estimation is probably the most used technique for item calibration. For the 1PL, 2PL, and 3PL models, the theory was developed by such authors as Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984, 1986), and computations can be made using the software package Bilog-MG (Zimowski, Muraki, Mislevy & Bock, 1996). MML estimation procedures are also available for IRT models with a multidimensional ability structure (see, for instance, Segall, this volume, chap. 3). Under the label "Full Information Factor Analysis", a multidimensional version of the 2PL and 3PL normal-ogive models was developed by Bock, Gibbons, and Muraki (1988) and implemented in TESTFACT (Wilson, Wood & Gibbons, 1991). A comparable model using a logistic rather than a normal-ogive representation was studied by

Reckase (1985, 1997) and Ackerman (1996a and 1996b). In this section, a general MML framework will be sketched, and then illustrated by its application to the 3PL model.

Let $\mathbf{u}_n$ be the response pattern of respondent n, $n = 1, \ldots, N$, and let $\mathbf{U}$ be the data matrix. In the MML approach, it is assumed that the possibly multidimensional ability parameters $\boldsymbol{\theta}_n$ are independent and identically distributed with density $g(\boldsymbol{\theta}; \boldsymbol{\lambda})$. Usually, it is assumed that ability is normally distributed with population parameters $\boldsymbol{\lambda}$ (which are the mean μ and the variance σ^2 for the unidimensional case, or the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Phi}$ for the multidimensional case). Item parameters $\boldsymbol{\beta}$ consist of discrimination parameters (a_i or $\mathbf{a}_i$ for the unidimensional and the multidimensional cases, respectively), item difficulties b_i, and guessing parameters c_i.

In applications of IRT to CAT, students seldom respond to all available items. In the calibration stage, a calibration design is used where samples of students respond to subsets of items, which are often called *booklets* . In the online stage, every student is administered a virtually unique test by the very nature of the item selection mechanism of CAT. Both of these test administration designs are captured by introducing a test administration vector $\mathbf{d}_n$, which has elements d_{in}, $i = 1, \ldots, I$, where I is the number of items in the item pool. The item administration variable d_{in} is equal to one if student n responded to item i, and zero otherwise. The design for all students is represented by an $m \times I$ design matrix $\mathbf{D}$. The definition of the response variable is extended: the vector $\mathbf{u}_n$ has I elements, which are equal to one if a correct response is observed, equal to zero if an incorrect response is observed, and equal to an arbitrary constant if no response is observed. In this context, it is an interesting question whether estimates can be calculated treating the design as fixed and maximizing the likelihood of the parameters conditional on $\mathbf{D}$. If so, the design is called *ignorable* (Rubin, 1976). Using Rubin's theory on ignorability of designs, this question is extensively studied by Mislevy and Wu (1996). They conclude that for the estimation of θ, in adaptive testing the administration design is ignorable. The consequences for item calibration using MML will be returned to in the next section.

MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to $\boldsymbol{\theta}$, rather than maximizing the joint log-likelihood of all person parameters $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\beta}$. Let $\boldsymbol{\eta}$ be a vector of all item and population parameters. Then the marginal likelihood of $\boldsymbol{\eta}$ is given by

$$\log L(\boldsymbol{\eta}; \mathbf{U}, \mathbf{D}) = \sum_n \log \int \ldots \int p(\mathbf{u}_n \mid \mathbf{d}_n, \boldsymbol{\theta}_n, \boldsymbol{\beta}_i) g(\boldsymbol{\theta}_n; \boldsymbol{\lambda}) d\boldsymbol{\theta}_n. \qquad (14.1)$$

The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of person parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that marginal maximum likelihood estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

The marginal likelihood equations for $\boldsymbol{\eta}$ can be easily derived using Fisher's identity (Efron, 1977; Louis 1982; also see, Glas, 1992, 1998). The first-order derivatives with respect to $\boldsymbol{\eta}$ can be written as

$$\mathbf{h}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}; \mathbf{U}, \mathbf{D}) = \sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) \,, \tag{14.2}$$

with

$$\boldsymbol{\omega}_n(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{u}_n, \boldsymbol{\theta}_n \mid \mathbf{d}_n, \boldsymbol{\eta}), \tag{14.3}$$

where the expectation is with respect to the posterior distribution $p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \mathbf{d}_n; \boldsymbol{\eta})$. The identity in (14.2) is closely related to the EM algorithm (Dempster, Laird & Rubin, 1977), which is an algorithm for finding the maximum of a likelihood marginalized over unobserved data. The present application fits this framework when the response patterns are viewed as observed data and the ability parameters as unobserved data. Together they are referred to as the complete data. The EM algorithm is applicable in situations where direct inference based on the marginal likelihood is complicated, and the complete data likelihood equations, i.e., equations based on $\boldsymbol{\omega}_n(\boldsymbol{\eta})$, are easily solved. Given some estimate of $\boldsymbol{\eta}$, say $\boldsymbol{\eta}^*$, the estimate can be improved by solving $\sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}^*) = 0$ with respect to $\boldsymbol{\eta}$. Then this new estimate becomes $\boldsymbol{\eta}^*$ and the process is iterated until convergence.

Application of this framework to deriving the likelihood equations of the structural parameters of the 3PL model proceeds as follows. The likelihood equations are obtained upon equating (14.2) to zero, so explicit expressions are needed for (14.3). Given the design vector $\mathbf{d}_n$, the ability parameter θ_n, and the item parameters of the 3PL model, the probability of response pattern $\mathbf{u}_n$ is given by

$$p(\mathbf{u}_n \mid \mathbf{d}_n, \theta_n, a_i, b_i, c_i) = \prod_i P_i(\theta_n)^{d_{in}u_{in}} (1 - P_i(\theta_n))^{d_{in}(1-u_{in})} \,,$$

where $P_i(\theta_n)$ is the probability of a correct response to item i, as defined in van der Linden and Pashley (this volume, chap. 2, formula 1.1). Define P_{in} and S_{in} by $P_{in} = c_i + (1 - c_i)S_{in}$, so S_{in} is the logistic part of the probability P_{in}. By taking first-order derivatives of the logarithm of this expression, the expressions for (14.3) are found as

$$\omega_n(a_i) = \frac{(u_{in} - P_{in})(1 - c_i)S_{in}(1 - S_{in})(\theta_n - b_i)}{P_{in}(1 - P_{in})}, \tag{14.4}$$

$$\omega_n(b_i) = \frac{(P_{in} - u_{in})(1 - c_i)S_{in}(1 - S_{in})a_i}{P_{in}(1 - P_{in})}, \tag{14.5}$$

and

$$\omega_n(c_i) = \frac{(u_{in} - P_{in})(1 - S_{in})}{P_{in}(1 - P_{in})}. \tag{14.6}$$

The likelihood equations for the item parameters are found upon inserting these expressions into (14.2) and equating the resulting expressions to zero. To derive the likelihood equations for the population parameters, the first-order derivatives of the log of the density of the ability parameters $g(\theta; \mu, \sigma)$ are needed. In the present case, $g(\theta; \mu, \sigma)$ is the well-known expression for the normal distribution with mean μ and standard deviation σ, so it is easily verified that these derivatives are given by

$$\omega_n(\mu) = \frac{(\theta_n - \mu)}{\sigma^2}$$

and

$$\omega_n(\sigma) = \frac{(\theta_n - \mu)^2 - \sigma^2}{\sigma^3}.$$

The likelihood equations are again found upon inserting these expressions in (14.2) and equating the resulting expressions to zero.

Also, the standard errors are easily derived in this framework: Mislevy (1986) points out that the information matrix can be approximated as

$$\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta}) \approx \sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})', \tag{14.7}$$

and the standard errors are the diagonal elements of the inverse of this matrix.

The basic approach presented so far can be generalized in two ways. First, the assumption that all respondents are drawn from one population can be replaced by the assumption that there are multiple populations of respondents. Usually, it is assumed that each population has a normal ability distribution indexed by a unique mean and variance parameter. Bock and Zimowski (1997) point out that this generalization together with the possibility of analyzing incomplete item administration designs provides a unified approach to such problems as differential item functioning, item parameter drift, nonequivalent groups equating, vertical equating, and matrix-sampled educational assessment. Item calibration for CAT also fits within this framework.

A second extension of this basic approach is Bayes modal estimation (the term "modal" refers to the mode of the posterior distribution). This approach is motivated by the fact that item parameter estimates in the 3PL model are sometimes hard to obtain because the parameters are poorly determined by the available data. In these instances, item-characteristic curves can be appropriately described by a large number of different item parameter values over the ability scale region where the respondents are located. As a result, the estimates of the three item parameters in the 3PL model are often highly correlated. To obtain "reasonable" and finite estimates, Mislevy (1986) considers a number of Bayesian approaches. Each of them entails the introduction of prior distributions on item parameters. Parameter estimates are then computed by maximizing the log-posterior density of $\boldsymbol{\eta}$, which is proportional to $\log L(\boldsymbol{\eta};\mathbf{U}) + \log p(\boldsymbol{\eta} \mid \boldsymbol{\zeta}) + \log p(\boldsymbol{\zeta})$, where $p(\boldsymbol{\eta} \mid \boldsymbol{\zeta})$ is the prior density of the $\boldsymbol{\eta}$, characterized by parameters $\boldsymbol{\zeta}$, which in turn follow a density $p(\boldsymbol{\zeta})$. In one approach, the prior distribution is fixed; in another approach, often labeled empirical Bayes, the parameters of the prior distribution are estimated along with the other parameters. In the first case, the likelihood equations in (14.1) change to $\partial \log L(\boldsymbol{\eta};\mathbf{U})/\partial\boldsymbol{\eta} + \partial \log p(\boldsymbol{\eta} \mid \boldsymbol{\zeta})/\partial\boldsymbol{\eta} = \mathbf{0}$. In the second case, in addition to these modified likelihood equations, the additional equations $\partial \log p(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta} = \mathbf{0}$ must also be solved. For details refer to Mislevy (1986). In the following sections, two methods for parameter drift in the framework of the 3PL model and MML estimation will be presented.

14.2.2 Impact of Violations of Ignorability on Item Parameter Estimation

In applications of IRT to CAT, students seldom respond to all available items. Every student is administered a virtually unique test by the very nature of the item selection mechanism of CAT. In the context of CAT, it is an interesting question whether estimates of item parameters can be calculated treating the design matrix $\mathbf{D}$ as fixed by maximizing the likelihood of the parameters conditional on $\mathbf{D}$. If so, the design is called ignorable (Rubin, 1976). In the present section, we assess a number of situations where ignorability is violated. Therefore, first the ignorability principle will be outlined in some detail. Let the potential responses be partitioned into the actually observed responses $\mathbf{u}_{obs}$ and the unobserved responses $\mathbf{u}_{mis}$. As above, the parameter of interest is denoted by $\boldsymbol{\eta}$, and it is assumed that the probability model for $\mathbf{u}_{mis}$ depends on parameters $\boldsymbol{\phi}$. The key concept in the theory of ignorability is “missing at random” (MAR). MAR holds if

$$p(\mathbf{D}|\mathbf{u}_{obs}, \mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X}) = p(\mathbf{D}|\mathbf{u}_{obs}, \boldsymbol{\phi}, \mathbf{X}),$$

where $\mathbf{X}$ are covariates that might play a role. So MAR holds, if the missing data indicators $\mathbf{D}$ do not depend on the missing data $\mathbf{u}_{mis}$, in fact, they only depend on the observed data $\mathbf{u}_{obs}$, and possibly on covariates $\mathbf{X}$. Then, there is a technical condition. In a frequentist framework, the condition is that $\boldsymbol{\phi}$ and η are distinct; that is, the space of $\boldsymbol{\phi}$ and η factorizes into a $\boldsymbol{\phi}$-space and a $\boldsymbol{\eta}$-space and the two sets of parameters have no mutual functional restrictions. In a Bayesian framework $\boldsymbol{\phi}$ and θ are distinct if $p(\boldsymbol{\phi}|\eta, \mathbf{X}) = p(\boldsymbol{\phi}|\mathbf{X})$, that is, if they have independent priors. Rubin (1976) proved the following:

Theorem

If $\boldsymbol{\phi}$ and η are distinct, and MAR holds,
then
in a frequentist framework $p(\mathbf{u}_{obs}, \mathbf{D} \mid \eta, \boldsymbol{\phi}, \mathbf{X}) \propto p(\mathbf{u}_{obs}, \mid \eta, \mathbf{X})$,
and in a Bayesian framework $p(\eta \mid \mathbf{u}_{obs}, \mathbf{D}, \mathbf{X}) \propto p(\theta \mid \mathbf{u}_{obs}, \mathbf{X})$.

The frequentist version implies that inferences such as maximum likelihood estimation can be based on the likelihood of the observed data, $p(\mathbf{u}_{obs}, \mid \boldsymbol{\eta}, \mathbf{X})$, and the process causing the missingness does not have to be taken into account. In the same manner, the Bayesian version implies that inferences can be based on a posterior $p(\boldsymbol{\eta}|\mathbf{u}_{obs}, \mathbf{X})$ that ignores the probability model for **D**. It should be noted that conditioning on **D** may produce an overestimate of the sample variability of the data, and consequently an underestimate of the standard error of the estimate of θ. Unbiased inferences on standard errors might be obtained if the data are also "observed at random", that is, if $p(\mathbf{D}|\mathbf{u}_{obs}, \mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X}) = p(\mathbf{D}|\mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X})$, so $\mathbf{u}_{obs}$ does not depend on **D**.

Ignorability in CAT directly follows from the theorem: In CAT the item selection process completely depends on the observed responses and is completely independent of the unobserved responses. Further, ignorability also holds when CAT data are used to calibrate the item and population parameters using maximum marginal likelihood (MML; see Bock and Aitkin, 1981, the impact of targeted designs on MML estimation was studied by Glas, 1988, and Mislevy and Chang, 2000).

In the present chapter, two cases are investigated where the observed data no longer determine the design **D**: the case where auxiliary information on the students' proficiency is used to select items and the case of item review where the original responses are no longer available. The impact of these violations on the estimates of the item parameters using CAT data in the calibration phase will be assessed using a simulation study.

Consider the response pattern of one student; the index i is dropped for convenience. In a situation of item review, the contribution to the log-likelihood given the original data $\mathbf{u}_{obs}$ and the reviewed data $\mathbf{u}_{mis}$ can be written as

$$\begin{aligned}
&\log p(\mathbf{u}_{obs}, \mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\eta}) \\
&\quad = \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\mathbf{u}_{mis}, \mathbf{D}; \theta, \boldsymbol{\beta}) g(\theta; \boldsymbol{\lambda}) d\theta \\
&\quad = \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) d\theta \\
&\quad = \log p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) \\
&\qquad + \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) d\theta.
\end{aligned}$$

Note that this contribution now consists of a term $\log p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$ and a term $\log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$. The former gives rise to a log-likelihood associated with a CAT design and if $\mathbf{u}_{mis}$ were observed, these data could be used to obtain consistent estimates of $\boldsymbol{\eta}$. The latter term is the expectation of the probability of $\mathbf{u}_{obs}$ with respect to the posterior distribution $p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$. However, if the missing data process is ignored, the expectation of $p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta})$ is considered with respect to $g(\theta; \boldsymbol{\lambda})$; that is, the log-likelihood then becomes a sum of terms

$$\log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) g(\theta; \boldsymbol{\lambda}) d\theta. \tag{14.8}$$

The effect is that $p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta})$ is averaged over the wrong proficiency distribution, that is, a distribution with a wrong location parameter and a wrong scale parameter. To assess the effect, consider two students, one with a high θ-value and one with a low θ-value. The first student is administered difficult items, and the second student is administered easy items. However, in (14.8) both their θ-values are assumed to be drawn from the same distribution, and as a result, the easy items are overestimated and the difficult items are underestimated. The effect is due to ignoring the covariates $\mathbf{u}_{mis}$ and $\mathbf{D}$. When the design is governed by auxiliary information about θ, say θ_0, the situation is essentially the same: when the covariate θ_0 is ignored, the proper posterior $p(\theta|\theta_0; \boldsymbol{\beta}, \boldsymbol{\lambda})$ is replaced with $g(\theta; \boldsymbol{\lambda})$, and the result is bias in the estimates of $\boldsymbol{\eta}$.

14.2.3 Simulated Examples

To assess the magnitude of the bias caused by ignoring covariates, simulation studies were conducted. A number of simulation studies were conducted to elucidate the two cases discussed above. The following eight conditions were introduced.

1. Random item selection. In this condition, for every simulee a new set of item parameters was randomly drawn from the standard normal distribution and responses to this randomly assembled test were generated. So this condition did not entail CAT; it was used as a baseline for reference.
2. Computerized adaptive testing.
3. Computerized adaptive testing with item review. In this condition, new responses were generated for all the selected items. So the condition is far more extreme than what can be expected in real-life testing situations.
4. Computerized adaptive testing with item review only for proficiency levels $\theta > 0.0$. In this first set of simulations, the results will just be a combination of the two previous conditions; the purpose of this condition will become apparent in the simulation studies pertaining to item calibration.
5. Computerized adaptive testing where the first half of the test items were chosen to be optimal at the true proficiency value.
6. Computerized adaptive testing where all items were chosen to be optimal at the true proficiency value.
7. Computerized adaptive testing where the first half of the test items were chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and a standard deviation equal to 1.0.
8. Computerized adaptive testing where the first half of the test items were chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and a standard deviation equal to 2.0.

Adaptive test data were generated for 1,000 simulees with parameters drawn from the standard normal distribution. The item bank consisted of 200 items equally spaced between -2.0 and 2.0, and the test length was 20 items. The one-parameter logistic model (1PLM) was used to avoid contamination of the results by the possibly poor identification of the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM). Unless indicated otherwise, the starting value of the proficiency estimate was equal to zero. The proficiency parameter was estimated by maximum likelihood and maximum information was used as a selection criterium. Using these adaptive test data, MML estimates of the item parameters were computed under the assumption that θ had a standard normal distribution.

In every condition reported below, 100 replications were made. In the condition of random item selection, the test of 20 items was resampled from the item bank for every simulee.

The results are shown in Table 14.1. For five items from the item bank, the last three columns give the bias, standard error, and mean of the estimates over the replications, respectively. The following conclusions can be drawn.

1. Comparing random item selection and CAT, it can be seen that the latter greatly reduced the standard error. In both cases, the bias was relatively small.
2. In all other conditions, the bias was substantial.
3. In CAT with item review, there is inward bias; that is, easy items are overestimated and difficult items are underestimated.
4. If only simulees with $\theta > 0$ review the items, the bias in the easy items vanishes.
5. Choosing the complete test to be optimal at the true θ completely contaminates the calibration in the sense that all item parameters shrink to zero.

14.3 Item Fit Analysis

14.3.1 *Lagrange Multiplier Tests*

The idea behind the Lagrange multiplier (LM) test (Aitchison & Silvey, (1958), and the equivalent efficient score test (Rao, 1947), can be summarized as follows. Consider some general parameterized model and a special case of the general model, the so-called restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by setting one or more parameters of the general model to constants. The LM test is based on evaluating a quadratic function of the partial derivatives of the log-likelihood function of the general model evaluated at the ML estimates of the restricted model. The LM test is evaluated using the ML estimates of the parameters of the restricted model. The unrestricted elements of the vector of the first-order derivatives are equal to zero because their values originate from solving

Table 14.1 Squared bias and standard errors for calibration of β

Item Selection Mode	β	Bias	S.E.	Mean
Random selection	−2.0	0.01	0.32	−1.96
	−1.0	0.05	0.20	−0.94
	0.0	0.01	0.23	−0.01
	1.0	0.01	0.29	1.01
	2.0	0.08	0.29	2.08
CAT	−2.0	0.02	0.19	−2.00
	−1.0	0.03	0.26	−1.03
	0.0	0.01	0.08	−0.01
	1.0	0.00	0.21	0.99
	2.0	0.00	0.19	2.00
CAT with item review	−2.0	0.64	0.22	−1.33
	−1.0	0.34	0.29	−0.65
	0.0	0.01	0.07	−0.01
	1.0	0.28	0.22	0.71
	2.0	0.60	0.18	1.39
CAT with item review	−2.0	0.07	0.21	−1.90
if $\theta > 0.0$	−1.0	0.15	0.29	−0.84
	0.0	0.01	0.07	0.01
	1.0	0.20	0.19	0.79
	2.0	0.52	0.22	1.47
50% optimal	−2.0	0.43	0.23	−1.54
at true θ	−1.0	0.38	0.25	−0.61
	0.0	0.00	0.17	−0.00
	1.0	0.33	0.22	0.66
	2.0	0.40	0.22	1.59
100% optimal	−2.0	1.92	0.39	−0.05
at true θ	−1.0	0.92	0.21	−0.07
	0.0	0.04	0.18	0.04
	1.0	0.93	0.22	0.06
	2.0	1.84	0.38	0.15
50% initial responses at	−2.0	0.21	0.20	−1.76
$\widehat{\theta}$ with s.d.$(\widehat{\theta}) = 1.0$	−1.0	0.17	0.19	−0.82
	0.0	0.00	0.17	0.00
	1.0	0.08	0.21	0.91
	2.0	0.24	0.20	1.75

the likelihood equations. The magnitude of the elements of the vector of first-order derivatives corresponding with restricted parameters determines the value of the statistic: the closer they are to zero, the better the model fits.

More formally, the principle can be described as follows. Consider a null hypothesis about a model with parameters η_0. This model is a special case of a general model with parameters η. In the case discussed here, the special model is derived from the general model by setting one or more parameters to zero. So if the parameter vector η_0 is partitioned as $\eta_0 = (\eta_{01}, \eta_{02})$, the null hypothesis entails $\eta_{02} = 0$. Let $\mathbf{h}(\eta)$ be the partial derivatives of the log-likelihood of the general model, so

$\mathbf{h}(\boldsymbol{\eta}) = \partial \log L(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in $\boldsymbol{\eta}$. The test will be based on the statistic

$$LM = \mathbf{h}(\boldsymbol{\eta}_{02})^t \boldsymbol{\Sigma}^{-1}\mathbf{h}(\boldsymbol{\eta}_{02}), \tag{14.9}$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{10}\boldsymbol{\Sigma}_{00}^{-1}\boldsymbol{\Sigma}_{01}$$

and

$$\boldsymbol{\Sigma}_{pq} = \sum_n \mathbf{h}_n(\boldsymbol{\eta}_{0p})\mathbf{h}_n(\boldsymbol{\eta}_{0q})^t.$$

The statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_{02}$ (Aitchison & Silvey, 1958; Rao, 1947).

Recently, the LM principle has been applied in the framework of IRT for evaluating differential item functioning (Glas, 1998) and the axioms of unidimensionality and local stochastic independence (Glas, 1999). Though originally presented in the framework of a fixed item administration design, these tests can also be applied in the framework of the stochastic design characteristics for CAT. However, the result with respect to the asymptotic distribution of the statistics does not automatically apply to the case of a stochastic design. The ignorability principle ensures consistency of estimators in a CAT design, but it does not apply to sample inferences, such as confidence intervals and the distributions of statistics for evaluation of model fit (Mislevy & Chang, 1998). Therefore, for the applications presented here, a power study under the null model will be part of the example to be presented. The results will show that the asymptotic distribution of the LM statistics is hardly affected by CAT.

14.3.2 An LM Test for the Fit of Item-Characteristic Curves

The idea of the LM test and modification index presented here will be to partition the latent ability continuum into a number of segments, and to evaluate whether an item's ICC conforms to the form predicted by the null model in each of these segments. However, the actual partitioning will take place on the observed number-correct scale rather than on the θ scale. Usually, the unweighted sum score and the associated estimate of θ will highly correlate. Let the item of interest be labeled k, and let the other items be labeled $i = 1, 2, \ldots, k-1, k+1, \ldots, K$. Let $r_n^{(k)}$ be the unweighted sum score on the response pattern of student n without item k. The possible scores $r_n^{(k)}$ will be partitioned into S disjoint subsets using boundary scores $r_0 < r_1 < r_2 \ldots < r_s < \ldots < r_S$, with $r_0 = 0$ and $r_S = K - 1$. Further, define

$$w_s\left(r_n^{(k)}\right) = \begin{cases} 1 & \text{if } r_{s-1} < r_n^{(k)} < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{14.10}$$

so $w_s\left(r_n^{(k)}\right)$ is an indicator function that assumes a value equal to one if the number-correct score without item k is in score range s. As an alternative model to the 2PLM and 3PLM, consider a model where the item discrimination and difficulty parameters are redefined as $a_n + \sum_s w_s\left(r_n^{(k)}\right)\delta_{1s}$ and $b_n + \sum_s w_s\left(r_n^{(k)}\right)\delta_{2s}$. The simultaneous hypothesis $\delta_{1s} = 0$ and $\delta_{2s} = 0$ ($s = 2, \ldots, S$; that is, $s = 1$ is used as a baseline) can be evaluated using an LM test. For respondents with a number-correct score in category s, it holds that

$$\omega_n(\delta_{1s}) = \theta_n(u_{in} - P_i(\theta_n)) \tag{14.11}$$

and

$$\omega_n(\delta_{2s}) = P_i(\theta_n) - u_{in}, \tag{14.12}$$

where $\omega_n(\delta_{1s})$ and $\omega_n(\delta_{2s})$ are defined as in (14.3). Using (14.2) it can be inferred that the elements of the vectors of first-order derivatives $h(\delta_1)$ and $h(\delta_1)$ are given by

$$\begin{aligned}&\sum_n w_s\left(r_n^{(k)}\right)E(\omega_n(\delta_{1s}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\\ &\quad= \sum_n w_s\left(r_n^{(k)}\right)E(\theta_i(u_{in} - P_i(\theta_n)) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\end{aligned} \tag{14.13}$$

and

$$\begin{aligned}&\sum_n w_s\left(r_n^{(k)}\right)E(\omega_n(\delta_{2s}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\\ &\quad= \sum_n w_s\left(r_n^{(k)}\right)E(P_i(\theta_n) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) - \sum_n w_s\left(r_n^{(k)}\right)u_{in}.\end{aligned} \tag{14.14}$$

Notice that (14.14) is the difference between the observed number of persons of sub-sample s with a correct score on item i, and its expected value. So (14.14) can be seen as a residual. A test for the simultaneous hypothesis $\delta_{1s} = 0$ and $\delta_{2s} = 0$, for $s = 1, \ldots, S-1$, can be based on a statistic with an asymptotic χ^2 distribution with $2(S-1)$ degrees of freedom, where the statistic defined by (14.9) is evaluated using MML estimates of the null model, that is, the 2PL model or 3PL model. It is also possible to define separate tests for $\delta_{1s} = 0$ or $\delta_{2s} = 0$ ($s = 1, \ldots, S-1$). These tests are based on LM statistics with $S-1$ degrees of freedom.

14.3.3 An LM Test for Parameter Drift

We noted earlier that parameter drift can be evaluated by checking whether pretest and online data can be properly described by the same IRT model. Consider G

groups labeled $g = 1, \dots, G$. It is assumed that the first group partakes in the pretesting stage, and the following groups partake in the online stage. The application of the LM tests to monitoring parameter drift is derived from the LM test for differential item functioning proposed by Glas (1998) for the 2PL model. This is a test of the hypothesis that the item parameters are constant over groups, that is, the hypothesis $a_{ig} = a_i$ and $b_{ig} = b_i$, for all g. To see the relation with the LM framework, consider two groups, and define a variable y_n that is equal to one if n belongs to the first group and zero if n belongs to the second group. Defining $a_{iy} = a_i + y_n\delta_1$ and $b_{iy} = b_i + y_n\delta_2$, the hypothesis given by $\delta_1 = 0$ and $\delta_2 = 0$ can be evaluated using the LM test. For more than two groups, more dummy variables y_n are needed to code group membership. This approach can of course also be used to monitor parameter drift in CAT. Further, generalization to the 3PL model entails adding $\delta_3 = 0$, with $c_{iy} = c_i + y_n\delta_3$, to the null hypothesis.

For actual implementation of this approach using adaptive testing data, the high correlation of estimates of the three item parameters discussed in the previous section must be taken into account. Another parameter estimation problem arises specifically in the context of adaptive testing. Guessing (which may be prominent in the calibration stage) may rarely occur in the online stage because items are tailored to the ability level of the respondents. Therefore, a test focused on all three parameters simultaneously often proves computationally unstable. In the present chapter, three approaches will be studied. In the first, the LM test will be focused on simultaneous parameter drift in a_i and b_i; in the second approach, the LM test will be focused on parameter drift in c_i. These two tests will be labeled $LM(a_i, b_i)$ and $LM(c_i)$, respectively. In the third approach, the guessing parameter will be fixed at some plausible constant, say, the reciprocal of the number of response alternatives of the items, and the LM statistic will be used to test whether this fixed guessing parameter is appropriate in the initial stage and remains so when the adaptive testing data are introduced. So the hypothesis considered is that $c_{ig} = c_i$ for all g. Using simulation studies, it will be shown that the outcomes of these three approaches are quite comparable.

14.3.4 A CUSUM Test for Parameter Drift

The CUSUM chart is an instrument of statistical quality control used for detecting small changes in product features during the production process (see, for instance, Wetherill, 1977). The CUSUM chart is used in a sequential statistical test, where the null hypothesis of no change is never accepted. In the present application, loss of production quality means that the item is becoming easier and less discriminating.

Contrary to the case of the LM test, the CUSUM test needs estimation of the item parameters for every group of students $g = 1, \dots, G$. As above, the first group partakes in the pretesting stage, and the following groups take an adaptive test. However, estimation of the guessing parameter is problematic in a CAT situation because, as already mentioned, guessing may be prominent in the calibration stage, while it may rarely occur in the online stage, where the items are tailored to the ability level of the respondents. Two possible solutions include fixing the guessing parameter to some plausible constant such as the reciprocal of the number of response options, or concurrent estimation of the item guessing parameter using all available data. In either approach, the null hypothesis is $a_{ig} - a_{i1} \geq 0$ and $b_{ig} - b_{i1} \geq 0$, for the respondent groups $g = 1, \ldots, G$. Therefore, a one-sided CUSUM chart will be based on the quantity

$$S_i(g) = \max\left\{S_i(g-1) + \frac{a_{i1} - a_{ig}}{Se(a_{ig} - a_{i1})} + \frac{b_{i1} - b_{ig}}{Se(b_{i1} - b_{ig} \mid a_{i1} - a_{ig})} - k, 0\right\}, \tag{14.15}$$

where $Se(a_{ig} - a_{i1}) = \sigma_a$ and $Se(b_{i1} - b_{ig} \mid a_{i1} - a_{ig}) = \sqrt{\sigma_b^2 - \sigma_{ab}^2/\sigma_a^2}$, with σ_a^2, σ_b^2, and σ_{ab} the appropriate elements of the covariance matrix of the parameter estimates given by (14.7). Further, k is a reference value determining the size of the effects one aims to detect. The CUSUM chart starts with $S_i(1) = 0$ and the null hypothesis is rejected as soon as $S_i(g) > h$, where h is some constant threshold value. The choice of the constants k and h determines the power of the procedure. In the case of the Rasch model, where the null hypothesis is $b_{ig} - b_{i1} \geq 0$, and the term involving the discrimination indices is lacking from (14.15), Veerkamp (1996) successfully uses $k = 1/2$ and $h = 5$. This choice was motivated by the consideration that the resulting test has good power against the alternative hypothesis of a normalized shift in item difficulty of approximately half a standard deviation. In the present case, one extra normalized decision variable is employed, namely, the variable involving the discrimination indices. So, for instance, a value $k = 1$ can be used to have power against a shift of one standard deviation of both normalized decision variables in the direction of the alternative hypothesis. However, there are no compelling reasons for this choice; the attractive feature of the CUSUM procedure is that the practitioner can choose the effect size k to meet the specific characteristics of the problem. Also, the choice of a value for h is determined by the targeted detection rate, especially by the trade-off between Type I and II errors. In practice, the values of h and k can be set using simulation studies. Examples will be given below.

14.4 Examples

In this section, the power of the procedures suggested above will be investigated using a number of simulation studies. Since all statistics involve an approximation of the standard error of the parameter estimates using (14.7), first the precision of the approximation will be studied by assessing the power of the statistics under the null model, that is, by studying the Type I error rate. Then the power of the tests will be studied under various model violations. These two topics will first be studied for the LM tests, then for the CUSUM test.

Table 14.2 Type I error rate of LM test

K	L	N_g	Percentage at 10% $LM(c_i)$	Significant $LM(a_i, b_i)$
50	20	500	8	9
		1000	10	10
	40	500	9	10
		1000	11	8
100	20	500	12	10
		1000	8	9
	40	500	10	12
		1000	10	10

In all simulations, the ability parameters θ were drawn from a standard normal distribution. The item difficulties b_i were uniformly distributed on $[-1.5, 1.5]$, the discrimination indices a_i were drawn from a log-normal distribution with a zero mean and a standard deviation equal to 0.25, and the guessing parameters were fixed at 0.20, unless indicated otherwise. In the online stage, item selection was done using the maximum information principle. The ability parameter was estimated by its expected a posteriori value (EAP); the initial prior was standard normal.

The results of eight simulation studies with respect to the Type I error rate of the LM test are shown in Table 14.2. The design of the study can be inferred from the first three columns of the table. It can be seen that the number of items K in the item bank was fixed at 50 for the first four studies and at 100 for the next four studies. In both the pretest stage and the online stages, test lengths L of 20 and 40 were chosen. Finally, as can be seen in the third column, the number of respondents per stage, N_g, was fixed at 500 and 1000 respondents. So summed over the pretest and online stage, the sample sizes were 1000 and 2000 respondents, respectively. For the pretest stage, a spiraled test administration design was used. For instance, for the $K = 50$ studies, for the pretest stage, five subgroups were used; the first subgroup was administered items 1 – 20, the second items 11 – 30, the third items 21 – 40 the fourth items 31 – 50, and the last group received the items 1 – 10 and 41 – 50. In this manner, all items drew the same number of responses in the pretest stage. For the $K = 100$ studies, for the pretest stage four subgroups administered 50 items were formed, so here the design was 1 – 50, 26 – 75, 51 – 100 and 1 – 25 and 76 – 100. For each study, 100 replications were run. The results of the study are shown in the last two columns of Table 14.2. These columns contain the percentages of $LM(c_i)$ and $LM(a_i, b_i)$ tests that were significant at the 10% level. It can be seen that the Type I error rates of the tests conform to the nominal value of 10%. These results support the adequacy of the standard error approximations for providing accurate Type I error rates.

The second series of simulations pertained to the power of the LM statistics under various model violations. The setup was the same as in the above study with $K = 100$ items in the item bank, a test length $L = 50$, $N_1 = 1000$ simulees in the pretest stage and $N_2 = 1000$ simulees in the online stages. Two model violations were simulated. In the first, the guessing parameter c_i went up in the online stage; in

Table 14.3 Power of LM test

Model Violation		Percentage at 10% $LM(a_i, b_i)$	Significant $LM(c_i)$
$c_i = 0.25$	Hits	25	15
	False alarm	08	10
$c_i = 0.30$	Hits	45	35
	False alarm	13	11
$c_i = 0.40$	Hits	95	85
	False alarm	17	20
$b_i = -0.20$	Hits	25	30
	False alarm	13	12
$b_i = -0.40$	Hits	55	70
	False alarm	15	20
$b_i = -0.60$	Hits	80	95
	False alarm	13	27

the second, the item difficulty b_i went down in the online stage. Six conditions were investigated: c_i rose from 0.20 to 0.25, 0.30, and 0.40, respectively, and b_i changed from the initial value by –0.20, –0.40 and –0.60, respectively. These model violations were imposed on the items 5, 10, 15, etc. So 20 out of the 100 items were affected by this form of parameter drift. 100 replications were made for each condition. Both the $LM(c_i)$ and $LM(a_i, b_i)$ tests were used. The results are shown in Table 14.3. This table displays both the percentage of "hits" (correctly identified items with parameter drift) and "false alarms" (items without parameter drift erroneously identified as drifting). Three conclusions can be drawn. Firstly, it can be seen that the power of the tests increases as the magnitude of the model violation grows. Secondly, the power of the test specifically aimed at a model violation is always a little larger than the power of the other test, but the differences are quite small. For instance, in the case $b_i = -0.60$, the power of $LM(a_i, b_i)$ is 0.95, while the power of $LM(c_i)$ is 0.85. The third conclusion that can be drawn from the table is that the percentage of "false alarms" is clearly higher than the nominal 10% error rate. A plausible explanation might be that the improper parameter estimates of the 20% items with parameter drift influence the estimates of the 80% non-affected items. Finally, it can be noted that the agreement between the two tests with respect to the flagged items was high; agreement between the two tests was always higher than 0.84.

As mentioned above, the power of the CUSUM procedure is governed by choosing an effect size k and a critical value h. A good way to proceed in a practical situation is to calibrate the procedure when the pretest data have become available. First, the practitioner must set an effect size k of interest. Then, assuming no parameter drift, online data can be simulated using the parameter estimates of the pretest stage. Finally, CUSUM statistics can be computed to find a value for h such that an acceptable Type I error rate is obtained. An example will be given using the same set-up as above: there were $K = 100$ items in the item bank, test length was $L = 50$, and the pretest data consisted of the responses of $N_1 = 1000$ simulees.

Table 14.4 Type I error rate of CUSUM test

Effect Size	$h = 2.5$	$h = 5.0$	$h = 7.5$	$h = 10.0$
$k = 0.50$	17	04	01	00
$k = 1.00$	09	06	01	00
$k = 2.00$	01	00	00	00

Then, four batches of responses of $N_g = 1000$ ($g = 2, \ldots, 5$) simulees were generated as online data, and CUSUM statistics $S_i(g)$ were computed for the iterations $g = 2, \ldots, 5$. This procedure was carried out for three effect sizes k and four thresholds h; the values are shown in Table 14.4.

In the table, the percentages items flagged in the fifth iteration ($g = 5$) of the procedure are shown for the various combinations of k and h. Since no parameter drift was induced, the percentages shown can be interpreted as Type I error rates. For an effect size $k = 0.50$, it can be seen that a value $h = 2.5$ results in 17% flagged items, which is too high. A value $h = 5.0$ results in 4% flagged items, which might be considered an acceptable Type I error rate. Also, for an effect size $k = 1.00$ a critical value $h = 5.0$ seems a good candidate. Finally, for $k = 2.00$, all four values of h produce low Type I error rates. So it must be concluded that, given the design and the sample size, detection of parameter drift with an effect size of two standard deviations may be quite difficult.

This result was further studied in a set of simulations where model violations were introduced. These studies used the setup $K = 100$, $L = 50$, and $N_g = 1000$, for $g = 1, \ldots, 5$. The model violations were similar to the ones imposed above. So in six conditions, the guessing parameter c_i rose from 0.20 to 0.25, 0.30, and 0.40, respectively, and b_i changed from the initial value by −0.20, −0.40, and −0.60, respectively. Again, for each condition, 20 of the 100 items were affected by the model violation. The results are shown in Table 14.5. For the simulation studies with effect sizes $k = 0.50$ and $k = 1.00$, a critical value $h = 5.0$ was chosen; for the studies with effect size $k = 2.00$, the critical value was $h = 2.5$.

For every combination of effect size and model violation, 20 replications were made. The last four columns of Table 14.5 give the percentages of "hits" (flagged items with parameter drift) and "false alarms" (erroneously flagged items per condition) for the iterations $g = 2, \ldots, 5$. The percentages are aggregated over the 20 replications per condition. As expected, the highest percentages of "hits" were obtained for the smaller effect sizes $k = 0.50$ and $k = 1.00$, and the larger model violations. The top is the combination $k = 1.00$ and $b_i = -0.60$, which, for $g = 5$, has an almost perfect record of 99% "hits". In this condition, the percentage of "false alarms" remained at a 10% level. The worst performances were obtained for combinations of $k = 0.50$ and $k = 2.00$ with small violations as $c_i = 0.25$, $c_i = 0.30$, and $b_i = -0.20$. These conditions both show a low "hit" rate and a "false alarm" rate of approximately the same magnitude, which is relatively high for a "false alarm" rate.

Table 14.5 Power of CUSUM test

Effect	Model		Iteration			
Size	Violation		$g = 2$	$g = 3$	$g = 4$	$g = 5$
$k = 0.50$	$c_i = 0.25$	Hits	00	00	05	15
		False alarm	00	04	05	13
	$c_i = 0.30$	Hits	00	05	10	20
		False alarm	00	03	05	06
	$c_i = 0.40$	Hits	00	30	75	85
		False alarm	00	00	01	03
$k = 1.00$	$c_i = 0.25$	Hits	15	25	30	45
		False alarm	05	13	17	21
	$c_i = 0.30$	Hits	15	35	55	50
		False alarm	03	03	03	06
	$c_i = 0.40$	Hits	30	75	90	85
		False alarm	03	04	06	09
$k = 2.00$	$c_i = 0.25$	Hits	00	05	15	15
		False alarm	00	00	03	00
	$c_i = 0.30$	Hits	05	15	15	20
		False alarm	03	01	04	04
	$c_i = 0.40$	Hits	15	30	55	60
		False alarm	00	01	01	01
$k = 0.50$	$b_i = -0.20$	Hits	00	00	10	15
		False alarm	00	00	06	05
	$b_i = -0.40$	Hits	00	15	45	60
		False alarm	01	06	09	15
	$b_i = -0.60$	Hits	05	35	65	80
		False alarm	00	00	04	04
$k = 1.00$	$b_i = -0.20$	Hits	00	20	40	35
		False alarm	00	01	03	05
	$b_i = -0.40$	Hits	25	50	55	65
		False alarm	01	04	06	09
	$b_i = -0.60$	Hits	20	75	95	99
		False alarm	03	06	10	10
$k = 2.00$	$b_i = -0.20$	Hits	00	00	05	05
		False alarm	00	01	03	03
	$b_i = -0.40$	Hits	05	10	30	35
		False alarm	00	00	03	01
	$b_i = -0.60$	Hits	00	25	75	75
		False alarm	01	03	04	03

14.5 Discussion

This chapter showed how to evaluate whether the IRT model of the pretest stage also fits the online stage. Two approaches were presented. The first was based on LM statistics. It was shown that the approach supports the detection of specific model violations and has the advantage of known asymptotic distributions of the statistics on which it is based. Two specific model violations were considered here, but the approach also applies to other model violations, such as violation of local independence and multidimensionality (see Glas, 1999). The second approach is based on CUSUM statistics. The distribution of these statistics is not known, but an appropriate critical value h can be found via simulation. An advantage, however, is that the practitioner can tune the procedure to the needs of the specific situation. When choosing h, the subjective importance of making "hits" and avoiding "false alarms" can be taken into account, and the effect size k can be chosen to reflect the magnitude of parameter drift judged relevant in a particular situation. Summing up, both approaches provide practical tools for monitoring parameter drift.

References

Ackerman, T. A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20,* 309–310.

Ackerman, T. A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–329.

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813–828.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R. D. & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer-Verlag.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B. 39,* 1–38.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Laird, and D. Rubin). *J. R. Statist. Soc. B., 39,* 1–38.

Fischer, G. H. & Scheiblechner, H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12,* 23–51.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1) (pp. 236–258). Norwood, NJ: Ablex Publishing Corporation.

Glas, C. A. W. (1988). The Rasch Model and multi-stage testing. *Journal of Educational Statistics, 13,* 45–52.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64,* 273–294.

Glas, C. A. W. & Suarez-Falcon, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27,* 87–106.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–903.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226–233.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Mislevy, R. J. & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65,* 149–156.

Mislevy, R. J. & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Neyman, J. & Scott, E. L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16*, 1–32.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Rigdon S. E. & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Stocking, M. L. (1993). *Controlling exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Sympson, J. B. & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.

Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing.* Unpublished doctoral thesis, Twente University, the Netherlands.

Wetherill, G. B. (1977). *Sampling inspection and statistical quality control* (2nd ed.). London: Chapman and Hall.

Wilson, D. T., Wood, R. & Gibbons, R. D. (1991) *TESTFACT: Test scoring, item statistics, and item factor analysis* (computer software). Chicago: Scientific Software International, Inc.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International, Inc.

Chapter 15
Estimation of the Parameters in an Item-Cloning Model for Adaptive Testing

Cees A.W. Glas, Wim J. van der Linden, and Hanneke Geerlings

15.1 Introduction

Item response theory (IRT) models with random person parameters have become a common choice among practitioners in the field of educational and psychological measurement. Though initially the choice for such models was motivated by an attempt to get rid of the statistical problems inherent in the incidental nature of the person parameters (Bock & Lieberman, 1970), the insight soon emerged that such models more adequately represent cases where the focus is not on the measurement of individual persons but on the estimation of characteristics of populations. Early examples of models with random person parameters in the literature are those proposed by Andersen and Madsen (1977) and Sanathanan and Blumenthal (1978), who were interested in estimates of the mean and variance in a population of persons, and by Mislevy (1991), who provided tools for inference from a response model with a regression structure on the person parameters introduced to account for sampling persons differing background variables.

In spite of the popularity of these models with random person parameters, the measurement literature shows only a recent interest in models with random item parameters. Nevertheless, such models have the potential of better representing testing formats that involve random selection of items, cases where sets of items can be considered as exchangeable once we know they belong to the same "group" or "class", or testing in which the item parameters can be expected to vary over different groups of respondents (de Boeck, 2008).

C.A.W. Glas (✉)
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden
CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA

H. Geerlings
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_15,

The most obvious case of measurement with random items has been a predecessor of computerized adaptive testing known as domain-referenced testing (e.g., Millman, 1973). In one version of this type of testing, items are randomly sampled from a pool until the test taker can be scored with satisfactory precision. The model originally used to guide domain-referenced testing programs with dichotomously scored items was the binomial error model (Lord & Novick, 1968, chap. 23), which is given by

$$\Pr\{X_n = x \mid k, \pi_n\} = \binom{k}{x} \pi_n^x (1 - \pi_n)^{k-x},$$

where X_n is the number of successes for person n on a test of size k sampled from the domain and π_n is the person's success parameter. Clearly, the success parameter in this model depends on both the person and the domain of test items. Attempts to decompose π_n into separate components for the person and the items led to the introduction of IRT models with random item and person parameters. One of the first models of this kind was offered by Albers, Does, Imbos, and Janssen (1989), who introduced an extension of the Rasch model with random item and person parameters to estimate the progress of learning in a longitudinal study with tests sampled from the same pool of items at different time points.

Another application of the idea of item sampling relevant to adaptive testing has become available through the introduction of computer-generated items in educational measurement. Using an item-cloning technique (see, for instance, Bejar, 1993, and Roid & Haladyna, 1982), it is no longer necessary to write each individual item in the domain. Instead, computer algorithms are used to generate families of items from a smaller set of "parent items". One of the more popular types of computer generation of items is based on so-called replacement set procedures (Millman & Westman, 1989), where the algorithm replaces elements in the parent item (e.g., key terms, relations, numbers, and distractors) with randomly sampled elements from well-defined sets of alternatives. The substitution introduces some random variation within a family of items derived from the same parent, and it becomes necessary to model the item parameters as random and include hyperparameters that describe the distributions of the item parameters within the families (Geerlings, van der Linden & Glas, 2009).

Adaptive testing based on item cloning involves a two-stage item selection process: first, a family of items optimal at the updated ability estimate is selected; and second, an item is sampled randomly from the family (Glas & van der Linden, 2001, 2003). Observe that this sampling is more general than the previous case of domain-referenced testing because we now consider sampling from alternative families in the same pool.

An example of the use of a model with random item parameters outside adaptive testing is given in Janssen, Tuerlinckx, Meulders, and de Boeck (2000). These authors are interested in the process of standard setting on a criterion-referenced test with sections of items in the test grouped under different criteria. As in the testlet model by Bradlow, Wainer, and Wang (1999), it is assumed that this grouping involves within-group dependency, and therefore an IRT model is chosen to have random item parameters with different distributions for different sections. A Bayesian argument in favor of this approach is that if the only thing known a priori about the items is their grouping under a common criterion, they are exchangeable given the criterion and can be treated as if they are a random sample. Another example is found in large-scale surveys, such as international educational surveys or national educational assessments. For instance, in such surveys as PISA, TIMSS, and PIRLS, it has proven useful to model cultural bias in background questionnaires using country-specific item parameters. An analogous approach in the field of consumer research was proposed by de Jong, Steenkamp, and Fox (2007).

It is the purpose of this chapter to give a statistical treatment of the problem of estimating the parameters in the item-cloning model developed for adaptive testing in Glas & van der Linden (2001, 2003); see also Johnson and Sinharay (2005) and Sinharay, Johnson, and Williamson (2003). The model has random item parameters and multiple families of items. The model allows not only for item properties that have traditionally been modeled using the three-parameter logistic model (item difficulty, discriminating power, and possibility to guess) but also for dependency between these features within item families (e.g., correlation between parameters for discriminating power and guessing). The model is hierarchical in the sense that it has hyperparameters that describe the distributions of the item parameters within their families. The hyperparameters will also be referred to as Level 2 item parameters. The relation between the actual items that are administered and the Level 1 and Level 2 parameters depends on the sampling model adopted for the application. In adaptive testing with item cloning, the actual items are at Level 1, but they are grouped under Level 2 units representing their families.

Two different estimation procedures are presented. The first procedure is fully Bayesian with sampling from the joint posterior distribution of all parameters using a Markov chain Monte Carlo (MCMC) simulation algorithm (i.e., a Gibbs sampler). In the second procedure, Bayes modal estimates are computed using a likelihood distribution marginalized over the incidental parameters. The Bayes modal estimates are derived in two versions: one based on the assumption that the items at Level 1 are unique, and one where it is assumed that at Level 1 the items are presented to a finite number of persons greater than one. These and other features of the sampling design critical to parameter estimation in the model will be discussed in considerable detail.

15.2 The Model

We will use family as a general name for a population of items at Level 2 from which the Level 1 items in the tests are realizations. The set of families from which they are obtained is denoted as $p = 1, \ldots, P$. The items in family p are labeled as $i_p = 1, \ldots, k_p$. Throughout this chapter, we will assume that P, is fixed. For ease of exposition, we will also assume that $P \leq k$, where k is the test length. As for the values of k_p, different options will be discussed below.

It proves convenient to introduce a sampling design variable d_{ni_p}, which assumes a value equal to one if person n responded to item i_p, and zero otherwise. Let X_{ni_p} be the response variable for person n and item i_p. If $d_{ni_p} = 1$, X_{ni_p} takes the value one for a correct response and a value zero for an incorrect response. If $d_{ni_p} = 0$, X_{ni_p} takes an arbitrary value r ($r \neq 0; r \neq 1$). Notice that with this definition the design variables are completely determined by the response variables; they are only introduced to facilitate the mathematical presentation.

15.2.1 Level 1 Model

The first-level model is the three-parameter normal-ogive (3PNO) model, which describes the probability of a correct response as

$$p(x_{ni_p} = 1 \mid d_{ni_p} = 1, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = c_{i_p} + (1 - c_{i_p})\Phi(a_{i_p}\theta_n - b_{i_p}), \quad (15.1)$$

where a_{i_p}, b_{i_p}, and c_{i_p} are the Level 1 item parameters, θ_n is a person parameter, and $\Phi(.)$ is the normal cumulative density function. The parameterization of the model in (15.1) is slightly different from the usual parameterization, which has $a_{i_p}(\theta_n - b_{i_p})$ as the argument of $\Phi(.)$. The only motivation for our choice is to simplify the presentation below.

The reason for considering the 3PNO model rather than the 3PL model is that the Bayesian estimation procedure that will be presented here is a generalization of the Bayesian estimation procedure developed by Albert (1992) for the 2PNO and Béguin and Glas (2001) for the 3PNO. However, as is well known, for an appropriately chosen scale factor, both models are numerically nearly indistinguishable and either model is expected to fit only if the other one does.

15.2.2 Level 2 Model

The values of the Level 1 item parameters $(a_{i_p}, b_{i_p}, c_{i_p})$ in (15.1) are considered as realizations of a random vector. We will use the transformation

$$\boldsymbol{\xi}_{i_p} = (a_{i_p}, b_{i_p}, \text{logit}\, c_{i_p}), \quad (15.2)$$

which gives the item parameters scales for which the following assumption of multivariate normality is reasonable:

$$\boldsymbol{\xi}_{i_p} \sim N(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p), \quad (15.3)$$

where $\boldsymbol{\mu}_p$ is the vector with the mean values of the item parameters for family p and $\boldsymbol{\Sigma}_p$ their covariance matrix. Observe that the hyperparameters ($\boldsymbol{\mu}_p$, $\boldsymbol{\Sigma}_p$) are allowed to vary across item families.

In the inferences below, we assume that θ_n has a standard normal distribution

$$\theta_n \sim N(0, 1). \tag{15.4}$$

This assumption holds if person n is from a population of exchangeable persons with a normal distribution of abilities. Persons and items are thus distributed independently; that is, we do not assume that the items are sampled dependently on the person abilities.

15.2.3 Prior for Hyperparameters

A conjugate prior distribution is chosen for the hyperparameters $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. Since the Level 1 item parameters are normally distributed, this leads to a normal-inverse-Wishart distribution for $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ (see, for instance, Box & Tiao, 1973, or Gelman, Carlin, Stern & Rubin, 1995). The prior follows from the specification

$$\begin{aligned}
\boldsymbol{\Sigma}_p &\sim \text{Inverse-Wishart}_{\nu_{p0}}(\boldsymbol{\Sigma}_{p0}), \\
\boldsymbol{\mu}_p \mid \boldsymbol{\Sigma}_p &\sim \text{MVN}(\boldsymbol{\mu}_{p0}, \boldsymbol{\Sigma}_p/\kappa_{p0})
\end{aligned}$$

and has a density given by

$$\begin{aligned}
p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) \propto{}& \left|\boldsymbol{\Sigma}_{p0}\right|^{-((\nu_{p0}+3)/2+1)} \\
& \exp\left(-\frac{1}{2}tr(\boldsymbol{\Sigma}_{p0}\boldsymbol{\Sigma}_p^{-1}) - \frac{\kappa_{p0}}{2}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_{p0})^t \boldsymbol{\Sigma}_p^{-1}(\boldsymbol{\mu}_p - \boldsymbol{\mu}_{p0})\right),
\end{aligned} \tag{15.5}$$

where $\boldsymbol{\Sigma}_{p0}$ and ν_{p0} are the scale matrix and degrees of freedom for the prior on $\boldsymbol{\Sigma}_p$ and $\boldsymbol{\mu}_{p0}$ and κ_{p0} are the mean and weight for the prior on $\boldsymbol{\mu}_p$, respectively. The weight expresses the information in the prior distribution as the number of prior measurements it can be equated to.

The goal of the prior is only to bind the distribution of the hyperparameters to a likely region of possible values. In practice, analysis may start with a mild common prior for all hyperparameters, that is, choosing the same values $\boldsymbol{\mu}_{p0}, \boldsymbol{\Sigma}_{p0}, \kappa_{p0}$, and ν_{p0} for all p. If the analyses show these priors to be less adequate, more stringent priors for some of the Level 2 item parameters should be chosen.

The choice of priors should always allow for the fact that the model can be poorly identified, if the values of the θ parameter are concentrated in a region for which the response functions are equally well approximated by different combinations of item parameter values. Effects of such trade-off relations between the model parameters on their maximum likelihood estimators have been described by authors as Wingersky and Lord (1984), Stocking (1989), and Patz and Junker (1999a). For example, Wingersky and Lord (1984, p. 355) observe that the sampling covariance between estimates of the discrimination and difficulty parameters tends to be positive for easy items and negative for difficult items. These authors also describe effects of trade-off relations between the other item parameters. The existence of such relations motivated our choice of values for the prior covariance matrix $\boldsymbol{\Sigma}_{p0}$ in the empirical examples below. For instance, a negative value was chosen for the covariance between the discrimination and the logit-guessing parameter because similar response functions are obtained if the value of the discrimination parameter goes down when the lower asymptote goes up. Likewise, the positive value for the covariance between the difficulty and discrimination parameters was chosen because a lower value for the former can be counterbalanced by a lower value for the latter, particularly if the respondents are highly proficient. More precise choices of the absolute values of the elements of the prior covariance matrices are possible if we have information on the distribution of the person parameters and use techniques from optimal design (Berger, 1997; van der Linden, 1994) to determine our sampling design.

15.2.4 Likelihood Function

The response vector of person n is denoted as $\mathbf{x}_n$, and defined by $\mathbf{x}'_n = (x_{ni_1}, \ldots, x_{ni_p}, \ldots, x_{nk_P})$. The ensemble of response data is collected in a data matrix $\mathbf{X}$, which has rows $\mathbf{x}'_n$, for $n = 1, \ldots, N$. Using the assumptions of (1) independence between persons, (2) independence between items and persons, and (3) local independence within persons, the portion of the likelihood function given the response data $\mathbf{X}$ for the model in (15.1) can be written as

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\xi}; \mathbf{X}) &= \prod_n p(\mathbf{x}_n \mid \mathbf{d}_n, \theta_n, \boldsymbol{\xi}) \\ &= \prod_n \prod_p \prod_i p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p}), \end{aligned} \tag{15.6}$$

where $\mathbf{d}_n$ is a vector with entries d_{ni_p}, $\boldsymbol{\theta}$ is a vector with entries θ_n, and $\boldsymbol{\xi}$ is a vector with entries $\boldsymbol{\xi}_{i_p}$, for all values of the indices n, p, and i_p. The convention will be followed that $p(x_{ni_p} \mid d_{ni_p} = 0, \theta_n, a_{i_p}, b_{i_p}, c_{i_p}) = 1$.

15.3 Sampling Design

A sampling design for a calibration study is a choice of values for the design variables (d_{ni_p}). It governs the sampling of the items and persons and thus controls how much response data we have for each realization of the item parameters as well as the family and person parameters in the model. As already indicated, the number of item families P is fixed and not larger than test length k. The size of the sample of persons is denoted as N.

The following two quantities of the sampling design are important for our treatment of the estimation problem: (1) the number of persons who respond to item i_p,

$$N_{i_p} = \sum_n d_{ni_p}, \tag{15.7}$$

(2) the number of persons who respond to an item from family p,

$$N_p = \sum_i N_{i_p}. \tag{15.8}$$

Since every item family p has nine hyperparameters (three in $\boldsymbol{\mu}_p$ and six in $\boldsymbol{\Sigma}_p$), it is assumed that $N_p \geq 9$. Because we sum over n in (15.8) and P is fixed, N_p always grows in the size of the sample, N. The assumption $N_p \geq 9$ is thus easily met in real-life applications. Also, we can always consider $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ as structural parameters in the sense of Neyman and Scott (1948): if the sample size N goes up, the dimension of these parameters remains fixed, while their number of observations N_p also goes up (see below).

15.3.1 Type of Design

As for the size of N_{i_p}, the following two cases are considered:

Case 1. N_{i_p} grows with N but k_p is fixed. This case arises if the number of items per family are fixed by design. Because the number of families and the length of the test are fixed, an increase in the number of persons can only be met if more persons get the same items. An example of this type of design is testing with item generation where a finite set of clones per family is generated before testing and each person gets one item sampled from the set for each family. The same type of design arises in educational assessments with a finite number of subpopulations.

Case 2. k_p grows with N but N_{i_p} is fixed. This case arises if the number of persons that get an item is fixed by design. As a consequence, an increase in the sample size can only be met if more items per family are generated. An example of this type of design is testing with item generation where the computer generates a new item from each family for each person ("item generation on the fly") or the number of persons per item is kept below a threshold for security reasons.

15.3.2 Structural and Incidental Parameters

It is of interest to discuss these two cases further in connection with the distinction between structural and incidental parameters in statistical models introduced by Neyman and Scott (1948). In an estimation problem, the number of structural parameters remains finite if the number of observations goes to infinity, whereas the number of incidental parameters goes to infinity too. The presence of incidental parameters causes problems for statistical inference; for instance, the solutions to the likelihood equations for the structural parameters may lose their consistency or asymptotic efficiency.

In the estimation problem addressed in this chapter, $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ are always structural parameters, but, dependent on the sampling design, the item parameters $\boldsymbol{\xi}_{i_p}$ can be structural (Case 1) or incidental (Case 2). However, if the item parameters $\boldsymbol{\xi}_{i_p}$ are structural, we need not necessarily be interested in them. For example, in computerized adaptive testing with item cloning (Glas and van der Linden, 2003), once the item families are calibrated, no matter the type of item cloning used during the test, we only use the hyperparameters $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ to select the families and score the persons.

On the other hand, in applications such as educational assessments with items behaving differently in different populations, we certainly are interested in $\boldsymbol{\xi}_{i_p}$. Estimates of these parameters may help us to score the persons in the assessment more accurately than scoring based only on the hyperparameters, particularly if they represent families with large variation. In fact, for this application the only reason to use the multilevel IRT model in this chapter is to get better item parameter estimates by “borrowing information” from the other items in the same family.

To obtain consistent estimates of structural parameters, Kiefer and Wolfowitz (1956) suggested marginalizing the likelihood function over the incidental parameters. In marginal maximum likelihood (MML) and Bayes modal estimation, marginalization is accomplished by numerical methods. In a fully Bayesian analysis using a MCMC algorithm, marginalization is accomplished by Monte Carlo integration.

15.3.3 Estimation Methods

We will discuss several different methods to estimate parameters in the multilevel model in this chapter. The first method is fully Bayesian and based on the Gibbs sampler. This method is indifferent as to the status of $\boldsymbol{\xi}_{i_p}$ as structural or incidental parameter. If parameters have to be removed from the problem because they are incidental or structural but there is no interest in them, they should just be ignored in the draws from the posterior distribution in the output of the computer program. However, as shown below, the method runs into identifiability problems if $N_{i_p} = 1$ for some i_p.

The other methods involve Bayes modal estimation. The first method, which will be summarized only, was presented in Glas and van der Linden (2003) and is based on the assumption that $N_{i_p} = 1$ for all items. The method estimates the hyperparameters $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ from a likelihood marginalized over $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$. Because we have a prior for these parameters, it seems obvious to estimate them as the mode of their posterior distribution rather than to compute an MML estimate. The marginalization has two possible advantages. First, the method works for $N_{i_p} = 1$. Second, if the item parameters are incidental by design, the estimators of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ retain the usual asymptotic properties of a maximum likelihood estimate. This method is thus an alternative to the fully Bayesian method if $N_{i_p} = 1$ for some of the items.

Because estimation based on a Gibbs sampler is time-intensive, it may seem attractive to use the same Bayes modal estimator for cases with $N_{i_p} \geq 2$. For this case, marginalization of the likelihood over the parameter $\boldsymbol{\xi}$ quickly leads to estimation equations that are intractable. We will show this for the case of $N_{i_p} = 2$. However, for problems with smaller numbers of items, Bayes modal estimation remains possible if we marginalize only over $\boldsymbol{\theta}$ and estimate the item parameters $\boldsymbol{\xi}$ along with the hyperparameters, which makes sense only if they are structural by design. Our conclusion from this case will be that if N_{i_p} increases for some of the items, the fully Bayesian method with the Gibbs sampler remains feasible and should be recommended.

15.4 Fully Bayesian Estimation (Gibbs Sampler)

The Gibbs sampler is a Markov chain Monte Carlo (MCMC) procedure for sampling from the joint posterior distribution of all items (Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and the components are sampled consecutively from their conditional posterior distributions given the last sampled values for all other components. This sampling scheme is repeated until the distribution of sampled values forms a stable estimate of the joint posterior distribution. Albert (1992) applies Gibbs sampling to estimate the parameters of the 2PNO model. A generalization to the 3PNO model is given by Béguin and Glas (2001). A more general introduction to MCMC for IRT models is found in Patz and Junker (1999a), whereas applications for models with multiple raters, multiple item types, and missing data are given in Patz and Junker (1999b), models with a multilevel structure on the ability parameters in Fox and Glas (2001) and multidimensional models in Shi and Lee (1998) and Béguin and Glas (2001).

15.4.1 Data Augmentation

Béguin and Glas (2001) introduce a data augmentation scheme for the 3PNO that will also be used here. This data augmentation scheme is based on the following interpretation of the 3PNO. Suppose that a person n either knows the correct answer to item i_p with probability $\Phi(\lambda_{ni_p})$, with $\lambda_{ni_p} = a_{i_p}\theta_n - b_{i_p}$, or does not know the correct answer with probability $1 - \Phi(\lambda_{ni_p})$. In the first case, a correct response is given with probability 1 and in the second case, the person guesses the correct response with probability c_{i_p}. Then the marginal probability of a correct response is equal to $\Phi(\lambda_{ni_p}) + c_{i_p}(1 - \Phi(\lambda_{ni_p}))$. Let

$$W_{ni_p} = \begin{cases} 1 \text{ if person } n \text{ knows the correct answer to item } i_p, \\ 0 \text{ if person } n \text{ doesn't know the correct answer to item } i_p. \end{cases} \tag{15.9}$$

So if $W_{ni_p} = 0$, person n will guess the response to item i_p, and if $W_{ni_p} = 1$, person n will know the answer and will give a correct response. Consequently, the conditional probability of $W_{ni_p} = w_{ni_p}$ given $X_{ni_p} = x_{ni_p}$ is given by

$$\begin{aligned} P(W_{ni_p} = 1 \mid X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) &\propto \Phi(\lambda_{ni_p}), \\ P(W_{ni_p} = 0 \mid X_{ni_p} = 1, \lambda_{ni_p}, c_{i_p}) &\propto c_{i_p}(1 - \Phi(\lambda_{ni_p})), \\ P(W_{ni_p} = 1 \mid X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) &= 0, \\ P(W_{ni_p} = 0 \mid X_{ni_p} = 0, \lambda_{ni_p}, c_{i_p}) &= 1. \end{aligned} \tag{15.10}$$

In addition to W_{ni_p}, following Albert (1992), the data are also augmented with latent data Z_{ni_p}, which are independent and normally distributed with mean $\lambda_{ni_p} = a_{i_p}\theta_n - b_{i_p}$ and standard deviation equal to 1. The latent data W_{ni_p} are considered as indicators of the sign of Z_{ni_p}; if $W_{ni_p} = 0$ or 1, Z_{ni_p} is negative or positive, respectively.

15.4.2 Posterior Distribution

The aim of the procedure is to simulate samples from the joint posterior distribution given by

$$p(\boldsymbol{\xi}, \boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{Z}, \mathbf{W} \mid \mathbf{X}) \propto p(\mathbf{Z}, \mathbf{W} \mid \mathbf{X}; \boldsymbol{\xi}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\xi} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \boldsymbol{\mu}_{p0}, \boldsymbol{\Sigma}_{p0}). \tag{15.11}$$

The right-hand-side probability (density) functions are given by (15.12) (see below), (15.4), (15.3), and (15.5), respectively.

15.4.3 Steps in the Gibbs Sampler

The steps of the Gibbs sampler are the following.

Step 1 The posterior $p(\mathbf{Z}, \mathbf{W} \mid \mathbf{X}; \boldsymbol{\xi}, \boldsymbol{\theta})$ is factored as $p(\mathbf{Z} \mid \mathbf{W}, \boldsymbol{\xi}, \boldsymbol{\theta}) p(\mathbf{W} \mid \mathbf{X}; \boldsymbol{\xi}, \boldsymbol{\theta})$. For the cases with $d_{ni_p} = 1$, the values of w_{ni_p} and z_{ni_p} are drawn in the following two substeps:

(a) w_{ni_p} is drawn from the conditional distribution of W_{ni_p} given the data $\mathbf{X}$, and $\boldsymbol{\xi}$, $\boldsymbol{\theta}$**,** which is given in (15.10).

(b) z_{ni_p} is drawn from the conditional distribution of Z_{ni_p} given $\mathbf{W}$, $\boldsymbol{\xi}$, and $\boldsymbol{\theta}$, which is defined as

$$Z_{ni_p} \mid \mathbf{W}, \boldsymbol{\xi}, \boldsymbol{\theta} \sim \begin{cases} N(\lambda_{ni_p}, 1) \text{ truncated at the left at 0,} \quad \text{if } w_{ni_p} = 1, \\ N(\lambda_{ni_p}, 1) \text{ truncated at the right at 0, if} w_{ni_p} = 0. \end{cases} \tag{15.12}$$

Step 2 The value of $\boldsymbol{\theta}$ is drawn from the conditional posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Z}$ and $\boldsymbol{\xi}$. The distribution is derived as follows. From the definition of the latent variables Z_{ni_p}, it follows that $Z_{ni_p} + b_{i_p} = a_{i_p}\theta_n + \varepsilon_{ni_p}$, with ε_{ni_p} being a standard normally distributed residual. Because (a_{i_p}, b_{i_p}) is fixed, the equality defines a linear model for the regression of $Z_{ni_p} + b_{i_p}$ on a_{i_p}, with regression coefficient θ_n, which has a normal prior with parameters $\mu = 0$ and $\sigma = 1$. Therefore, the posterior of θ_n is also normal (this is a well-known result from Bayesian regression analysis; see, for instance, Box & Tiao, 1973). That is,

$$\theta_n \sim N\left(\frac{\hat{\theta}_n/\nu + \mu/\sigma^2}{1/\nu + 1/\sigma^2}, \frac{1}{1/\nu + 1/\sigma^2}\right), \tag{15.13}$$

where

$$\hat{\theta}_n = \left[\sum_p \sum_{i_p} d_{ni_p}\alpha_{i_p}(z_{ni_p} + b_{i_p})\right] / \left[\sum_p \sum_{i_p} d_{ni_p}a_{i_p}^2\right]$$

and

$$\nu = 1/\left[\sum_p \sum_{i_p} d_{ni_p}a_{i_p}^2\right].$$

Step 3 The vector of random item parameters $\boldsymbol{\xi}_{i_p}$ is partitioned into $\boldsymbol{\delta} = (\boldsymbol{\delta}_{i_p}) = (a_{1_1}, b_{1_1}, \ldots, a_{i_p}, b_{i_p}, \ldots)$ and $\mathbf{c} = (c_{1_1}, \ldots, c_{i_p}, \ldots)$. Hence, their conditional posterior density factors as

$$p(\boldsymbol{\xi}_{i_p} | \mathbf{x}_{i_p}, \mathbf{z}_{i_p}, \mathbf{w}_{i_p}, \boldsymbol{\theta}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) = p(\text{logit}\, c_{i_p} | \mathbf{x}_{i_p}, \mathbf{w}_{i_p}, \boldsymbol{\mu}_p^{c|\delta}, \boldsymbol{\Sigma}_p^{c|\delta}) \\ \times p(\boldsymbol{\delta}_{i_p} | \mathbf{z}_{i_p}, \boldsymbol{\theta}, \boldsymbol{\mu}_p^{\delta}, \boldsymbol{\Sigma}_p^{\delta}),$$

where $\boldsymbol{\mu}_p^{c|\delta}$ and $\boldsymbol{\Sigma}_p^{c|\delta}$ are the expectation and variance of logit c_{i_p} conditional on $\boldsymbol{\delta}_{i_p}$. Furthermore, $\boldsymbol{\mu}_p^{\delta}$ and $\boldsymbol{\Sigma}_p^{\delta}$ are the elements of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$ corresponding to the discrimination and difficulty parameters. Then the following two substeps are made:

(a) The value of $\boldsymbol{\delta}_{i_p}$ is drawn from the conditional posterior distribution of the parameters of $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, $\mathbf{z}_{i_p}$, $\boldsymbol{\mu}_p^{\delta}$, and $\boldsymbol{\Sigma}_p^{\delta}$. The distribution is derived as follows: parameters $\boldsymbol{\delta}_{i_p}$ can be viewed as coefficients of the regression of $\mathbf{z}_{i_p} = (z_{ni_p})$, on $\mathbf{X} = (\boldsymbol{\theta}, -\mathbf{1})$, with $-\mathbf{1}$ being a column vector with entries –1. So we have $\mathbf{z}_{i_p} = \mathbf{X}\boldsymbol{\delta}_{i_p} + \boldsymbol{\varepsilon}_{i_p}$. Only persons responding to item i_p are considered here. Further, $\boldsymbol{\delta}_{i_p}$ has a normal prior with mean $\boldsymbol{\mu}_p^{\delta}$ and variance $\boldsymbol{\Sigma}_p^{\delta}$. Define $\widehat{\boldsymbol{\delta}}_{i_p} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{z}_{i_p}$, $\mathbf{d} = \mathbf{X}^t\mathbf{X}\widehat{\boldsymbol{\delta}}_{i_p} + \left(\boldsymbol{\Sigma}_p^{\delta}\right)^{-1}\boldsymbol{\mu}_p^{\delta}$, and define $\mathbf{D} = \left(\mathbf{X}^t\mathbf{X} + \left(\boldsymbol{\Sigma}_p^{\delta}\right)^{-1}\right)^{-1}$. Again, using the result from Bayesian regression analysis, mentioned in Step 2,

$$\boldsymbol{\delta}_{i_p} \mid \boldsymbol{\theta}, \mathbf{z}_{i_p}, \mathbf{X}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p \sim N(\mathbf{Dd}, \mathbf{D}). \tag{15.14}$$

(b) The value of c_{i_p} is sampled from the conditional posterior distribution given $\mathbf{x}_{i_p}, \mathbf{w}_{i_p}, \boldsymbol{\mu}_p^{c|\delta}$, and $\boldsymbol{\Sigma}_p^{c|\delta}$. Let t_{i_p} be the number of persons who do not know the correct answer to item i_p and guess the response. For the probability of a correct response of a person n on item i_p given $w_{ni_p} = 0$, it thus holds that $P(X_{ni_p} = 1 \mid W_{ni_p} = 0) = c_{i_p}$. The number of correct responses obtained by guessing, S_{i_p}, say, has a binomial distribution with parameters c_{i_p} and t_{i_p}. Since logit c_{i_p} has a normal prior with parameters $\boldsymbol{\mu}_p^{c|\delta}$ and $\boldsymbol{\Sigma}_p^{c|\delta}$, the procedure for sampling in a generalized linear model with a logit-link and a normal prior (see Gelman, Carlin, Stern & Rubin, 1995, sects 9.9 and 10.6) can be used.

Step 4 Values for $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ are drawn from the conditional posterior distribution given $\boldsymbol{\xi}$. The number of items sampled from family p is equal to k_p. The prior distribution in (15.5) is the conjugate for $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$. Hence, the posterior distribution is also normal-inverse-Wishart, with parameters

$$\boldsymbol{\mu}_{pt} = \frac{\kappa_{p0}}{\kappa_{pt}}\boldsymbol{\mu}_{p0} + \frac{k_p}{\kappa_{pt}}\overline{\boldsymbol{\xi}}_p, \tag{15.15}$$

$$\boldsymbol{\Sigma}_{pt} = \boldsymbol{\Sigma}_{p0} + \mathbf{S} + \frac{\kappa_{p0}k_p}{\kappa_{pt}}(\overline{\boldsymbol{\xi}}_p - \boldsymbol{\mu}_{p0})(\overline{\boldsymbol{\xi}}_p - \boldsymbol{\mu}_{p0})^t, \tag{15.16}$$

where $\kappa_{pt} = \kappa_{p0} + k_p$, $\nu_{pt} = \nu_{p0} + k_p$, $\mathbf{S} = \sum_{i_p}^{k_p}(\boldsymbol{\xi}_{i_p} - \overline{\boldsymbol{\xi}}_p)(\boldsymbol{\xi}_{i_p} - \overline{\boldsymbol{\xi}}_p)^t$, and $\overline{\boldsymbol{\xi}}_p = \sum_{i_p}^{k_p}\boldsymbol{\xi}_{i_p}$.

The corresponding posterior distribution is, therefore, given by

$$\begin{aligned}
\boldsymbol{\Sigma}_p \mid \boldsymbol{\xi}_p &\sim \text{Inverse-Wishart}_{\nu_{pt}}(\boldsymbol{\Sigma}_{pt}^{-1}),\\
\boldsymbol{\mu}_p \mid \boldsymbol{\Sigma}_p, \boldsymbol{\xi}_p &\sim N(\boldsymbol{\mu}_{pt}, \boldsymbol{\Sigma}_p/\kappa_{pt}).
\end{aligned} \tag{15.17}$$

The procedure thus amounts to iterative generation of parameter values using the above four steps. Multiple MCMC chains can be started from different points to evaluate convergence by comparing the between- and within-sequence variance. Another approach is to generate a single Markov chain and to evaluate convergence by dividing the chain into subchains and comparing between- and within-subchain variance (see, for instance, Robert & Casella, 1999, p. 366). In the examples given below, the latter procedure was used because it proved less wasteful in the number of iterations needed.

15.4.4 Identifiability Problems

As already discussed, the procedure breaks down if N_{i_p} becomes too small. This point can now be illustrated using the above steps in the Gibbs sampler. For example, Step 3a is based on a normal linear model $\mathbf{z}_{i_p} = \mathbf{X}\boldsymbol{\delta}_{i_p} + \boldsymbol{\varepsilon}_{i_p}$. However, if $N_{i_p} = 1$ and item i_p is administered to one person, $\mathbf{z}_{i_p}$ has only one entry, and it is not possible to estimate two regression coefficients from one observation. The same problem happens for the generalized linear model in Step 3b.

15.5 Bayes Modal Estimation ($N_{i_p} = 1$)

Glas and van der Linden (2003) present a Bayes modal procedure for the estimation of the hyperparameters $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ that can be used if $N_{i_p} = 1$. The procedure is summarized in this section.

The fact that every person is administered a unique item and, hence, the item parameters are unique for every person will be made explicit by adding index n to the item parameters and writing $\boldsymbol{\xi}_{ni_p}$. The likelihood is marginalized over $\boldsymbol{\xi}_{ni_p}$ and θ_n. These parameters are stacked in vectors $\boldsymbol{\xi}_p$ and $\boldsymbol{\theta}$, respectively. The parameters we estimate are in the vector $\boldsymbol{\eta} = (\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p, \ldots, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$.

The marginal probability of observing response pattern $\mathbf{x}_n$ is given by

$$\begin{aligned} p(\mathbf{x}_n; \mathbf{d}_n, \boldsymbol{\eta}) &= \int \ldots \int \prod_{i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{ni_p}) p(\boldsymbol{\xi}_{ni_p} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\theta_n) d\boldsymbol{\xi}_{ni_p} d\theta_n \\ &= \int \left[\prod_{i_p} \int \ldots \int p(x_{ni_p} | \, d_{ni_p}, \theta_n, \boldsymbol{\xi}_{ni_p}) p(\boldsymbol{\xi}_{ni_p} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) d\boldsymbol{\xi}_{ni_p} \right] \\ &\quad \times p(\theta_n) d\theta_n. \end{aligned} \tag{15.18}$$

Notice that (15.18) entails a multiple integral over $\boldsymbol{\xi}_{ni_p}$.

Glas and van der Linden (2003) show that Bayes modal estimation equations can be derived by taking the expectations of (15.15) and (15.16) with respect to the posterior distribution of $\boldsymbol{\xi}_{ni_p}$ given the response pattern $\mathbf{x}_n$. That is, we now have

$$\boldsymbol{\mu}_p = \frac{\kappa_{p0}}{\kappa_p}\boldsymbol{\mu}_{p0} + \frac{k_p}{\kappa_p}\widetilde{\boldsymbol{\xi}}_p \tag{15.19}$$

and

$$\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_{p0}+\sum_n E\left[(\boldsymbol{\xi}_{i_p} - \widetilde{\boldsymbol{\xi}}_p)(\boldsymbol{\xi}_{i_p} - \widetilde{\boldsymbol{\xi}}_p)^t \mid \mathbf{x}_n, \boldsymbol{\eta}\right]+\frac{\kappa_0 k_p}{\kappa_p}(\widetilde{\boldsymbol{\xi}}_p-\boldsymbol{\mu}_{p0})(\widetilde{\boldsymbol{\xi}}_p-\boldsymbol{\mu}_{p0})^t, \tag{15.20}$$

with

$$\widetilde{\boldsymbol{\xi}}_p = \frac{1}{k_p}\sum_n E(\boldsymbol{\xi}_p \mid \mathbf{x}_n, \boldsymbol{\eta}).$$

These equations can be solved using an EM or Newton–Raphson algorithm (Bock & Aitkin, 1981; Mislevy, 1986).

15.6 Bayes Modal Estimation ($N_{i_p} \geq 2$)

The parameters to be estimated are in a vector $\boldsymbol{\eta} = (\boldsymbol{\xi}_{1_1}, \ldots, \boldsymbol{\xi}_{k_P}, \ldots, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$. The marginal log-likelihood function for this vector becomes

$$\log L(\boldsymbol{\eta}; \mathbf{x}) = \sum_p \sum_n \left[\log p(\mathbf{x}_n \mid \mathbf{d}_n, \boldsymbol{\xi}_p) + \sum_{i_p} \log p(\boldsymbol{\xi}_{i_p} \mid d_{ni_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)\right] + \log p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\mu}_{p0}, \boldsymbol{\Sigma}_{p0}), \tag{15.21}$$

where the probability of observing response pattern $\mathbf{x}_n$ given $\boldsymbol{\xi}_p$ is obtained as

$$p(\mathbf{x}_n \mid \mathbf{d}_n, \boldsymbol{\xi}_p) = \int \prod_{i_p} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{i_p}) p(\theta_n) d\theta_n.$$

As before, we use the convention that $p(x_{ni_p} \mid d_{ni_p} = 0, \theta_n, bf\boldsymbol{\xi}_{i_p}) = 1$.

The marginal estimation equations for $\boldsymbol{\eta}$ can be easily derived from (15.21) using Fisher's identity (Efron, 1977; Louis, 1982). Application of the identity in an IRT framework amounts to taking the first-order derivatives as if the ability parameters θ_n were observed and then taking the expectation with respect to the conditional posterior distribution of θ_n given the data $\mathbf{x}_n$ (see Glas, 1992, 1998). For the present case, the first-order derivatives with respect to $\boldsymbol{\eta}$ are found as

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}; \mathbf{x}) = \sum_p \sum_n E\left(\frac{\partial}{\partial \boldsymbol{\eta}} \log f_{p,n}(\boldsymbol{\eta}, \theta_n; \mathbf{x}_n) \mid \mathbf{x}_n, \boldsymbol{\eta}\right) = \mathbf{0}, \tag{15.22}$$

where the complete-data log-likelihood $\sum_p \sum_n \log f_{p,n}(\boldsymbol{\eta}, \theta_n; \mathbf{x}_n)$, which would be the log-likelihood if θ_n were observed, is equal to

$$\begin{aligned}
&\sum\nolimits_p \sum\nolimits_n \log f_{p,n}(\boldsymbol{\eta}, \theta_n; \mathbf{x}_n)\\
&\quad = \sum\nolimits_p \sum\nolimits_n \Big[\log p(\mathbf{x}_n \mid \mathbf{d}_n, \theta_n, \boldsymbol{\xi}_p) + \log p(\theta_n)\\
&\qquad + \sum\nolimits_{i_p} \log p(\boldsymbol{\xi}_{i_p} \mid d_{ni_p}, \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)\Big] + \log p(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p | \boldsymbol{\mu}_{p0}, \Sigma_{p0}).
\end{aligned}$$

Let Φ_{ni_p} and P_{ni_p}be shorthand notations for $\Phi(\lambda_{ni_p})$, with $\lambda = a_{i_p}\theta_n - b_{i_p}$, and $P_{ni_p} = p(x_{ni_p} = 1 \mid d_{ni_p} = 1, \theta_n, \boldsymbol{\xi}_{i_p}) = c_{i_p} + (1 - c_{i_p})\Phi_{ni_p}$, respectively. Further, let g_{ni_p} be the normal density evaluated at the point λ_{ni_p}. Using Fisher's identity, Bayes modal equations for the parameters $\xi_{i_p u}$, $u = 1, \ldots, 3$, are found as

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(x_{ni_p} - P_{ni_p})(1 - c_{i_p}) g_{ni_p} \theta_n}{P_{ni_p}(1 - P_{ni_p})} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + \frac{(a_{i_p} - \mu_{p1})}{\sigma_{p1}} = 0, \quad (15.23)$$

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(P_{ni_p} - x_{ni_p})(1 - c_{i_p}) g_{ni_p}}{P_{ni_p}(1 - P_{ni_p})} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + \frac{(b_{i_p} - \mu_{p2})}{\sigma_{p2}} = 0, \quad (15.24)$$

and

$$\sum_{n|d_{ni_p}=1} E\left(\frac{(x_{ni_p} - P_{ni_p})(1 - \Phi_{ni_p})}{P_{ni_p}(1 - P_{ni_p})} \middle| \mathbf{x}_n, \boldsymbol{\eta} \right) + \frac{(\operatorname{logit} c_{i_p} - \mu_{p3})}{\sigma_{p3}} = 0, \quad (15.25)$$

where the sums range over all persons n for which $d_{ni_p} = 1$. These expressions are a straightforward generalization of the usual likelihood equations for the 3PNO; for details, refer to Glas (2000).

It is easily verified that the Bayes modal equations for the hyperparameters $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ are analogous to those for the previous case given in (15.19) and (15.20).

The estimation equations can be solved using an EM or Newton–Raphson algorithm. If in practical applications the number of parameters becomes large, use of these algorithms is not feasible.

Expressions for confidence intervals can also be derived using Fisher's identity (Louis, 1982; Mislevy, 1986; Glas, 1998). However, the computation of the asymptotic covariance matrix of the estimates also involves the inversion of a matrix of second-order derivatives (information matrix). In the application presented below, only the within-family information matrix was inverted; that is, the covariances between the families were assumed to be zero. This approximation resulted in confidence intervals that were larger than the confidence intervals that would have been obtained if the complete information matrix were inverted.

15.6.1 Discussion

The complications involved in Bayes modal estimation with marginalization both over $\boldsymbol{\theta}$ and the item parameters $\boldsymbol{\xi}$, as in the previous case for $N_{i_p} = 1$, for a sampling design with items shared by persons can be illustrated as follows. Assume that each item is given to two respondents, say n and m. The responses of both respondents now depend on the same item parameter; this dependency will be made explicit by labeling these parameters as $\boldsymbol{\xi}_{nmi_p}$.

The complete-data likelihood used in (15.22) should now be written as

$$p(\mathbf{x} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\eta}) = \prod_{(n,m)} \prod_{p} \prod_{i} p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{nmi_p}) \\ \times p(x_{mi_p} \mid d_{mi_p}, \theta_m, \boldsymbol{\xi}_{nmi_p}) p(\boldsymbol{\xi}_{nmi_p} \mid \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) p(\theta_n) p(\theta_m),$$

where the first product ranges over all pairs of respondents (n, m) with a common item.

If we marginalized this likelihood over $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the result would be

$$p(\mathbf{x}; \boldsymbol{\eta}) = \prod_{(n,m)} \int \cdots \int \prod_{p} \prod_{i} \int \cdots \int p(x_{ni_p} \mid d_{ni_p}, \theta_n, \boldsymbol{\xi}_{nmi_p}) \\ \times p(x_{mi_p} \mid d_{mi_p}, \theta_m, \boldsymbol{\xi}_{nmi_p}) p(\boldsymbol{\xi}_{nmi_p} | \boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p) d\boldsymbol{\xi}_{nmi_p} \\ \times p(\theta_n) p(\theta_m) d\theta_n d\theta_m.$$

The integral in this likelihood cannot be factored any further. In fact, as the number of respondents receiving the same item increases, we are quickly left with a multiple integral that cannot be computed by the usual Gauss–Hermite procedure (see, for instance, Glas, 1992). Fortunately, the fully Bayesian procedure with the Gibbs sampler discussed above does not have this problem.

15.7 Some Numerical Examples

A number of studies were conducted to assess the feasibility of the procedures in practical situations. In some practical situations, the number of responses per family of items might be quite low and the number of item parameters might be quite high. In such cases, the convergence of the MCMC or the EM algorithm to realistic parameter estimates is not a priori obvious. For that matter, in a Bayesian framework, the computation of estimates can be supported by a sensible choice of priors.

The study consisted of two stages. In the first stage, two real data sets were analyzed to obtain some idea of the covariance between the item parameters. Then, in the second stage, the estimates obtained in the first stage were used in a number of simulation studies aimed at assessing the quality of parameter recovery.

In the first stage, two different data sets were used. The first data set consisted of the responses of 429 students to $P = 10$ families of multiple-choice items in a computer-based test for a course on Naval Architecture at Ngee Ann Polytechnic in Singapore. The data were collected in 1999 and 2000. Each family was generated by randomly changing the numerical information in the item stem and the response alternatives every time an item was administered. The second data set consisted of the responses of a sample of 4,000 students from the population participating in the 1991 central examination on French-language comprehension in Secondary Education in the Netherlands. In this case, the test was a traditional paper-and-pencil test. Students were clustered in 116 schools. Because of differences among the schools in test administration circumstances, time points, curricula, etc., it was assumed that the item parameters might vary. It was expected that the item-parameter variance might be high in the first example and low in the second example.

For the French-language comprehension data set, Bayes modal estimates of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p (p = 1, \ldots, 40)$ were obtained by marginalizing over all incidental parameters $\boldsymbol{\xi}$ and $\boldsymbol{\theta}$. Also, for both data sets, joint estimates of $\boldsymbol{\xi}$, $\boldsymbol{\theta}$, $\boldsymbol{\mu}_p$, and $\boldsymbol{\Sigma}_p$ were obtained using the MCMC method run with 13,000 iterations, 3,000 of which were burn-in iterations. We report expected a posteriori (EAP) estimates as point estimates. In the second procedure, Bayes modal (MAP) estimates of $\boldsymbol{\xi}$, $\boldsymbol{\mu}_p$, and $\boldsymbol{\Sigma}_p$ were obtained by marginalizing only over $\boldsymbol{\theta}$. The computations were carried out using an EM algorithm.

For both data sets, the same prior covariance matrix $\boldsymbol{\Sigma}_{p0}$ was used for all families p. The values in $\boldsymbol{\Sigma}_{p0}$ are shown in Table 15.1; our considerations for the choice of this prior were given above. The prior for the family parameters was chosen equal to $\boldsymbol{\mu}_{p0} = (1.0, 0.0, \text{logit}(0.25))$. To obtain convergence in the analysis of the language comprehension data, it turned out that the parameters in the normal-inverse-Wishart prior for $(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ had to be set equal to $\nu_{p0} = 10$ and $\kappa_{p0} = 10$, respectively. Since $k_p = 116$, this choice results in a slightly informative prior. An uninformative prior sufficed for the Naval Architecture data.

The averages of the point estimates of the covariance matrices are shown in Table 15.1 (first three columns), together with their average confidence intervals (last three columns). For the EAP estimates, the posterior standard deviation is reported; for the MAP estimates, the values computed using the normal approximation are shown. It can be seen that both the posterior variance of the item discrimination and difficulty parameters were generally lower than expected. Also, the standard errors of the MAP estimates were smaller than those of the EAP estimates. This effect is consistent with the findings of Glas, Wainer, and Bradlow (2000). They argue that posterior distributions of bounded parameters, such as a variance or a discrimination parameter, are skewed. The standard error of the MAP estimate used here was based on an assumption of asymptotic normality, which, in turn, was based on a Taylor expansion of the likelihood with terms of order greater than two ignored. The fact that only the within-family information matrices were used to obtain the standard errors did not nullify the effect.

In the second stage of this study, we assessed the quality of parameter recovery. Since the difference in the covariances obtained for the two examples given above

Table 15.1 Prior and posterior item covariance matrices

Prior Covariance Matrix					
0.200					
0.100	1.000				
−0.050	0.050	0.100			
Posterior Covariance Matrix					
French-Language Comprehension					
EAP Estimate			Standard Error		
0.102			0.017		
0.031	0.208		0.017	0.033	
−0.018	0.010	0.116	0.018	0.020	0.039
Posterior Covariance Matrix					
French Language Comprehension					
MAP Estimate			Standard Error		
0.098			0.014		
0.029	0.199		0.012	0.025	
−0.018	0.006	0.107	0.015	0.016	0.037
Posterior Covariance Matrix					
Naval Architecture					
EAP Estimate			Standard Error		
0.120			0.032		
0.027	0.122		0.030	0.051	
0.001	0.002	0.110	0.022	0.023	0.073

Note: EAP is an expected posterior estimate using MCMC and MAP is a Bayes modal estimate using numerical integration.

was not dramatically different, it was decided to examine two conditions each realized in a different simulation study. In the first study, the prior parameters $\boldsymbol{\mu}_{p0}$ and $\boldsymbol{\Sigma}_{p0}$ were the same as in the examples presented above. The family parameters, $\boldsymbol{\mu}_p$, were drawn from a normal distribution indexed by $\boldsymbol{\mu}_{p0}$ and $\boldsymbol{\Sigma}_{p0}$, and $\boldsymbol{\Sigma}_p$ was set equal to $\boldsymbol{\Sigma}_{p_0}$. Then, for each family, 10 items were randomly drawn from a normal distribution with parameters $\boldsymbol{\mu}_p$ and $\boldsymbol{\Sigma}_p$. To produce realistic data, family and item discrimination parameters drawn below 0.5 were truncated to 0.5. The responses to the items were generated for simulees with an ability parameter randomly drawn from a standard normal distribution. Every simulee responded to 20 random items from 20 different families. The total data matrix consisted of 1,000 responses. The sampling design was thus an instance of Case 1 (k_p fixed and N_{i_p} growing in N) discussed earlier. The Gibbs sampler was run for 13,000 iterations, including 3,000 burn-in iterations. To obtain convergence, the parameters in the normal-inverse-Wishart prior had to be set equal to $\nu_{p0} = 2$ and $\kappa_{p0} = 2$, respectively. Since $k_p = 10$, this choice entails a rather informative prior.

The second simulation had a similar setup. The average of the EAP estimates of the mean and covariance matrix obtained using the French-language examination was used as $\boldsymbol{\mu}_{p0}$ and $\boldsymbol{\Sigma}_{p0}$. Further, the number of families was equal to 40, the number of items per family was equal to 20, and the number of responses to each item was 200. In this study, the total number of responses was equal to 4,000. The data were analyzed using the Bayes modal method with marginalization only over θ.

Results from the two simulations are presented Table 15.2. The results are averaged over 10 replications and all items. The rows labeled a, b and logit c are for the item parameters; all other rows for the item-family parameters. The column labeled EAP relates to EAP estimates obtained using the Gibbs sampler, the column labeled MAP relates to Bayes modal estimates. Both columns give the mean absolute error of the estimates, averaged over items and replications. Especially in the case $P = 40$, the estimates of the covariance matrices seemed much more precise than the estimates of the item parameters. This result, however, may be explained by the

Table 15.2 Mean absolute error in EAP and MAP estimates

P	k_p	N_{i_p}	Parameter	True	EAP	MAP
20	10	100				
			Level 1			
			a	1.00	0.40	0.41
			b	0.00	0.51	0.43
			logit c	−1.10	0.33	0.32
			Level 2			
			μ_a	1.00	0.31	0.28
			μ_b	0.00	0.49	0.37
			$\mu_{logit\,c}$	−1.10	0.21	0.19
			σ_a^2	0.20	0.08	0.09
			σ_b^2	1.00	0.29	0.28
			$\sigma_{logit\,c}^2$	0.10	0.38	0.48
			$\sigma_{a,b}$	0.10	0.09	0.07
			$\sigma_{a,logit\,c}$	−0.05	0.05	0.05
			$\sigma_{b,logit\,c}$	0.05	0.07	0.09
40	20	200				
			Level 1			
			a	0.95	0.39	0.42
			b	0.19	0.36	0.33
			logit c	−0.98	0.31	0.25
			Level 2			
			μ_a	0.96	0.30	0.26
			μ_b	0.18	0.32	0.20
			$\mu_{logit\,c}$	−1.00	0.20	0.16
			σ_a^2	0.10	0.04	0.04
			σ_b^2	0.21	0.10	0.08
			$\sigma_{logit\,c}^2$	0.12	0.01	0.01
			$\sigma_{a,b}$	0.03	0.00	0.03
			$\sigma_{a,logit\,c}$	−0.02	0.01	0.01
			$\sigma_{b,logit\,c}$	0.01	0.02	0.02

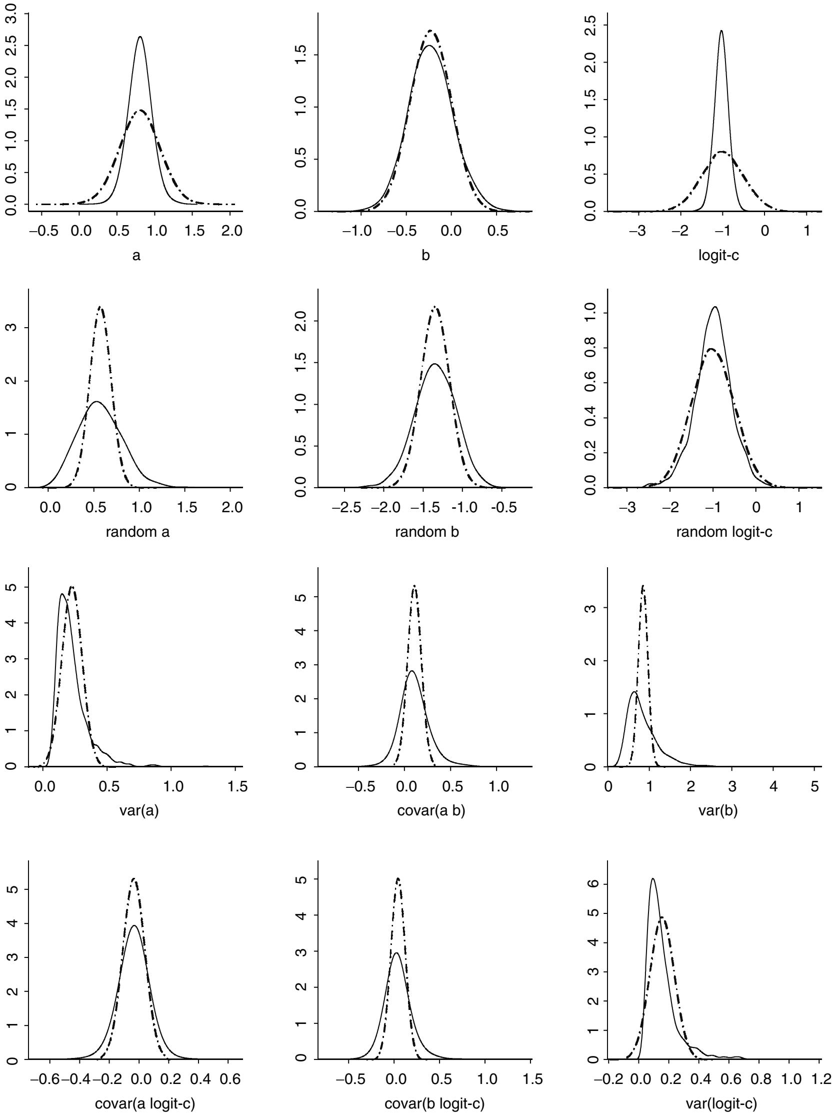


Fig. 15.1 Posterior densities generated via MCMC (solid line) and normal approximations computed via numerical marginalization (dotted line)

fact that the covariance matrices were chosen equal to their prior values. Further inspection of the results shows that the MAEs of the MAP estimates were somewhat smaller than those of the corresponding EAP estimates.

Figure 15.1 shows the posterior distributions of a typical set of parameters for a run with $P = 20$. The three graphs in the first row are the posterior distributions of the three elements of $\boldsymbol{\mu}_p$ for a typical item family p. The three graphs in the next row show the posterior distributions of the three parameters of an arbitrarily chosen item i_p. The last two rows give the posterior distributions of the elements of $\mathbf{\Sigma}_p$, for the same item-family p. The dotted curves in the graphs are the asymptotic distributions computed using the normal approximation described above. It can be seen that the latter approximations were not always realistic. The normal approximation of the variance of logit c_{i_p}, for instance, gave discernible larger positive weight to negative values. The actual posterior distributions of several elements of $\mathbf{\Sigma}_p$ are notably skewed to the right.

Convergence of the Gibbs sampler is usually evaluated by dividing the generated chain into batches and comparing the within and between batch variance of the generated values. Figure 15.2 shows the convergence of the Gibbs sampler for the same 12 parameters. The plot is based on the 2,000 draws taken equally spaced from the 10,000 draws following the burn-in iterations. From inspection of the plots it can be concluded that the chain converged properly. In practice, visual inspection of the convergence plots of all parameters is not very practical. However, convergence can also be evaluated by standard analysis of variance methods.

Figures 15.3 and 15.4 give a scatter plot of the generating values (x-axis) and the EAP-estimates (y-axis) of the family and item parameters for two replications of both simulation studies. The truncation of the discrimination parameters at 0.5 was caused by the generation strategy described above. It can be seen from the plots that the relation between the generated and recovered parameters was quite good; in fact, all correlations were above 0.80. We could not prepare similar plots for logit c_{i_p} and its mean and the elements of the covariance matrices, because the variance in the generating values was too low and zero, respectively.

15.8 Final Remarks

In some areas of measurement item parameters should not be modeled as fixed but as random. Examples of such areas are item sampling, computerized item generation, surveys with substantial variability of item parameters over subgroups of respondents, measurement with substantial estimation error in the item-parameter estimates, and grouping of items under a common stimulus or in a common context. A model for multiple item families with random parameters was discussed and Bayesian estimation methods for such models were outlined. The model differed from the multilevel IRT model for testlets in Wainer, Bradlow, and Du (2000) in that the latter only has a random interaction parameter between persons and items but fixed parameters a_i, b_i, and c_i. The statistical approach to parameter estimation for the models is the same, however; these authors also use an MCMC framework. The same holds for the model introduced in Janssen, Tuerlinckx, Meulders, and de Boeck (2000), which is a two-parameter version of the one in (15.1) obtained by setting $c_i = 0$. Their second-level model specifies independent normal distributions for a_i, and b_i and is thus a special case of (15.3) with $\mathbf{\Sigma}_p$ reduced to a 2 × 2 identity matrix. These authors use an MCMC framework with uninformative priors

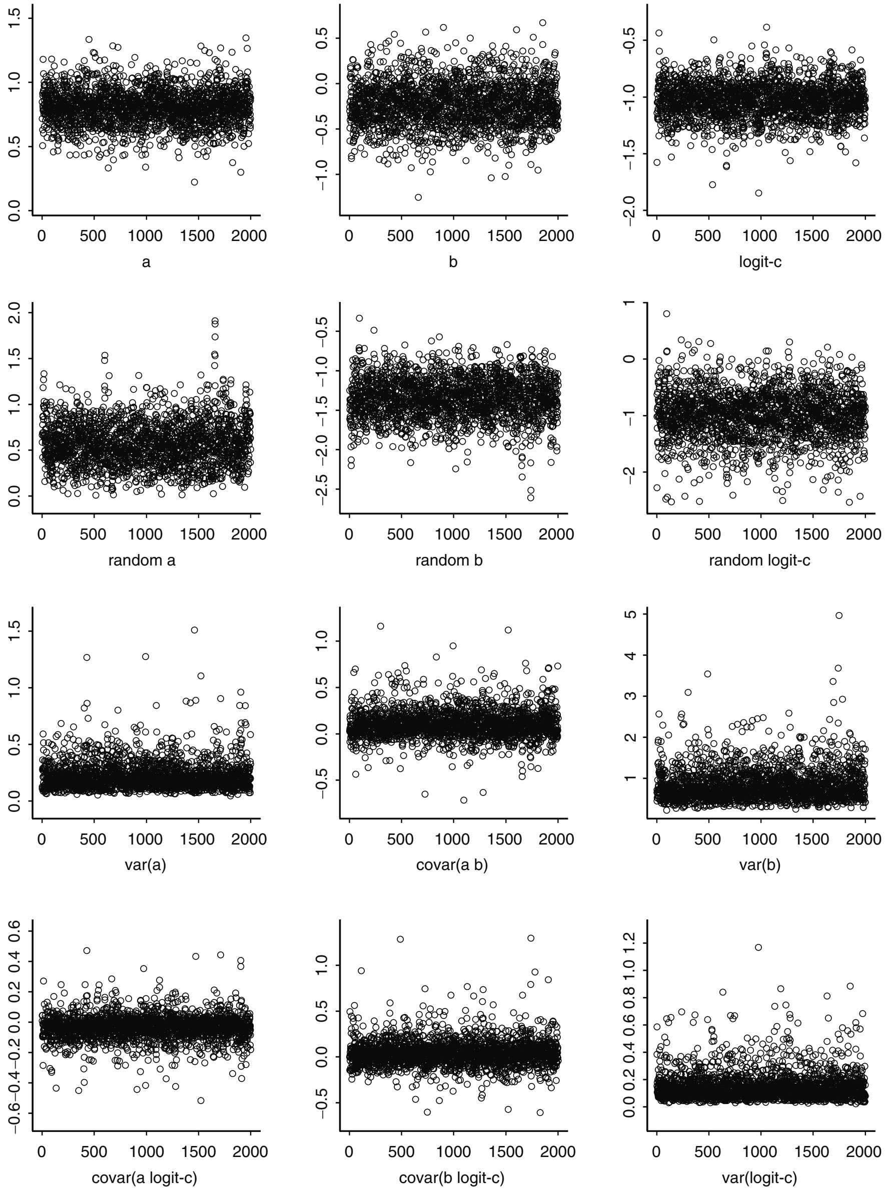


Fig. 15.2 Convergence of the Gibbs sampler; 2,000 draws sampled equally spaced from 10,000 MCMC draws

for (μ_a, μ_b) rather than the full prior in (15.5). Finally, Albers, Does, Imbos, and Janssen (1989) propose a one-parameter version of the normal-ogive model, i.e., the model in (15.1) with $a_i = 1$ and $c_i = 0$, but added a growth parameter for each

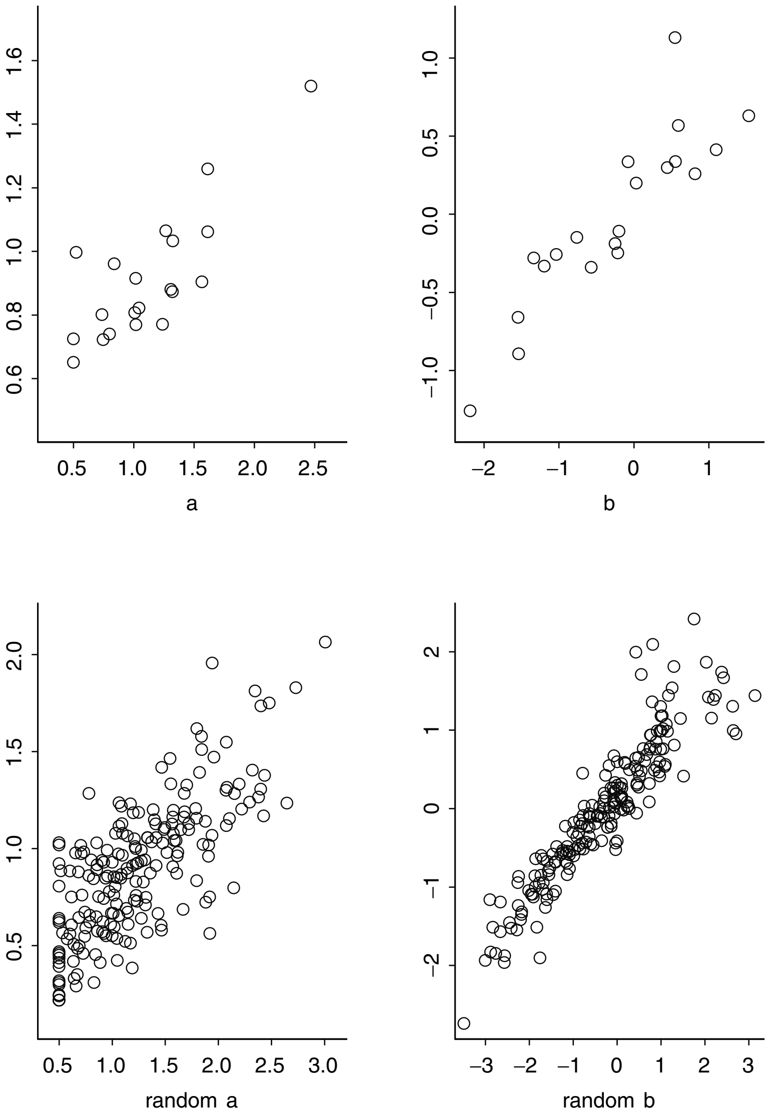


Fig. 15.3 Generating (x-axis) values and parameter estimates (y-axis) for $K = 20, k_{i_p} = 10$

person that is assumed to increase linearly over time. The statistical treatment of this model was entirely within the maximum-likelihood framework.

A goal of this chapter was to show that the sampling design is a crucial factor in the choice between estimation procedures. If every item is responded to by a sufficient number of persons, Bayesian methods using the Gibbs sampler can be used. If only one response is given to some of the items, this approach breaks down

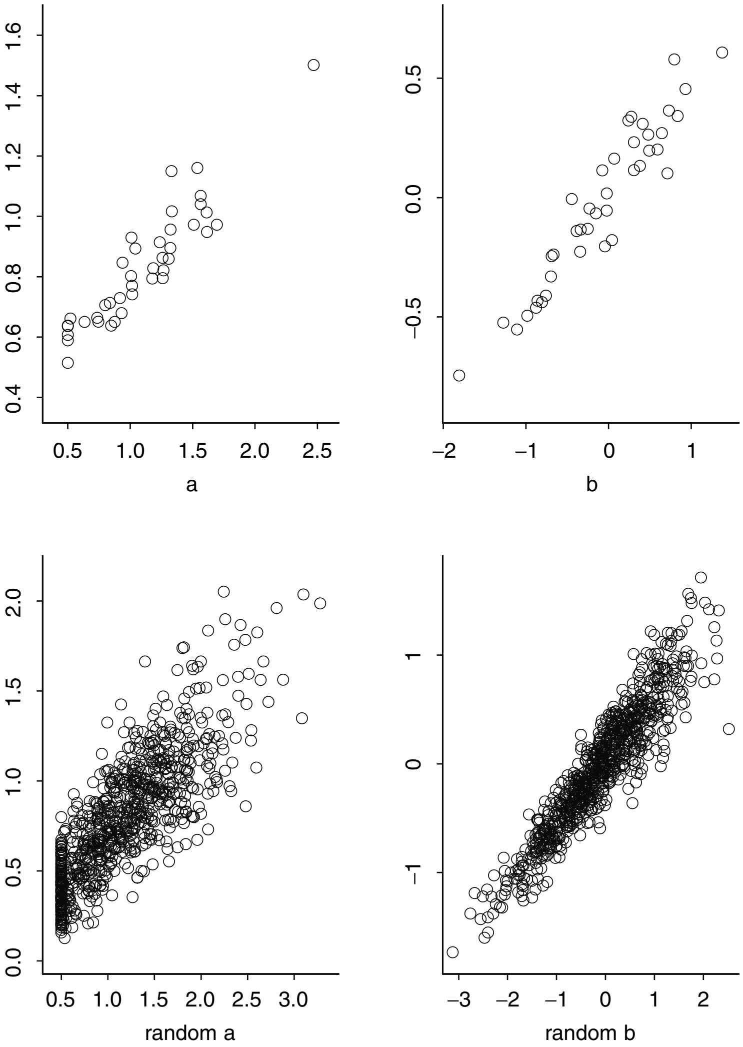


Fig. 15.4 Generating (x-axis) values and parameter estimates (y-axis) for $K = 40, k_{i_p} = 20$

because of identifiability problems. However, in this case, a Bayes modal estimation procedure using a posterior distribution marginalized with respect to the ability and item parameters can be used to estimate the means and covariance matrices of the item-family parameters.

Finally, a Bayes modal estimation procedure was derived in which the likelihood is marginalized only with respect to the ability parameters and both the item and family parameters are estimated. The numerical examples showed that these estimates were not substantially different from the estimates obtained in the MCMC estimation procedure.

References

Albers, W., Does, R. J. M. M., Imbos, T. & Janssen, M. P. E. (1989). A stochastic growth model applied to repeated tests of academic knowledge. *Psychometrika, 54*, 451–466.

Albert, J. H. (1992). Bayesian estimation of normal-ogive item response curves using Gibbs sampling. *Journal of Educational and Behavioral Statistics, 17*, 251–269.

Andersen, E. B. & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357–374.

Béguin, A. A. & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541–562.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Berger, M. P. F. (1997). Optimal designs for latent variable models: A review. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 71–79). Münster, Germany: Waxmann.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D. & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

de Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*, 533–559.

de Jong, M. G., Steenkamp, J. B. E. M. & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34,* 260–278.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. P. Demster, N. M. Laird and D. B. Rubin). *Journal of the Royal Statistical Society* (Series B)*, 39,* 1–38.

Fox, J. P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271–288.

Geerlings, H., van der Linden, W. J. & Glas, C. A. W. (2009). *Modeling rule-based item generation*. Manuscript submitted for publication.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85*, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman and Hall.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1; pp. 236–258). Norwood, NJ: Ablex Publishing Corporation.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8*, 647–667.

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 183–199). Boston: Kluwer-Nijhof Publishing.

Glas, C. A. W. & van der Linden, W. J. (2001). *Modeling variability in item parameters in item response models.* (Research Rep. 01-11). Enschede, the Netherlands: University of Twente.

Glas, C. A. W. & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27*, 247–261.

Glas, C. A. W., Wainer, H. & Bradlow, E. T. (2000). MML and EAP estimates for the testlet response model. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston: Kluwer-Nijhof Publishing.

Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Johnson, M. S. & Sinharay, S. (2005). Calibration of polytomous item families using Bayesian hierarchical modeling. *Applied Psychological Measurement, 29,* 369–400.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–906.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226–233.

Millman, J. (1973). Passing score and test lengths for domain-referenced measures. *Review of Educational Research, 43,* 205–216.

Millman, J. & Westman, R. S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement, 26,* 177–190.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Pychometrika, 51*, 177–195.

Mislevy, R. J. (1991). Randomization-based inferences about latent variables from complex samples. *Pychometrika, 56*, 177–196.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16,* 1–32.

Patz, R. J. & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Patz, R. J. & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.

Robert, C. P. & Casella, G. (1999). *Monte Carlo statistical methods.* New York: Springer-Verlag.

Roid, G. & Haladyna, T. (1982). *A technology for test-item writing*. New York: Academic Press.

Sanathanan, L. & Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association, 73,* 794–799.

Shi, J. Q. & Lee, S. Y. (1998). Bayesian sampling based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51,* 233–252.

Sinharay, S., Johnson, M. S. & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*, 295–313.

Stocking, M. L. (1989). *Empirical estimation errors in item response theory as a function of test properties* (Research Report 89-5). Princeton, NJ: Educational Testing Service.

van der Linden, W. J. (1994). Optimum design in item response theory: Test assembly and item calibration. In G. H. Fischer and D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 305–318). New York: Springer-Verlag.

Wainer, H., Bradlow, E. T. & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer-Nijhof Publishing.

Wingersky, M. S. & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8,* 347–364.

Chapter 16
Detecting Person Misfit in Adaptive Testing

Rob R. Meijer and Edith M.L.A. van Krimpen-Stoop

16.1 Introduction

An examinee's test score does not reveal the operation of undesirable influences of test-taking behavior such as faking on biodata questionnaires and personality tests, guessing, or knowledge of the correct answers due to test preview on achievement tests. These and other influences may result in inappropriate test scores, which may have serious consequences for practical test use, for example, in job and educational selection, where classification errors may result. In the context of item response theory (IRT) modeling, several methods have been proposed to detect item score patterns that are not in agreement with the expected item score pattern based on a particular test model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level (θ). Research with respect to methods that provide information about the fit of an individual item score pattern to a test model is usually referred to as appropriateness measurement or person fit measurement. Most studies in this area are, however, in the context of paper-and-pencil (p&p) tests. As will be argued below, the application of person fit theory presented in the context of p&p tests cannot simply be generalized to a computerized adaptive test (CAT). In this chapter we introduce and review the existing literature on person fit in the context of a CAT.

Before we introduce person fit research, it is important to realize that not all types of aberrant behavior affect individual test scores. For example, a person may guess the correct answers of some of the items but also guess wrongly on some of the other items and, as the result of the stochastic nature of guessing, this process may not result in substantially different test scores under most IRT models to be discussed below. Whether aberrant behavior leads to nonfitting item score patterns depends on numerous factors such as the type and amount of aberrant behavior.

Furthermore, it may be noted that all methods discussed in this chapter can be used to detect nonfitting item score patterns, but several of these methods do not

R.R. Meijer (✉) and E.M.L.A. van Krimpen-Stoop
Heymans Institute, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen,
The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_16,

allow the recovery of the mechanism that created the deviant item score patterns. Other methods explicitly test against specific violations of a test model assumption or against particular types of deviant item score patterns. The latter group of methods therefore may facilitate the interpretation of nonfitting item score patterns.

16.2 Practical and Theoretical Relevance of Person Fit Analysis in a CAT

There are a number of causes why an examinee may not respond according to the IRT model in a p&p test and in a CAT. Misfitting item score patterns may be the result of

- low-ability persons who copy answers from more able neighbors (see e.g., Levine and Rubin, 1979),
- exceptionally creative persons who discover novel interpretations for some items (e.g., Levine and Drasgow, 1988),
- persons who are randomly guessing the correct answers to all the items in the test because they only take the test to become familiar with the items that are administered, or
- persons who have preknowledge of (some of the) items in the test and as a result give correct answers to relatively difficult items.

In the context of a CAT, some of the above-mentioned causes are less likely. For example, copying answers from more able neighbors is very unlikely because in a CAT each examinee is administered different items in a different order. Also, alignment errors are unlikely in adaptive testing, because the answers are given via the keyboard and the next item is only shown after an answer to the current item is given.

When the exam is administered by means of a computer, misfit may be the result of

- persons who are unfamiliar with a computer or may have trouble settling in or warming-up. For these persons the earliest items are more likely to be answered incorrectly than later items;
- unmotivated persons in pretesting situations. For these persons the answers may be unrelated to the item difficulty and a pattern may be generated that seems the result of random response behavior;
- persons who have preknowledge of (some of the) items in the test, and as a result give correct answers to subareas of the test. For example, a group of test takers (sources) are administered a CAT and memorize the items presented. Then the sources make a list of their items and provide other test takers (memorizers) the opportunity to memorize these items. If these memorizers are successful, they may use their item preknowledge when administered a test. An alternative is that examinees may have extra training in various subareas of the exam, which may result in multidimensionality in a CAT that is assumed to measure a single ability.

From a practical point of view it is important to identify examinees that behave in a deviant way because the estimated latent trait ($\hat{\theta}$) is inadequate to obtain an accurate impression of someone's trait level. The question of what to do when an examinee is suspected of, for example, memorizing items is not so easy to answer. It probably depends on the purpose and circumstances under which the test or examination is administered. A rigorous decision is to let the examinee redo the exam, although this decision cannot be based on statistical information only. Statistical evidence should not be used as the only evidence for accusing an examinee of cheating. Interesting in this respect is a remark in a court case in the United States:

> Statistics are not irrefutable; they come in infinitive variety and, like any other kind of evidence, they may be rebutted. In short, their usefulness depends on all of the surrounding facts and circumstances. (*Teamster v. U.S.* : Good, 2001, p. 13)

Surrounding facts and circumstances in a testing situation may be observations of cheating or very large differences in total scores between two administrations of the test. McLeod and Lewis (1999) suggested that the first priority is to continue the test administration. They discussed three options. The first option is to continue testing using highly secure items with known characteristics. A second option they suggested is to administer a few items that the examinee will probably answer correctly. First then an examinee's θ should be estimated at for example, $\theta = 1.5$, and then items at, for example, the $\theta = -1$ range are administered. If the examinee has attained his or her score through prior knowledge, some of these items may prove difficult. A drawback of this strategy is that the memorizer should not have a true high θ level. After an analysis of the examinee, the CAT could continue to be administered if the responses indicated that the examinee was responding consistently. A third option is to stop the CAT mode of testing and continue using a secure linear form. The test can be administered so that the examinee is unaware of the change.

Although originally developed for ability testing, a CAT can also be successfully applied in the typical performance domain. For example, Waller and Reise (1989) showed that an application of computerized adaptive testing using the Multidimensional Personality Questionnaire can save as much as 50% of test items with little loss of accuracy. When applied in personality testing, CAT-based person fit statistics may be used to tackle problems with aberrant score patterns linked to this type of psychological testing, such as variable response inconsistency, that is, responses to items for which a particular pattern of responding is semantically inconsistent (e.g., answering true to both "My sleep is fitful and disturbed" and "I wake up fresh and rested most mornings"). Reise and Waller (1993) explored the use of person fit in personality measurement and noted that person fit may be used for the detection of variation due to inappropriateness of the personality trait measured by the test for describing several examinees.

From a theoretical point of view it is also interesting to investigate whether it is possible to design statistics that are suited to detect misfitting item score patterns. As we discuss below, a CAT has characteristics that make it difficult to apply person fit statistics in a CAT that are designed for p&p tests.

16.3 Review of Existing Literature

16.3.1 Person Fit in Paper-and-Pencil Testing

Several statistics have been proposed to investigate the fit of an item score pattern to an IRT model.

In IRT, the probability of obtaining a correct answer on item i $(i = 1, ..., n)$ is explained by an examinee's latent trait value θ and the characteristics of the item (Hambleton & Swaminathan, 1985). Let U_i denote the binary (0, 1) response to item i, a_i the item discrimination parameter, b_i the item difficulty parameter, and c_i the item guessing parameter. The probability of correctly answering an item according to the three-parameter logistic IRT model (3PLM) is defined by

$$P(U_i = 1|\theta) \equiv P_i(\theta) = c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \tag{16.1}$$

When $c_i = 0$, the 3PLM becomes the two-parameter logistic IRT model (2PLM):

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \tag{16.2}$$

Most person fit research has been conducted using fit statistics that are designed to investigate the probability of an item score pattern under the null hypothesis of fitting response behavior. A general form in which most person fit statistics for binary scoring can be expressed is

$$W(\theta) = \sum_{i=1}^{n}[U_i - P_i(\theta)]^2 v_i(\theta), \tag{16.3}$$

where $v_i(\theta)$ is a suitable weight. The expected value of the statistic equals 0 and often the variance is taken into account to obtain a standardized version of the statistic. For example, Wright and Stone (1979) proposed a person fit statistic based on standardized residuals

$$V(\theta) = \sum_{i=1}^{n} \frac{[U_i - P_i(\theta)]^2}{nP_i(\theta)[1 - P_i(\theta)]}. \tag{16.4}$$

V can be interpreted as the corrected mean of the squared standardized residuals based on n items; relatively large values of V indicate deviant item score patterns.

Most studies in the literature have been conducted using some suitable form of the log-likelihood function

$$l(\theta) = \sum_{i=1}^{n}\{U_i \ln P_i(\theta) + [1 - U_i]\ln[1 - P_i(\theta)]\}. \tag{16.5}$$

This statistic, first proposed by Levine and Rubin (1979), was further developed by Drasgow, Levine, and Williams (1985), who proposed a standardized version of l, denoted as l_z, that was less confounded with the trait level. Statistics like V or l_z can only be used to investigate the probability of an item score pattern under the null hypothesis of normal response behavior. In general, let t be the observed value of a person fit statistic T. Then, the significance probability or probability of exceedance is defined as the probability under the sampling distribution that the value of the test statistic is smaller than the observed value: $p* = P(T \leqslant t)$, or larger than the observed value: $p* = P(T \geq t)$, depending on whether low or high values of the statistic indicate aberrant item score patterns. An alternative is to test this null hypothesis against an a priori specified alternative model of aberrant response behavior. Levine and Drasgow (1988) proposed a method for the identification of aberrant item score patterns that is statistically optimal; that is, no other method can achieve a higher rate of detection at the same Type I error rate. They calculated a likelihood-ratio statistic that provides the most powerful test for the null hypothesis that an item score pattern is normal versus the alternative hypothesis that it is aberrant. In this test, the researcher has to specify a model for normal behavior and for a particular type of aberrant behavior in advance. Klauer (1995) followed the same strategy and used uniformly most powerful tests in the context of the Rasch model to test against person-specific item discrimination as well as violations of local independence and unidimensionality. For a review of person fit statistics in a p&p context, see Meijer & Sijtsma (1995, 2001).

16.3.2 Person Fit in Computerized Adaptive Testing

To investigate the fit of an item score pattern in a CAT one obvious option is to use one of the statistics proposed for p&p tests. However, person fit research conducted in this context has shown that this is not straightforward. Nering (1997) evaluated the first four moments of the distribution of l_z for a CAT. His results were in concordance with the results using p&p tests: the variance and the mean were smaller than expected and the null distributions were negatively skewed. As a result, the normal approximation in the tails of the null distribution was inaccurate. Van Krimpen-Stoop and Meijer (1999) simulated the distributions of l_z and l_z^* , an adapted version of l_z in which the variance was corrected for $\hat{\theta}$ according to the theory presented in Snijders (2001). They simulated item scores with a fixed set of administered items and item scores generated according to a stochastic design, where the choice of the administered items depended on the responses to the previous items administered. Results indicated that the distribution of l_z and l_z^* differed substantially from the standard normal distribution although the item characteristics and the test length determined the magnitude of the difference. Empirical Type I errors were too small compared to nominal Type I errors.

Glas, Meijer, and van Krimpen-Stoop (1998) adapted the person fit statistics discussed by Klauer (1995) to the 2PL model and investigated the detection rate of these statistics in a CAT. They found small detection rates for most simulated types of aberrant item score patterns: the rates varied between 0.01 and 0.24 at significance level $\alpha = 0.10$ (one-sided). Robin (2002) also found low power for l_z to detect aberrant responses in a CAT.

A possible explanation for these results is that the characteristics of CATs are generally unfavorable for the assessment of person fit using existing person fit statistics. The first problem is that CATs contain relatively few items compared to p&p tests. Because the detection rate of a person fit statistic is sensitive to test length, and longer tests result in higher detection rates (e.g., Meijer, Molenaar, & Sijtsma, 1994), the detection rate in a CAT will, in general, be lower than that of a p&p test. A second problem is that almost all person fit statistics assume a spread in the item difficulties: generally speaking, aberrant response behavior consists of many 0 scores for easy items and many 1 scores for difficult items. In a CAT, the spread in the item difficulties is relatively modest: in particular, toward the end of the test when $\hat{\theta}$ is close to θ, items with similar item difficulties will be selected and as a result it is difficult to distinguish normal from aberrant item scores. This results in an underdispersion of the assumed null distribution.

Person-fit statistics that are especially designed for a CAT may be more powerful than “conventional” person fit statistics, and the statistical properties of the former statistics should be less susceptible to the characteristics of a CAT.

16.3.3 Person Fit Statistics Designed for a CAT

Although there are a plethora of person fit statistics proposed for p&p tests (see Meijer and Sijtsma, 2001), the number of statistics that has been designed for a CAT is relatively modest and there has not been much experience applying these statistics to empirical data. Below we describe methods that are proposed to detect person misfit in a CAT.

Statistics for Item Memorization

McLeod and Lewis (1999) proposed a statistic Z_c that was designed to detect item score patterns that result from memorization. Before the statistic can be calculated, the item bank is divided into three parts: easy items, items of medium difficulty, and difficult items. Let $\text{Easy}[P_i(\theta) - U_i]$ denote the mean residual for the easy items and $\text{Diff}[P_i(\theta) - U_i]$ the mean residual for the most difficult items in an administered CAT; then Z_c is given by

$$Z_c = \frac{\text{Easy}[P_i(\theta) - U_i] - \text{Diff}[P_i(\theta) - U_i]}{\sqrt{\text{Var}_{\text{easy}} + \text{Var}_{\text{diff}}}} \qquad (16.6)$$

with $\mathrm{Var}_{\mathrm{easy}} = \left\{\sum_{\mathrm{Easy}}\{P_i(\theta)Q_i(\theta)\}/n^2_{\mathrm{Easy}}\right\}$, and $\mathrm{Var}_{\mathrm{diff}} = \sum_{\mathrm{diff}}\{P_i(\theta)Qi(\theta)\}/n^2_{\mathrm{Diff}}\}$ and $Q_i(\theta) = 1 - P_i(\theta)$. Z_c is positive when an examinee answered the easy items incorrectly and the difficult items correctly.

As McLeod and Lewis (1999) discussed, a drawback of this index is the need for two ranges of items. If an examinee does not receive at least one easy item and one difficult item, the index cannot be computed. For a better estimate, several items from each category are necessary. A possible solution is to design a CAT to administer at least one item from each category. A second solution is to design a CAT algorithm that administers at least one easy item to each examinee. This would allow computation of Z_c for those examinees who have a greater chance of inflating their test scores, namely, receiving many difficult items. Furthermore, note that Z_c does not weight the residuals in terms of the probability of correct responses. This index uses the relative difficulty as compared with all other items in the item bank to assign the relative weights. Therefore, unlike l_z, the relative weights are not a function of the probability of a correct response and are only indirectly influenced by $\hat{\theta}$. Applying this statistic to an operational Graduate Record Examination Quantitative CAT bank with 14% memorized items resulted, however, in low detection rates.

McLeod, Lewis, and Thissen (2003) proposed a Bayesian posterior log odds-ratio index for detecting item preknowledge. The log odds ratio index is given by

$$\log_{10}\left[\frac{p(s|U_1,\ldots,U_k)/[1-p(s|U_1,\ldots,U_k)]}{p(s)/[1-p(s)]}\right], \tag{16.7}$$

where $p(s|U_1,\ldots,U_k)$ denotes the probability that an examinee has memorized items after k administered items and $p(s)$ denotes the prior probability to memorize items in advance. Thus, for example, assume that $k = 25$; then a final log odds ratio of 0 implies that after 25 items, there is no more suspicion that the examinee is using item preknowledge than there was before the 25 items were administered. A final log odds ratio of 1 implies that there is 10 times more suspicion that an examinee is cheating than there was before the 25 items were administered. A final log odds-ratio of -1 indicates that there is 10 times less suspicion than there was before the 25 items were administered.

In this approach to person fit, the estimated probability that each examinee has preknowledge of the items is updated after each item response. These probabilities are based on IRT parameters, a model specifying the probability that each item has been memorized, and the examinee's item responses. McLeod et al. (2003) applied the 3PLM to model normal response behavior and a modified 3PLM to model item preknowledge. For this model, the probability of a correct response is the combination of (a) the probability of answering an item correctly based on an examinee's preknowledge of the item and (b) the probability of answering the item correctly based on the examinee's θ level when the examinee did not have preknowledge of the item. If an examinee has preknowledge of the item (has memorized the item), the item will be answered correctly. If a test taker has not memorized the item the probability of a correct response is equivalent to Equation 16.1. The quantity that must be specified is the probability that an item has been memorized.

McLeod et al. (2003) distinguished among three approaches. In the first approach, the probability that an item has been memorized is a constant probability for all items in the item pool. In the second approach this probability is a function of the item's difficulty and in the third approach the probability of memorization is a function of the specific item bank and item selection algorithm used to generate the CAT. The probability that an item has been memorized is computed using simulations in which some number of simulees memorize their tests. Using Bayes' theorem, the probabilities of a correct, incorrect, and a prior probability that an examinee is using preknowledge are combined with each response to give the posterior probability that a test taker has had the opportunity of item preknowledge. Simulation results showed some promise for the use of the odds ratio index.

Person-Fit Based on CUSUM Techniques

Both Bradlow, Weiss, and Cho (1998) and van Krimpen-Stoop and Meijer in a series of articles (2000, 2001, 2002) proposed person fit statistics that use the property of a CAT that a fitting item score pattern will consist of an alternation of correct and incorrect responses, especially at the end of the test when $\hat{\theta}$ comes closer to θ. Therefore, a string of consecutive correct or incorrect answers may be the result of aberrant response behavior. Sums of consecutive negative or positive residuals $[U_i - P_i(\theta)]$ can be investigated using a cumulative sum procedure (CUSUM; Page, 1954). CUSUM procedures are sensitive to strings of positive and negative values of a statistic. Person-fit statistics are often defined in terms of the difference between observed and expected scores. A commonly used statistic is V, the mean of the squared standardized residuals based on n items (Equation 16.4). One of the drawbacks of V is that negative and positive residuals cannot be distinguished. The distinction is of interest in a CAT because a string of negative or positive residuals may indicate aberrant behavior. For example, suppose an examinee with an average θ value responds to a test and the examinee has preknowledge of the items in the last part of the test. As a result, in the first part of the test the responses will alternate between zero and one, whereas in the second part of the test more and more items will be correctly answered due to item preknowledge; thus, in the second part of the test, consecutive positive differences will tend to occur.

The usefulness of a CUSUM procedure can be explained as follows. A CAT can be viewed as a multistage test, where each item is a stage and each stage can be seen as a point in time; at each stage a response to one item is given. Let i_k denote the kth item in the CAT; that is, k is the stage of the CAT. Further, let the statistic T_k be a function of the residuals at stage k, n the final test length. Below, an example of a statistic T is given. For each examinee, at each stage k of a CAT, the CUSUM procedure can be determined as

$$C_k^+ = \max[0, T_k + C_{k-1}^+], \tag{16.8}$$

$$C_k^- = \min[0, T_k + C_{k-1}^-], \quad \text{and} \tag{16.9}$$

$$C_0^+ = C_0^- = 0, \tag{16.10}$$

where C^+ and C^- reflect the sum of consecutive positive and negative residuals, respectively. Let UB and LB be some appropriate upper and lower bounds, respectively. Then, when $C^+ > UB$ or $C^- < LB$, the item score pattern can be classified as not fitting the model, otherwise; the item score pattern can be classified as fitting the model.

Let S_k denote the set of items administered as the first k items in the CAT and $R_k = \{1, \ldots, I\} \backslash S_{k-1}$ the set of remaining items in the pool; from R_k the kth item in the CAT is administered. A principle of CAT is that θ is estimated at each stage k based on the responses to the previously administered items, that is, the items in set S_{k-1}. Let $\hat{\theta}_{k-1}$ denote the estimated θ at stage $k - 1$. Thus, based on $\hat{\theta}_{k-1}$, the item for the next stage, k, is selected from R_k. The probability of correctly answering item i_k, according to the 2PL model, evaluated at $\hat{\theta}_{k-1}$ can be written as

$$P_{i_k}\left(\hat{\theta}_{k-1}\right) = \frac{\exp\left[a_{i_k}\left(\hat{\theta}_{k-1} - b_{i_k}\right)\right]}{1 + \exp\left[a_{i_k}\left(\hat{\theta}_{k-1} - b_{i_k}\right)\right]}. \tag{16.11}$$

In van Krimpen-Stoop and Meijer (2000, 2001), different statistics T_k were proposed that are all functions of the residual between the observed and expected item scores on item i_k. For example, a simple statistic is

$$T_k\left(\hat{\theta}\right) = \frac{U_{i_k} - P_{i_k}\left(\hat{\theta}\right)}{\sqrt{P_{i_k}\left(\hat{\theta}\right)\left[1 - P_{i_k}\left(\hat{\theta}\right)\right]}}.$$

In a CUSUM procedure, in general, normally distributed statistics are used and theoretical critical values can be used to classify item score patterns as fitting or misfitting. However, because the statistics in van Krimpen-Stoop and Meijer (2000) were not normally distributed, critical values were determined by means of a simulation study. Critical values were found to be stable across θ, and thus one critical value could be used to classify an individual score pattern as fitting or misfitting.

By examining the plot of the values of C^+ and C^- against the stage of the CAT, it is possible to find out "what went wrong". Suppose, for example, an examinee only takes an exam to obtain knowledge about the type of questions that are being asked. Then it is plausible that the CUSUM passes the lower bound about halfway the CAT. On the other hand, when an examinee has preknowledge of a number of difficult items, the CUSUM may pass the upper bound after the responses to these items. Bradlow et al. (1998) suggested that by careful inspection of the boundaries, ordering of the observations, and modifications to the definition of the CUSUM, this methodology can be used to identify different types of outliers. For example, to detect warm-up outliers, we leave the items in administration order and use the inspecting sums of consecutive negative residuals. To find tiring outliers, we reverse the administration order of the items and apply the same methodology as for the warm-up outliers.

To identify subexperts, van Krimpen-Stoop and Meijer (2001) divided the item score pattern into disjoint subsets of items. For each subset of items a statistic

$$Z\left(\hat{\theta}\right) \frac{\sum\left[U_{i_k} - P_{i_k}\left(\hat{\theta}\right)\right]}{\sqrt{\sum P_{i_k}\left(\hat{\theta}\right)\left[1 - P_{i_k}\left(\hat{\theta}\right)\right]}}$$

was determined, where the sum is across the items in the subset. Based on simulation studies and some additional experience, they found that for subsets of items consisting of more than 10 items, $Z(\hat{\theta})$ was sensitive to misfitting item score at $\alpha = 0.01$, especially for θ values between $[-2, -1]$ and $[1, 2]$.

As an alternative, van Krimpen-Stoop (2001, chap. 4) suggested using an exponentially weighted moving average procedure. In this procedure an exponentially decreasing weight for each observation of statistic T is taken, where the largest weight is given to the most recent observation, and the weights given to previous observations are decreasing geometrically from the most recent to the first. This may be an advantage over the CUSUM, because during test administration the item selection becomes more accurate; thus, the responses to the items that were selected on the most accurate θ estimates are assigned the largest weight, whereas in the CUSUM all items were assigned an equal weight. Future research should point out whether this procedure leads to an improvement over the CUSUM procedure.

Analogously to a CAT with dichotomous items, van Krimpen-Stoop and Meijer (2002) proposed a CUSUM procedure for CATs with polytomous items. In polytomous CATs, however, the statistic T is determined slightly different: the observed score U_{ij} can obtain the values $j = 0, 1, \ldots, m$ and the expected score equals $\sum_{j=0}^{m} jP_{ij}(\hat{\theta}_k)$, where $P_{ij}(\hat{\theta}_k)$ is the probability of obtaining score j on item i. A simple statistic then equals

$$T_k = 1/k\left[U_{i_k} - \sum jP_{ij}\left(\hat{\theta}_k\right)\right]. \qquad (16.12)$$

Although different functions of the residual T can be defined, van Krimpen-Stoop and Meijer (2002) used only an unstandardized residual T_k. Simulation results showed that detection rates were reasonably high and that when items were disclosed in the first part of the CAT, the CUSUM had higher detection rates compared with disclosed items in the last part of the CAT.

Statistics Using Response Times

An interesting alternative approach to detect aberrant response patterns in a CAT was proposed by van der Linden and van Krimpen-Stoop (2003). To counter the problems raised by the characteristics of a CAT (short tests and reduced spread in the item difficulties), they suggested complementing checks on unexpected item responses with checks on examinee's response times. In a CAT the response time can be recorded, and in high-stakes testing it is realistic to assume that the response time to produce an answer to a particular item reflects the time needed to process the item. Unexpected response times are indicative of specific types of aberrant response behavior. For example, examinees with preknowledge to some of the items in the item bank may answer these items with a shorter response time than expected. To mention another example in the educational context, unmotivated persons in pretesting situations may also answer most items more quickly than expected because there is nothing at stake.

There is some empirical evidence in psychological research that response time may be used to detect aberrant response behavior. For example, Holden (1998; see also Knowles & Condon, 1999) compared item response times and a traditional validity scale for their relative abilities to identify fakers on a personnel inventory. Unemployed persons actively seeking work were randomly assigned either to respond honestly or to fake well. Item response times performed as well in detecting fakers as the best traditional validity index did. Item response times correctly identified over 64% of individuals as either responding honestly or faking. Results were consistent with previous studies that used this model of item response dissimulation. Other evidence was provided by Rammsayer (1999), who investigated response times in CATs. Using perceptual and cognitive discrimination task he found that response times are significantly longer for incorrect than for correct answers. Furthermore, there was no indication that longer response times for incorrect answers can be interpreted as an artifact of higher task difficulty. Finally, he suggested that timing behavior may represent an independent personality trait as suggested by the construct of personal tempo.

Van der Linden and van Krimpen-Stoop (2003) used a loglinear model to model response time given by

$$\ln T_{ij} = \mu + \delta_i + \tau_j + \varepsilon_{ij}, \tag{16.13}$$

where δ_i is a parameter for the response time required by item i, τ_j is a parameter for the slowness of examinee j, μ is a parameter indicating general response time level for the population of examinees and pool of items, and ε_{ij} a normally distributed residual or interaction term for item i and examinee j with mean 0 and variance σ. Item responses were modeled using the 3PLM. In a simulation study based on item parameters from the Arithmetic Reasoning test in the Armed Services Vocational Aptitude Battery (ASVAB), they found that incorporating response time to detect aberrant response behavior to detect preknowledge and speededness resulted in more power than using only information about unexpected responses.

16.4 Discussion

Because few studies on person fit in computerized adaptive testing have been conducted, it is difficult to compare the relative power of the different methods that have been discussed in this chapter. Most studies used data where aberrant response patterns were simulated. The relevant question is, of course, how realistic these simulations really are and what the configuration is of "real" aberrant response patterns. Thus, future research should concentrate on the application of person fit statistics in real testing applications. Furthermore, studies are needed analyzing empirical data together with background variables to obtain extra information about the type of misfit. Inspecting plots of the sum of residuals against the item order in the CUSUM procedure may help the interpretation of the type of misfit. For example, large residuals at the start of the CAT may indicate warm-up effect, whereas at the end of the CAT it may point to fatigue. Research is also needed to distinguish between examinees with item score patterns for whom an inappropriate IRT model is used and those whose item score patterns can be explained using additional information. Another potential direction may be the use of person fit statistics in combination with item content. In, for example, diagnostic testing, students have all sorts of different problems in keeping up with the curriculum, and it is sometimes very difficult for a teacher to determine content areas in which a student might be having problems. A statistic that would pinpoint areas of difficulty would be extremely useful for a teacher who wishes to individualize instruction in the classroom.

In Meijer (2004) a CAT was analyzed that consisted of five different subtest areas. In this CAT the examinee gets a mixture of different subtest areas. For the test agency, it is interesting to know which subtest areas an examinee masters or does not master. Therefore, for each person the total score on each subtest is reported. Additionally, it is interesting to investigate if some persons may generate unexpected combinations of subtest scores. Any score combination that lies inside some predefined critical area can be classified as misfitting the model. The combination of global testing, graphical inspection, and local testing may help to better diagnose misfit.

A possibility is, for example, to calculate the probability of a combination of total scores (see Rosa et al., 2001 and Meijer, 2004). Let, again, the score on item i be denoted by u_i, let the item score vector be denoted by $\mathbf{u}$, and let the sum score for a set of items be denoted by x. The likelihood for any summed score is

$$L_x(\theta) = \sum_{(U_i)=x} L(\mathbf{u}|\theta), \tag{16.14}$$

where the summation is over all response patterns that contain x correct responses. That is, given θ, the likelihood of a summed score is obtained as the sum of the likelihoods of all response patterns that have that summed score. The probability of each score x is then

$$P_x = \int L_x(\theta)\phi(\theta)d\theta, \tag{16.15}$$

where $\phi(\theta)$ is the population density. An algorithm to compute $L_x(\theta)$ was proposed by Lord and Wingersky (1984). This algorithm assumes that the individual $P_i(\theta)$ are estimated under a specified IRT model.

To investigate unexpected sum scores on subtests of items, a generalization of (16.14) can be used. Assume that there are two subtests x and x' . The likelihood of a combination of sum scores can be calculated by

$$L_{xx'}(\theta) = L_x(\theta)L_{x'}(\theta), \tag{16.16}$$

and the probability of the response pattern of the summed scores $\{x, x'\}$ equals

$$P_{xx'} = \int L_{xx'}(\theta)\phi(\theta)d\theta. \tag{16.17}$$

If a score combination is very unlikely, values of $P_{xx'}$ can be calculated for each score combination x and x' and plotted in a diagram (Rosa et al., 2001) and $P_{xx'}$ can then be used to construct a $(1-\alpha)100\%$ "highest density region" (HDR) for the response combinations. It is important to note that the values of $P_{xx'}$ cannot be interpreted as reflecting likely or unlikely events in any absolute sense because the magnitude of the individual $P_{xx'}$ depends on the number of row and column score points. To construct the HDR, first the cells should be ordered from largest to smallest $P_{xx'}$. The 95% HDR can then be determined by considering all cells that contribute to the first 95% of the cumulative total of $P_{xx'}$. According to the model, 95% of the examinees should obtain score combinations in that list of cells. Cells that are outside this region represent score combinations that are thus unlikely given the model.

Finally, the use of response times for computerized adaptive testing in the personality domain may be promising. Quick response times may in an educational context point at preknowledge; in typical performance testing it may be linked to random response behavior or even to particular personality traits. For example, Farrow et al. (2003) found significant correlations between truthful response times to auditorily presented questions and neuroticism scores. These preliminary data suggest that personality variables may play a part in response times. Statistics that are based on typical response times may so be used to identify personality traits. Auxiliary information from earlier testing, personality characteristics, personal history, and socioeconomic background may further enhance the interpretation of test performance.

References

Bradlow, E. T., Weiss, R. E., Cho, M. (1998). Identification of outliers in computerized adaptive testing. *Journal of the American Statistical Association*, *93*, 910–919.

Drasgow, F., Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology 38*, 67–86.

Farrow, T. F. D., Reilly, R., Rahman, T. A., Herford, A. E.,Woodruff, P. W. R. & Spence, S. A. (2003). Sex and personality traits influence the difference between time taken to tell the truth or lie. *Perceptual and Motor Skills*, *97*, 451–460.

Glas C. A. W., Meijer, R. R. & van Krimpen-Stoop, E. M. L A. (1998). *Statistical tests for person-misfit in computerized adaptive testing*. Technical Report RR 98-01. Enschede, the Netherlands, University of Twente.

Good, P. I. (2001). *Applying statistics in the courtroom.* Boca Raton, FL: Chapman & Hall.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Holden, R. R. (1998). Detecting fakers on a personnel test: Response latencies versus standard validity scales. *Journal of Social Behavior and Personality*, *13*, 387–398.

Klauer, K. C.(1995). The assessment of person fit. In G. F. Fischer & I. W. Molenaar (eds.), *Rasch models: foundations, recent developments, and applications* (pp. 97–110). New York: Springer-Verlag.

Knowles, E. S. & Condon, C. A. (1999). Why people say "Yes": A dual-process theory of acquiesence. *Journal of Personality and Social Psychology*, *77*, 379–386.

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161–176.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Lord, F. M. & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 453–461.

McLeod, L. D. & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, *23*, 147–160.

McLeod, L. D., Lewis, C. & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121–137.

Meijer, R. R. (2004). Using patterns of summed scores in paper-and-pencil tests and computer-adaptive tests to detect misfitting item score patterns. *Journal of Educational Measurement*, *41*, 119–136.

Meijer, R. R., Molenaar, I. W. & Sijtsma, K. (1994). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*, 107–135.

Meijer, R. R. & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review of recent developments. *Applied Measurement in Education, 8*, 261–272.

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115–127.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41*, 100–115.

Rammsayer, T. (1999). Timing behavior in computerized testing: Response times as a function of correct and incorrect answers. *Diagnostica*, *45*, 178–183.

Reise, S. P. & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143–151.

Robin, F. (2002). *Investigating the relationship between test response behavior, measurement and person fit*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Rosa, K., Swygert, K. A., Nelson, L. & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items-scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.) *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum.

Snijders, T. A. B. (2001). Asymptotic distribution of person fit statistics with estimated person parameters. *Psychometrika, 66,* 331–342.

van der Linden, W. J. & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

van Krimpen-Stoop, E. M. L. A. (2001). *Detection of misfitting item-score patterns in computerized adaptive testing,* unpublished doctoral dissertation. University of Twente, the Netherlands.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (1999). Simulating the null distribution of person fit statistics for conventional and adaptive tests. *Applied Pychological Measurement, 23,* 327–345.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (2000). Detecting person misfit in adaptive testing using statistical process control techniques. In W. J. van der Linden & C. A. W. Glas: *New developments in computerized adaptive testing: theory and practice* (pp.201–219). Boston: Kluwer-Nijhoff Publishing.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (2001). CUSUM-based person fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics*, *26*, 199–217.

van Krimpen-Stoop, E. M. L. A. & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement*, *26*, 164–180.

Waller, N. G. & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. *Journal of Personality and Social Psychology*, *57*, 1071–1058.

Wright, B. D. & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago: Mesa Press.

Chapter 17
The Investigation of Differential Item Functioning in Adaptive Tests

Rebecca Zwick

17.1 Introduction

Differential item functioning (DIF) refers to a difference in item performance between equally proficient members of two demographic groups. From an item response theory (IRT) perspective, DIF can be defined as a difference between groups in item response functions. The classic example of a DIF item is a mathematics question containing sports jargon that is more likely to be understood by men than by women. An item of this kind would be expected to manifest DIF against women: They are less likely to give a correct response than men with equivalent math ability. In reality, the causes of DIF are often far more obscure. Camilli and Shepard (1994) and Holland and Wainer (1993) provide an excellent background in the history, theory, and practice of DIF analysis.

There are several reasons that DIF detection may be more important for computerized adaptive tests (CATs) than it is for nonadaptive tests. Because fewer items are administered in a CAT, each item response plays a more important role in the examinees' test scores than it would in a nonadaptive testing format. Any flaw in an item, therefore, may be more consequential. Also, an item flaw can have major repercussions in a CAT because the sequence of items administered to the examinees depends in part on their responses to the flawed item. Finally, administration of a test by computer creates several potential sources of DIF that are not present in conventional tests, such as differential computer familiarity, facility, and anxiety, and differential preferences for computerized administration. Legg and Buhr (1992) and Schaeffer, Reese, Steffen, McKinley and Mills, (1993) report ethnic and gender group differences in some of these attributes; Powers and O'Neill (1993) review the literature on this topic.

The investigation of DIF in CATs can be conducted using several different administration schemes: First, the items to be assessed for DIF can be administered adaptively. Second, the items to be assessed can be "seeded" throughout

R. Zwick (✉)
Department of Education, University of California, 2216 Phelps Hall, Santa Barbara,
CA 93106–9490, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics
for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_17,

the exam and administered nonadaptively. Finally, the potential DIF items can be administered in an intact nonadaptive section. This chapter focuses on the first situation. DIF analysis for adaptively administered items involves two major technical challenges: First, an appropriate matching variable for DIF analysis must be determined. Clearly, the number-right score, often used in large-scale applications of DIF analyses by major testing companies, is not appropriate; on the other hand, matching on a scale score based on an IRT model is not entirely straightforward. Second, a method is needed that can provide stable results in small samples: Even if the total number of examinees for a given CAT is large, the number of responses for some items may be very small.

This chapter presents the methods that have been developed to analyze DIF in CATs, along with results of applications to simulated data. In the final section, fruitful directions for future research are outlined.

17.2 Methods for Assessing DIF in CATs

Formal discussion of DIF procedures for CATs appears to have begun in the late 1980s. Steinberg, Thissen, and Wainer (1990) recommended the application of a likelihood-ratio test approach that involves determining whether the fit of an IRT model to the data is impaired by constraining item parameters to be the same for two groups of examinees. While this approach has become a well-established DIF analysis method for nonadaptive tests (e.g., see Thissen, Steinberg & Wainer, (1993), it does not appear to have been applied to CATs, possibly because of the complexities introduced by the incomplete data that result from CAT administration (See Section 17.3 for further discussion).

In another early proposal for assessing DIF in CATs, Holland suggested comparing examinee groups in terms of item percents correct, basing the analysis on only those test-takers who received the item late in their CATs (Holland & Zwick, 1991). However, analyses of simulated data (Zwick, Thayer & Wingersky, 1993, 1994a) did not support the assumption underlying this procedure – that “examinees” who receive a particular item late in the CAT will be well-matched in ability. This key assumption would likely be violated even more severely in actual CATs, which involve many nonpsychometric constraints on item selection.

In addition to recommending the IRT-likelihood-ratio approach, Steinberg et al. (1990) suggested that DIF assessment procedures for CATs might be developed by matching examinees on expected true score and then applying existing DIF methods. The CAT DIF methods of Zwick, Thayer and Wingersky, (ZTW; 1994a, 1994b, 1995) and the CAT version of the empirical Bayes DIF method of Zwick, Thayer and Lewis (ZTL; 1997, 1999, 2000, Zwick & Thayer, 2002, 2003) are consistent with this recommendation; these methods are discussed in the subsequent sections. The only other main CAT DIF method that appears in the literature – the CATSIB procedure of Nandakumar and Roussos (2001, 2004; Roussos, 1996) – is also discussed below. In addition to the publications that propose specific methods, Miller (1992) and Way (1994) have addressed the general data analysis issues involved in conducting DIF analyses in CATs (also see Section 17.3 for new developments).

The ZTW methods are modifications of the Mantel–Haenszel (MH; 1959) DIF procedure of Holland and Thayer (1988) and of the standardization method of Dorans and Kulick (1986); the ZTL approach, originally developed for nonadaptive tests and later modified for CATs, is an enhancement of the MH method; and CATSIB is a modification of the SIBTEST procedure of Shealy and Stout (1993a, 1993b). The original "nonadaptive" versions of the Mantel–Haenszel, standardization, and SIBTEST methods are reviewed in the next section; the CAT analogues are then described.

17.2.1 A Review of the Mantel–Haenszel, Standardization, and SIBTEST Procedures

In the MH procedure of Holland and Thayer (1988), which is widely used by testing companies for DIF screening, a $2 \times 2 \times K$ table of examinee data is constructed based on item performance (right or wrong), group membership (the *focal group*, which is of primary interest, or the *reference group*), and score on an overall proficiency measure with K levels, used to match examinees. The two examinee groups are then compared in terms of their odds of answering the item correctly, conditional on the proficiency measure. The odds ratio is assumed to be constant over all levels of the proficiency measure.

Assume that there are T_k examinees at the kth level of the matching variable. Of these, n_{Rk} are in the reference group and n_{Fk} are in the focal group. Of the n_{Rk} reference group members, A_k answered the studied item correctly while B_k did not. Similarly, C_k of the n_{Fk} matched focal group members answered the studied item correctly, whereas D_k did not. The MH measure of DIF can then be defined as

$$MH\ D\text{-}DIF = -2.35 \ln(\hat{\alpha}_{MH}), \tag{17.1}$$

where $\widehat{\alpha}_{MH}$ is the Mantel–Haenszel (1959) conditional odds-ratio estimator given by

$$\hat{\alpha}_{MH} = \frac{\sum_k A_k D_k / T_k}{\sum_k B_k C_k / T_k}. \tag{17.2}$$

In Equation (17.1), the transformation of $\hat{\alpha}_{MH}$ places $MHD\text{-}DIF$ (which stands for "Mantel–Haenszel delta difference") on the ETS delta scale of item difficulty (Holland & Thayer, 1985). The effect of the minus sign is to make $MHD\text{-}DIF$ negative when the item is more difficult for members of the focal group than it is for comparable members of the reference group. Phillips and Holland (1987) derived an estimated standard error for $\ln(\hat{\alpha}_{MH})$; their result proved to be identical to that of Robins, Breslow and Greenland, (1986).

The Mantel–Haenszel chi-square test provides an approximation to the uniformly most powerful unbiased test of the null hypothesis of no DIF (common odds ratio equal to one) versus the hypothesis of constant DIF (common odds ratio not equal to one). Rejection of the null hypothesis suggests that item performance and group membership are associated, conditional on the matching variable.

The results of an MH DIF analysis typically include *MH D-DIF* (or some equivalent index based on the estimated odds ratio), along with its estimated standard error. In making decisions about whether to discard items or flag them for review, however, testing companies may rely instead on categorical ratings of the severity of DIF. Several testing companies have adopted a system developed by ETS for categorizing the severity of DIF based on both the magnitude of the DIF index and the statistical significance of the results (see Zieky, 1993). According to the original version of this classification scheme, a "C" categorization, which represents moderate to large DIF, requires that the absolute value of *MH D-DIF* be at least 1.5 and be significantly greater than 1 (at $\alpha = 0.05$). A "B" categorization, which indicates slight to moderate DIF, requires that *MH D-DIF* be significantly different from zero (at $\alpha = 0.05$) and that the absolute value of *MH D-DIF* be at least 1, but not large enough to satisfy the requirements for a C item. Items that do not meet the requirements for either the B or C categories are labeled "A" items, which are considered to have negligible DIF. Items that fall in the C category are subjected to further scrutiny and may be eliminated from tests. For most purposes, it is useful to distinguish between negative DIF (DIF against the focal group, by convention) and positive DIF (DIF against the reference group). This distinction yields a total of five DIF classifications: C–, B–, A, B+, and C+. (The rules for assigning items to DIF categories have been modified slightly over time, but the version outlined here was used in the ZTW and ZTL research described in this chapter.)

In the standardization DIF procedure (Dorans and Kulick, 1986), data are organized the same way as in MH DIF analysis. The standardization index, often called *STDP-DIF* (which stands for "standardized proportion difference"), compares the item proportions correct for the reference and focal groups, after adjusting for differences in the distribution of members of the two groups across the levels of the matching variable.

More specifically,

$$STD\ P\text{-}DIF = \sum w_k \hat{p}_{Fk} - \sum w_k \hat{p}_{Rk}, \tag{17.3}$$

where w_k is a weight associated with the kth level of the matching variable. In typical applications of *STD P-DIF*, including those described below,

$$w_k = \frac{n_{Fk}}{n_F}, \tag{17.4}$$

where $n_F = \sum_k n_{Fk}$ is the total number of examinees in the focal group. Under this weighting scheme the term before the minus sign in (17.3) is simply the proportion of the focal group that answers the studied item correctly, and the term following the minus sign is an adjusted proportion correct for the reference group. Although a standard error formula for *STDP-DIF* was developed by Holland (see Dorans & Holland, 1993) and two alternative formulations were derived by Zwick, (1992; see Zwick & Thayer, 1996), *STD P-DIF* is usually used as a descriptive measure and not as the basis for a formal hypothesis test.

The original versions of the MH and standardization DIF procedures involve matching examinees from two groups on the basis of observed test score – typically, the number correct. Under the classical test theory model, it can be shown that reference and focal group members who are matched in terms of observed scores will not, in general, be matched in terms of true score (see also Sheały and Stout, 1993a, 1993b; Zwick, 1990; Zwick, Thayer and Lewis, 1997). The measurement error problem vanishes under certain Rasch model conditions because of the sufficiency of the number-correct score for θ, but, except in that special case, the severity of the problem increases as the difference between the reference and focal group ability distributions increases and as the test reliability decreases. To address this problem, the SIBTEST procedure developed by Shealy and Stout, 1993a, 1993b) matches examinees on an estimated true score obtained by applying a "regression correction" to the observed score. The SIBTEST measure of DIF, $\hat{\beta}$, can be defined as follows:

$$\hat{\beta} = \sum_k w_k \left(\overline{Y}^*_{Fk} - \overline{Y}^*_{Rk}\right), \tag{17.5}$$

where $\overline{Y}^*_{Fk}$ and $\overline{Y}^*_{Rk}$ are adjusted mean scores (described below) on the studied item for the focal and reference groups, respectively, and w_k is a weight. Although the weight can in principle be defined as in (17.4), Shealy and Stout recommend defining it as

$$w_k = (n_{Rk} + n_{Fk})/N, \tag{17.6}$$

where N is the total sample size. Note that SIBTEST can be applied to a set of studied items simultaneously, in which case the $\overline{Y}^*_k$ values in (17.5) are the adjusted means for the set of items.

The steps involved in obtaining the adjusted means used in the original version of SIBTEST, which are described in detail by Shealy and Stout, 1993a), are as follows: (1) Assuming a classical test theory model for the regression of true score on observed score (in this case, the number-right score on the matching items, excluding the studied item), obtain the expected true score for each group at each of the K levels of the matching variable. For each group, the slope of this regression is the reliability of the set of matching items in that group (Shealy and Stout, 1993a, pp. 190–193). This adjustment is equivalent to the correction proposed by T. L. Kelley (1923) as a means of adjusting an observed test score for measurement error. (2) For each of the K levels of the matching variable, average the expected true score for the reference and focal groups and regard that average as the true score corresponding to the kth level. (3) For each level of the matching variable, estimate the expected item score, given the true score, for each group, assuming that the regression of item score on true score is locally linear. This expected item score is the adjusted item mean $\overline{Y}^*_k$ for that group. (Newer versions of SIBTEST formulate the regression correction somewhat differently; see Jiang & Stout, 1998.)

If the weighting function in (17.4) rather than (17.6) is chosen, and if the test is either very reliable or yields similar score distributions and reliabilities for the two groups, then the value of $\hat{\beta}$ will be close to that of *STDP-DIF* (Equation 17.3). The SIBTEST test statistic, which is obtained by dividing $\hat{\beta}$ by its standard error, is approximately standard normal under the null hypothesis of no DIF. Under some conditions, SIBTEST has been shown to provide better Type I error control than the MH (Roussos & Stout, 1996).

17.2.2 A Modification of the MH and Standardization Approaches for CATs (ZTW)

The ZTW CAT DIF approach requires that IRT item parameter estimates be available for all items. After responding to the CAT, examinees are matched on the expected true score for the entire CAT pool, and the MH or standardization procedures applied. Specifically, the matching variable is

$$\textit{Expected true score on CAT} = \sum_{i=1}^{I} \hat{p}_i \left(\hat{\theta}_{CAT} \right), \tag{17.7}$$

where $\hat{p}_i\,(\hat{\theta}_{CAT})$ is the estimated item response function for item i, evaluated at $\hat{\theta}_{CAT}$, the maximum likelihood estimate (MLE) of ability based on responses to the set of items received by the examinee, and I is the number of items in the pool. In the original ZTW studies, one-unit intervals of expected true score were used for matching; in our more recent application to very sparse data (Zwick & Thayer, 2002, 2003), two-unit intervals were found to work better.

In the initial ZTW simulation study, (1993, 1994a) that evaluated the performance of these methods, the pool consisted of 75 items, 25 of which were administered to each examinee using an information-based CAT algorithm. Item responses were generated using the three-parameter logistic (3PL) model, in which the probability of a correct response on item i in group G ($G = R$ or F, denoting the reference or focal group) can be represented as

$$p_{iG}(\theta) = c_i + (1 - c_i)\{1 + \exp[-(1.7a_i(\theta - b_{iG})]\}^{-1}, \tag{17.8}$$

where θ is the examinee ability parameter, a_i is the discrimination parameter for item i, c_i is the probability of correct response for a very low-ability examinee (which was constant across items in our simulation), and b_{iG} is the item difficulty in group G. The focal group difficulty, b_{iF}, is equal to $b_{iR} - d_i$. Hence, d_i is the difference between reference and focal group difficulties.

A simple relation between item parameters and MH DIF exists only in the Rasch model (Fischer, 1995; Holland and Thayer, 1988; Zwick, 1990), in which the *MHD-DIF* statistic provides an estimate of $4a_i d_i$ under certain assumptions (see Donoghue, Holland & Thayer, 1993). Even when the Rasch model does not hold, however, *MHD-DIF* tends to be roughly proportional to $a_i d_i$ (ZTW, 1993, 1994a). Therefore, in this study, we used $a_i d_i$ as an index of the magnitude of DIF present in item i.

In practice, the true item parameters are, of course, unavailable for estimating abilities and calculating item information within the CAT algorithm. To produce more realistic predictions about the functioning of the DIF methods in applications to actual examinee data, item parameter estimates, rather than the generating parameters, were used for these purposes. (This simulation design issue is discussed further in a later section.) A calibration sample was generated that consisted of 2,000 simulated examinees who responded to all 75 items in the pool under non-DIF conditions. Item calibration, based on the 3PL model, was conducted using LOGIST (Wingersky, Patrick & Lord, 1988).

The main simulation included 18 conditions. In half the conditions, the number of examinees per group was 500, while in the other half, the reference group had 900 members and the focal group had 100. The simulation conditions also varied in terms of focal group ability distribution (same as or different from reference group) and pattern of DIF. A detailed analysis of the accuracy of the CAT DIF estimates and of the classification of items into the A, B, and C categories (ZTW, 1993, 1994a) showed that the methods performed well. A small portion of the validity evidence is included here.

17.2.3 Correlations among CAT DIF Statistics, Nonadaptive DIF Statistics, and Generating DIF

For six of the 18 simulation conditions (all of which included reference and focal sample sizes of 500), two nonadaptive versions of the *MH D-DIF* and *STD P-DIF* statistics were computed for comparison to the CAT results. For both nonadaptive approaches, all 75 pool items were "administered" to all examinees. In the first procedure (referred to as "$\hat{\theta}$-75"), examinees were matched on an expected true score calculated using the MLE of ability based on all 75 responses. That is, the matching variable in Equation (17.7) was replaced by

$$\textit{Expected true score based on all pool items} = \sum_{i=1}^{I} \hat{p}_i\left(\hat{\theta}_I\right), \tag{17.9}$$

where $\hat{\theta}_I$ is the MLE of ability based on all $I = 75$ items. The second nonadaptive approach ("Number Right") was a conventional DIF analysis in which examinees were matched on number-right score. Correlations (across the 75 items in the pool) among the CAT-based DIF statistics, the DIF statistics based on nonadaptive administration, and the DIF magnitude index, $a_i d_i$, are presented in Table 17.1.

Because of a complex estimation procedure used only for the CAT-based DIF analyses, the CAT DIF statistics were much more precisely determined than were the DIF statistics for the other two matching variables. (This estimation procedure was used only within the context of the simulation and is not involved in ordinary applications; see ZTW, 1994.) To avoid giving a spuriously inflated impression of the performance of the CAT analyses, correlations were corrected for unreliability (see ZTW, 1993, 1994a for details). These corrected correlations (which occasionally exceed one) provide a more equitable way of comparing the three sets of DIF statistics than do the uncorrected correlations (also shown).

Table 17.1 shows that the CAT, $\hat{\theta} - 75$, and Number Right analyses produced results that were highly correlated with each other and with the DIF magnitude index. In particular, the two analyses based on all 75 item responses produced virtually identical results. (The similarity between these approaches may be substantially less for shorter tests.) The median (over conditions) of the corrected correlations with the DIF magnitude index were very similar for the CAT, $\hat{\theta} - 75$, and Number Right analyses, which is somewhat surprising since the CAT DIF approach uses ability estimates based on only 25 item responses. Correlations with the DIF magnitude index tended to be slightly higher for *MH D-DIF* than for *STD P-DIF*, a finding that is probably an artifact of the metric of the DIF magnitude index (i.e., the index is roughly proportional to the quantity estimated by *MH D-DIF*, whereas *STD P-DIF* is in the proportion metric).

Several extensions of the initial ZTW research were conducted. In one study (ZTW, 1995), we examined the effect on ability and DIF estimation of applying the

Table 17.1 Correlations between DIF estimates and DIF magnitude index values (from Zwick, Thayer & Wingersky, 1993, 1994a)

Type of DIF Measure		Type of Correlation	Median Correlation	
			MH D	STD P
$\hat{\theta} - CAT$	$\hat{\theta} - 75$	Uncorrected	0.89	0.86
		Corrected	0.99	0.95
$\hat{\theta} - CAT$	Number Right	Uncorrected	0.88	0.87
		Corrected	0.99	0.95
$\hat{\theta} - CAT$	Magnitude Index	Uncorrected	0.96	0.96
		Corrected	0.97	0.97
$\hat{\theta} - 75$	Number Right	Uncorrected	0.99	0.98
		Corrected	>1.00	>1.00
$\hat{\theta} - 75$	Magnitude Index	Uncorrected	0.87	0.87
		Corrected	0.97	0.95
Number Right	Magnitude Index	Uncorrected	0.88	0.87
		Corrected	0.98	0.95

Note: The two leftmost columns refer to the variables used to compute the correlations, e.g., "$\hat{\theta} - CAT$" refers to the DIF statistics (*MHD-DIF* or *STDP-DIF*) obtained after matching examinees on the CAT-based ability estimate. The DIF magnitude index is defined as $a_i d_i$ (see text). "Corrected" correlations are Pearson correlations that have been corrected for attenuation due to unreliability. The two rightmost columns give median correlations over six simulation conditions.

Rasch model to data that were generated using the 3PL model. Although the DIF statistics were highly correlated with the generating DIF, they tended to be slightly smaller in absolute value than in the 3PL analysis, resulting in a lower probability of detecting items with extreme DIF. This reduced sensitivity appeared to be related to a degradation in the accuracy of matching. In another study (ZTW, 1994b), we addressed the question of how to assess DIF in nonadaptively administered pretest items that have not yet been calibrated. A simple procedure that involved matching on the sum of the CAT-based expected true score (Equation 17.7) and the score on the pretest item (0 or 1) was found to work as well as more sophisticated matching procedures that required calibration of the pretest items. In another spin-off of the ZTW research, we discovered that adaptive administration has a systematic effect on the standard errors of DIF statistics. For fixed group sample sizes, adaptive administration tends to lead to smaller standard errors for *MH D-DIF* and larger standard errors for *STD P-DIF* than does nonadaptive administration. Although this phenomenon seems counterintuitive at first, it appears to be related to the fact that item proportions correct are closer to 0.5 in adaptive than in nonadaptive tests; this has opposite effects on the standard error of *MH D-DIF*, which is in the logit metric, and the standard error of *STD P-DIF*, which is in the proportion metric (see Zwick,1997; ZTW, 1994b).

17.2.4 An Empirical Bayes (EB) Enhancement of the MH Approach (ZTL)

Zwick, Thayer and Mazzeo, (1997, 1999, 2000) developed an empirical Bayes (EB) approach to Mantel–Haenszel DIF analysis that yields more stable results in small samples than does the ordinary MH approach and is therefore well suited to adaptive testing conditions. The computations, which involve only the *MHD-DIF* indexes and their standard errors, are detailed in the original references.

The model can be expressed as follows. Because ln $(\hat{\alpha}_{MH})$ has an asymptotic normal distribution (Agresti, 1990), it is reasonable to assume that

$$MH_i \,|\omega_i \sim N\left(\omega_i, \sigma_i^2\right), \tag{17.10}$$

where MH_i denotes the *MHD-DIF* statistic for item i, $E\left(MH_i\right) = \omega_i$ represents the unknown parameter value corresponding to MH_i, and σ_i^2 is the sampling variance of MH_i.

The following prior distribution is assumed for ω_i:

$$\omega_i \sim N\left(\mu, \tau^2\right), \tag{17.11}$$

where μ is the across-item mean of ω_i and τ^2 is the across-item variance. The parameters of the prior are estimated from the data. The posterior distribution of ω_i, given the observed *MH D-DIF* statistic, can be expressed as

$$f\left(\omega_i \,|MH_i\right) \propto f\left(MH_i \,|\omega_i\right) f\left(\omega_i\right). \tag{17.12}$$

Standard Bayesian calculations (see, e.g., Gelman, Carlin, Stern & Rubin, 1995) show that this distribution is normal with mean $W_i MH_i + (1 - W_i)\mu$ and variance $W_i \sigma_i^2$, where

$$W_i = \frac{\tau^2}{\sigma_i^2 + \tau^2}. \tag{17.13}$$

The posterior distribution of DIF parameters in (17.12) is used as the basis for DIF inferences. (An alternative version of the EB DIF method allows estimation of the distribution of the item's DIF *statistic* in future administrations.) The posterior distribution can be used to probabilistically assign the item to the A, B, and C DIF categories described in an earlier section. In addition, the posterior mean serves as a point estimate of the DIF parameter for that item. The posterior mean can be regarded as a *shrinkage* estimator of Mantel–Haenszel DIF: The larger the MH standard error, σ_i^2, the more the EB estimation procedure "shrinks" the observed MH_i value toward the prior mean (which is usually close to zero because MH statistics must sum to approximately zero in typical applications). On the other hand, as σ_i^2 approaches zero, the EB DIF estimate approaches MH_i.

In the initial phase of research (ZTL, 1997, 1999), the EB methods were extensively investigated through simulation study and were applied experimentally to data from paper-and-pencil tests, including the Graduate Record Examinations (GRE). Subsequent work involved an elaboration of the method that was based on the use of loss functions for DIF detection (ZTL, 2000). The EB DIF methods have been used by the U.S. Department of Defense to investigate DIF in the CAT version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB; see Krass & Segall, 1998) and have been experimentally applied to NCLEX, the computerized adaptive licensure exam of the National Council of State Boards of Nursing. Also, Miller and Fan (1998) compared the ZTL approach to a method identical to the MH version of the ZTW procedure and concluded that the ZTL approach was more promising for the detection of DIF in high-dimensional CATs. The most recent EB DIF research (Zwick & Thayer, 2002, 2003), sponsored by the Law School Admission Council (LSAC), was an investigation of the applicability of these methods to a large-scale computerized adaptive admissions test. Some findings from this study, which was part of an investigation (Pashley, 1997) of the feasibility of a computerized adaptive Law School Admission Test (LSAT), are described in the subsequent sections.

17.2.5 LSAT Simulation Study

In developing a modification of the EB DIF methods for the LSAT CAT context, we needed to accommodate LSAC's interest in CATs that are adaptive on the testlet level rather than the item level. To test the EB CAT procedure, therefore, we designed a simulation involving testlet-based CAT administration. The CAT pool consisted of 10 five-item testlets at each of three difficulty levels — a total of

150 items. The simulation included several conditions that varied in terms of focal group ability distribution (same as or different from reference group) and in terms of sample size (3,000 per group or 1,000 per group). In the large-*n* conditions, item-level sample sizes (within a group) ranged from 86 to 842; for the small-*n* conditions, the range was from 16 to 307. The data were generated using the 3PL model (Equation 17.8). As in our previous simulation studies of the EB method, we defined true DIF as follows in the LSAC research:

$$\mathit{True\ DIF} \;=\; -2.35\int \ln\left\{\frac{p_{iR}(\theta)/q_{iR}(\theta)}{p_{iF}(\theta)/q_{iF}(\theta)}\right\} f_R(\theta)d\theta, \tag{17.14}$$

where $p_{iG}(\theta)$ is the item response function for group G, given by Equation (17.8), $q_{iG}(\theta) = 1 - p_{iG}(\theta)$, and $f_R(\theta)$ is the reference group ability distribution. Pommerich, Spray and Parshall, (1995) proposed similar indexes in other contexts (see Roussos, Schnipke & Pashley, (1999) for discussion). This quantity can be viewed as the true MH value, unaffected by sampling or measurement error (see ZTL, 1997).

We matched examinees for DIF analysis on the basis of the expected true score for the entire item pool, as in the ZTW (1994a, 1995) studies; this seemed most consistent with available LSAC scoring plans. Our procedures for estimating the parameters of the prior, which had been developed for nonadaptive tests, needed some modification for application to CATs (see Zwick & Thayer, 2002, 2003). A major goal of the study was to determine whether the EB method, previously tested on samples no smaller than 200 examinees for the reference group and 50 for the focal group, could be applied successfully with even smaller samples.

17.2.6 Properties of EB DIF Estimates

How close were the EB DIF values to the target values given by Equation (17.14), and how did their accuracy compare to that of the non-Bayesian version of the MH statistics? We compared these two types of DIF estimates using root mean-square residuals (*RMSR*s), defined for each item as follows:

$$\mathit{RMSR} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{D}_r - \mathit{True\ DIF}\right)^2}, \tag{17.15}$$

where r indexes replications, R is the number of replications, $\hat{D}_r$ is either the *MH D-DIF* statistic or the EB posterior mean from the rth replication, and *True DIF* is the appropriate value from equation 17.14. The $RMSR$ represents the average departure, in the MH metric, of the DIF estimate from the *True DIF* value. If these *True DIF* values are regarded as the estimands for the DIF statistics, then these *RMSR* values give estimates of the mean-square error (the average distance between the parameter estimate and the parameter) for the DIF statistics.

17.2.7 RMSRs of EB and MH Point Estimates in the No-DIF Case

We first investigated the performance of the EB method when DIF was absent (i.e., the *True DIF* value for each item was zero). Abilities for both examinee groups were drawn from a standard normal distribution. A large-sample case, with 3,000 examinees per group and 200 replications, and a small-sample case, with 1,000 examinees per group and 600 replications, were considered. (In this portion of the study, item sample sizes per group ranged from about 290 to 800 in the large-sample condition and from about 80 to 300 in the small-sample condition.) The top panel of Table 17.2 gives, for each of the two sample sizes, the 25th, 50th, and 75th percentiles of the distribution of *RMSR* values across the 150 items. The difference in the performance of the EB DIF approach and that of the non-Bayesian MH statistic is quite striking: The median *RMSR* for the MH method was roughly 10 times the median $RMSR$ for the EB approach in both sample-size conditions. The EB DIF statistic departed from its target value of zero by an average of about 0.03 in the large-sample case and 0.07 in the small-sample case; the corresponding values for MHD-*DIF* were 0.37 and 0.68. However, these results for the no-DIF case might be regarded as "stacked" in favor of the EB method since the estimate of the prior mean in (17.11) will be close to zero, which is the target value for the DIF statistics. As noted earlier, this property of the estimated prior mean results from the fact that, in standard applications, the sum of the MH statistics over items is constrained to be near zero. It is useful, therefore, to separately examine the *RMSRs* for the EB and MH DIF estimates for the case in which DIF is present.

Table 17.2 *RMSR* results for EB and MH DIF statistics in LSAT simulation study (Zwick & Thayer,2002, 2003)

	Initial Group $n = 1,000$		Initial Group $n = 3,000$	
	EB	MH	EB	MH
	DIF Absent; Reference $N(0,1)$, Focal $N(0,1)$			
25th %ile	0.068	0.543	0.031	0.298
Median	0.072	0.684	0.034	0.365
75th %ile	0.078	0.769	0.037	0.417
	DIF Present; Reference $N(0,1)$, Focal $N(0,1)$			
25th %ile	0.460	0.565	0.284	0.317
Median	0.509	0.713	0.341	0.390
75th %ile	0.542	0.787	0.380	0.444
	DIF Present; Reference $N(0,1)$, Focal $N(-1,1)$			
25th %ile	0.464	0.585	0.302	0.322
Median	0.517	0.641	0.361	0.366
75th %ile	0.560	1.190	0.442	0.594

Note: Each *RMSR* summarizes results across replications (600 in the small-n condition and 200 in the large-n condition). The results above are summaries over the 150 items.

17.2.8 RMSRs of EB and MH Point Estimates in the DIF-present Case

In the conditions for which DIF was present, the *True DIF* values in this study (see Equation 17.14) ranged from –2.3 to 2.9 in the MH metric, with a standard deviation of about one. Here, as in the no-DIF conditions, we compared the EB point estimates of DIF to the *MH D-DIF* statistics using root mean-square residuals, defined in Equation (17.15). The bottom two panels of Table 17.2 summarize the results for the 150 items in four simulation conditions. The table gives, for each condition, the 25th, 50th, and 75th percentiles of the distribution of *RMSR* values across the 150 items. In the two small-*n* simulation conditions, the *RMSR* tended to be substantially smaller for the EB estimate than for *MH D-DIF*. In the large-*n* conditions, the advantage of the EB estimates was greatly reduced, which is to be expected, since the MH standard errors are small when samples are large, causing the EB DIF estimate to be close to the MH values.

The small-*n* results were also examined separately for easy, medium, and hard items. The smallest sample sizes occurred for the 50 hard items when the focal group ability distribution was $N(-1, 1)$, implying that it was centered more than two standard deviations lower than the mean difficulty of the items, which was 1.27. Here, reference group sample sizes ranged from 80 to 151, with a mean of 117; focal group sample sizes ranged from 16 to 67, with a mean of 40. These sample sizes are substantially smaller than is ordinarily considered acceptable for application of the MH procedure. Table 17.3 summarizes the *RMSR* results for these items, as well as the number of *RMSR* values exceeding 1 (about 1 SD unit in the *True DIF* metric). While only two of the 50 values exceeded 1 for the EB method, all 50 *RMSR*s for the MH procedure were greater than one. The median *RMSR* for the EB method for these items was 0.53, compared to 1.25 for the MH. It is interesting to note that, in a different subset of the results (not shown) for which the MH *RMSR* had a median of 0.53 (medium-difficulty items, $N(-1, 1)$ focal group distribution), the sample sizes averaged about 240 per group. Roughly, speaking, then, the EB procedure achieved the same stability for samples averaging 117 and 40 reference and focal group members, respectively, as did the MH for samples averaging 240 per group.

Table 17.3 Distribution of *RMSR*s for the 50 hard items in the small-sample condition (Zwick & Thayer, 2002, 2003)

	EB	MH
25th %ile	0.514	1.190
Median	0.532	1.252
75th %ile	0.558	1.322
Number > 1	2	50

Note: The range of item sample sizes across the 50 items and 600 replications was from 80 to 151, with a mean of 117 for the reference group and from 16 to 67, with a mean of 40 for the focal group.

17.2.9 Bias of EB and MH Point Estimates in the DIF Case.

The generally smaller *RMSR* values for the EB estimates are consistent with theory. Such estimates have smaller mean-square error than their non-Bayesian counterparts. They are not, however, unbiased; in fact, the bias of these estimates is greatest for the extreme parameter values. Table 17.4 shows the results of an analysis of the bias of the EB and MH estimates conducted by Zwick & Thayer, (2002, 2003) for the same simulation conditions displayed in Table 17.3. The squared *RMSR* for each item can be decomposed into two terms — the variance and the squared bias, B^2. In the present context, these components are defined as follows:

$$Variance = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{D}_r - \overline{\widehat{D}}\right)^2, \tag{17.16}$$

where $\overline{\widehat{D}}$ is the across-replication average of the R DIF statistics $\widehat{D}_r$ and

$$B^2 = \left(\overline{\widehat{D}} - TrueDIF\right)^2. \tag{17.17}$$

Table 17.4 shows the 25th, 50th, and 75th percentiles of the distribution (across the 150 items) of the variance and squared bias of the EB and MH estimates in the four simulation conditions. In the large-sample conditions, the EB and MH estimates showed similar amounts of bias, and the variances of the MH statistics tended to be larger than those of the EB statistics. In the small-sample conditions, the EB bias tended to be greater, particularly when the reference and focal group distributions differed.

Table 17.4 Variance (Var) and Squared Bias (B^2) results for EB and MH DIF Statistics in LSAT simulation study (Zwick & Thayer, 2002, 2003)

	Initial Group $n = 1,000$				Initial Group $n = 3,000$			
	EB		MH		EB		MH	
	DIF present; Reference $N(0,1)$, Focal $N(0,1)$							
	Var	B^2	*Var*	B^2	*Var*	B^2	*Var*	B^2
25th %ile	0.195	0.001	0.316	0.001	0.079	0.000	0.095	0.001
Median	0.238	0.007	0.498	0.004	0.108	0.001	0.141	0.003
75th %ile	0.259	0.035	0.592	0.018	0.127	0.019	0.166	0.015
	DIF present; Reference $N(0,1)$, Focal $N(-1,1)$							
	Var	B^2	*Var*	B^2	*Var*	B^2	*Var*	B^2
25th %ile	0.191	0.004	0.335	0.000	0.084	0.001	0.103	0.000
Median	0.210	0.027	0.402	0.002	0.100	0.003	0.127	0.001
75th %ile	0.242	0.088	1.402	0.013	0.182	0.016	0.339	0.011

Note: Var and B^2 are defined in (17.16) and (17.17).The results above are summaries over the 150 items.

A possible concern about the EB method is that extreme DIF items tend to be most affected by the biasedness of the EB estimates. In the present study, however, the relative performance of the EB method was quite good even for items with large DIF. Several items for which EB bias was substantial still had smaller *RMSR* values for EB than for MH, showing that the EB statistics were, on the average, closer to their target values than the MH statistics. For example, the largest B^2 value for the EB approach (1.58) occurred in the small-n, different-distributions condition for an item that had a *True DIF* value of 2.4. The EB variance value was 0.16, resulting in an *RMSR* of 1.32. The MH statistic, by contrast, had a B^2 value of 0.69 and a variance term of 2.34, resulting in an *RMSR* of 1.74. For only a very few items (six in the small-n, different-distributions condition, two in the small-n, same-distribution condition, and none in the remaining two conditions) did the EB $RMSR$ values exceed the MH *RMSR* values by more than 0.1.

17.2.10 Probabilistic Classification of DIF Results

In addition to offering an alternative point estimate of DIF, the EB method provides a probabilistic version of the A, B, and C DIF classification system. Two related problems associated with the traditional classification approach are that (1) when sample sizes are small, the DIF category is unstable and may vary substantially from one test administration to another and (2) attaching an A, B, or C label to an item may convey the mistaken notion that an item's DIF category is deterministic. The EB approach yields an estimate of the *probability* that the true DIF for an item falls into the A, B, and C categories, based on an estimate of the posterior distribution of DIF parameters (see ZTL,1997, 1999 for details). The estimated A, B, and C probabilities can be regarded as representing our state of knowledge about the true DIF category for the item.

A possible advantage of the EB method of probabilistic DIF classification is that it may convey information about the sampling variability of DIF results in a more comprehensible way than do the current procedures. This alternative way of representing the variability of DIF findings lends itself well to graphical display. Pie charts can be used effectively to represent the posterior probabilities associated with the A, B, and C categories, as shown in Figure 17.1. The displayed item, which had an actual status of A (*True DIF* near zero), was incorrectly classified as a C+ using the standard ETS procedure. According to the EB approach however, the estimated probability of A status was 0.65. The EB methods can be modified easily if the current rules used to assign items to categories are adjusted (e.g., see Miller & Fan 1998) or, if other hypothesis-testing approaches are substituted, for the Mantel–Haenszel procedure.

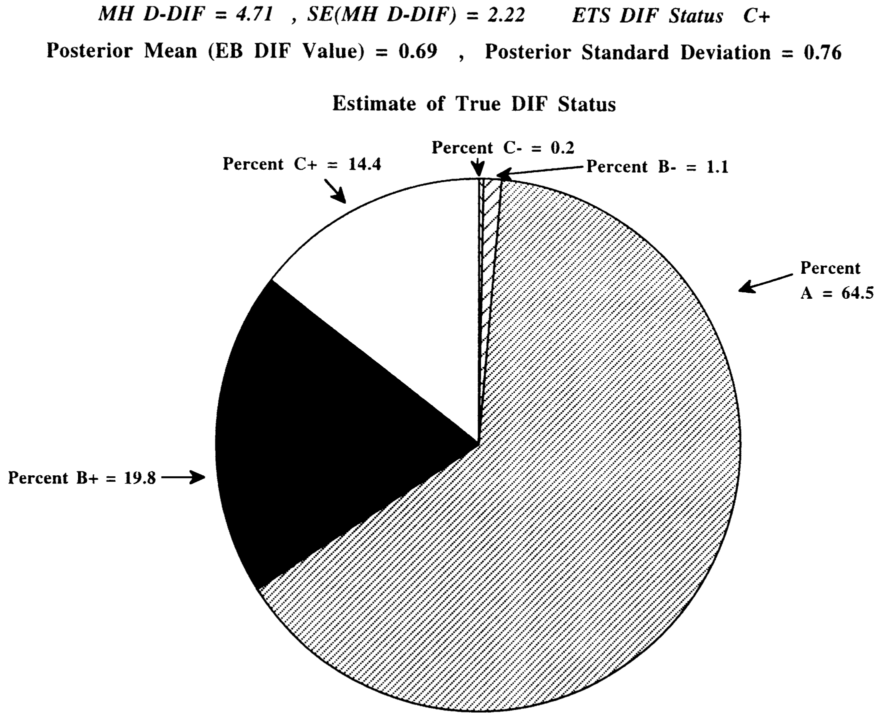


Fig. 17.1 EB DIF results for a CAT simulation item with true classification "A". Reference group: $N(0, 1), n = 101$; focal group: $N(-1, 1), n = 23$. (Adapted from Zwick & Thayer, 2002, 2003)

17.2.11 CATSIB: A Modification of the SIBTEST Method of Shealy and Stout

In CATSIB, the modification of SIBTEST for CAT (Nandakumar and Roussos, 2001, 2004; Roussos, 1996), examinees are matched on a regression-corrected version of an IRT-based ability estimate (grouped into intervals). Nandakumar and Roussos (2001, 2004) applied CATSIB to simulated "pretest" items that had been "administered" nonadaptively to at least 250 members in each examinee group. CAT administration affected only the items used to match examinees for DIF analysis (25 items per examinee, out of a pool of 1,000), not the 16 suspect items themselves, all of which were administered to each examinee. Although Nandakumar and Roussos 2004, pp. 179–180) suggest that their study is the only one to propose DIF analysis methods for pretest items in the CAT context, the design of their study resembles the earlier pretest DIF study of ZTW (1994b).

The CATSIB simulation findings on DIF parameter estimation, Type I error, and power were quite favorable; a partial summary of the Type I error and power results from Nandakumar and Roussos, (2001, 2004) is given in Table 17.5 (for $= 0.05$, with combined reference and focal sample sizes ranging from 500 to 1,000). When

Table 17.5 Summary of two-tailed rejection rates from CATSIB simulation study (adapted from Nandakumar & Roussos, 2001.)

	No DIF	DIF = 0.05	DIF = 0.10
Ref. & Focal θ means same			
$n_R = 250, n_F = 250$	0.051	0.275	0.728
$n_R = 500, n_F = 250$	0.044	0.336	0.840
$n_R = 500, n_F = 500$	0.057	0.478	0.940
Focal θ mean 0.5 SD lower			
$n_R = 250, n_F = 250$	0.046	0.260	0.715
$n_R = 500, n_F = 250$	0.045	0.328	0.811
$n_R = 500, n_F = 500$	0.051	0.462	0.935
Focal θ mean 1 SD lower			
$n_R = 250, n_F = 250$	0.050	0.235	0.638
$n_R = 500, n_F = 250$	0.057	0.294	0.719
$n_R = 500, n_F = 500$	0.058	0.397	0.894

Note: The nominal Type I error rate was 0.05. Results are averaged across six items in the no-DIF condition and across five items in each of the two DIF conditions. In all conditions, results are also averaged across 400 replications. The DIF metric here represents the average difference between the probabilities of correct response for matched reference and focal group examinees.

DIF was equal to 0.05 in the SIBTEST metric (i.e., probabilities of correct response for matched reference and focal group examinees differed by an average of 0.05), CATSIB's power ranged from 0.24 to 0.48; when DIF was equal to 0.10, the power ranged from 0.64 to 0.94. Power decreased as the difference between reference and focal group ability distributions increased. Type I error was quite well controlled overall. It is difficult to compare the CATSIB results to the ZTW and ZTL results for several reasons. First, the item sample sizes and administration mode (adaptive versus nonadaptive) are different. Also, Nandakumar and Roussos, (2001, 2004) included estimation of Type I error rates and power, while the ZTW and ZTL research focused on parameter estimation and probabilistic DIF classification. Yet another factor that makes CATSIB and the other CAT DIF procedures hard to compare is a particular feature of the simulation procedures used in the CATSIB studies. Most aspects of the Nandakumar and Roussos, (2001, 2004) simulation were carefully designed to be realistic. For example, great care was taken in choosing the properties of the generating item parameters so as to produce data resembling actual test results.

Another strong feature of the simulation is that it involved administration of 25 CAT items from a pool of 1000. This ratio of pool size to test length is more realistic than that used in ZTW and ZTL; the exposure control features implemented in the Nandakumar and Roussos simulation were also more elaborate. However, as the authors themselves mentioned, the simulated CAT administration and DIF analyses departed in one major way from actual practice and from the ZTW and ZTL studies: The true item parameters – those used in data generation – were used in all computations. In the CATSIB context, this means that true, rather than estimated parameters were involved in three major aspects of the simulation and analysis: the assignment of items to examinees via the CAT algorithm, the computation of the regression correction, and the calculation of examinee ability estimates, which are used for DIF matching. Nandakumar and Roussos, (2001, 2004) noted that their future research will use item parameter estimates in applying the CATSIB method, a change that should lead to a more realistic assessment of the utility of CATSIB. In summary, the CATSIB procedure seems promising and warrants further study. One recent CATSIB report describes efforts to eliminate the DIF estimation bias that occurred for certain items in the Nandakumar and Roussos study (Roussos, Nandakumar and Banks, 2006); another report presents a kernel-smoothed version of CATSIB (Nandakumar, Banks & Roussos, 2006).

17.3 Recent Developments

Recently, Lei, Chen, and Yu, (2006) published the results of a simulation study that compared three competing procedures in terms of their effectiveness in detecting DIF in seeded "pretest" items. The pretest items were administered to all test-takers, while the remaining items were adaptively administered. The three methods were CATSIB (Nandakumar and Roussos, 2001, 2004), a modified version of the logistic regression DIF procedure of Swaminathan and Rogers, (1990) in which an IRT-based ability estimate was substituted for the number-right score, and a modification of the IRT-based likelihood-ratio test (IRT-LRT) approach (Thissen, Steinberg & Wainer, 1993). In the modified IRT-LRT procedure, item responses that were missing due to CAT administration were imputed so that responses to a subset of the CAT items, which were assumed to be DIF-free, could be used as an anchor test.

Overall, the IRT-LRT method performed best, demonstrating adequate Type I error control and generally good power. The generalizability of this result is in question, however, because the imputation procedure that was implemented to accommodate CAT data could not be used in practice: Not only was the form of the true IRT model (3PL) assumed to be known, but responses that were missing due to CAT administration were imputed using the true item parameters (i.e., those used in data generation).

The logistic regression method exhibited poor Type I error control when the reference and focal groups differed in average ability; surprisingly, the Type I error inflation was worse when group sample sizes were equal. CATSIB showed poor Type I error control when the two groups had unequal sample sizes, particularly when average abilities also differed. The inflated error rates for CATSIB were attributed to problems in matching "test-takers" under these conditions. Also, although CATSIB showed some power advantages in detecting uniform DIF, it was not as effective as the other two methods in detecting group differences in item discrimination. The authors recommended that a similar investigation be conducted using a CAT modification of Crossing SIBTEST (a SIBTEST variant intended for detecting nonuniform DIF; Li & Stout, 1996) be conducted.

17.4 Future Research

A number of important questions remain to be addressed in future CAT DIF research:

1. How does the performance of existing CAT DIF methods compare when simulation design and analysis features are held constant? As noted earlier, the existing studies used differing simulation approaches and assessed different aspects of the DIF procedures. It would be useful to create a common simulation data set on which all existing methods could be applied. A common set of criteria could then be used to evaluate the results, including the performance of the DIF methods in terms of parameter estimation, Type I error rate, power, and DIF classification.
2. Can automated test assembly (ATA) procedures be used effectively to reduce DIF? Some work has been conducted that is relevant to this question. The Defense Department, in the early stages of the development of its ATA algorithms for paper-and-pencil tests, considered using these algorithms to regulate the amount of DIF. The focus, however, was on balancing DIF across forms rather than reducing the presence of DIF (Gary Thomasson, personal communication, September 4, 1998). Stocking, Jirele, Lewis and Swanson, (1998) explored the feasibility of using an ATA algorithm to reduce score differences between African-American and White examinees, and between male and female examinees, on the SAT I Mathematical Reasoning test. The goal of impact reduction was incorporated into a previously developed ATA algorithm, which was designed to "select items from a pool ... in such a way as to minimize the weighted sum of deviations from constraints reflecting desirable test properties ..." (Stocking et al., 1998, p. 203). If DIF information from a pretest were available, a similar approach could be used to minimize DIF, rather than impact. Furthermore, although the SAT application involved a paper-and-pencil test, the DIF reduction feature could be incorporated into a CAT algorithm. Exploration of ATA-based approaches to DIF minimization would be fruitful.
3. What sources of DIF are of particular concern in CATs, and how can they be reduced? An entirely different, but extremely important type of investigation that needs to be undertaken is field research to study the sources of DIF (as well as other threats to validity) that are of particular concern in CATs, and to determine how they can be reduced. It is not difficult to imagine situations in which CAT administration could introduce DIF into an item that was DIF-free in its paper-and-pencil incarnation. Suppose, for example, that for most items on a math test, computer experience has little effect on the probability of correct response, but that, on complex figural response items that require examinees to use a mouse to point to a graphical display, those who are computer-savvy have an advantage. Now suppose that computer familiarity (given a particular level of math ability) is more likely to occur in certain demographic groups, a conjecture that appears quite plausible (e.g., see Legg and Buhr, 1992; Wenglinsky, 1998). This phenomenon would create DIF on the figural response items. Another interesting

hypothesis of this kind is the following: Suppose that nonnative speakers of English rely on the ability to make notes directly on the test booklet, perhaps consisting of a partial translation of the item. If this type of note-taking were particularly important on certain types of items, computer administration could result in DIF.

In summary, a research effort that includes both simulation-based technical investigations and a program of field studies is needed to further our understanding of DIF in CATs and, more generally, to help us evaluate the fairness of computerized adaptive tests.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Donoghue, J. R., Holland, P. W. & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel–Haenszel and standardization measures of differential item functioning. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Hillsdale, NJ: Erlbaum.

Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel–Haenszel and standardization. In P. W. Holland and H. Wainer (Eds.), *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Erlbaum.

Dorans, N. J. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355–368.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian data analysis*. London: Chapman & Hall.

Holland, P. W. & Thayer, D. T.(1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Research Report No. 85–43). Princeton, NJ: Educational Testing Service.

Holland, P. W. & Thayer, D.T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale, NJ: Erlbaum.

Holland, P. W. & Wainer, H. (Eds.), (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Holland, P. W. & Zwick, R. (1991). *A simulation study of some simple approaches to the study of DIF for CAT's* (Internal memorandum). Princeton, NJ: Educational Testing Service.

Jiang, H. and Stout, W. F. (1998). Improved Type I error control and reduced estimation bias for DIF detection using SIBTEST. *Journal of Educational and Behavioral Statistics, 23,* 291–322.

Kelley, T. L. (1923). *Statistical methods*. New York: Macmillan.

Krass, I. & Segall, D. (1998). *Differential item functioning and online item calibration* (Draft report). Monterey, CA: Defense Manpower Data Center.

Legg, S. M. & Buhr, D. C. (1992). Computerized adaptive testing with different groups. *Educational Measurement: Issues and Practice, 11,* 23–27.

Lei, P.-W., Chen, S.-Y. & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43,* 245–264.

Li, H.-H. & Stout,W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61,* 647–677.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22,* 719–748.

Miller, T. R. (1992, April). *Practical considerations for conducting studies of differential item functioning in a CAT environment.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Miller, T. R. & Fan, M. (1998, April). *Assessing DIF in high dimensional CATs.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Nandakumar, R., Banks, J. C. & Roussos, L. (2006). *Kernel-smoothed DIF detection procedure for computerized adaptive tests* (Computerized testing report 00-08). Newtown, PA: Law School Admission Council.

Nandakumar, R. & Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Research report). Newtown, PA: Law School Admission Council.

Nandakumar, R. & Roussos, L. A. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29,* 177–200.

Pashley, P. J. (1997). *Computerized LSAT research agenda: Spring 1997 update* (LSAC report). Newtown, PA: Law School Admission Council.

Phillips, A. & Holland, P. W. (1987). Estimation of the variance of the Mantel–Haenszel log-odds-ratio estimate. *Biometrics*, *43,* 425–431.

Pommerich, M., Spray, J. A. & Parshall, C. G. (1995). *An analytical evaluation of two common-odds ratios as population indicators of DIF* (ACT Report 95-1). Iowa City: American College Testing Program.

Powers, D. E. & O'Neill, K. (1993). Inexperienced and anxious computer users: Coping with a computer-administered test of academic skills. *Educational Assessment, 1,* 153–173.

Robins, J., Breslow, N. & Greenland, S. (1986). Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics, 42,* 311–323.

Roussos, L. (1996, June). *A Type I error rate study of a modified SIBTEST DIF procedure with potential application to computerized-adaptive tests.* Paper presented at the annual meeting of the Psychometric Society, Banff, Alberta, Canada.

Roussos, L., Nandakumar, R. & Banks, J. C. (2006). *Theoretical formula for statistical bias in CATSIB estimates due to discretization of the ability scale* (Computerized testing report 99-07). Newtown, PA: Law School Admission Council.

Roussos, L. A., Schnipke, D. L. & Pashley, P. J. (1999). A generalized formula for the Mantel–Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24,* 293–322.

Roussos, L. & Stout, W. F. (1996). Simulation studies of effects of small sample size and studied item parameters on SIBTEST and Mantel–Haenszel Type I error performance. *Journal of Educational Measurement, 33,* 215–230.

Schaeffer, G., Reese, C., Steffen, M., McKinley, R. L. & Mills, C. N. (1993). *Field test of a computer-based GRE general test* (ETS Research Report No. RR 93-07). Princeton, NJ: Educational Testing Service.

Shealy, R. & Stout, W.F. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

Shealy, R. & Stout, W. F. (1993b). An item response theory model for test bias and differential test functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Erlbaum.

Steinberg, L., Thissen, D. & Wainer, H. (1990). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 187–231). Hillsdale, NJ: Erlbaum.

Stocking, M. L., Jirele, T., Lewis, C. & Swanson, L. (1998). Moderating possibly irrelevant multiple mean score differences on a test of mathematical reasoning. *Journal of Educational Measurement, 35,* 199–222.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Way, W. D. (1994). *A simulation study of the Mantel–Haenszel procedure for detecting DIF for the NCLEX using CAT* (Internal technical report). Princeton, NJ: Educational Testing Service.

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics* (ETS Policy Information Center report). Princeton, NJ: Educational Testing Service.

Wingersky, M. S., Patrick, R. & Lord, F. M. (1988). *LOGIST user's guide: LOGIST Version 6.00.* Princeton, NJ: Educational Testing Service.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.

Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics, 15,* 185–197.

Zwick, R. (1992). *Application of Mantel's chi-square test to the analysis of differential item functioning for functioning for ordinal items* (Technical memorandum). Princeton, NJ: Educational Testing Service.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel–Haenszel measure of differential item functioning. *Educational and Psychological Measurement, 57,* 412–421.

Zwick, R. & Thayer, D. T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics, 21,* 187–201.

Zwick, R. & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel–Haenszel DIF analysis to computer-adaptive tests. *Applied Psychological Measurement, 26,* 57–76.

Zwick, R. & Thayer, D. T. (2003, August). *An empirical Bayes enhancement of Mantel–Haenszel DIF analysis for computer-adaptive tests* (Computerized Testing Report No. 98-15). Newtown, PA: Law School Admission Council.

Zwick, R., Thayer, D. T. & Lewis, C. (1997) *An investigation of the validity of an empirical Bayes approach to Mantel–Haenszel DIF analysis* (ETS Research Report No. 97-21). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T. & Lewis, C. (1999). An empirical Bayes approach to Mantel–Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1–28.

Zwick, R., Thayer, D. T. & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25,* 225–247.

Zwick, R., Thayer, D. T. & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing DIF in polytomous items. *Applied Measurement in Education, 10,* 321–344.

Zwick, R., Thayer, D. T. & Wingersky, M. (1993). *A simulation study of methods for assessing differential item functioning in computer-adaptive tests.* (ETS Research Report 93-11). Princeton, NJ: Educationl Testing Service.

Zwick, R., Thayer, D. T. & Wingersky, M. (1994a) A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement, 18,* 121–140.

Zwick, R., Thayer, D. T. & Wingersky, M. (1994b) *DIF analysis for pretest items in computer-adaptive testing* (ETS Research Report 94-33). Princeton, NJ: Educational Testing Service.

Zwick, R., Thayer, D. T. & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341–363.

Part V
Multistage and Mastery Testing

Chapter 18
Multistage Testing: Issues, Designs, and Research

April Zenisky, Ronald K. Hambleton, and Richard M. Luecht

18.1 Introduction

Just as traditional computerized adaptive testing (CAT) involves adaptive selection of individual items for sequential administration to examinees as a test is in progress, multistage testing (MST) is an analogous approach that uses sets of items as the building blocks for a test. In MST terminology, these sets of items have come to be termed *modules* (Luecht & Nungester, 1998) or *testlets* (Wainer & Kiely, 1987) and can be characterized as short versions of linear test forms where some specified number of individual items are administered together to meet particular test specifications and provide a certain proportion of the total test information. The individual items in a module may be all related to one or more common stems (such as passages or graphics) or be more generally discrete from one another, per the content specifications of the testing program for the test in question. These self-contained, carefully constructed, fixed sets of items are the same for every examinee to whom each set is administered, but any two examinees may or may not be presented with the same sequence of modules, nor even the same modules.

Thus, the "stage" in multistage testing is an administrative division of the test that facilitates the adapting of the test to the examinee. Each examinee is administered modules for a minimum of two stages, where the exact number of stages is a test-design decision affected by the extent of desired content coverage and measurement precision. In each stage, an examinee receives a module that is targeted in difficulty to the examinee's provisional ability estimate computed from performance on modules administered during the previous stage(s). Within a stage, there are typically two or more modules that vary from one another on the basis of average difficulty. Because the modules vary in this way, the particular sequence of item sets that any one examinee is presented with is adaptively chosen based on the examinee's ability estimate. After an examinee finishes each item set, his or her ability estimate is

A. Zenisky (✉) and R.K. Hambleton
Center for Educational Assessment, University of Massachusetts, Amherst, MA 01002, USA

R.M. Luecht
ERM Department, University of North Carolina at Greensboro, Greensboro, NC 26170, USA

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_18,

updated to reflect the new measurement information obtained about that examinee's ability, and the next module is chosen to provide an optimal level of measurement information for a person at that computed proficiency level. High-performing examinees receive modules of higher average difficulty, while less able examinees are presented with modules that are comparatively easier.

Generally, implementation of MST is very flexible, but the exact structure of the MST and how it works during the operational delivery of the test involve a series of critical decisions with consequences for the relative efficiency of the test. For example, more modules per stage may make a test more adaptable to a wider range of examinee proficiency levels, but then, more easy items and hard items are needed to build the MST modules. The items must also be selected to simultaneously meet all requisite content specifications for any test route taken by the examinee. This is often very challenging when the modules at a given stage must be matched on content and also span a fairly wide range of item difficulty. Similarly, fewer items per stage likewise encourages the use of more adaptation, but can result in routing decisions being made on smaller and smaller slices of the content domain. The amount of precision that is "spent" at each of the MST stages is also a consideration. For example, a decision to use the most discriminating items at the outset of testing facilitates better routing at the beginning of the MST, but may compromise the quality of proficiency estimation (and hence, routing) at later test stages when a more precise matching between difficulties of modules and proficiency levels is possible and the discriminating powers of test items can be fully capitalized upon. Clearly, there are many trade-offs to consider in choosing a particular MST design.

An overview of MST is provided in this chapter with particular focus on (1) variations in approaches to implementing the MST model in operational testing given various measurement and practical considerations, and (2) summarizing the tradition of research into MST with relevant research findings. We will conclude this chapter by highlighting several emerging areas for future research.

18.2 Fundamentals of the MST Design

The idea of set-based tests with mechanical branching rules, independent of IRT, administered via paper and pencil can be traced to studies by Angoff and Huddleston (1958), Cronbach and Gleser (1965), and Linn, Rock and Cleary, (1969), among others. Then, with the fundamental tenets of modern item response theory (IRT) outlined by Lord and Novick (1968), Lord (1971) was the first researcher to provide the framework and measurement justification for adaptive-by-stage testing with IRT. Two-stage testing was described there as a method of obtaining improved measurement for not only typical examinees but also, and most importantly, those at the extremes of the ability distribution. To this day, investigation into the alternative test designs within the broad heading of MST in many contexts and domains is ongoing (see Mead, 2006).

In terms of implementing MST, as with CAT, there are many design variables and development procedures that come together and impact what the finished ’test form’ looks like under MST, and thus it is a highly customizable approach to adaptive testing. At the same time, some of the practical issues that arise with CAT [as inventoried by Green, Bock, Humphreys, Linn and Reckase (1984), Mills and Stocking (1996), and Wise and Kingsbury (2000)] are relevant in one form or another to the implementation of MST. There are, however, aspects of the development and operational use of MST that are different enough from CAT to warrant a review of the design variables present in MST, the methods used in developing multistage tests, and the operational issues that must be reckoned with.

For example, in providing an overview of two-stage testing using IRT, Lord (1980) outlined a number of design considerations that he identified as impacting the nature and quality of ability estimation from tests using a two-stage procedure. His ideas, as abstracted below, can be generalized to a test of n stages:

- Total number of items in the test
- Number of items in the initial and each n-stage module
- Difficulty of the initial module
- Number (and difficulty) of alternative modules in each stage
- Cut-points for routing examinees to modules, and
- Method for scoring stages and each nth-stage test.

While Lord suggested that it was not possible to identify truly statistical optimal designs for each and every operational testing context, different combinations of these variables can provide high-quality results as needed for a particular test’s use or the interpretations to be made based on the test scores. To Lord’s (1980) list can be added several additional considerations that have emerged through MST research, including the number of stages, the ability distribution of the examinee population, the extent of target information overlap for modules within stages, whether random module selection (at an appropriate difficulty level) or panel-based administration is used, whether content balancing is done at the module or total test level, the choice of method for automated test assembly, the size and quality of the item bank, how test information is distributed across stages, the placement of cut-scores for pass–fail decisions, the issue of item review, and item-exposure levels. We will consider next many of these variables and highlight some research studies that have explored their effects on measurement quality.

18.3 Structuring a Multistage Test

The total number of items in the test, the total number of stages in the test, and how many items to include per stage are some first considerations that arise in the process of developing a multistage test. An often-cited benefit of adaptive testing is the opportunity to shorten tests in terms of the number of items presented to each examinee by targeting tests to examinee ability (thereby reducing testing time), but considerations of domain coverage and measurement precision must still be balanced. Research in MST for credentialing exams has examined a wide range of test lengths, including studies with over 150 items administered to examinees over six stages (Luecht & Nungester, 1998) and with 35 items (two stages), as found in some information-technology testing applications (Xing & Hambleton, 2004). Jodoin, Zenisky and Hambleton (2006) found that a 40-item, two-stage test performed nearly as well as a 60-item, three-stage test (as represented by decision accuracy, coefficient kappa, and correlations between true and estimated abilities from each design). Ultimately, multistage tests can provide some reduction in testing time, although this depends on a host of other factors, including the desired level of measurement precision and the complexity of the content constraints to be implemented.

While much of the MST research to date has focused on two- and three-stage tests in which all examinees receive the same number of stages, there are exceptions, of course. Computerized mastery testing (CMT), which is a variation on the basic MST approach, involves variable-length mastery tests where different examinees may receive different numbers of modules, and a four-stage test was the focus of a study by Luecht, Nungester and Hadadi (1996; see also Luecht & Nungester, 1998). The number of stages is also affected by policy considerations: For example, in a high-stakes context, stakeholders may not be comfortable using a two-stage test due to a perception of some examinees being unable to recover or "pass" if their true abilities are at or above passing and they are routed to a lower-difficulty module in the second stage. Clearly, measurement efficiency is not the only consideration taken under advisement in the process of deciding the appropriate number of stages to include.

Considering module length, some recent studies (Jodoin, Zenisky and Hambleton, 2006; Hambleton & Xing, 2006) have implemented modules consisting of 20 items in each of three stages, while Luecht & Nungester (1998) worked with three-stage tests composed of modules that were 60 items in length (for a total test length of 180 items). Alternatively, work by other researchers has explored other configurations of items, such as longer first-stage tests (Xing & Hambleton, 2004) or tests with more items in the stage(s) after the first (Reese, Schnipke & Luebke, 1999; Schnipke & Reese, 1999; Reese & Schnipke, 1999; Kim & Plake, 1993).

Patsula (1999) defined the rationale for longer first stages as relating to the need for more accurate measurement in the first stage prior to routing (the "Routing Test" strategy), while extending the length of subsequent stages may be justified by the thinking that since the tests are more closely aligned with examinee ability at later points in the test, providing more items tailored to estimated ability in those stages is capitalizing on the information obtained from examinees after some routing has been done (the "Higher Stage" strategy).

Once examinees have been administered the first stage, an issue discussed by Lord (1980) is the number and the relative difficulty of the modules in each and every subsequent stage. A multistage test is often represented schematically as having two or three modules varying by difficulty in each of the second and third stages (Figure18.1). The design process for these modules in stages subsequent to the first

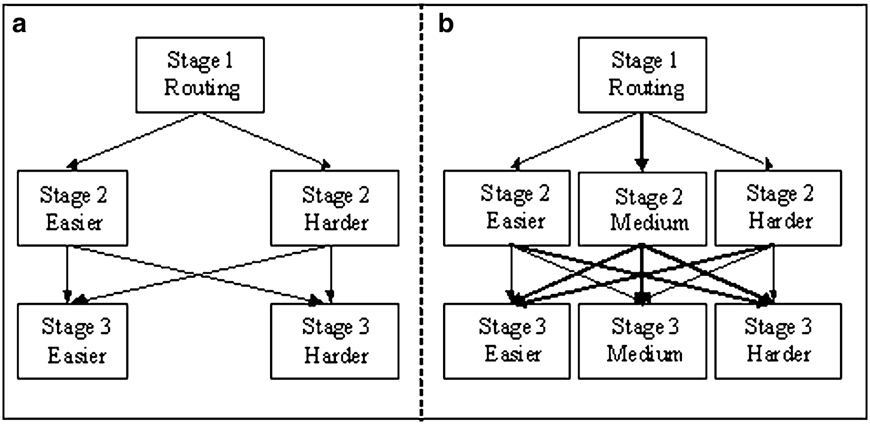


Fig. 18.1 Three-stage tests with (**a**) two and (**b**) and three modules in the second and third stages

is contingent on several points, including the level of routing precision desired by the testing program, the depth and breadth of the item bank, and the extent to which such modules should be discrete or can overlap. In Figure 18.1, notice that the modules differ by difficulty: for example, relatively easy, medium, and relatively hard, they are generally aligned relative to the ability continuum of examinees, such that lower-ability examinees should be routed to the easier modules in each stage while more proficient examinees would be presented with more difficult modules.

However, to the extent that “easy”, “medium”, and “hard” are relative terms in their meaning, these modules are actually referenced by the ability scale (which in IRT generally has a mean set equal to zero and a standard deviation of scores equal to one). An example in the case of a stage with two levels might involve using test assembly procedures to target the two modules at 0.0 and 0.5, respectively (see, for example, Breithaupt & Hare, 2007). In the process of constructing such modules, a testing program might want to make the modules more distinct from one another, such as in recent studies by Xing & Hambleton (2004) where the easy and hard modules were centered as much as one full standard deviation apart.

However, as module difficulty is generally defined by average b-parameter estimates, such averages can be obtained in two ways. Lord (1980) referred to these as either *peaked* or *nonpeaked* distributions of items within modules. Peaked modules are those in which items are all of approximately equal difficulty, while nonpeaked modules contain more variation and so the average difficulty is arrived at via a more heterogeneous assemblage of items.

An additional consideration for implementation of MST (and other test designs) is whether or not to permit examinees to return to previously administered items during the course of testing (see Hadadi, Luecht, Swanson & Case, 1998). In CAT, item review and changes following an initial item choice selection during testing are generally limited as a matter of policy: Items in item-level CAT are chosen one by one on the basis of updated proficiency estimates, and so to allow examinees to go back and change one or more answers may have the effect of decreasing the efficiency of the adaptive algorithm because the items administered after the changed answer might no longer be the most optimal to select at that point in the test session (Wainer, 1993). In a fixed-length test, this has the practical consequence of potentially increasing the final estimate of the standard error of measurement (SEM), and in cases where a test is to conclude by reaching a threshold for the size of the SEM, the test length may need to be extended, perhaps unreasonably.

Also, revising the answers in CAT earlier in the test may be of greater significance than those later in the test. In most MST applications, examinees may be permitted to complete items in almost any sequence of items within a module, though the decision to permit review and revision between modules encounters the same obstacle as is found between items in CAT. For this reason, review within stages is generally permitted, but not across stages. Ultimately, research seems to suggest that the primary benefit of item review is related to a psychological comfort factor, and in the context of certification and licensure using MST, the option to review within stages may be sufficient to alleviate anxiety for many examinees (Patsula, 1999).

Lord (1980) cited the issue of strategies and cut-scores for routing examinees to modules as a particularly critical one in MST, as the quality of the method by which examinees are routed to certain modules as opposed to others defines the usefulness of an adaptive, multistage administration. Some of the options cited in the literature for routing examinees to modules between stages include using number-correct (NC) scoring, cumulative weighted NC, and IRT-based provisional proficiency scores such as maximum-likelihood estimates (MLE) or estimated a priori (EAP) estimates (Luecht, 2000). Other approaches also considered in the literature include using maximum testlet information and Wald's (1947) sequential probability ratio test (Luecht, Nungester and Hadadi, 1996). To implement NC scoring, Luecht, Brumfield & Breithaupt (2006) suggested incremental computation of upper and lower bounds for NC scoring of various combinations of routings through the panel structure. Location of routing points can be done using either the approximate maximum information (AMI) or defined population intervals (DPI) approach. The AMI method uses cumulative TIFs to identify optimal decision points for module selection, while the DPI structure is used to specify proportional routings through the panel and module structure.

Lord (1980) suggested that the difficulty levels of the modules should match the estimated ability levels of the examinees who are routed to them. Schnipke & Reese (1999) used NC scoring and a simulation study methodology, in which they tried an approach to minimizing the mean-square error (MSE) of ability estimates from simulated examinees administered easy, medium, and hard modules in order to figure out at which NC value the MSE was lowest between low and medium modules and medium and hard modules. Dodd and Fitzpatrick (2002) advanced a routing method that is both NC- and information-based involving computing NC ability estimates and then selecting modules based on information at that estimate.

Kim & Plake (1993) used a simple comparison procedure in which examinees were routed to the module whose average difficulty most closely matched their estimated ability on the ability scale. Hambleton & Xing (2006) chose to implement strategies anchored to the proficiency scale (related to the DPI method suggested by Luecht, Brumfield & Breithaupt 2006). Here, approximately equal numbers of examinees were routed to each second-level module. A suggested variation on this approach is to have examinees within two standard errors of the value that the MST is targeted at routed to the middle difficulty module; examinees on either side of those cutoff values are routed to the easy or hard modules as appropriate.

Another aspect of routing concerns the possible pathways for routing (Luecht & Nungester, 1998). To the extent that examinees are routed between modules from stage to stage, the number of possible pathways for routing is a variable that can also be controlled by the testing program. In some testing applications, examinees might not be permitted to move from the easiest module in one stage to the hardest module in the immediately subsequent stage since such a move would most likely reflect the role of measurement error. Such dramatic changes in estimation of ability between later stages are not likely under normal testing conditions, and may well be considered a flag for score review in some testing programs.

Closely connected to the methods for routing are the methods for scoring modules and the entire test. Lord (1980) suggested that in a situation with statistically equivalent items, simple NC scoring could be appropriate. In the psychometric literature, while relatively few studies have focused directly on this aspect of the design, scoring in the context of MST has involved Bayesian analysis, approaches based on maximum-likelihood estimation, the testlet models of Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (2000), and more extensive methods based on NC scoring.

Schnipke & Reese (1999) authored an important study that explored the use of NC routing and Bayes modal estimates of ability in the context of two-stage, multistage, and maximum-information testlet-based designs. Thissen (1998) obtained EAP ability estimates for examinees based on a pattern of two or more summed scores, and also developed a method for using Gaussian approximation to EAP ability estimation that is in essence a weighted linear combination of such estimates from separate summed scores, which allows for the estimation of ability from raw-score patterns obtained through MST. A recent study by Zenisky (2004) found few differences in either ability estimation or decision consistency and accuracy among four methods of routing [NC, defined population intervals, proximity (with mean b-value differences), and random assignment].

An important distinction among various MST designs and implementation strategies involves the degree to which the modules themselves are structured as larger test administration units. That is, MST modules can be adaptively administered as unique components in real time or housed within preconstructed test administration units called “panels”. Luecht & Nungester (1998; also see Luecht, Brumfield & Breithaupt, 2006, as well as Luecht & Burgin, 2003) argued that the highly structured panel approach to MST design offers many operational advantages, including tight quality controls over test assembly, strong controls over item and module exposure, and simplified scoring and data management.

18.4 Automated Test Assembly

The consideration of the average difficulty of modules and the differentiation of modules by difficulty brings up the issues involved in how modules for MST are assembled. While these modules could be assembled by hand, the complexity of the task and the volume of modules needed in most large-scale testing contexts lead most programs to choose the automated test assembly (ATA) route instead (see Luecht & Nungester, 1998; Luecht, 2000; Breithaupt, Ariel & Veldkamp, 2005; Breithaupt & Hare, 2007; Luecht, 2006; Luecht, Brumfield & Breithaupt, 2006; van der Linden, 2005). In large part, the literature on methods for automated assembly of modules and tests for MST builds on the extensive psychometric research that exists for item selection and test assembly for CAT, but ATA in an MST context is an aspect of the design that contributes substantially to differentiating MST from the other test designs. The issues for automatically assembling multistage tests are many. As described by Luecht & Nungester (1998), Luecht (2000), Luecht, Brumfield & Breithaupt, (2006), and van der Linden, (2005), the challenges include item bank size, the potential to have the algorithm meet an objective function (i.e., an objective function is a test specification such as “the test should closely match a target information function”), the possibility of different specifications for different modules, and the need for multiple replications to ensure module security and minimize item exposure.

ATA software (e.g., CASTISEL, ConTEST) is designed to implement optimization algorithms or heuristics (or both) to satisfy certain content or statistical goals and explicit and implicit rules about test fairness and test content (see, for example, Breithaupt & Hare, 2007; Luecht, 2000, 2006), and it is all done in advance of testing, which permits human review of the modules if desired. This systematization allows for the process of module development to be more standardized, particularly with respect to difficulty and test information, and reduces the labor-intensive task of hand-assembling the numerous modules needed for a large-scale, operational, high-stakes MST testing program. ATA software requires that the constraints and goals of the modules to be built be specified as a mathematical optimization model to be maximized or minimized (Luecht, 1998, 2000, 2006; Luecht & Nungester 1998; van der Linden, 2005), and the task for the software is to solve that model using integer programming, network-flow, or some other such approach.

To implement MST in practice, item bank considerations are critical to ensure that the “demands” (constraints and statistical objective functions) can be met by the “supply” (the item bank). Automated test assembly facilitates the process of selecting items for a particular MST design, but still requires that the item pool be of a depth and breadth to support such construction (see, for example, van der Linden, Ariel & Veldkamp, 2006). Such item bank considerations were a focus for recent studies by Xing (2001) and Xing & Hambleton (2004).

In the Xing study, (2001), varying conditions of item bank size and quality and placement of passing score were compared. Of the 72 possible conditions in the study (4 computer-based test designs × 2 levels of bank size × 3 levels of item quality × 3 levels of passing score), it was found that as item quality improved, so did both decision accuracy (DA) and decision consistency (DC). Xing also noted that the benefit of larger item banks came in the form of greater ability to meet statistical targets such as test information functions and automated test assembly constraints. A subsequent study further exploring variations in item bank size and item quality (Xing & Hambleton, 2004) found little difference among different test designs for a credentialing exam (linear forms, two-stage MST, and CAT), but the quality and size of the item bank did make a practically significant different in the results.

Similarly, deciding how to distribute test information across stages involves weighing efficiency and using test design to maximize the information to be obtained. This notion of using test information in the development of tests in a panel-based structure has been described by Luecht (2000) as a way to provide consistent control over error variance of estimated scores at various regions of the proficiency scale, in contrast to CAT, where the "target" for the test information function (TIF) can be understood as the overall maximum information possible after the last item is administered to an individual examinee (for maximizing score precision). For MST, however, modules can be viewed as intermediate administration structures of the test, and thus TIFs are specified for each module. The issue in this attribute of the MST design focuses on the partitioning of the target test information function across stages: Is it better to obtain greater test information early on in the test for better module selection, or hold off and wait until some later point in test administration when the matching of examinees and the difficulties of modules can capitalize on the higher discriminating powers of items? This is an important area for research.

With respect to the state of ATA research, one particularly promising approach is the normalized weighted absolute deviations heuristic (NWADH; Luecht, 1998, 2000; Luecht, Nungester and Hadadi, 1996; Luecht & Nungester, 1998), which uses item-level information functions to manage need and availability of items in the bank to assemble modules and/or panels as specified by constraints. Other work by Armstrong et al. (2000) and Reese, Schnipke & Luebke (1999) has invoked a weighted-deviations model in a process that involves the selection of items at random from the item bank to create modules. Berger's (1994) work on building optimal modules used test information in an item-selection methodology predicated on estimating ability as efficiently as possible. This technique is, however, limited by the ability-level-specific meaning of optimal, in that what is optimal for one ability level (range) is clearly not for a different level.

van der Linden and Adema, (1998) presented another method for ATA using 0–1 linear programming (LP) where they conceptualized a multiple-form assembly problem instead as a series of two-form assemblies. 0–1 LP was also the subject of an earlier study by Adema (1990) in which a variation on LP was referred to as mixed-integer programming (MIP). Such MIP models, as noted by Adema, are comprised of both integer and continuous decision variables. In this paper, Adema also used a 0–1 linear programming approach for assembling an MST. van der Linden 2000) presented several alternative methods for ATA based on mixed-integer programming for assembling tests from a bank with an item-set structure. These methods were evaluated using mathematical programming feasibility and expected solution times. An example of software appropriate for this purpose is CPLEX, which solves integer programming and extensive linear programming problems (see www.ilog.com for more information).

Luecht (1997), Vos (2000), and Vos and Glas (2001) have also studied another aspect of ATA for MST: the case of building tests or modules with multidimensional constraints. As multidimensional IRT (MIRT) is increasingly being studied for eventual use in operational testing, its application to MST is a logical extension of previous research. As reported by Luecht, in the multidimensional case, TIFs are needed not only for total test or modules but also for separate content areas in which subscores are to be reported.

With so many approaches to ATA, finding a methodology that aligns with the goals of different testing programs is possible. Ideally, however, with respect to MST, these automated test assembly algorithms not only need to be flexible enough to develop modules for various MST designs but also should be capable of creating multiple panels that control the overlap of items or modules between panels (Luecht, 2003; Luecht & Burgin, 2003; Luecht, Brumfield & Breithaupt, 2006). For test development, such an approach can improve efficiency with respect to the basic assembly of modules and permit great attention to be paid to those aspects of test assembly that are not so easily automated. There are qualitative concerns (for example, sensitivity and fairness issues) that are not so easily managed via automation, and those aspects of a test or module clearly benefit from careful review by test developers.

Another consideration in the specification of constraints for MST ATA is whether domain coverage should be achieved within stages or across the whole test (Luecht & Nungester, 1998; Folk & Smith, 2002). To meet elaborate content specifications within stages can require more items at each stage, while meeting test specifications across an entire test provides greater flexibility in terms of test assembly. One difficulty in content balancing across the entire test, however, is that test users may not consider it appropriate to route examinees through a limited number of stages when each of the stages is not reasonably representative of the domain of interest. In other words, if the set of items an examinee is given only covers a portion of the test specifications, should decisions about the rest of the test to be presented be based on data that are incomplete in that respect, from a fairness perspective? Research is not clear on this point, but it may be that stages with fewer items in relatively constrained domains of interest (i.e., reading comprehension) may be perfectly appropriate for content balancing within stages whereas more content-based and/or cognitively complex domains may require more items within a stage to accomplish the same goal. In some testing applications, resolving this dilemma may result in the administration of more items than are strictly necessary for precise ability estimation (Folk & Smith, 2002).

18.5 Comparative Studies of MST

Clearly, given the design considerations detailed previously, what is generically referred to as “the MST design” in fact comprises an enormous range of theoretical and practical alternatives for implementation. While these variations do correspond to a high level of complexity for implementation, this design also represents tremendous flexibility for individual testing agencies. With such an accommodating design, MST is a very customizable approach to obtaining measurement precision for examinees along an ability continuum. However, the measurement properties associated with the many possible MST variations are not yet well understood, and so comparative studies into applications of MST using IRT have continued.

First, many studies have taken an outcomes-oriented approach with particular focus on the effects of various test structures and different implementation strategies, particularly with respect to the dimensions suggested by Lord (1977, 1980). Comparing results from simulation studies of MST and other test designs with respect to ability estimation and classification of individuals into pass–fail categories (see Luecht, 2006, for example) provides this information. The second direction for MST research to this point has been on investigating research into modules constructed around sets of items with common stems such as passages or graphics.

18.6 Evaluating MST Relative to Other Test Designs

Numerous studies of MST involve examination of the quality of ability estimates with respect to the entire continuum of examinee ability, where criteria such as root mean-square error (RMSE), bias, and relative efficiency are used to compare true and estimated values for simulated examinee ability. In the work of Reese & Schnipke (1999), where the efficiency of a two-stage testlet design was compared with CAT and a paper-and-pencil linear test, ability estimation was evaluated using RMSE and bias. Across the entire ability distribution, the CAT naturally exhibited the lowest RMSE and the least bias, although the most carefully constructed two-stage tests were actually the most error-free in the ability range from -2.0 to 2.0.

A subsequent study by Reese, Schnipke & Luebke (1999) that focused on strategies for optimal assembly of testlets found that a carefully constructed and content-balanced two-stage test outperformed the CAT and the paper-and-pencil test in the middle portion of the ability scale with respect to both bias and RMSE, even though the statistical constraints for assembly were not strictly met. An additional study authored by Schnipke & Reese (1999) found that several testlet-based designs (including a basic two-stage design, a two-stage design with the possibility of changing second-stage levels if misrouting was suspected, and a multistage test with four stages and a 1-3-4-5 design of modules) resulted in improved measurement precision as defined by RMSE and bias relative to paper-and-pencil testing. The quality of the measurement from those MST designs was almost as good as that observed with the CAT designs under study as well.

Studies by Kim (1993) and Kim & Plake (1993) also focused on two-stage testing. The purpose of the former study was to compare an IRT-based, two-stage test to an individualized CAT. The results from this study indicated that a fixed-length CAT provided superior measurement precision for ability estimation to IRT-based two-stage tests of equivalent length. In the Kim & Plake (1993) study, which was an extension of the Kim (1993) work, it was found that the structure and attributes of the routing test most substantially influenced measurement precision, but in most cases CAT again provided more accurate ability estimates than any of the two-stage designs under consideration in this study. The best of the two-stage designs was the one with a rectangular distribution of items in the routing test and an odd number of second-stage modules.

In Patsula (1999), 12 different MST designs were considered, also relative to CAT and paper-and-pencil. These designs varied with respect to the number of stages (2 or 3), the number of modules in each second- and third-stage test (either 3 or 5), and the number of items in each stage (between 6 and 24 in Stage 1, between 12 and 24 in Stage 2, and between 6 and 18 in Stage 3). As evaluated on the basis of RMSE, bias, and relative efficiency, the errors in ability estimation decreased as more stages and/or modules per stage were added, though changes in the number of items per stage seemed to have little impact on the quality of ability estimation.

However, for credentialing examinations, while individual proficiency estimates are important, the primary outcome of consequence is the classification of examinees into pass–fail categories on the basis of such scores. Thus, the second approach taken in studies of MST designs has focused more purposefully on the making of those binary pass–fail decisions using item response theory and different test designs including MST (Luecht, 2000; Luecht & Nungester, 1998). These results have generally been evaluated in terms of decision accuracy (DA) and decision consistency (DC). DA indicates whether a decision made about a examinee (e.g., pass or fail) from a test reflects the truth or is consistent with an external criterion, in that it is computed as a proportion of decisions that are consistent with the true decision classifications or classifications based on a measure that is external to the test itself over all examinees. Similarly, DC reflects the consistency or stability of decisions for individual examinees made over parallel forms. The kappa coefficient is also helpful in this type of research, in that it measures the agreement between the decision based on truth (in simulation studies, truth is known), and on estimated ability, adjusted for agreement that might be expected to be due to chance factors alone.

Xing (2001) found that the three CBT designs (linear parallel forms, MST, and CAT) provided essentially comparable results (as defined by DA, DC, and kappa) in a simulation study investigating the effects of item quality, bank size, and placement of the passing score (based on content considerations in the study). Within each design, enhancing item quality and enlarging the item bank resulted in significant improvements in terms of the criteria of interest for pass–fail decision-making. In a follow-up study by Xing & Hambleton (2004), choice of test design was again found to be far less of a factor in terms of minimizing Type I and Type II classification errors than were bank size and item quality. These authors suggest that when the pass–fail decisions are the primary objective of an examination, the complexity and effort associated with adaptive test designs may not be entirely justified from a resource-allocation perspective: It may be as or more effective for test developers to administer a linear test and instead focus development on mechanisms for improving the item bank. Hambleton & Xing (2006) then explored optimal and nonoptimal designs for linear parallel forms and MST, where optimal and nonoptimal are defined as relative to higher measurement precision in either the region of the cut-score for passing or in the region of the proficiency scale where many of the examinees are located. It was found that the distinction made little practical significance, in that all of the designs investigated provided measurement results that were better than random item selection.

In a recent study by Jodoin, Zenisky and Hambleton (2006), a 60-item, three-stage MST was compared with a 40-item, two-stage test as well as several 60-item, linear-on-the-fly (LOFT) forms and the original, 60-item, operationally-used, linear test forms. While the results from all test designs were by and large comparable with respect to DC and DA, the three-stage MST and the LOFT forms provided results that were only minimally better than the original operational tests. This was in part due to the difficulty encountered by the ATA software in meeting the target information functions for the multistage and LOFT designs due to stringent content constraints. Interestingly, however, the results for the two-stage MST (which, at 40 items, was two-thirds as long as the 60-item, three-stage MST) were only very slightly lower than those observed for the three-stage MST.

18.7 Modules with Common-Stem Items

The second specialized area of MST research is focused on a particular module structure, specifically the case where the items within the module are not conditionally independent of one another (Wainer & Kiely, 1987; Wainer & Lewis, 1990). In that situation, the appropriateness of some IRT models for adapting the modules to examinee ability is directly called into question due to violations of assumptions about local independence and unidimensionality. These assumptions are related because in the case where the local independence assumption is violated, something other than examinee ability is influencing responses.

Such dependence is a problem in the context of IRT-based MST, where the modules are composed of sets of items linked in some way such as a passage or graphic, because research has demonstrated that in such cases reliability of the test composed of such sets of items tends to be overestimated, resulting in overconfidence in the precision of examinee scores (Sireci, Thissen & Wainer, 1991; Zenisky, Hambleton & Sireci, 2002). Lee and Frisbie (1999) also developed an approach to estimating the reliability of such modules using generalizability theory. In addition to problems with estimation of reliability, item sets based on a common stem have also been investigated for the presence of differential testlet functioning (a generalization on studies of differential item functioning; see Wainer Sireci & Thissen, 1991).

In dealing with such testlets with respect to estimating examinee scores, the common approach has involved scoring methods using polytomous IRT models (e.g., Thissen Steinberg & Mooney, 1989). While polytomous models may be useful in that conditional independence between the item sets can be retained, the use of polytomous models also results in a net loss of item information because not all parameters are estimated for each dichotomously scored item within the polytomous item set. For example, with the graded response model of Samejima (1969), a single discrimination parameter for the polytomous item is computed, along with a threshold value for each score point.

Recent research efforts have been directed toward alternative methods for conceptualizing and analyzing modules with items that have dependencies and can still facilitate adaptive testing. This is an important emerging area of research for MST. Work by Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du, (2000) in what has come to be described as testlet response theory has brought about the development of modifications to the two- and three-parameter logistic IRT models, which allow for on-the-fly construction of item sets that appropriately meet constraints including the minimization of local dependence. The model from Bradlow, Wainer, and Wang (1999) includes an extra parameter to represent the interaction effect between an examinee and a given testlet, while the second study is a further generalization of the previous work, but due to added complexity in the three-parameter logistic model, this methodology is more intensive computationally. Further work in this regard has also been done by Vos and Glas (this volume, chap. 20) and Glas, Wainer and Bradlow (2000).

18.8 Conclusions

As the stakes associated with educational and psychological testing results continue to increase, more attention is being paid to issues such as the role of measurement errors and misclassifications. For testing programs, particularly in the area of certification and licensure (where agencies have the dual responsibilities of providing fairness for examinees and protecting the public), obtaining highly precise scores and associated decision accuracy are critical aspects of establishing test score validity. This is particularly the case in CBT applications such as MST and CAT where technologies for administration and test development are changing and being updated with incredible speed. In that regard, the goal of trying to identify the single "best" approach or design structure in MST for practice is not a practically viable one. However, efforts to ascertain general psychometric properties associated with various design variables of an MST can be useful as agencies interested in the use of MST go about the process of designing feasibility studies and assessing the costs and benefits (both measurement and otherwise) for their testing programs associated with instituting a computer-based multistage test.

For professional credentialing assessment, multistage testing can be viewed as an effort to capitalize on the efficiency of CAT and the test form assembly controls of linear testing to maximize the accuracy of the pass–fail decision to be made for each examinee. Through this review of the MST literature, it is particularly clear that the relative benefits of MST are very much dependent on the characteristics, needs, and goals of individual testing programs. Issues such as (but not limited to) the depth and breadth of the item bank, the selection of automated test assembly algorithms, the specific design structure implemented, and the placement of the cut-scores for making the critical pass–fail decisions are just a few of the essential variables that must be deliberated upon during the process of developing such a test.

Among these variables, several have emerged as potentially having a great deal of practical significance on results for test-takers. The choice of design, the amount and distribution of test information, and the test length are all variables with such promise. In addition, routing methodologies are an important and relatively understudied aspect of MST. To date, the focus of MST research has been toward the "front end of development," specifically toward the more structural variables and the test development aspects. Given that MST is not a widely used, operational test design (for an important exception, see Luecht, Brumfield & Breithaupt, 2006; Breithaupt, Ariel & Veldkamp, 2005), attention to this aspect of the approach can be understood as the next logical direction for research attention. Only a relatively few strategies have been tried, including routings based on NC scoring and population distributions, and the literature does not seem to contain many studies that have empirically compared any of the proposed strategies for either accuracy of ability estimation or classification. While the methods used presently seem to work sufficiently, it seems clear that the measurement effectiveness of the design is predicated on the nature and defensibility of the routing decisions, and as such it is only with additional research efforts in this design aspect that high-stakes decisions can be made on the basis of scores from a multistage test.

To this end, it appears that a number of research topics remain for continued study of MST. A multistage test is a highly complex and variable test design, but as noted previously, such variability can be viewed as an advantage in terms of design flexibility. If a multistage test can be built to greater resemble a CAT in terms of measurement precision and accuracy, it may be preferred because the design strikes a balance among adaptability, practicality, measurement accuracy, and control over test forms. As the relational effects between different design variables are delineated, and more dedicated automated test assembly software becomes available, the potential exists for MST to take on an increasingly significant role as a viable alternative for testing agencies involved with the important task of assessment in a variety of measurement contexts.

References

Adema, J. J. (1990). The construction of customized two-stage tests. *Journal of Educational Measurement, 27,* 241–253.

Angoff, W. & Huddleston, E. (1958). *The multi-level experiment: A study of a two-level testing system for the College Board Scholastic Aptitude Test* (Statistical Report No. SR-58-21). Princeton, NJ: Educational Testing Service.

Armstrong, R., Jones, D., Koppel, N. & Pashley, P. (2000, April). *Computerized adaptive testing with multiple forms structures.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Berger, M. P. F. (1994). A general approach to algorithmic design of fixed-form tests, adaptive tests, and testlets. *Applied Psychological Measurement, 18,* 141–153.

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.

Breithaupt, K., Ariel, A. & Veldkamp, B. P. (2005). Automated simultaneous assembly for multistage testing. *International Journal of Testing, 5,* 319–330.

Breithaupt, K. & Hare, D. R. (2007), Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67,* 5–20.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Dodd, B. G. & Fitzpatrick, S. J. (2002). Alternatives for scoring CBTs. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 215–236). Mahwah, NJ: Lawrence Erlbaum Associates.

Folk, V. G. & Smith, R. L. (2002). Models for delivery of computer-based tests. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 41–66). Mahwah, NJ: Lawrence Erlbaum Associates.

Glas, C. A. W., Wainer, H. & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* Boston: Kluwer-Nijhof Publishing.

Green, B. F., Jr., Bock, R. D., Humphreys, L. G., Linn, R. B. & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360.

Hadadi, A., Luecht, R. M., Swanson, D. B. & Case, S. M. (1998, April). *Study 1: Effects of modular subtest structure and item review on examinee performance, perceptions and pacing.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego.

Hambleton, R. K. & Xing, D. (2006). Optimal and non-optimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education, 19,* 221–239.

Jodoin, M. G., Zenisky, A. L. & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams. *Applied Measurement in Education, 2006, 19,* 203–220.

Kim, H. (1993). *Monte Carlo simulation comparison of two-stage testing and computer adaptive testing.* Unpublished doctoral dissertation, University of Nebraska, Lincoln.

Kim, H. & Plake, B. (1993, April). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing.* Paper presented at the meeting of the National Council on Measurement in Education, Atlanta.

Lee, G. & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12,* 237–255.

Linn, R., Rock, D. & Cleary, T. (1969). The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement, 29,* 129–146.

Lord, F. M. (1971). A theoretical study of two-stage testing. *Psychometrika, 36,* 227–242.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 227–238.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luecht, R. M. (1997, March). *An adaptive sequential paradigm for managing multidimensional content.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22,* 224–236.

Luecht, R. M. (2000, April). *Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M. (2003, April). *Exposure control using adaptive multistage item bundles.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Luecht, R. M. (2006). Designing tests for pass-fail decisions using item response theory. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 575–596). Mahwah, NJ: Lawrence Erlbaum Associates.

Luecht, R. M., Brumfield, T. & Breithaupt, K. (2006). A testlet-assembly design for adaptive multistage tests. *Applied Measurement in Education, 19,* 189–202.

Luecht, R. M. & Burgin, W. (2003, April). *Test information targeting strategies for adaptive multistage testing designs.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago.

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35(3),* 229–249.

Luecht, R. M., Nungester, R. J. & Hadadi, A. (1996, April). *Heuristic-based CAT: Balancing item information, content and exposure.* Paper presented at the meeting of the National Council on Measurement in Education, New York.

Mead, A. (2006). An introduction to multistage testing [Special Issue]. *Applied Measurement in Education, 19,* 185–260.

Mills, C. N. & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287–304.

Patsula, L. N. (1999). *A comparison of computerized-adaptive testing and multi-stage testing.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Reese, L. M. & Schnipke, D. L. (1999). *An evaluation of a two-stage testlet design for computerized testing* (Computerized Testing Report 96-04). Newtown, PA: Law School Admission Council.

Reese, L. M., Schnipke, D. L. & Luebke, S. W. (1999). *Incorporating content constraints into a multi-stage adaptive testlet design.* (Law School Admissions Council Computerized Testing Report 97-02). Newtown, PA: Law School Admission Council.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Schnipke, D. L. & Reese, L. M. (1999). *A comparison of testlet-based test designs for computerized adaptive testing* (Law School Admissions Council Computerized Testing Report 97-01). Newtown, PA: Law School Admission Council.

Sireci, S. G., Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Thissen, D. (1998, April). *Scaled scores for CATs based on linear combinations of testlet scores.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego.

Thissen, D., Steinberg, L. & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247–260.

van der Linden, W. J. (2000). Optimal assembly of tests with item sets. *Applied Psychological Measurement, 24,* 225–240.

van der Linden, W. J. (2005). *Models for optimal test design.* New York: Springer-Verlag.

van der Linden, W. J. & Adema, J. J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35,* 185–198.

van der Linden, W. J., Ariel, A. & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31,* 81–99.

Vos, H. J. (2000, April). *Adaptive mastery testing using a multidimensional IRT model and Bayesian sequential decision theory.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Vos, H. J. & Glas, C. A. W. (2001, April). *Multidimensional IRT based adaptive sequential mastery testing.* Paper presented at the meeting of the National Council in Measurement in Education, Seattle.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12(1),* 15–20.

Wainer, H., Bradlow, E. T. & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Boston: Kluwer-Nijhof Publishing.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–201.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*, 1–14.

Wainer, H., Sireci, S. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197–219.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wise, S. L. & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21,* 135–155.

Xing, D. (2001). *Impact of several computer-based testing variables on the psychometric properties of credentialing examinations.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Xing, D. & Hambleton, R. K. (2004). Impact of test design, item quality, and item bank size on the psychometric properties of computer-based credentialing exams. *Educational and Psychological Measurement, 64,* 5–21.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Zenisky, A. L., Hambleton, R. K. & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39,* 1–16.

Chapter 19
Three-Category Adaptive Classification Testing

Theo J.H.M. Eggen

19.1 Introduction

Educational and psychological testing can have many practical purposes. From the perspective of test users, a distinction among selection, placement, certification, licensuring, monitoring of progress in proficiency, and diagnostic testing can be made. From the measurement point of view, it generally suffices to distinguish between estimation and classification.

In estimation, the goal is to get an estimate of the ability or proficiency of a person on a well-defined domain on a one-dimensional scale. Traditionally, computerized adaptive tests (CATs) (Wainer, 2000) are designed to achieve this goal as quickly and/or as precisely as possible. In classification, the goal is to determine to which of a limited number of competency or proficiency categories a person belongs. In this case, one or more cutting points are set on the proficiency scale to define the number of categories. It is not the precise estimate of the proficiency of the person that is important, but the correct classification in a category.

Testing with classification as the main goal already has a long tradition. In personnel measurement, Cronbach and Gleser (1965) already expressed the opinion that the main purpose of testing was to make qualitative classification decisions. In educational measurement, classification tests were mainly developed as a form of criterion-referenced-measurement (Hambleton, Swaminathan, Algina & Coulson, 1978), in which on the basis of a score on a test a decision is made whether a certain criterion or standard is met. Initially, several approaches were developed for making decisions in one of two categories. Examples are the decision of mastery or nonmastery of a domain of interest, the pass-fail decision on an exam, or the decision to certify someone or not. There are, however, many classification problems in which a decision in one of more than two categories is desired. For example, in state assessments in the USA it is common to report that the performance of students is at a basic, proficient, or advanced level. In mastery testing sometimes masters, partial masters and nonmaster are distinguished (Vos, 1999). In placement testing,

T.J.H.M. Eggen (✉)
Cito Institute for Educational Measurement, P.O. Box 1034, 6801 MG Arnhem, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_19,

the result on the test can contribute to the decision on the placement of a student in a course that is offered in three or more levels (Eggen and Straetmans, 2000). And as a final example, the everyday practice of assigning grades to students is mentioned (Weiss & Kingsbury, 1984).

In this chapter, the main emphasis will be on decisions in one of three categories, which is the closest extension from the standard two-category decision problem. These decisions are made in the context of a computerized adaptive test that uses an item bank that is calibrated with an item response theory (IRT) model. Although CAT algorithms designed for the efficient estimation of proficiency have also been shown to be useful for classification, several researchers have shown that specific algorithms designed for classification do perform better. Because these algorithms are aimed at making classification decisions, tests with these algorithms are sometimes called *computerized classification tests* (Parshall, Spray, Kalohn & Davey, 2002).

In this chapter, the approach based on the application of the sequential probability ratio test (SPRT; Wald, 1947) will be presented. This approach does not need estimates of a person's proficiency during testing since the decisions are taken on the basis of conducting (combinations of) statistical tests. Special attention will be paid to the selection of items based on the Kullback–Leibner information (Cover & Thomas, 1991), which conceptually has a strong relationship to statistical testing. First, a short review of the different approaches in classification testing will be given. Although most approaches have been developed for two-category classification problems, the generalization to more categories is almost always feasible.

19.2 Overview of Approaches to Classification Testing

In classification testing, minimizing of the number of incorrect decisions on the basis of a test is a major goal. Initially, psychometric theory for classification testing was developed for linear, nonadaptive tests with a fixed length. The psychometric theory has the following two basic elements. Firstly, a psychometric model relating the probability of a correct response of a person to his or her unknown true proficiency. Secondly, a specification of a loss structure evaluating the costs and benefits for each possible combination of decision outcome and true level of proficiency.

At first, only a simple binomial model specifying that given the true level of proficiency, the (same) probability of answering correctly for all items was used as the psychometric model. Later, item response models were used. In the approach, optimal decision rules are developed, specifying the classification decision to be taken for each possible observed outcome of the test. The optimal rules, minimizing expected losses, are obtained by either Bayesian or minimax decision theory (DeGroot, 1970).

If the length and the content of the test are not fixed, the goal is to maximize the probability of making correct classification decisions together with the minimization of the length of the test. Two main approaches can be distinguished here: adaptive classification testing, and sequential classification testing.

In adaptive classification testing, the selection of items and the stopping rule are adapted to the observed results during testing. In sequential classification testing, however, only the stopping rule is adaptive and the items are selected at random from an available item bank. In sequential classification testing, optimal decision rules are given for the classification decision to be taken and the number of items to be administered. These rules are derived using Bayesian sequential decision theory (e.g., Smith & Lewis, 1995) or minimax sequential decision theory (e.g., Vos, 2002).

The approach presented here has its roots in sequential classification testing. One of the first applications of sequential classification testing, dating back to Ferguson (1969), uses Wald's sequential probability ratio test (SPRT), which will be described later in detail. Ferguson (1969) used the binomial test model, with all items, given the true level of proficiency, of equal difficulty. Reckase (1983) was the first to apply a modification of the SPRT, allowing for different probabilities of correct answers to items, which used random selection of items from an IRT calibrated item bank.

In adaptive classification testing, the random selection of items is replaced by the selection of items adapted to the performance of the person during testing. In item selection, either Bayesian item selection criteria (van der Linden, 1998) or maximum information criteria (Kingsbury & Weiss, 1983) are used. Using one of these criteria for item selection during testing, estimates of the proficiency of the person are made on the basis of the responses to the items and confidence intervals of the person's proficiency are constructed for taking a decision or continuing testing.

In the approach that will be treated extensively in this chapter, the selection of items is based on maximum information criteria, but the decisions are not made on the basis of a statistical estimation procedure, that is, confidence intervals of the person's proficiency, but on the basis of an application of the SPRT statistical testing procedure.

19.3 Basic Elements of Adaptive Testing

Computerized adaptive tests assume the availability of an item bank that is calibrated with an item response model. Confining ourselves to item banks with items that are dichotomously scored, logistic item response models are commonly used. In these models a specification is given of the relationship between the proficiency of a person, θ, and the probability of a score on an item. The score on item i, $X_i = x_i$, is either correct $x_i = 1$ or wrong $x_i = 0$, and the probability of a correct answer is given by

$$p_i(\theta) = \mathrm{P}(X_i = 1 \mid \theta) = c_i + (1 - c_i)\frac{\exp(a_i(\theta - b_i))}{1 + \exp(a_i(\theta - b_i))}. \tag{19.1}$$

In (19.1) the three-parameter logistic model (3PL) is given, with a_i, b_i, c_i respectively being the discrimination, the difficulty, and the guessing parameter. In operational CATs, very often the simpler two-parameter logistic model (2PL), where for all i, $c_i = o$ in (19.1), or the one-parameter logistic model (1PL), where for all i, $c_i = o$ and $a_i = a$ in (19.1), is used. In CATs, the parameters of the IRT model are always assumed to be estimated with such a precision that they can be considered to be known. In the inference of a person, the likelihood function of the person's proficiency, θ, plays a central role. Given the scores on k items $x_i, i = 1, \ldots, k$, and the parameters of the items, this function is

$$L(\theta; \underline{x}_k) = L(\theta; x_1, \ldots, x_k) = \prod_{i=1}^{k} L(\theta; x_i) = \prod_{i=1}^{k} p_i(\theta)^{x_i}(1 - p_i(\theta))^{1-x_i}. \tag{19.2}$$

In CATs where the main aim is the efficient estimation of the proficiency of a person, this likelihood function (19.2) is the basis for estimating the proficiency and also for the selection of items. The maximum likelihood (ML) estimate of the proficiency after administering k items follows from an iterative maximization procedure:

$$\widehat{\theta}_k = \max_{\theta} L(\theta; \underline{x}_k). \tag{19.3}$$

A good alternative for proficiency estimation is the weighted maximum likelihood (WML) method proposed by Warm (1989). WML estimates are less biased than ML estimates, especially in CATs where operating on an item bank of a moderate size is an important issue. WML estimates follow from

$$\widehat{\theta} = \max(w(\theta).L(\theta; \underline{x}_k)). \tag{19.4}$$

In the 2PL model as well as in the 1PL model, the weight function (19.4) takes a simple form:

$$w(\theta) = \left[\sum I_i(\theta)\right]^{1/2}. \tag{19.5}$$

In (19.5) $I_i(\theta)$ is the Fisher information function of item i, which is defined as

$$I_i(\theta) = E\left(\frac{\frac{\partial}{\partial\theta} L(\theta; x_i)}{L(\theta; x_i)}\right)^2. \tag{19.6}$$

The information function is commonly used for item selection in CATs where estimation is the main goal of testing: an item is selected if it gives maximum information at the current proficiency estimate. This proceeds as follows: if the current proficiency estimate after administering k items is $\widehat{\theta}_k$, then the next item to be selected from the item bank is the item j for which $\max_j I_j(\widehat{\theta}_k)$. In a classification problem with two categories, Spray and Reckase (1994) report that it is more efficient to select the items that have maximum Fisher information at the cutting point θ_0 rather than at the current proficiency estimate.

For a test consisting of k items, the test information function is the sum of the information of the items in the test $I(\theta) = \sum_{i=1}^{k} I_i(\theta)$. Selecting items with maximum information maximizes the contribution to the test information. The usefulness of this is readily understood if an estimate of the proficiency of a person is wanted, especially when the maximum likelihood estimator (MLE) (19.3) is used. In this case, the standard error of the MLE (19.3) is estimated by $\mathrm{SE}(\widehat{\theta}_k) = 1/\sqrt{I(\widehat{\theta}_k)}$. Therefore, by selecting items having maximum information, the contribution to decreasing of the standard error is greatest. Furthermore, from the definition in (19.6) it can be seen that maximizing the information is the same as maximizing the contribution of an item to the expected relative rate of change of the likelihood function. As Chang and Ying (1996) have pointed out, the greater this change rate at a given value of θ, the better this value can be distinguished from points near this value and the better this value can be estimated.

19.4 The SPRT in CAT

The likelihood function of a person's proficiency θ (19.2) is also the main source of inference if the sequential probability ratio test is used in adaptive testing. In this case, the likelihood function is used differently in the testing algorithm than in the case of statistical estimation. This will become clear in the following description of the statistical testing procedure. From the application of the SPRT it is inferred whether another item is to be administered and which decision is made when testing is stopped. The use of the SPRT first will be described for a decision in one of two categories, and thereafter generalized to a decision in one of three categories.

19.4.1 Classification in Two Categories

On the latent proficiency scale, a decision or cutting point θ_c between, for example, a person denoted as a master or a nonmaster, or between an examinee passing and an examinee failing an exam, is given. A small region on both sides of this point, a so-called indifference zone, is selected. The widths of these regions are δ_1 and δ_2. The indifference interval expresses the fact that, due to measurement errors, making the right decision about persons very near the cutting point can never be guaranteed. One could also say that the interval expresses the indifference of a decision maker of the classification of the persons who are that close to the cutting point. With sequential testing, the classification problem is formulated in terms of statistical hypotheses:

$$\mathrm{H0}\colon \theta \leqslant \theta_c - \delta_1 = \theta_1 \text{ against } \mathrm{H1}: \theta \geqslant \theta_c + \delta_2 = \theta_2. \tag{19.7}$$

Next, acceptable decision error probabilities are specified as follows:

$$\mathrm{P}(\text{accept H0} |\text{ H0 is true}) \geqslant 1 - \alpha, \tag{19.8}$$

$$\mathrm{P}(\text{accept H0} |\text{ H1 is true}) \leqslant \beta. \tag{19.9}$$

in which α and β are small constants. The test meeting these decision error rates can be carried out using the SPRT (Wald, 1947). The test statistic being used is the ratio between the values of the likelihood function (19.2) at the point on the proficiency scale of the alternative hypothesis and at the point of the null hypothesis:

$$LR(\theta_2,\theta_1;\underline{x}_k)=\frac{L(\theta_2;\underline{x}_k)}{L(\theta_1;\underline{x}_k)}=\prod_{i=1}^{k}\frac{L(\theta_2;x_i)}{L(\theta_1;x_i)}=\prod_{i=1}^{k}\frac{p_i(\theta_2)^{x_i}(1-p_i(\theta_2))^{1-x_i}}{p_i(\theta_1)^{x_i}(1-p_i(\theta_1))^{1-x_i}}. \tag{19.10}$$

It will be clear that high values of this ratio indicate that the person is more likely to have a proficiency above the cutting point, and small values support the decision that the person's proficiency is below the cutting point. That is, the test meets the error rates if the following procedure is used:

Continue sampling if

$$\frac{\beta}{1-\alpha}<LR(\theta_2,\theta_1;\underline{x}_k)<\frac{1-\beta}{\alpha}, \tag{19.11}$$

Accept H0 if

$$LR(\theta_2,\theta_1;\underline{x}_k)\leqslant\frac{\beta}{1-\alpha}, \tag{19.12}$$

Reject H0 if

$$LR(\theta_2,\theta_1;\underline{x}_k)\geqslant\frac{1-\beta}{\alpha}. \tag{19.13}$$

The expression in (19.11) is called the *critical inequality* of the test. Although Wald (1947) showed that eventually a decision will be made with probability 1 with the SPRT, in practice the test is truncated usually by specifying a maximum test length, $k_{\max}$. At this test length, a forced decision, the most obvious decision, is taken: H0 is rejected if the test statistic is larger than the midpoint of the critical inequality interval; otherwise, it is accepted.

19.4.2 Classification in Three Categories

The above testing procedure is readily generalized to cases of classification in one of three categories. Classification in one of three categories involves two cutting points on the latent scale, θ_{c1} and θ_{c2}, by which three different levels of proficiency are distinguished. An indifference zone is identified around each cutting point. Schematically, the situation can be presented as follows.

Two pairs of hypotheses are formulated:

$$\text{H0}_1:\theta\leqslant\theta_{c1}-\delta_{11}=\theta_{11}\text{ against H1}_1:\theta\geqslant\theta_{c1}+\delta_{21}=\theta_{12} \tag{19.14}$$

$$\text{H0}_2:\theta\leqslant\theta_{c2}-\delta_{12}=\theta_{21}\text{ against H1}_2:\theta\geqslant\theta_{c2}+\delta_{22}=\theta_{22} \tag{19.15}$$

The SPRT test described in (19.11), (19.12), and (19.13) is applied for each pair of hypotheses. In the specification of the acceptable decision errors, as in (19.8), the small constants α_1 and β_1, α_2 and β_2, respectively, are used.

$$\text{P(accept H0}_1 \mid \text{H0}_1 \text{ is true)} \geqslant 1-\alpha_1 \text{ and} \tag{19.16}$$

$$\text{P(accept H0}_1 \mid \text{H1}_1 \text{ is true)} \leqslant \beta_1, \tag{19.17}$$

$$\text{P(accept H0}_2 \mid \text{H0}_2 \text{ is true)} \geqslant 1-\alpha_2 \text{ and} \tag{19.18}$$

$$\text{P(accept H0}_2 \mid \text{H1}_2 \text{ is true)} \leqslant \beta_2. \tag{19.19}$$

Next the two SPRTs are combined into one procedure. Hereafter, the decisions to assign a person to a certain category, based on a combination of two SPRTs, is given.

	decision test 1	
decision test 2	H0_1: $\theta \leqslant \theta_{11}$	H1_1: $\theta \geqslant \theta_{12}$
H0_2: $\theta \leqslant \theta_{21}$	1	2
H1_2: $\theta \geqslant \theta_{22}$		3

So when, for instance, the first null hypothesis (19.14) is rejected, which is equivalent to deciding category 2 or 3, and the second null hypothesis (19.15) is accepted, which is equivalent to deciding category 1 or 2, in the combined procedure the decision is category 2. In Eggen (1999) it is shown that when the 2PL model is used, the simultaneous occurrence of accepting the null hypothesis in (19.14), category 1, and rejecting the null hypothesis in (19.15), category 3, cannot occur. Although this cannot be proven, when using the 3PL model, in practice this will not cause any problems because it is very unlikely to occur. And if it occurs, no final decision will be made and another item will be administered.

The combined procedure operates as follows:

Make decision 1 if

$$LR(\theta_{12},\theta_{11};\underline{x}_k) \leqslant \frac{\beta_1}{1-\alpha_1} \quad \text{and} \quad LR(\theta_{22},\theta_{21};\underline{x}_k) \leqslant \frac{\beta_2}{1-\alpha_2}, \tag{19.20}$$

Make decision 2 if

$$LR(\theta_{12},\theta_{11};\underline{x}_k) \geqslant \frac{1-\beta_1}{\alpha_1} \quad \text{and} \quad LR(\theta_{22},\theta_{21};\underline{x}_k) \leqslant \frac{\beta_2}{1-\alpha_2}, \tag{19.21}$$

Make decision 3 if

$$LR(\theta_{12},\theta_{11};\underline{x}_k) \geqslant \frac{1-\beta_1}{\alpha_1} \quad \text{and} \quad LR(\theta_{22},\theta_{21};\underline{x}_k) \geqslant \frac{1-\beta_2}{\alpha_2}, \tag{19.22}$$

Continue testing

$$\text{in all other cases.} \tag{19.23}$$

19.4.3 Evaluation of the Test Statistics

In practice, the test-critical inequalities, e.g., (19.11), in the SPRT with an IRT model are evaluated as follows. First, the logarithms are taken. Then we get

$$\ln\left(\frac{\beta}{1-\alpha}\right) < \ln(LR(\theta_2, \theta_1; \underline{x}_k)) < \ln\left(\frac{1-\beta}{\alpha}\right). \tag{19.24}$$

In (19.24) the logarithm of the likelihood ratio is equal to

$$\sum_{i=1}^{k} x_i \ln\left(\frac{p_i(\theta_2)(1-p_i(\theta_1))}{p_i(\theta_1)(1-p_i(\theta_2))}\right) + \sum_{i=1}^{k} \ln\left(\frac{(1-p_i(\theta_2))}{(1-p_i(\theta_1))}\right). \tag{19.25}$$

In (19.25) the second term is evaluated straightforwardly because it consists only of simple operations on known constants. The first term consists of the score obtained on the items weighted by the log odds of the events specified in hypotheses H1 and H0 in (19.7). It is easily shown that in the 3PL model the log odds can be written as

$$\ln\left(\frac{c_i + \exp(a_i(\theta_2 - b_i))}{c_i + \exp(a_i(\theta_1 - b_i))}\right), \tag{19.26}$$

which shows that the test statistics can be evaluated quite straightforwardly.

It is worthwhile noting that in case the 2PL model is used, and also in the case of the 1PL model, the log odds in (19.26) are equal to

$$a_i(\theta_2 - b_i) - a_i(\theta_1 - b_i) = a_i(\theta_2 - \theta_1) = a(\delta_2 + \delta_1), \tag{19.27}$$

which means that the critical inequality of the statistical test problem (19.7) can be written as follows:

$$\frac{\ln\left(\frac{\beta}{1-\alpha}\right) - \sum_{i=1}^{k} \ln\left[\frac{1-p_i(\theta_2)}{1-p_i(\theta_1)}\right]}{\delta_2 + \delta_1} < \sum_{i=1}^{k} a_i x_i < \frac{\ln\left(\frac{1-\beta}{\alpha}\right) - \sum_{i=1}^{k} \ln\left[\frac{1-p_i(\theta_2)}{1-p_i(\theta_1)}\right]}{\delta_2 + \delta_1}. \tag{19.28}$$

In (19.28) testing involves only the observed weighted score and known constants, which can easily be implemented. From this it is also easily seen that if the width of the indifference interval, $\theta_2 - \theta_1 = \delta_2 + \delta_1$, increases, the width of the critical interval gets smaller, which indicates that shorter tests can be used to make a decision.

19.5 Item Selection in CAT with the SPRT

An important part of a CAT algorithm is the item selection procedure, which determines during testing the choice of the items being administered. The item selection procedure is a major part of any CAT: the gain in efficiency in testing and the adaptation to the performance of examinees are established in this part of the algorithm. In adaptive testing using statistical testing in the algorithm, four main approaches for item selection can be distinguished:

1. Random item selection, which is traditionally mostly used in combination with the SPRT.
2. Selecting on Fisher information.
3. Bayesian item selection criteria, which are also used in adaptive testing with estimation. These criteria, discussed by van der Linden (1998), will not be considered here.
4. Selecting on Kullback–Leibler (K-L) information.

Next, item selection procedures will be described that are based on Kullback–Leibler information. It will be shown that the item's Kullback–Leibler information expresses the expected contribution of an item to the discriminatory power between two hypotheses. Conceptually, K-L information fits the statistical testing algorithm more closely than Fisher information.

19.5.1 Kullback-Leibler Information

The use of the relative entropy or K-L information (Cover & Thomas, 1991) for selecting items in an SPRT CAT was introduced by Eggen (1999). This information concept will be described next.

K-L information is a measure of the distance between two distributions:

$$K(f_2 \parallel f_1) = E_{f_2} \ln\left(\frac{f_2(x)}{f_1(x)}\right), \tag{19.29}$$

which is the expected information in an observation of X for discriminating between two hypotheses: H0: $f(x) = f_1(x)$ against H1: $f(x) = f_2(x)$. The larger this information, the more efficient the statistical test will be.

The definition in (19.29) can be directly applied to the SPRT application in adaptive testing: H0 is the hypothesis for which we have a distribution (likelihood) with parameter value $\theta = \theta_1$, and under H1 the distribution has parameter $\theta = \theta_2$. Then the K-L information is

$$K(\theta_2 \parallel \theta_1) = E_{\theta_2} \ln\left(\frac{L(\theta_2; \underline{x}_k)}{L(\theta_1; \underline{x}_k)}\right) = E_{\theta_2} \ln\left(\prod_{i=1}^{k} \frac{L(\theta_2; x_i)}{L(\theta_1; x_i)}\right) \tag{19.30}$$
$$= \sum_{i=1}^{k} E_{\theta_2} \ln\left(\frac{L(\theta_2; x_i)}{L(\theta_1; x_i)}\right) = \sum_{i=1}^{k} K_i(\theta_2 \parallel \theta_1).$$

It is seen that the K-L test information (k items) can be written as the sum of the K-L information of the items. The K-L item information, $K_i(\theta_2 \parallel \theta_1)$, is defined for any pair $\theta_2 > \theta_1$ and is a positive real number and, consequently, an eligible item information index. Applying an item selection procedure based on having maximum K-L information is useful, since this procedure will maximize the contribution to the K-L test information. When the K-L test information is maximized, the expected difference between the log-likelihoods under both hypotheses is maximized. This is the same as making the likelihood ratio more extreme, which is, in turn, expected to minimize the number of items needed to make a decision because the test statistic is the likelihood ratio (see, e.g., (19.11)).

If an IRT model for dichotomously scored items is used, the K-L item information index (19.30) can be written as:

$$K_i(\theta_2 \parallel \theta_1) = p_i(\theta_2) \ln\left(\frac{p_i(\theta_2)}{p_i(\theta_1)}\right) + (1 - p_i(\theta_2)) \ln\left(\frac{1 - p_i(\theta_2)}{1 - p_i(\theta_1)}\right). \tag{19.31}$$

In the case of the 3PL IRT model, this becomes

$$K_i(\theta_2 \parallel \theta_1) = p_i(\theta_2). \ln\left(\frac{c_i + \exp(a_i(\theta_2 - b_i))}{c_i + \exp(a_i(\theta_1 - b_i))}\right) + \ln\left(\frac{(1 - p_i(\theta_2))}{(1 - p_i(\theta_1))}\right) \tag{19.32}$$

and in the 2PL model the even simpler expression

$$K_i(\theta_2 \parallel \theta_1) = p_i(\theta_2) a_i(\theta_2 - \theta_1) + \ln\left(\frac{(1 - p_i(\theta_2))}{(1 - p_i(\theta_1))}\right). \tag{19.33}$$

Note that the expressions in (19.32) and (19.33) are the same as in the corresponding test statistic or likelihood ratios in (19.25), – (19.27). By this it is easily understood that selecting items with maximum K-L information maximizes the discrepancy between the likelihood under H0 and H1.

19.5.2 K-L Information in the Three-Category Problem

In the case of a classification problem in two categories, the K-L item information can be used directly in a straightforward way for item selection. The K-L item information will be computed in two points of the hypotheses tested: that item will be selected for which $\max_i K_i(\theta_2 \parallel \theta_1)$. In the three-way classification for K-L item selection, there are more choices; two of which are generally applicable.

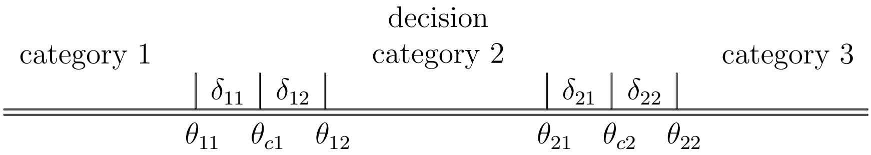


Fig. 19.1 Schematic representation of the classification problem with three categories

The first is a naive, but simple method: an item is selected that maximizes the K-L information at two fixed points. In the case of three-category problem, possible choices are (see Figure 19.1): $K_i(\theta_{21} \parallel \theta_{12}), K_i(\theta_{c2} \parallel \theta_{c1}), K_i(\theta_{22} \parallel \theta_{11})$, or the K-L information computed in any two other points of the proficiency scale. In Eggen (1999) it is shown that the specific choice of the two points hardly influences the performance of the test. Obviously, varying the specific points in the computation of the K-L item information does not have a large impact on the ordering of the items on this K-L information. We will therefore consider only the selection method that uses the K-L item information in the two cutting points:

$$K_i(\theta_{c2} \parallel \theta_{c1}). \tag{19.34}$$

A better performing, alternative approach to this simple method was also proposed. This alternative approach looks more precisely at the progress of hypothesis testing: as long as none of the pairs of hypotheses has led to a decision, items are chosen with maximum K-L information between the two cutting points θ_{c1} and θ_{c2}, but if one of the pairs of hypotheses has led to a decision while the other has not, items will be chosen that have maximum K-L information around the cutting point corresponding to the test that has not yet led to a decision. This means an item i is selected for which

$$\text{if } LR(\theta_{12}, \theta_{11}; \underline{x}_k) \geqslant \frac{1-\beta_1}{\alpha_1} \text{ then } \max_i K_i(\theta_{22} \parallel \theta_{21}), \tag{19.35}$$

$$\text{if } LR(\theta_{22}, \theta_{21}; \underline{x}_k) \leqslant \frac{\beta_2}{1-\alpha_2} \text{ then } \max_i K_i(\theta_{12} \parallel \theta_{11}), \tag{19.36}$$

$$\text{else : } \max_i K_i(\theta_{c2} \parallel \theta_{c1}). \tag{19.37}$$

In the next section, results of the performance of this selection method will be presented.

19.6 Performance of Three-Category Classification CAT with the SPRT

Research (e.g., Spray and Reckase 1996; Eggen and Straetmans, 2000) has shown that the performance of adaptive classification tests with the SPRT is generally very good. In particular, in two-category problems, the performance level is in general at least the same as or better than other (estimation based) adaptive testing procedures. With respect to the comparison of the item selection procedure in two-category problems, Eggen's (1999) conclusion is that there are almost no differences between the item selection methods based on an information criterion when used in combination with the SPRT. He compared the selection of items with maximum Kullback–Leibler information, items with maximum Fisher information at the current proficiency estimate, and items with maximum Fisher information at the cutting point. His finding was that there are hardly any differences between the three selection methods, although there seems to be a slight tendency that the selection of items with maximum Fisher information at the cutting point is better than the other methods with respect to the average number of items needed to make a decision. This result is in line with earlier findings of Spray and Reckase (1996).

In a three-category problem, the proper choice of the available item selection strategy is more important. Next, the performance will be illustrated with the results of a simulation study of a CAT operating on a mathematics item bank. In this example, the following item selection methods were compared:

R: Random item selection;

F1: Maximum Fisher information at the current proficiency estimate;

F2: Maximum Fisher information at the cutting point nearest to the current proficiency estimate; this method is comparable to the well performing selection method with only one cutting point;

K1: The simple maximum K-L information selection (19.34);

K2: The advanced maximum K-L information selection (19.35)–(19.37).

19.6.1 Simulation Example

The performances of the item selection procedures were evaluated and compared to each other by means of simulation studies. For the simulation studies, an operational item bank was used. This item bank contains 250 dichotomously scored mathematics items that are used in adult education to place students in one of three course levels and to measure the progress at these levels. The items were shown to fit the 2PL model. The scale was fixed by restrictions on the item parameters. The mean item difficulty is 0, and the mean item discrimination is 3.09. On this scale, the distribution of the proficiency in the population was estimated to be normal, with a mean of 0.294 and a standard deviation of 0.522.

For the classification problem in three categories, the cutting points were $\theta_{c1} = -0.13$, and $\theta_{c2} = 0.33$. The SPRT adaptive testing procedures were conducted for three different sets of error rates: $\alpha_1 = \beta_2 = 2\beta_1 = 2\alpha_2$ were respectively 0.05, 0.075, and 0.1. Halving β_1 and α_2 compared to α_1 and β_2 has the effect that it is expected that all three decisions will have the same error rate. The widths of the indifference zones are all equal to $\delta_{11} = \delta_{12} = \delta_{21} = \delta_{22} = 0.1$, and the maximum test length was $k_{\max} = 25$.

The simulations were conducted as follows. A proficiency of a simulee was randomly drawn from $N(0.294, 0.522)$. Three relatively easy items were selected first and subsequently items were selected using one of the investigated item selection methods. The simulee's response to an item was generated according to the 2PL IRT model and this procedure was repeated for 5,000 simulees.

19.6.2 Results

For the three decision error rates, the item selection procedures were compared on the mean number of items required (k) to make a decision and the classification accuracy, that is, the percentages of correct decisions ($\%$). The results are presented in Table 19.1.

It can be seen that for every selection method, there is an expected decrease in the mean number of required items if the acceptable error rates are increased. The differences hardly vary if the selection methods are compared. Doubling the acceptable error rate gives a decrease of, on average, about 2 items. Increasing the error rates has little effect on the percentages of correct decisions. A comparison of the selection methods shows that the differences between the methods are consistent over the different error rates.

The most eye-catching result is the bad performance of the random item selection method compared to any other selection method. Any psychometric information measure in the item selection in CAT with the SPRT gives a significant increase in the number of correct decisions and a decrease in the required number of items. This result could be expected and is in line with known results in two-category decision problems. On the other hand, the comparable method in the three-category problem (F2) of the best-performing method in the two-category problem is clearly almost the worst-performing selection method. This finding, confirming that of Eggen and Straetmans (2000), may be explained by the fact that the current estimate of the proficiency, especially in the beginning of the test, is so inexact that it is sometimes nearer to the wrong cutting point than the cutting nearest to the true value of the proficiency of the person. The two best-performing selection methods are the selection on Fisher information at the current proficiency estimate (F1) and the selection with the advanced K-L information (K2). The results of both methods do not show

Table 19.1 Mean number of items required (k) for a decision and percentage of correct decisions

	Error Rates					
	0.05		0.075		0.1	
Selection	k	$\%$	k	$\%$	k	$\%$
Random	23.1	83.1	22.1	83.0	21.7	82.5
F1	16.7	89.9	15.6	89.2	14.6	89.1
F2	21.8	87.0	20.5	87.7	19.4	87.4
K1	18.4	88.4	17.0	88.0	16.3	88.6
K2	17.0	89.2	15.6	89.2	14.2	89.4

any significant differences, and both are better than the simple K-L information selection method (K1). On average, the K1 method needs an average of 1.5 items more to reach the same percentage of correct decisions.

19.7 Concluding Remarks

In this chapter, it was demonstrated that the application of the SPRT to adaptive classification testing in with three-category problems is a straightforward generalization of the application in two-category problems. Good performance of the procedure was shown when the item selection was based on the Fisher information at the current proficiency estimate (F1) or on the K-L information (K2). Item selection based on the K-L information has conceptually as strong a relationship to statistical testing, as Fisher information has to statistical estimation. Both methods perform about equally well, but the K-L information item selection is computationally much easier, because during testing an estimate of the current proficiency is not needed.

In modern CAT applications, practical considerations play an important role. By putting constraints on the item selection, demands with respect to the content of the test and with respect to the exposure to items are realized. Eggen (1999) showed that by implementing a content control constraint in an SPRT CAT, selecting on Fisher information or on K-L information has in both methods a small negative effect on the performance. It would be worthwhile to study the effects of applying exposure control in the selection methods. It is to be expected that controlling for overexposure of items has a greater negative impact on the performance with selection based on K-L information than with selection based on Fisher information. This is easily seen in the two-category classification problem, because then the preference of the items based on K-L information is the same for all persons tested. So measures against overexposure will directly affect the performance of the testing procedure. In the three-category classification problem, in selecting on K-L information, we do not have one fixed preference for all persons tested, but also here only a limited number of preferences in the items will appear in selection on the basis of K-L information.

In this chapter, the application of the SPRT in adaptive classification testing was limited to the situation in which the items were dichotomously scored, and calibrated with a logistic test model. The generalization to another test model, to polytomously scored items and to more than 3 classification categories is expected to be possible without major problems.

References

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213–229.

Cover, T. M. & Thomas, J. A. (1991). *Elements of information theory.* New York: Wiley.

Cronbach, L. J. & Gleser, G. C. (1965). *Psychological tests and personnel decisions.* (2nd ed.). Urbana, IL: University of Illinois Press.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

Eggen, T. J. H. M. (1999). Item selection with adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249–261.

Eggen, T. J. H. M. & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 66,* 713–734.

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1–47.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 257–286). New York: Academic Press.

Parshall, C. G., Spray, J. A., Kalohn, J. C. & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In: D. J. Weiss (Ed.), *New horizons in testing* (pp. 237–255). New York: Academic Press.

Smith, R. L. & Lewis, C. (1995, April). *A Bayesian computerized mastery model with multiple cut scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Spray, J. A. & Reckase, M. D. (1994, April). *The selection of test items for decision making with a computer adaptive test.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Spray, J. A. & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21,* 405–414.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201–216.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24,* 271–292.

Vos, H. J. (2002). Applying the minimax principle to sequential mastery testing. In A. Ferligoj and A. Mrvar (Eds.), *Developments in social science methodology.* Metodoloski zvezki, 18, Ljubljana: FDV.

Wainer, H. (Ed.), (2000). *Computerized adaptive testing. A primer* (2nd ed.). Hilsdale, NJ: Lawrence Erlbaum Associates.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Warm, T. A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika,54,* 427–450.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21,* 361–375.

Chapter 20
Testlet-Based Adaptive Mastery Testing

Hans J. Vos and Cees A.W. Glas

20.1 Introduction

In mastery testing, the problem is to decide whether a test taker must be classified as a master or a nonmaster. The decision is based on the test taker's observed test score. Well-known examples of mastery testing include testing for pass-fail decisions, licensure, and certification. A mastery test can have both fixed-length and variable-length forms. In a fixed-length mastery test, the performance on a fixed number of items is used for deciding on mastery or nonmastery. Over the last few decades, the fixed-length mastery problem has been studied extensively by many researchers (e.g., De Gruijter & Hambleton, 1984; van der Linden, 1990). Most of these authors derived, analytically or numerically, optimal rules by applying (empirical) Bayesian decision theory (e.g., DeGroot, 1970; Lehmann, 1986) to this problem. In the variable-length form, in addition to the action of declaring mastery or nonmastery, the action of continuing to administer items is available also (e.g., Kingsbury and Weiss, 1983; Lewis & Sheehan, 1990; Sheehan and Lewis, 1992; Spray & Reckase, 1996).

In either case, items may be administered one at a time or in batches of more than one item. If it is plausible that the responses to items within a batch are more strongly related than the responses to items of different batches, these batches are usually referred to as *testlets*. The main advantage of variable-length mastery tests as compared to fixed-length mastery tests is that the former offer the possibility of providing shorter tests for those test takers who have clearly attained a certain level of mastery (or nonmastery) and longer tests for those for whom the mastery decision is not as clear-cut (Lewis & Sheehan, 1990). For instance, Lewis & Sheehan (1990) showed in a simulation study that average test lengths could be reduced by half without sacrificing classification accuracy.

H.J. Vos
Department of Research Methodology, Measurement, and Data Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands

C.A.W. Glas (✉)
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_20,

Two approaches to variable-length mastery testing can be distinguished. The first approach involves specification of the costs of misclassifications and the cost of test administration, with the decision to continue testing being guided by expected losses. Item and testlet selection are no issues in this approach; items (or testlets) are randomly selected and administered. In this case, the stopping rule (i.e., termination criterion) is adaptive, but the item selection procedure is not. This type of variable-length mastery testing is known as *sequential mastery testing*, and, in the sequel, it will be referred to as SMT. The procedure is usually modeled using sequential Bayesian decision theory. In the second approach, item (or testlet) selection is tailored to the test taker's estimated proficiency level. Kingsbury and Weiss (1983) denote this type of variable-length mastery testing as *adaptive mastery testing* (AMT). This approach, however, does not involve specification of the cost of misclassifications and the cost of test administration. The approach is usually modeled using an IRT model, and the process of testing continues until the difference between the test taker's proficiency θ and a cut-off point θ_c on the latent continuum can be estimated with a certain precision.

The purpose of the present chapter is to present a combination of sequential and adaptive mastery testing which will be referred to as *adaptive sequential mastery testing (*ASMT). The strong points of both approaches are combined; that is, the selection as well as the stopping rule are adaptive and the cost per observation is explicitly taken into account. To support decision making and adaptive item (or testlet) selection, the 1PL model (Rasch, 1960) and the 3PL model (Birnbaum, 1968) will be used to describe the relationship between proficiency level and observed responses.

The chapter is organized as follows. First, a concise review of the existing literature and earlier approaches to the variable-length mastery problem will be presented. Second, the combined approach of sequential and adaptive mastery testing will be described. Then this general approach will be applied to the 1PL and 3PL models, and to the 3PL testlet model by Wainer, Bradlow, and Du (2000; see also Bradlow, Wainer & Wang, 1999). Following this, a number of simulation studies will be presented that focus on the gain of a sequential procedure over a fixed-length test and the gain of an adaptive sequential test over a classical sequential test. Gain will be defined in terms of the average loss, the average number of items administered, and the percentage of correct decisions. Further, as in the previous chapter, the impact of ignoring the testlet structure on the performance of the procedure will be studied. The chapter concludes with some new lines of research.

20.2 Earlier Approaches to the Variable-Length Mastery Problem

In this section, earlier approaches to the variable-length mastery problem will be briefly reviewed. First, the application of the sequential probability ratio test (SPRT; Wald, 1947) to SMT is considered. Next, IRT-based adaptive mastery testing strategies will be reviewed. Finally, contributions of Bayesian decision theory to sequential mastery testing will be presented.

20.2.1 Contributions of SPRT to Variable-Length Mastery Testing

The application of Wald's SPRT, originally developed as a statistical quality control test in a manufacturing setting, to SMT dates back to Ferguson (1969). In this approach, a test taker's responses to items were assumed to be binomially distributed. Reckase (1983) proposed alternative sequential procedures within an SPRT framework, which, in contrast to Ferguson's approach, did not assume that items have equal difficulty but allowed them to vary in difficulty and discrimination by using an IRT model instead of a binomial distribution (also see Spray & Reckase, 1996).

20.2.2 IRT-Based Item Selection Strategies Applied to Adaptive Mastery Testing

Two IRT-based item selection strategies have been primarily used in implementing AMT. In the first approach, Kingsbury and Weiss (1983) proposed selecting the item that maximizes the amount of information at the test taker's last proficiency estimate. In the second approach, a Bayesian item selection strategy, the item to be administered next is the one that minimizes the posterior variance of the test taker's last proficiency estimate. A prior distribution for the test taker's proficiency level must be specified in this approach before the administration of the first item. As pointed out by Chang and Stout (1993), it may be noted that the posterior variance converges to the reciprocal of the test information when the number of items goes to infinity. Therefore, the two methods of IRT-based item selection strategies should yield similar results when the number of administered items is large.

20.2.3 Sequential Mastery Testing Based on Bayesian Decision Theory

As mentioned before, most researchers in this area have applied (empirical) Bayesian decision theory to the fixed-length mastery problem. Within a Bayesian decision-theoretic framework, the following two basic elements must be specified: a psychometric model for the probability of answering an item correctly given a test taker's proficiency level (i.e., the item response function), and a loss structure evaluating the total costs and benefits of all possible decision outcomes. These costs may reflect all relevant psychological, social, and economic aspects involved in the decision. The Bayesian approach allows the decision maker to incorporate into the decision process the costs of misclassifications. Furthermore, a prior distribution must be specified representing prior knowledge of the test taker's proficiency level. Finally, a cut-off point on the latent proficiency scale separating masters and nonmasters must be specified in advance by the decision maker using a method of standard setting (e.g., Angoff, 1971). In a Bayesian approach, optimal rules are obtained by minimizing the posterior expected loss associated with each possible decision outcome.

Lewis & Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995) have applied Bayesian sequential decision theory to the variable-length mastery problem. In addition to the elements needed in the previous Bayesian decision-theoretic approach, the cost per observation is explicitly specified in this framework. The cost of administering one additional item (or testlet) can be considered as an extension of the loss structure for the fixed-length mastery problem to the variable-length mastery problem. Posterior expected losses associated with nonmastery and mastery decisions can now be calculated at each stage of testing. The posterior expected loss associated with continuing to test is determined by averaging the posterior expected loss associated with each of the possible future decision outcomes relative to the probability of observing those outcomes (i.e., the posterior predictive probability). Analogous to the fixed-length mastery problem, the optimal sequential rule is found by selecting the action (i.e., mastery, nonmastery, or to continue testing) that minimizes the posterior expected loss at each stage using techniques of dynamic programming (i.e., backward induction). This technique makes use of the principle that at each stage of an optimal procedure, the remaining portion of the procedure is optimal when considered in its own right. As indicated by Lewis & Sheehan (1990), the action selected at each stage of testing is optimal with respect to the entire sequential mastery testing procedure.

Vos (1999) also applied Bayesian sequential decision theory to SMT. Like Smith and Lewis (1995), he assumed three classification categories (i.e., nonmastery, partial mastery, and mastery). However, as in Ferguson's (1969) SPRT approach, for the conditional probability of a correct response given the test taker's proficiency level, the binomial distribution instead of an IRT model is considered. This modeling of response behavior corresponds to the assumption that all items have equal difficulty or are sampled at random from a large (real or hypothetical) pool of items. Assuming that prior knowledge about the test taker's proficiency can be represented by a beta prior $B(\alpha, \beta)$ (i.e., its natural conjugate), it is shown that the number-correct score is sufficient to calculate the posterior expected losses at future stages of the mastery test.

20.3 Bayesian Sequential Decision Theory Applied to Adaptive Mastery Testing

In this section, the approach of applying Bayesian sequential decision theory to ASMT will be described. Before doing so, some necessary notation will be introduced and the general variable-length mastery problem will be formalized. Then, the loss function assumed will be discussed. Next, it will be shown how IRT models can be incorporated into ASMT.

20.3.1 Formalization of the Variable-Length Mastery Problem

In the following, it will be assumed that the variable-length mastery problem consists of S ($S \geq 1$) stages labeled $s = 1, \ldots, S$ and that at each stage one of the available testlets can be given. At each stage, one or more items labeled i are administered and the observed item response will be denoted by a discrete random variable U_i, with realization u_i. It is assumed that U_i takes the values 1 and 0 for a correct and incorrect response, respectively. Let $\mathbf{u}_s$ be the response to the sth testlet. For $s = 1, \ldots, S$, the decisions will be based on a statistic $\mathbf{w}_s$, which is a function of the response patterns $\mathbf{u}_s$, that is, $\mathbf{w}_s = f(\mathbf{u}_1, \ldots, \mathbf{u}_s)$. In many cases, $\mathbf{w}_s$ will be the response pattern $\mathbf{u}_1, \ldots, \mathbf{u}_s$ itself. However, below it will become clear that some computations are feasible only if the information about the complete response pattern is aggregated. At each stage s ($s = 1, \ldots, S - 1$), a decision rule $d(\mathbf{w}_s)$ can be defined as

$$d(\mathbf{w}_s) = \begin{cases} m, & \text{test taker is judged a master,} \\ n, & \text{test taker is judged a nonmaster,} \\ c, & \text{testing is continued.} \end{cases} \tag{20.1}$$

At the final stage, only the two mastery classification decisions m and n are available. Mastery will be defined in terms of the latent proficiency continuum of the IRT model. Therefore, let θ and θ_c denote the test taker's proficiency level and some prespecified cut-off point on the latent continuum, respectively. Examinees with proficiency θ below this cut-off point are considered nonmasters, while test takers with proficiency θ above this cut-off point are considered masters.

20.3.2 Linear Loss

As noted before, a loss function evaluates the total costs and benefits for each possible combination of action and test taker's proficiency θ. Unlike Lewis & Sheehan (1990), Sheehan and Lewis (1992), and Smith and Lewis (1995), threshold loss will not be adopted here. The reason is that this loss function, although frequently used in the literature, may be less realistic in some applications. An obvious disadvantage of threshold loss is that it does not depend on the distance between θ and θ_c. It seems more realistic to assume that loss is an increasing function of θ for nonmasters and a decreasing function of θ for masters. Moreover, the threshold loss function is discontinuous; at the cut-off point θ_c this function "jumps" from one constant value to another. This sudden change seems unrealistic in many decision-making situations. In the neighborhood of θ_c, the losses for correct and incorrect decisions should change smoothly rather than abruptly (van der Linden, 1981).

To overcome these shortcomings, van der Linden and Mellenbergh (1977) proposed a continuous loss function for the fixed-length mastery problem that is a linear function of the test taker's proficiency level θ (see also Huynh, 1980; van der Linden & Vos, 1996; Vos, 1997a, 1997b, 1999).

For the variable-length mastery problem, the piecewise linear loss functions for the master and nonmaster decision can be restated at each stage as

$$L(m, \theta) = \max\{sC, sC + A(\theta - \theta_c)\}, \tag{20.2}$$

with $A < 0$, and

$$L(n, \theta) = \max\{sC, sC + B(\theta - \theta_c)\}, \tag{20.3}$$

with $B > 0$; C is the cost of delivering one testlet, sC is the cost of delivering s testlets. For the sake of simplicity, following Lewis & Sheehan (1990), these costs are assumed to be equal for each decision outcome as well as for each sample. The above-defined function consists of a constant term and a term proportional to the difference between the test taker's proficiency level θ and the specified cut-off point θ_c. The conditions $A < 0$ and $B > 0$ are equivalent to the statement that for action m the loss is a decreasing function of the latent proficiency θ, whereas the loss for action n is assumed to be increasing in θ. The definitions (20.2) and (20.3) guarantee that the losses are at least sC. Unlike the specification of loss in van der Linden and Mellenbergh (1977), this specific formulation of linear loss is chosen because in many problems it has been convenient to work with nonnegative loss functions (see, for instance, DeGroot, 1970, p. 125).

The loss parameters A, B, and C have to be either theoretically or empirically assessed. For assessing loss functions empirically, most texts on decision theory propose lottery methods (e.g., Luce & Raiffa, 1957, chap. 2). In general, these methods use the notion of desirability of outcomes to scale the consequences of each pair of actions and the test taker's proficiency level.

At stage s, the decision as to whether the test taker is a master or a nonmaster, or whether another testlet will be administered, is based on the expected losses of the three possible decisions given the observation $\mathbf{w}_s$. The expected losses of the first two decisions are computed as

$$E(L(m, \theta) \mid \mathbf{w}_s) = sC + A \int_{-\infty}^{\theta_c} (\theta - \theta_c) p(\theta \mid \mathbf{w}_s) d\theta \tag{20.4}$$

and

$$E(L(n, \theta) \mid \mathbf{w}_s) = sC + B \int_{\theta_c}^{\infty} (\theta - \theta_c) p(\theta \mid \mathbf{w}_s) d\theta, \tag{20.5}$$

where $p(\theta \mid \mathbf{w}_s)$ is the posterior density of θ given $\mathbf{w}_s$. The expected loss of the third possible decision is computed as the expected risk of continuing to test. If the expected risk of continuing to test is smaller than the expected loss of a master or a nonmaster decision, testing will be continued. The expected risk of continuing to test is defined as follows. Let $\{\mathbf{w}_{s+1} | \mathbf{w}_s\}$ be the range of $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$. Then, for $s = 1, \ldots, S - 1$, the expected risk of continuing to test is defined as

$$E(R(\mathbf{w}_{s+1}) \mid \mathbf{w}_s) = \sum_{\{\mathbf{w}_{s+1} \mid \mathbf{w}_s\}} p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) R(\mathbf{w}_{s+1}), \tag{20.6}$$

where the posterior predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ is given by

$$p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) = \int p(\mathbf{w}_{s+1} \mid \theta, \mathbf{w}_s) p(\theta \mid \mathbf{w}_s) d\theta, \tag{20.7}$$

and risk is defined as

$$\begin{aligned} R(\mathbf{w}_{s+1}) = \min\{&E(L(m,\theta) \mid \mathbf{w}_{s+1}),\\ &E(L(n,\theta) \mid \mathbf{w}_{s+1}), E(R(\mathbf{w}_{s+2}) \mid \mathbf{w}_{s+1})\}. \end{aligned} \tag{20.8}$$

The risk associated with the last testlet is defined as

$$R(\mathbf{w}_S) = \min\{E(L(m,\theta) \mid \mathbf{w}_S), E(L(n,\theta) \mid \mathbf{w}_S)\}. \tag{20.9}$$

So, given an observation $\mathbf{w}_s$, the expected distribution of $\mathbf{w}_{s+1}, \mathbf{w}_{s+2}, \ldots, \mathbf{w}_S$ is generated and an inference about future decisions is made. Based on these inferences, the expected risk of continuation in (20.6) is computed and compared with the expected losses of a mastery or nonmastery decision. If the risk of continuation is less than these two expected losses, testing is continued. If this is not the case, the classification decision with the lowest expected loss is made.

Notice that the definitions (20.6)–(20.9) imply a recursive definition of the expected risk of continuing to test. In practice, the computation of the expected risk of continuing to test can be done by backward induction as follows. First, the risk of the last testlet is computed for all possible values of $\mathbf{w}_S$. Then the posterior predictive distribution $p(\mathbf{w}_S \mid \mathbf{w}_{S-1})$ is computed using (20.7), followed by the expected risk $E(R(\mathbf{w}_S) \mid \mathbf{w}_{S-1})$ defined in (20.6). This, in turn, can be used for computing the risk $R(\mathbf{w}_{S-1})$ for all $\mathbf{w}_{S-1}$ using (20.8). The iterative process continues until s is reached and the decision can be made to administer testlet $s+1$ or to decide on mastery or nonmastery.

20.3.3 The Rasch Model

In the Rasch model, the probability of a response pattern $\mathbf{u}$ on a test of K items is given by

$$\begin{aligned} p(\mathbf{u} \mid \theta, \mathbf{b}) &= \prod_{i=1}^{K} \frac{\exp(u_i(\theta - b_i))}{1 + \exp(\theta - b_i)} \\ &= \exp(t\theta) \exp(-\mathbf{u}'\mathbf{b}) P_0(\theta), \end{aligned} \tag{20.10}$$

where $\mathbf{b} = (b_1, \ldots, b_K)$ is a vector of item parameters, $\mathbf{u}'\mathbf{b}$ is the inner product of $\mathbf{u}$ and $\mathbf{b}$, t is the sum score $t = \sum_i u_i$, and

$$P_0(\theta) = \prod_{i=1}^{K} (1 + \exp(\theta - b_i))^{-1}. \tag{20.11}$$

Notice that t is the minimal sufficient statistic for θ. Further, it is easily verified that $P_0(\theta)$ is the probability, given θ, of a response pattern with all item responses equal to zero. The probability of observing t given θ is given by

$$\begin{aligned} p(t \mid \theta) &= \sum_{\{\mathbf{u}|t\}} p(\mathbf{u} \mid \theta) \\ &= \sum_{\{\mathbf{u}|t\}} \exp(t\theta - \mathbf{u}'\mathbf{b}) P_0(\theta) \\ &= \gamma_t(\mathbf{b}) \exp(t\theta) P_0(\theta), \end{aligned}$$

with $\gamma_t(\mathbf{b})$ a function defined by $\gamma_t(\mathbf{b}) = \sum_{\{\mathbf{u}|t\}} \exp(-\mathbf{u}'\mathbf{b})$, and where $\{\mathbf{u} \mid t\}$ stands for the set of all possible response patterns resulting in a sum score t. Let $g(\theta)$ be the prior density of θ. Usually the prior is taken to be standard normal. An important feature is that the posterior distributions of θ given $\mathbf{u}$ and t are the same, that is,

$$\begin{aligned} p(\theta \mid \mathbf{u}) &= \frac{\exp(t\theta - \mathbf{u}'b) P_0(\theta) g(\theta)}{\int \exp(t\theta - \mathbf{u}'b) P_0(\theta) g(\theta) d\theta} \\ &= \frac{\exp(t\theta) P_0(\theta) g(\theta)}{\int \exp(t\theta) P_0(\theta) g(\theta) d\theta} \\ &= \frac{\gamma_t(\mathbf{b}) \exp(t\theta) P_0(\theta) g(\theta)}{\int \gamma_t(\mathbf{b}) \exp(t\theta) P_0(\theta) g(\theta) d\theta} = p(\theta \mid t). \end{aligned}$$

At this point, an assumption will be introduced that may not be completely realistic. It will be assumed that local independence simultaneously holds within and between testlets, that is, all item responses are independent given θ. So at this point, no attempt is made here to model a possible dependence structure of testlet responses. This point will be addressed in the following section.

Applying the general framework of the previous section to the Rasch model entails choosing the minimal sufficient statistics for θ, that is, the unweighted sum scores, for the statistics $\mathbf{w}_s$. Let $\mathbf{t}_s, \mathbf{t}_s = (t_1, \ldots, t_s)$ be the score pattern on the first s testlets, and define $r_s = \sum_{d=1}^{s} t_d$. Let $p(\theta \mid r_s)$ stand for the posterior density of proficiency given r_s. Then the expected losses (20.4) and (20.5) and the expected risk (20.6) can be written as $E(L(m, \theta) \mid r_s)$, $E(L(n,\theta) \mid r_s)$, and $E(R(r_{s+1}) \mid r_s)$. More specifically, the expected risk is given by

$$E(R(r_{s+1}) \mid r_s) = \sum_{r_{s+1}|r_s} p(r_{s+1} \mid r_s) R(r_{s+1}), \tag{20.12}$$

and (20.7) specializes to

$$\begin{aligned} p(r_{s+1} \mid r_s) &= \int p(r_{s+1} \mid \theta, r_s) p(\theta \mid r_s) d\theta \\ &= \int \gamma_{t_{s+1}} \ \exp(t_{s+1}\theta) P_{0(s+1)}(\theta) \ p(\theta \mid r_s) d\theta, \end{aligned} \tag{20.13}$$

where $t_{s+1} = r_{s+1} - r_s$, $\gamma_{t_{s+1}}$ is a shorthand notation for the elementary symmetric function of the item parameters of testlet $s + 1$, and $P_{0(s+1)}(\theta)$ is equal to (20.11) evaluated using the item parameters of testlet $s + 1$. That is, $P_{0(s+1)}(\theta)$ is equal to the probability of a zero-response pattern on testlet $s + 1$, given θ. Since elementary functions can be computed very quickly and with a high degree of precision (Verhelst, Glas & van der Sluis, 1984), the risk functions can be explicitly computed.

20.3.4 The 3PL Model and the 3PL Testlet Model

A testlet is a subset of items related to some common context. Haladyna (1994) refers to context-dependent item sets. Usually, these sets take the form of a number of multiple-choice items organized under or within some text. Haladyna (1994) gives examples of comprehension-type items sets and problem-solving type item sets. When a test consists of a number of testlets, both the within-testlets and between-testlets dependences between the item responses play a role. One approach to analyze testlet data is to ignore the dependence structure and analyze the test as a set of atomistic items. This generally leads to an overestimate of measurement precision and bias in the item parameter estimates (Sireci, Wainer & Thissen, 1991; Yen, 1993; Wainer & Thissen, 1996). Another approach is to aggregate the item scores within the testlet to a testlet score and analyze the testlet scores using an IRT model for polytomously scored items. This approach discards part of the information in the item responses, which will lead to some loss of measurement precision. However, this effect seems to be small (Wainer, 1995). A rigorous way to solve the problem is to model the within and between dependence explicitly. Wainer, Bradlow, and Du (2000) introduce a generalization of the 3PLM given by

$$P_i(\xi_{s(i)}) = p(U_i = 1|\xi_{s(i)}) = c_i + (1 - c_i)\frac{\exp(a_i(\xi_{s(i)} - b_i))}{1 + \exp(a_i(\xi_{s(i)} - b_i)}, \tag{20.14}$$

where $s(i)$ is the testlet to which item i belongs and $\xi_{s(i)}$ is a personparameter depending on the specific testlet. For every testlet, every person independently draws

ξ_s from a normal distribution with mean θ and a variance σ_s^2. The density of ξ_s will be denoted $h(\xi_s \mid \theta, \sigma_s^2)$. The probability of the complete response pattern is given by

$$p(\mathbf{u}) = \left[\prod_s p(\mathbf{u}_s \mid \xi_s) h(\xi_s \mid \theta, \sigma_s^2)\right]$$

and the probability of the response pattern on testlet s is

$$p(\mathbf{u}_s \mid \xi_s) = \prod_i P_i(\xi_{s(i)})^{u_i} (1 - P_i(\xi_{s(i)}))^{1-u_i}.$$

The parameters in the model can be estimated in a Bayesian framework using *Markov chain Monte Carlo* (MCMC; Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, and Du, 2000) or in a frequentist framework using MML (Glas, Wainer & Bradlow, 2000).

Unlike the Rasch model, the 3PL model has no minimal sufficient statistic for θ. Therefore, one approach of applying the general framework for sequential testing to the 3PL model would be to substitute the complete response pattern $(\mathbf{u}_1, \ldots, \mathbf{u}_s)$ for $\mathbf{w}_s$. For the testlets where the responses are already known, say the testlets $1, \ldots, s^*$, this presents no problem. But for evaluation of $E(R(\mathbf{w}_{s+1}) \mid \mathbf{w}_s)$, $s \geq s^*$, however, this entails a summation over the set of all possible response patterns on the future testlets, and exact computation of this expected risk generally presents a major problem. One of the approaches to this problem is approximating (20.6) using Monte Carlo simulation techniques, that is, simulating a large number of draws from $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ to compute the mean of $R(\mathbf{w}_{s+1})$ over these draws. However, this approach proves quite time-consuming and is beyond the scope of the present chapter. The approach adopted here assumes that the unweighted sum score contains much of the relevant information provided by the testlets $s+1, \ldots, S$ with respect to θ. This is motivated by the fact that the expected number-right score $\sum_i P_i(\theta)$ is monotonically increasing in θ (see, for instance, Lord, 1980, pp. 46–49). Therefore, the following procedure is used.

Suppose that s^* testlets have been administered. Then $\mathbf{w}_{s^*,s}$ will be the observed response patterns on the s^* testlets and the, as yet, unobserved sum score on the testlets $d = s^* + 1, \ldots, s$. So let t_d be the sum score on testlet d, that is, $t_d = \sum_i u_{di}$, and let $r_{s^*,s}$ be the sum over the scores of the testlets $d = s^* + 1, \ldots, s$, that is, $r_{s^*,s} = \sum_{d=s^*+1}^{s} t_d$. Then $\mathbf{w}_s$ is defined as $\mathbf{w}_s = (\mathbf{u}_1, \ldots, \mathbf{u}_{s^*}, r_{s^*,s})$. Using these definitions, formulas (20.4)–(20.9) are evaluated with response patterns to support the computation of the posterior proficiency distribution given the observations, and sum scores as summary statistics for future response behavior.

The probability of a sum score t_s is computed by summing over the set of all possible response patterns $\mathbf{u}_s$ resulting in a sum score t_s, denoted by $\{\mathbf{u}_s \mid t_s\}$, that is,

$$p(t_s \mid \xi_s) = \sum_{\{\mathbf{u}_s \mid t_s\}} p(\mathbf{u}_s \mid \xi_s).$$

The recursion formulas needed for the computation of these probabilities can, for instance, be found in Kolen and Brennan (1995, pp. 181–183). The probability of t_s conditional on θ is given by

$$p(t_s \mid \theta; \sigma_{\xi s}) = \int p(t_s \mid \xi_s) h(\xi_s \mid \theta; \sigma_{\xi s}) d\xi_s.$$

Finally, let $\{t_{s^*+1}, \ldots, t_s \mid r_{s^*,s}\}$ be the set of all testlet scores $t_{s^*+1}, \ldots, t_s$ compatible with a total score $r_{s^*,s}$. Then

$$p(r_{s^*,s} \mid \theta; \sigma_{\xi s}) = \sum_{\{t_{s^*+1}, \ldots, t_s \mid r_{s^*,s}\}} p(t_s \mid \theta; \sigma_{\xi s})$$

and the posterior distribution of θ given $\mathbf{w}_s$ is given by $p(\theta \mid \mathbf{w}_s; \sigma_{\xi s}) \propto p(\mathbf{u}_1, \ldots, \mathbf{u}_{s^*} | \theta; \sigma_{\xi s})\ p(r_{s^*,s} \mid \theta; \sigma_{\xi s})\ g(\theta)$, with

$$p(\mathbf{u}_1, \ldots, \mathbf{u}_{s^*} \mid \theta; \sigma_{\xi s}) = \prod_{d=1}^{s^*} \int p(\mathbf{u}_s \mid \xi_s) h(\xi_s \mid \theta; \sigma_{\xi s}) d\xi_s.$$

Inserting these definitions into (20.4) – (20.9) defines the sequential mastery testing procedure for the 3PL testlet model.

20.3.5 Adaptive Sequential Mastery Testing

One of the topics addressed in this chapter is how the sequential testing procedure can be optimized when a large testlet bank is available. The question is which testlets must be administered next upon observing $\mathbf{w}_s$. Three approaches will be considered. The first two are taken directly from the framework of non-Bayesian adaptive mastery testing (see, for instance, Kingsbury and Weiss, 1983, Weiss & Kingsbury, 1984). Both are based on the maximum information criterion; the first approach entails choosing items or testlets with maximum information at θ_c, and the second one chooses items or testlets with maximum information at $\hat{\theta}_s$, which is an estimate of θ at stage s. The third approach relates to a distinct difference between the non-Bayesian and Bayesian approaches. In the former approach, one is interested in a point estimate of θ or in whether θ is below or above some cut-off point. In the latter approach, however, one is primarily interested in minimizing possible losses due to misclassifications and the costs of testing. This can be directly translated into a selection criterion for the next testlet. In a Bayesian framework for traditional computerized adaptive testing, one might be interested in the posterior expectation of θ. One of the selection criteria suited for optimizing testlet administration is choosing the testlet with the minimum expected posterior variance. If $\mathbf{w}_s$ is some function of the observed response pattern, and $\{\mathbf{w}_{s+1} | \mathbf{w}_s\}$ is the set of all possible values $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$, one may select the testlet where

$$\sum_{\{\mathbf{w}_{s+1}|\mathbf{w}_s\}} \mathrm{var}(\theta \mid \mathbf{w}_{s+1})p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$$

is minimum (see, for instance, van der Linden, 1998). In a sequential mastery testing framework, however, one is interested in minimizing possible losses, so as a criterion for selection of the next testlet, the minimization of

$$\sum_{\{\mathbf{w}_{s+1}|\mathbf{w}_s\}} \mathrm{var}(L(m,\theta) - L(n,\theta) \mid \mathbf{w}_{s+1})p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) \tag{20.15}$$

will be considered. That is, a testlet is chosen such that the expected reduction in the variance of the difference between the losses of the mastery and nonmastery decision is maximized. This criterion focuses on the posterior variance of the difference between the losses $L(m,\theta)$ and $L(n,\theta)$ given $\mathbf{w}_{s+1}$, and the criterion entails that the sum over all possible response patterns $\mathbf{w}_{s+1}$ of this posterior variance weighted by its posterior predictive probability $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ is minimal. In the case of the Rasch model, (20.15) is relatively easy to compute because in that case sum scores can be substituted for $\mathbf{w}_{s+1}$ and $\mathbf{w}_s$.

20.4 Performance of Sequential and Adaptive Sequential Mastery Testing

20.4.1 The 1PL Model

The main research questions addressed in this section will be whether, and under what circumstances, sequential testing improves upon a fixed test and whether, and under what circumstances, adaptive sequential testing improves upon sequential testing. The design of the studies will be explained using the results of the first study, reported in Table 20.1.

The study concerns the 1PL model, 40 items, and a cut-off point θ_c equal to 1.00. The 13 rows of the table represent 13 simulation studies of 2,000 replications each. For every replication, a true θ was drawn from a standard normal distribution. In the first simulation study, every simulee was presented one test with a fixed length of 40 items. For every simulee, the item parameters were drawn from a standard normal distribution. Also, the prior distribution of θ was standard normal. The remaining 12 rows relate to a two-factor design, the first factor being the test administration design, the second the selection method. The test administration design is displayed in the first two columns of the table. The three designs used were 4 testlets of 10 items, 10 testlets of 4 items, and 40 testlets of one item each. Four selection methods were studied. In the studies labeled "Sequential" the Bayesian SMT procedure was used. The studies labeled "Cut-off Point", "EAP Estimate" and "Min Variance" used ASMT procedures. The label "Cut-off Point" refers to studies

Table 20.1 Relationship between selection method and loss in the 1PL model

Maximum Number Testlets	Number Items Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
1	40	Fixed Test	1.00	0.94	0.4171
4	10	Sequential	0.28	0.89	0.1622
4	10	Cut-off Point	0.27	0.92	0.1555
4	10	EAP Estimate	0.27	0.89	0.1630
4	10	Min Variance	0.29	0.90	0.1623
10	4	Sequential	0.17	0.91	0.1094
10	4	Cut-off Point	0.19	0.91	0.1211
10	4	EAP Estimate	0.17	0.90	0.1151
10	4	Min Variance	0.16	0.91	0.1068
40	1	Sequential	0.09	0.89	0.0996
40	1	Cut-off Point	0.10	0.90	0.0899
40	1	EAP Estimate	0.10	0.90	0.0997
40	1	Min Variance	0.10	0.89	0.1028

where testlets were selected with maximum information at the cut-off point, “EAP Estimate” refers to studies with a selection procedure based on maximum information at the EAP estimate of proficiency, and “Min Variance” refers to studies with adaptive testlet selection using the Bayesian criterion defined by (20.15). For all simulations, the parameters of the loss functions (20.2) and (20.3) were equal to $A = -1.00$, $B = 1.00$ and $C = 0.01k_t$, where k_t stands for the number of items in a testlet. The motivation for this choice of C is keeping the total cost of administering 40 items constant.

For the SMT condition, the item parameters of the first testlet were all equal to zero and the item parameters of all other testlets were randomly drawn from a standard normal distribution. In the ASMT conditions, it was also the case that the first testlet had all item parameters equal to zero. The reason for starting both the SMT and ASMT procedures with testlets with similar item parameters was to create comparable conditions in the initial phase of the procedures. The following testlets were chosen from a bank of 50 testlets that was generated as follows.

For every simulee, $50k_t$ item parameters were drawn from the standard normal distribution first. Then these $50k_t$ item parameters were ordered in magnitude from low to high. The first k_t items comprised the first testlet in the bank, the second k_t items comprised the second testlet, etc. In this way, 50 testlets were created that were homogeneous in difficulty and attained their maximum information at distinct points of the latent proficiency scale. In the “EAP Estimate” condition, at stage $s, s = 1, ..., S - 1$, an expected a posteriori estimate of proficiency was computed and the expected risk of a “Continue Testing” decision was computed using the $S - s$ testlets with highest information at this estimate. If a “Continue Testing” decision was made, the next testlet administered was the most informative testlet of the $S - s$ testlets initially selected. The procedure in the “Min Variance” condition was roughly similar, only here the minimum variance criterion defined by (20.15) was used. Finally, in the "Cut-off Point" condition, testlets were selected from the testlet bank described above that were most informative at the cut-off point θ_c. The last three columns of Table 20.1 give the average proportion of testlets administered, the proportion of correct decisions, and the mean loss over 2,000 replications for each of the 13 conditions, where the loss in every replication was computed using (20.2) or (20.3) evaluated at the true value of θ, with s the number of testlets actually administered.

The study described in the previous paragraph was carried out for three total test lengths, $K = 10$, 20 and 40, and two choices of cut-off points, $\theta_c = 1.00$ and 0.10. The results for the combination $K = 40$ and $\theta_c = 1.00$ are given in Table 20.1. Notice that the mean loss in the SMT and ASMT conditions was much lower than in the fixed test condition, and mean loss decreased as a function of the number of testlets. Further, it can be seen that the decrease of mean loss was mainly due to a dramatic reduction in the proportion of testlets given. The number of correct classifications remained stable. Finally, it can be seen that there were no systematic or pronounced differences between SMT and ASMT. This picture also emerged in the $K = 10$ and 20 studies. Choosing a cut-off point $\theta_c = 0.10$ resulted in increased mean loss. For instance, for $K = 40$, the mean loss rose from 0.4171 to 0.4299 for the fixed test condition and from 0.0980 to 0.1541 for the 40-testlet condition, with mean loss averaged over the 4 selection methods. The reason for this increase is that moving the cut-off point closer to the mean of the proficiency distribution increases the number of test takers near the cut-off point. In summary, it was found that sequential mastery testing did indeed lead to a considerable decreased mean loss, mainly due to a significant decreased number of testlets administered. Across studies, ASMT did only fractionally better than SMT, and again across studies, the minimum variance criterion (20.15) and selection of testlets with maximum information near the cut-off point θ_c produce the best results, but the difference between them and the maximum information criterion is very small.

20.4.2 The 3PL Model and the 3PL Testlet Model

This section focuses on the question of whether the picture that emerged in the previous section for the 1PL model also holds for the 3PL mode l. In addition, the impact of the testlet structure will be studied. The simulation studies generally have the same setup as the studies of the previous section. The testlet bank was generated as above, with the difference that besides drawing the item difficulties from a standard normal distribution, item discrimination parameter values were drawn from a log-normal distribution with mean 0.00 and variance 0.25. The guessing parameter value was equal to 0.25 for all items. As above, the testlets were composed in such a way that they attained their maximum information at distinct points on the latent proficiency scale.

Since the differences among the three selection procedures for ASMT in the 1PL studies were very small, and the minimum variance criterion is quite time-consuming to compute, the latter selection criterion was not included in these studies. The results for a study with $K = 40$, $\theta_c = 1.00$, and no within-person variation of proficiency, that is, with $\sigma_{\xi s} = 0.00$, are shown in Table 20.2. It can be seen that the overall conclusion from the 1PL model studies still holds: there is a considerable decrease in mean loss as the number of testlets increases, and the decrease is not bought at the expense of an increased proportion of incorrect decisions. However, contrary to the results in Table 20.1, it can be observed that these studies showed a clear effect of adaptive testlet selection in terms of a decrease in mean loss. The magnitude of this decrease was positively related to the maximum number of testlets given.

In Table 20.3, analogous results are given for a situation where $\sigma_{\xi s} = 1.00$, and this within-person variance is explicitly taken into account in the testlet selection and decision procedure. Comparing the results to the results in Table 20.2, it can be seen that increasing the within-person variance resulted in an increase in mean loss. This increase is due to the addition of within-person variance that acts as a random

Table 20.2 Relationship between selection method and loss in the 3PL model

Maximum Number Testlets	Number Items Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
1	40	Fixed Test	1.00	0.93	0.4278
4	10	Sequential	0.32	0.91	0.1699
4	10	Cut-off Point	0.36	0.93	0.1730
4	10	EAP Estimate	0.36	0.92	0.1748
10	4	Sequential	0.29	0.91	0.1526
10	4	Cut-off Point	0.26	0.93	0.1324
10	4	EAP Estimate	0.25	0.92	0.1322
40	1	Sequential	0.39	0.91	0.1922
40	1	Cut-off Point	0.11	0.90	0.0990
40	1	EAP Estimate	0.13	0.96	0.0645

Table 20.3 Relationship between selection method and loss in the 3PL testlet model

Maximum Number Testlets	Number Items Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
1	40	Fixed Test	1.00	0.88	0.4795
4	10	Sequential	0.36	0.88	0.2106
4	10	Cut-off Point	0.36	0.88	0.2089
4	10	EAP Estimate	0.35	0.87	0.2233
10	4	Sequential	0.19	0.89	0.1442
10	4	Cut-off Point	0.17	0.88	0.1339
10	4	EAP Estimate	0.17	0.89	0.1381
40	1	Sequential	0.20	0.90	0.1332
40	1	Cut-off Point	0.16	0.90	0.1080
40	1	EAP Estimate	0.18	0.92	0.1100

Table 20.4 Relation between selection method and loss in the 3PL model when ignoring the testlet structure

Maximum Number Testlets	Number Items Testlet	Selection Method	Proportion Testlets Given	Proportion Correct Decisions	Mean Loss
1	40	Fixed Test	1.00	0.82	0.5564
4	10	Sequential	0.31	0.85	0.2209
4	10	Cut-off Point	0.36	0.88	0.2202
4	10	EAP Estimate	0.37	0.88	0.2199
10	4	Sequential	0.28	0.88	0.1773
10	4	Cut-off Point	0.25	0.90	0.1477
10	4	EAP Estimate	0.23	0.89	0.1506
40	1	Sequential	0.37	0.90	0.1955
40	1	Cut-off Point	0.12	0.87	0.1386
40	1	EAP Estimate	0.15	0.92	0.1120

error component. However, the positive effects of increasing the number of testlets and adaptive testlet selection remained evident.

Finally, in Table 20.4, results are given for a setup where the responses follow a testlet model with $\sigma_{\xi s} = 1.00$ but decisions and testlet selection were governed by the standard 3PL model, that is, a model with $\sigma_{\xi s} = 0.00$. In other words, for the computation of losses and making decisions, the testlet structure was not taken into account. It can be seen that mean loss is further inflated in all conditions. However, the advantage of SMT over using a fixed test and the advantage of ASMT over SMT were still apparent.

20.5 Discussion

In this chapter, a general theoretical framework for adaptive sequential mastery testing (ASMT) based on a combination of Bayesian sequential decision theory and item response theory was presented. It was pointed out how IRT-based sequential mastery testing (SMT) could be generalized to adaptive item and testlet selection rules, that is, to the case where the choice of the next item or testlet to be administered is optimized using the information from previous responses. The impact of IRT-based sequential and adaptive sequential mastery testing on average loss, proportion of correct decisions, and proportion of testlets given was investigated in a number of simulations using the 1PL as well as 3PL models. Two different dependence structures of testlet responses were introduced for the 3PL testlet model. In the first approach, it was assumed that local independence simultaneously holds within and among testlets; that is, all item responses are independent given the test taker's proficiency level. In the second approach, a hierarchical IRT model was used to describe a greater similarity of responses to items within than between testlets.

As far as the 1PL model is concerned, the results of the simulation studies indicated that the average loss in the SMT and ASMT conditions decreased considerably compared to the fixed test condition, while the proportion of correct decisions hardly changed. This result could mainly be ascribed to a significant decrease in the number of testlets administered. With the 3PL model, ASMT produced considerably better results than SMT, while with the 1PL model, the results of ASMT were only fractionally better. When testlet response behavior was simulated by a hierarchical IRT model with within-person proficiency variance, the average loss increased. Ignoring the within-person variance in the decision procedure resulted in a further inflation of losses.

In summary, the conclusion is that the combination of Bayesian sequential decision theory and modeling response behavior by an IRT model provides a sound framework for adaptive sequential mastery testing where both the cost of test administration and the distance between the test taker's proficiency and the cut-off point are taken into account.

The general approach sketched here can be applied to several other IRT models, for instance, to multidimensional IRT models (see, for instance, McDonald, 1997, or Reckase, 1997). The loss structure involved must allow for both conjunctive and compensatory testing strategies in this case. In decision theory, much work has already been done in this area under the name of "multiple-objective decision making" (Keeney & Raiffa, 1976). How the results reported there could be applied to the problems of ASMT in the case of multidimensional IRT models, still needs to be examined.

Another point of further study is the adoption of minimax sequential decision theory instead of Bayesian sequential decision theory (e.g., DeGroot, 1970; Lehmann, 1986). Optimal rules are found in this approach by minimizing the maximum expected losses associated with all possible decision rules. As pointed out by van der Linden (1981), the minimax principle assumes that it is best to prepare for the worst and establish the maximum expected loss for each possible decision rule. Minimax rules can therefore can be characterized as either conservative or pessimistic (Coombs, Dawes & Tversky, 1970). Analogous to Bayesian sequential decision theory, the cost of test administration is also explicitly taken into account in this approach.

References

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L.Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council of Education.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Chang, H.-H. & Stout, W. F. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika, 58,* 37–52.

Coombs, C. H., Dawes, R. M. & Tversky, A. (1970). *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

De Gruijter, D. N. M. & Hambleton, R. K. (1984). On problems encountered using decision theory to set cutoff scores. *Applied Psychological Measurement, 8*, 1–8.

Ferguson, R. L. (1969). *The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction*. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Glas, C. A. W., Wainer, H. & Bradlow, E. T. (2000). MML and EAP estimates for the testlet response model. In W. J. van der Linden & C. A. W.Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 271–287). Boston: Kluwer-Nijhoff Publishing.

Haladyna, T. M. (1994). *Developing and validating multiple- choice test items*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Huynh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika, 45,* 167–182.

Keeney, D. & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value trade-offs*. New York: John Wiley and Sons.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.): *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 257–283). New York: Academic Press.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating*. New York: Springer-Verlag.

Lehmann, E. L. (1986). *Testing statistical hypothesis. (*2nd ed.*)*. New York: Wiley.

Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14*, 367–386.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luce, R. D. & Raiffa, H. (1957). *Games and decisions*. New York: John Wiley and Sons.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York: Springer-Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–255). New York: Academic Press.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R.K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 271–286). New York: Springer-Verlag.

Sheehan, K. & Lewis, C. (1992). Computerized mastery testing with non-equivalent testlets. *Applied Psychological Measurement, 16*, 65–76.

Sireci, S. G., Wainer, H. & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237–247.

Smith, R. L. & Lewis, C. (1995). A Bayesian computerized mastery model with multiple cut scores. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Spray, J. A. & Reckase, M. D. (1996). Comparison of SPRT and sequential Bayes procedures for classifying examinees into two categories using a computerized test. *Journal of Educational and Behavioral Statistics, 21*, 405–414.

van der Linden, W. J. (1981). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4*, 469–492.

van der Linden, W. J. (1990). Applications of decision theory to test-based decision making. In R. K. Hambleton & J. N. Zaal (Eds.), *New developments in testing: Theory and applications* (pp. 129–155). Boston: Kluwer-Nijhof Publishing.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216.

van der Linden, W. J. & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1*, 593–599.

van der Linden, W. J. & Vos, H. J. (1996). A compensatory approach to optimal selection with mastery scores. *Psychometrika, 61,* 155–172.

Verhelst, N.D., Glas, C. A. W. & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1*, 245–262.

Vos, H. J. (1997a). Simultaneous optimization of quota-restricted selection decisions with mastery scores. *British Journal of Mathematical and Statistical Psychology, 50,* 105–125.

Vos, H. J. (1997b). A simultaneous approach to optimizing treatment assignments with mastery scores. *Multivariate Behavioral Research, 32*, 403–433.

Vos, H. J. (1999). Applications of Bayesian decision theory to sequential mastery testing. *Journal of Educational and Behavioral Statistics, 24*, 271–292.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*, 157–187.

Wainer, H., Bradlow, E. T. & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–270). Boston: Kluwer-Nijhof Publishing.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361–375.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187–213.

Chapter 21
Adaptive Mastery Testing Using a Multidimensional IRT Model

Cees A.W. Glas and Hans J. Vos

21.1 Introduction

Mastery testing concerns the decision to classify a student as a master or as a nonmaster. In the previous chapter, adaptive mastery testing (AMT) using item response theory (IRT) and sequential mastery testing (SMT) using Bayesian decision theory were combined into an approach labeled adaptive sequential mastery testing (ASMT). This approach is based on the one-parameter logistic model (1PLM; Rasch, 1960) and three-parameter logistic model (3PLM; Birnbaum, 1968). In the present chapter, ASMT is applied to a multidimensional IRT (MIRT) model.

In AMT (Weiss, 1983; Wainer, 1990), the available decisions are to classify a student as a master or a nonmaster, or to continue testing and administer another item or testlet (consisting of one or more items). Adaptive mastery tests are designed to maximize the proportion of correct classification decisions while minimizing the total test length. In a simulation study, Lewis and Sheehan (1990) showed that average test lengths could be reduced by a half without sacrificing classification accuracy. In AMT, response behavior is usually modeled by an IRT model. Test takers with a proficiency estimate that is sufficiently far above or below the cut-off point are classified as a master or a nonmaster, whereas those with an estimate within a prespecified region around the cut-off point are presented with another testlet. Further, the testlet selection mechanism is adaptive in the sense that the next testlet is selected in such a way that the expected gain in the precision of the proficiency estimate is maximized. Several generalizations of AMT to MIRT models have been proposed. For example, Segall (1996, this volume, chap. 3) proposed several applications of MIRT relative to the Armed Services Aptitude Battery (ASVAB). The procedures entail estimation of simultaneous scores on multiple, correlated dimensions via minimization of the joint posterior variance in a multidimensional space. Similarly,

C.A.W. Glas (✉)
Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

H.J. Vos
Department of Research Methodology, Measurement, and Data Analysis, P.O. Box 217, 7500 AE Enschede, The Netherlands

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_21,

Luecht (1996) compared various adaptive selection strategies for the United States Medical Licensing Examination, using both multidimensional and unidimensional composites (also see Luecht & Nungester, (1998, 2000) and van der Linden and Reese (1998) considered AMT with testlets in relation to the Law School Admission Test (LSAT).

In the present chapter, a version of ASMT is presented where response behavior is modeled by a multidimensional IRT model. Both conjunctive (i.e., minimal requirements for each proficiency dimension) and compensatory (i.e., low performance on one proficiency dimension can be compensated by high performance on another proficiency dimension) loss functions will be considered. Simulation studies will be used to evaluate the gain of SMT over fixed-length mastery testing and the gain of ASMT over SMT. Finally, the feasibility of the methods will be shown using a real data example.

21.2 Definition of the Decision Problem

As in the previous chapter, it will be assumed that the variable-length mastery problem consists of S ($S \geq 1$) stages labeled $s = 1, \ldots, S$ and at each stage a testlet can be administered. This testlet consists of one or more items indexed with i and the observed item responses for a randomly sampled student will be denoted by a discrete random variable U_i, with realization u_i. Let the vector of item responses $\mathbf{u}_s$ be the response pattern on the sth testlet. The decisions will be based on a statistic $\mathbf{w}_s$, which is a function of the response patterns up to stage s, that is, $\mathbf{w}_s = f(\mathbf{u}_1, \ldots, \mathbf{u}_s)$. In many cases, $\mathbf{w}_s$ will be the response pattern $\mathbf{u}_1, \ldots, \mathbf{u}_s$ itself. At each stage of testing, a decision rule $d(\mathbf{w}_s)$ is defined as

$$d(\mathbf{w}_s) = \begin{cases} m, & \text{mastery decision,} \\ n, & \text{nonmastery decision,} \\ c, & \text{testing is continued.} \end{cases} \tag{21.1}$$

At the final stage of testing, stage S, only the two classification decisions m and n are available. Mastery will be defined in terms of the latent proficiency continuum of the IRT model.

21.3 Multidimensional IRT Models

Multidimensional IRT models are IRT models for response behavior where the responses depend on more than one latent proficiency. Multidimensional IRT models for dichotomously scored items were first presented by McDonald (1967) and Lord & Novick (1968). These authors used a normal-ogive to describe the probability of a correct response. McDonald (1967, 1997) developed an estimation procedure based on an expression for the association between pairs of items derived from a polynomial expansion of the normal-ogive. The procedure is implemented in NOHARM (normal-ogive harmonic analysis robust method; Fraser & McDonald, (1988). An alternative approach using all information in the data, and therefore labeled "Full Information Factor Analysis", was developed by Bock, Gibbons, and Muraki (1988). This approach is a generalization of the marginal maximum likelihood (MML) and Bayes modal estimation procedures for unidimensional IRT models (see Bock & Aitkin, 1981, Mislevy, 1986), and has been implemented in TESTFACT (Wilson, Wood & Gibbons, 1991). A Bayesian estimation procedure using a Markov chain Monte Carlo (MCMC) technique was presented by Béguin and Glas (2001).

A comparable model using a logistic rather than a normal-ogive representation was studied by Andersen (1985), Glas (1992), Reckase (1985, 1997), and Ackerman (1996a, 1996b). In the present chapter, the logistic version of the model for dichotomous items will be used. In this version, the probability of a correct response is given by

$$p(U_i = 1 \mid \theta_1, \ldots, \theta_Q, a_{i1}, \ldots, a_{iQ}, b_i, c_i) = c_i + (1-c_i)\frac{\exp(\sum_q a_{iq}\theta_q - b_i)}{1 + \exp(\sum_q a_{iq}\theta_q - b_i)}, \tag{21.2}$$

where $\theta_1, \ldots, \theta_Q$ are proficiency parameters, $a_{i1}, \ldots, a_{iQ}$ are factorloadings, b_i is the item difficulty and c_i is the guessing parameter. The probability of a response pattern is given by

$$p(\mathbf{u} \mid \mathbf{a}, \mathbf{b}, \mathbf{c}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int, \ldots, \int p(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c}) g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \partial \boldsymbol{\theta}, \tag{21.3}$$

where $p(\mathbf{u} \mid \boldsymbol{\theta}, \mathbf{a}, \mathbf{b}, \mathbf{c})$ is the probability of a response pattern given $\boldsymbol{\theta}$, which is derived from (21.2) using the assumption of local independence, and $g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the Q-variate normal distribution. The latent scale is usually identified by imposing the restriction $\boldsymbol{\mu} = \mathbf{0}$.

Below, it will be assumed that local independence holds both between and within testlets. However, sometimes the association between responses to items within a testlet is stronger than the association between the responses to items of different testlets. In the previous chapter, it was shown that in that case dependence of the responses within a testlet can be modeled by treating the proficiency or item parameters as random effects (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow & Du, 2000; Glas, Wainer & Bradlow, 2000).

21.4 Compensatory and Conjunctive-Disjunctive Loss Functions

The first question that needs to be answered when defining a loss function is the definition of the distinction between masters and nonmasters. Following Coombs and Kao (1955; also see Coombs, 1960) we make a distinction between compensatory and conjunctive-disjunctive mastery models. In a compensatory mastery model, a low value on one dimension can be compensated for by a high value on another dimension. In a conjunctive mastery model, the proficiency must be above a certain cut-off point in all dimensions, and in a disjunctive model, the proficiency must be above a cut-off point in at least one dimension. Given this distinction, loss functions can be defined as follows.

21.4.1 Compensatory Loss Functions

First, an example of a linear compensatory loss function will be given. The generalization to nonlinear loss functions will be returned to in the last section of this chapter. Consider two dimensions. Let θ_{1c} and θ_{2c} denote prespecified cut-off points in the latent space. Consider a line in the two-dimensional proficiency space defined by $f_1(\theta_1, \theta_2) = A_1(\theta_1 - \theta_{1c}) + A_2(\theta_2 - \theta_{2c}) = 0$. The line $f_1(\theta_1, \theta_2)$ divides the latent space into two subspaces. Persons with a proficiency in one subspace are masters, while persons with a proficiency in the other subspace are nonmasters. The loss functions for the mastery and nonmastery decision are given by

$$L(m, \theta_1, \theta_2) = \max\{sC, sC + f_1(\theta_1, \theta_2)\} \tag{21.4}$$

with $A_1, A_2 < 0$, and

$$L(n, \theta_1, \theta_2) = \max\{sC, sC + f_2(\theta_1, \theta_2)\}, \tag{21.5}$$

with $f_2(\theta_1, \theta_2) = B_1(\theta_1 - \theta_{1c}) + B_2(\theta_2 - \theta_{2c})$ $(B_1, B_2 > 0)$, respectively. C is the cost of delivering one testlet, and sC is the cost of delivering s testlets. To ensure that $f_2(\theta_1, \theta_2) = 0$ defines the same line as $f_1(\theta_1, \theta_2) = 0$, the additional constraint $A_1/A_2 = B_1/B_2$ is imposed. Notice that the loss structure is compensatory in the sense that a proficiency below a cut-off point on one dimension can be compensated for by a proficiency above a cut-off point on the other dimension.

In Q dimensions, the loss function becomes

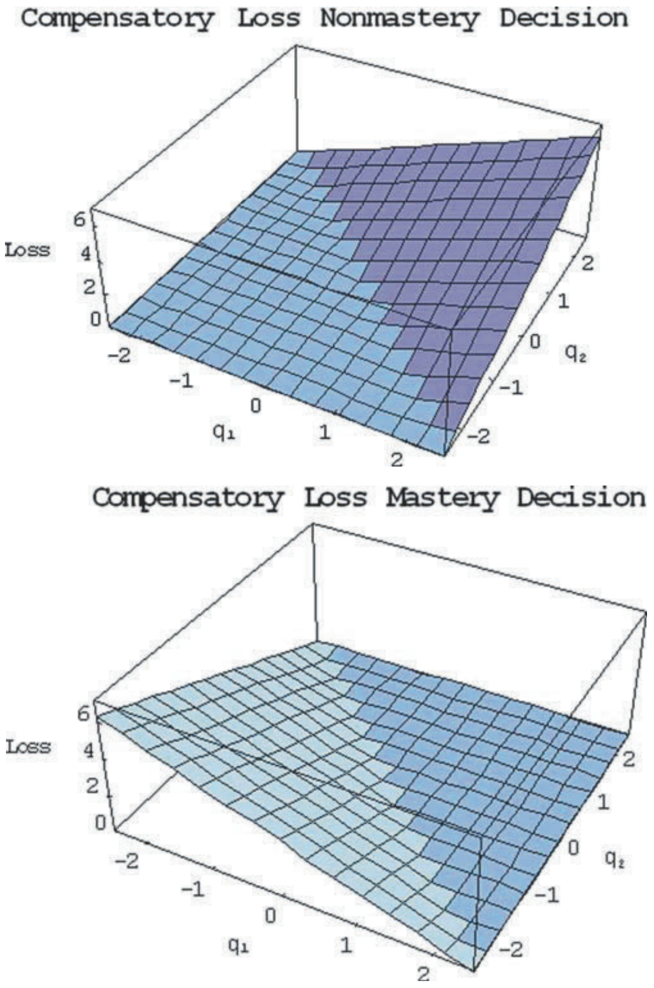
$$L(m, \boldsymbol{\theta}) = \max\{sC, sC + \mathbf{A}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c)\} \tag{21.6}$$

and

$$L(n, \boldsymbol{\theta}) = \max\{sC, sC + \mathbf{B}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c)\}, \tag{21.7}$$

where $\mathbf{A}$ and $\mathbf{B}$ are vectors of weights with all elements negative and positive, respectively, and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$ are the proficiency vector and a vector of cut-off points, respectively. An additional constraint is that $\mathbf{A}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c) = 0$ and $\mathbf{B}'(\boldsymbol{\theta} - \boldsymbol{\theta}_c) = 0$ define the same $(Q-1)$-dimensional linear subspace.

An example of the loss function for the nonmastery and mastery decision is given in Figure 21.1. The parameters A_1 and A_2 are equal to –1.25, the parameters B_1 and


Compensatory Loss Nonmastery Decision

Compensatory Loss Mastery Decision

Fig. 21.1 Compensatory loss functions

B_2 are equal to 1.25, $\theta_{1c} = \theta_{2c} = 0$, and C is equal to 0 for convenience. For the nonmastery decision (top panel), the loss function increases in θ_1 and θ_2 right of the line $\theta_1 + \theta_2 = 0$; that is, the loss associated with an incorrect nonmastery decision increases with proficiency. The bottom panel gives an analogous but opposite pattern for the mastery decision.

The choice of the actual values of the parameters of the loss function requires a comparison between the cost of observations and the cost of incorrect decisions. Consider a unidimensional latent proficiency variable θ and a threshold loss function $L(m, \theta) = sC + AI(\theta, \theta_c)$, where $I(\theta, \theta_c)$ is an indicator function that assumes a value equal to one if $\theta < \theta_c$ and zero otherwise. If, for instance, $A = -1$ and $C = 0.02$, this reflects the position that an incorrect mastery decision has the same cost as administering 50 items. Elaborating on such notions, Lewis and Sheehan (1990) used simulation studies for estimating the cost of item administration relative to the costs associated with incorrect decisions. The simulation study was performed such that decision rules with desirable operating characteristics (average test length, expected passing rate, and expected proportions of false mastery and nonmastery decisions) resulted. In the case of linear loss functions and multidimensional proficiencies, analogous, but slightly more complex, methods may be used. In such cases, a number of target proficiency levels may be chosen for which the cost of a miss-classification is defined in terms of the administration of a number of items. Furthermore, simulation studies can also be used here to determine whether the resulting operating characteristics are acceptable. Several examples of such simulation studies will be given below.

21.4.2 Conjunctive Loss Functions

In a conjunctive loss function, a student is considered a master if the proficiency is above a cut-off point on all dimensions, and is considered a nonmaster if the proficiency is below a cut-off point on at least one dimension. In two dimensions, this can be translated into the following loss function. Define

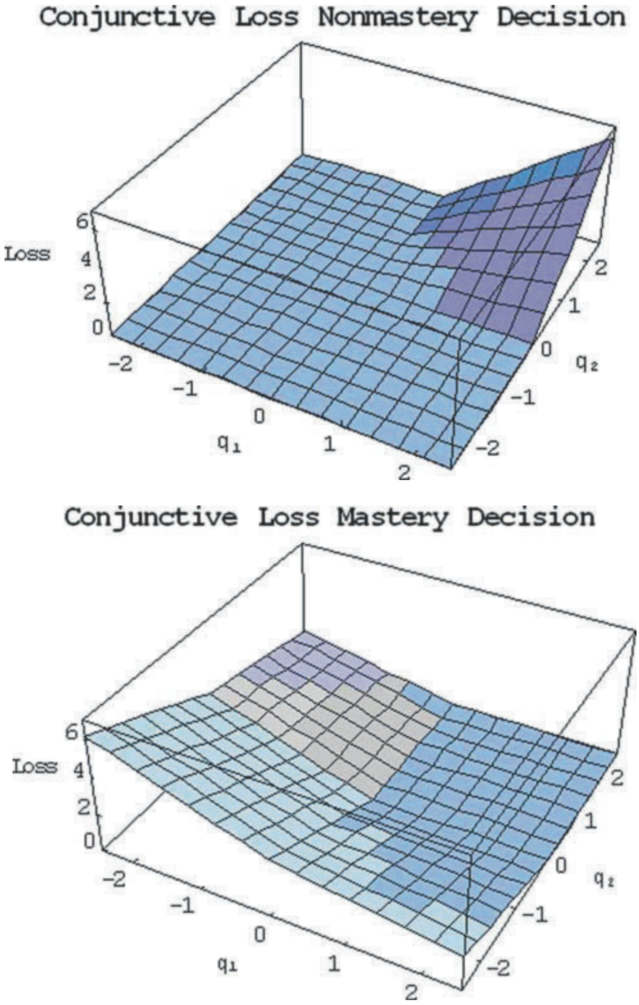
$$L(m,\theta_1,\theta_2) = \begin{cases} sC + A_1(\theta_1-\theta_{1c}) + A_2(\theta_2-\theta_{2c}) \\ \qquad\qquad\qquad \text{if } \theta_1 \le \theta_{1c} \text{ and } \theta_2 \le \theta_{2c}, \\ sC + A_2(\theta_2-\theta_{2c}) + A_3(\theta_1-\theta_{1c})(\theta_2-\theta_{2c}) \\ \qquad\qquad\qquad \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 < \theta_{2c}, \\ sC + A_1(\theta_1-\theta_{1c}) + A_4(\theta_1-\theta_{1c})(\theta_2-\theta_{2c}) \\ \qquad\qquad\qquad \text{if } \theta_1 < \theta_{1c} \text{ and } \theta_2 > \theta_{2c}, \\ sC \\ \qquad\qquad\qquad \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 > \theta_{2c}, \end{cases} \tag{21.8}$$

and

$$L(n,\theta_1,\theta_2) = \begin{cases} sC + (\theta_1-\theta_{1c})^{B_1}(\theta_2-\theta_{2c})^{B_2} & \text{if } \theta_1 > \theta_{1c} \text{ and } \theta_2 > \theta_{2c}, \\ sC & \text{otherwise,} \end{cases} \tag{21.9}$$

with $A_1, A_2 < 0$ and $B_1, B_2 > 0$. Both loss functions are continuous, and $L(n,\theta_1,\theta_2)$ is strictly positive and increasing on the space where $L(m,\theta_1,\theta_2)$ is equal to sC. In the same manner, $L(m,\theta_1,\theta_2)$ is strictly positive and decreasing on the space where $L(n,\theta_1,\theta_2)$ is sC. Notice that $L(m,\theta_1,\theta_2) = sC + A_1(\theta_1-\theta_{1c})$ on the line $\theta_2 = \theta_{2c}$, and $L(m,\theta_1,\theta_2) = A_2(\theta_2-\theta_{2c})$ on the line $\theta_1 = \theta_{1c}$.

An example of the loss function for the nonmastery and the mastery decision is given in Figure 21.2. The parameters A_1, A_2, A_3, and A_4 are equal to −1.25, −1.25,


Conjunctive Loss Nonmastery Decision

Conjunctive Loss Mastery Decision

Fig. 21.2 Conjunctive loss functions

0.2, and 0.2, respectively, and the parameters B_1 and B_2 are equal to 1.25. Again, $\theta_{1c} = \theta_{2c} = 0$ and $C = 0$ for convenience.

Coombs and Kao (1955) showed that conjunctive and disjunctive models are isomorphic and only one mathematical model needs to be developed for the analysis of the problem. In the present case, it is easily verified that choosing Equation (21.8) as the definition for $L(n, \theta_1, \theta_2)$, choosing Equation (21.9) for the definition of $L(m, \theta_1, \theta_2)$, and setting $A_1, A_2 > 0$ and $B_1, B_2 < 0$ defines the loss structure for the disjunctive case.

21.5 Computation of Expected Loss and Risk Using Backward Induction

At stage s, the decision whether the respondent is a master or a nonmaster, or whether another testlet will be administered, is based on the expected losses of the three possible decisions given the observation $\mathbf{w}_s$. The expected losses of the first two classification decisions are computed as

$$E(L(m,\boldsymbol{\theta}) \mid \mathbf{w}_s) = \int, \ldots, \int L(m,\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) \partial \boldsymbol{\theta} \tag{21.10}$$

and

$$E(L(n,\boldsymbol{\theta}) \mid \mathbf{w}_s) = \int, \ldots, \int L(n,\boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) \partial \boldsymbol{\theta}, \tag{21.11}$$

where $p(\boldsymbol{\theta} \mid \mathbf{w}_s)$ is the posterior density of $\boldsymbol{\theta}$ given $\mathbf{w}_s$. The expected loss of the third possible decision is computed as the expected risk of the decision to continue testing. If the expected risk of the decision to continue testing is smaller than the expected loss of a master or a nonmaster decision, testing will be continued. The expected risk of the decision to continue testing is defined as follows.

Let $\{\mathbf{w}_{s+1} \mid \mathbf{w}_s\}$ be the range of $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$. Then, for $s = 1, \ldots, S-1$, the expected risk of the decision to continue testing is defined as

$$E(R(\mathbf{w}_{s+1}) \mid \mathbf{w}_s) = \sum_{\{\mathbf{w}_{s+1} \mid \mathbf{w}_s\}} R(\mathbf{w}_{s+1}) p(\mathbf{w}_{s+1} \mid \mathbf{w}_s), \tag{21.12}$$

where the so-called posterior predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$ is given by

$$p(\mathbf{w}_{s+1} \mid \mathbf{w}_s) = \int, \ldots, \int p(\mathbf{w}_{s+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{w}_s) \partial \boldsymbol{\theta}, \tag{21.13}$$

and the risk at stage $s+1$ is inductively defined as

$$\begin{aligned} R(\mathbf{w}_{s+1}) = \min\{&E(L(m,\boldsymbol{\theta}) \mid \mathbf{w}_{s+1}), \\ &E(L(n,\boldsymbol{\theta}) \mid \mathbf{w}_{s+1}), E(R(\mathbf{w}_{s+2}) \mid \mathbf{w}_{s+1})\}. \end{aligned} \tag{21.14}$$

The risk associated with the last testlet is defined as

$$R(\mathbf{w}_S) = \min\{E(L(m,\boldsymbol{\theta}) \mid \mathbf{w}_S), E(L(n,\boldsymbol{\theta}) \mid \mathbf{w}_S)\}. \tag{21.15}$$

So, given an observation $\mathbf{w}_s$, the expected distribution of $\mathbf{w}_{s+1}, \mathbf{w}_{s+2}, \ldots, \mathbf{w}_S$ is generated and an inference about future decisions is made. Based on these inferences, the expected risk of continuation as defined in Equation (21.12) is computed and compared with the expected losses of a mastery or nonmastery decision. If the risk of continuation is smaller than these two expected losses, testing is continued. If this is not the case, the classification decision with the smallest expected loss is made.

Notice that the definitions (21.12) – (21.15) imply a recursive definition of the expected risk of continuation. Computation of the expected risk of the decision to continue testing is done by techniques of dynamic programming (i.e., backward induction; see, for instance, Bellman, 1957; DeGroot, 1970; Ferguson, 1967; Lindgren, 1976; Winston, 1994). First, the risk of the last testlet in the sequence, which was labeled testlet S, is computed for all possible values of $\mathbf{w}_S$. Then the posterior predictive distribution $p(\mathbf{w}_S \mid \mathbf{w}_{S-1})$ is computed using formula (21.13), followed by the expected risk $E(R(\mathbf{w}_S) \mid \mathbf{w}_{S-1})$ defined in formula (21.12). This, in turn, can be used for computing the risk $R(\mathbf{w}_{S-1})$, for all $\mathbf{w}_{S-1}$, using formula (21.14), and this iterative process continues until s is reached and the decision can be made whether to administer testlet $s+1$, or to decide on mastery or nonmastery.

In the Bayesian principle outlined here, it is assumed that prior knowledge about the student's proficiency level can be characterized by a prior distribution, say a Q-variate normal distribution $g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$. This prior distribution represents our best prior beliefs concerning the student's proficiency before any testlet has been administered. The prior might be specified as either an empirical (i.e., empirical Bayes approach) or subjective prior. In the first approach (e.g., Robbins, 1964), empirical data from other students in the group to which the individual student belongs (i.e., "comparable group") are used as collateral data. The obvious approach is to use the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ obtained with the estimates of the item parameters. Further, van der Linden (1999) presented an empirical initialization procedure, where the prior is enhanced by collateral background information on the student. That is, the prior $g(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is enhanced with the linear regression model $\boldsymbol{\mu} = \boldsymbol{X\beta}$, where $\mathbf{X}$ are the student's values on a number of background variables and $\boldsymbol{\beta}$ are regression coefficients. This approach is closely related to the approach for implying collateral information in item calibration suggested by Mislevy (1988). In the second approach, using a subjective prior, prior knowledge about the student's proficiency is specified by subjective assessment. Although the precise stipulation of prior beliefs is not easy to accomplish, fortunately, extensive aids are available to help a decision maker specify a prior distribution (e.g., Novick & Jackson, 1974).

21.6 The Compound Multidimensional Rasch Model

The theory presented thus far is applicable to the broad class of IRT models defined above. The theory of adaptive sequential mastery testing will now be worked out in detail for a special case of the general model, the so-called compound multidimensional Rasch model (Glas, 1992). The model was further elaborated by Adams, Wilson, and Wu (1997) and Adams, Wilson, and Wang (1997), and can be estimated using the computer program ConQuest (Wu, Adams & Wilson, 1997). In this model, it is assumed that the complete test, or, in the present case, the complete testlet, consists of Q subtests, where every subtest relates to a specific proficiency θ_q, $q = 1,\ldots,Q$. Further, it is assumed that the ensemble of person parameters $\theta_1,\ldots,\theta_Q$ has a Q-variate normal distribution with a mean equal to zero and a covariance matrix $\boldsymbol{\Sigma}$.

Given $\theta_1,\ldots,\theta_Q$, the probability of a response pattern $\mathbf{u}_1,\ldots,\mathbf{u}_Q$ is given by

$$\begin{aligned} p(\mathbf{u}_1,\ldots,\mathbf{u}_Q \mid \theta_1,\ldots,\theta_Q) &= \prod_{q=1}^{Q}\prod_{i=1}^{K_q}\frac{\exp(u_{qi}(\theta_q - b_{qi}))}{1+\exp(\theta_q - b_{qi})} \\ &= \prod_{q=1}^{Q}\exp(t_q\theta_q)\exp(-\mathbf{u}_q'\mathbf{b}_q)P_{q0}(\theta_q), \quad (21.16) \end{aligned}$$

where $\mathbf{b}_q = (b_{1q},\ldots,b_qK_q)'$ is a vector of item parameters, $\mathbf{u}_q'\mathbf{b}_q$ is the inner product of $\mathbf{u}$ and $\mathbf{b}_q$, t_q is the sum score $t_q = \sum_i u_{qi}$, and

$$P_{q0}(\theta_q) = \prod_{i=1}^{K_q}(1+\exp(\theta_q - b_{qi}))^{-1}. \quad (21.17)$$

Notice that t_q is the minimal sufficient statistic for θ_q. Further, it is easily verified that $P_{q0}(\theta_q)$ is the probability, given θ_q, of a response pattern with all item responses equal to zero. The probability of observing t_q, given θ_q, is given by

$$\begin{aligned} p(t_q \mid \theta_q) &= \sum_{\{\mathbf{u}_q|t_q\}} p(\mathbf{u}_q \mid \theta_q) \\ &= \sum_{\{\mathbf{u}_q|t_q\}} \exp(t_q\theta_q - \mathbf{u}_q'\mathbf{b_q})P_{q0}(\theta_q) \\ &= \gamma_{t_q}(\mathbf{b}_q)\exp(t_q\theta_q)P_{q0}(\theta_q), \end{aligned}$$

where $\gamma_{t_q}(\mathbf{b}_q)$ is the so-called elementary symmetric function defined by

$$\gamma_{t_q}(\mathbf{b}_q) = \sum_{\{\mathbf{u}_q|t_q\}} \exp(-\mathbf{u}_q'\mathbf{b_q}),$$

and where $\{\mathbf{u}_q \mid t_q\}$ stands for the set of all possible response patterns resulting in a sum score t_q.

Given $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_Q)$, the probability of a response pattern $\mathbf{u} = (\mathbf{u}_1,\ldots,\mathbf{u}_Q)$ is given by

$$\begin{aligned} p(\mathbf{u} \mid \boldsymbol{\theta}) &= \prod_{q=1}^{Q}\exp(t_q\theta_q)\exp(-\mathbf{u}_q'\mathbf{b}_q)P_{q0}(\theta_q) \\ &= \exp(\mathbf{t}'\boldsymbol{\theta})\exp(-\mathbf{u}'\boldsymbol{b})P_0(\boldsymbol{\theta}), \end{aligned}$$

where $\mathbf{b} = (\mathbf{b}_1, \ldots, \mathbf{b}_Q)$ is a vector of item parameters, $\mathbf{t} = (t_1, \ldots, t_Q)$, and

$$P_0(\boldsymbol{\theta}) = \prod_{q=1}^{Q} P_{q0}(\theta_q).$$

The probability of observing $\mathbf{t}$, given $\boldsymbol{\theta}$, is given by

$$p(\mathbf{t} \mid \boldsymbol{\theta}) = \Gamma_{\mathbf{t}}(\mathbf{b}) \exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta})$$

where $\Gamma_{\mathbf{t}}(\mathbf{b})$ is the product of the elementary symmetric functions $\gamma_{t_q}(\mathbf{b}_q)$ for $q = 1, \ldots, Q$. Below, $\Gamma_{\mathbf{t}}(\mathbf{b})$ will be referred to as a compound elementary symmetric function.

Usually, the prior $\boldsymbol{\theta}$ is standard normal, so let $g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})$ be the normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Then

$$p(\boldsymbol{\theta} \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})}{p(\mathbf{t})} = \frac{\exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma})}{\int, \ldots, \int \exp(\mathbf{t}'\boldsymbol{\theta}) P_0(\boldsymbol{\theta}) g(\boldsymbol{\theta} \mid \boldsymbol{\Sigma}) \partial\boldsymbol{\theta}}.$$

Notice that $\Gamma_{\mathbf{t}}(\mathbf{b})$ cancels from the numerator and the denominator.

Applying the general framework of the previous section to the Rasch model boils down to choosing the minimal sufficient statistics for $\boldsymbol{\theta}$, that is, the unweighted sum scores for the statistics $\mathbf{w}_s$. So let t_{sq} be the score pattern on the qth subtest for the sth occasion. Further, define $\mathbf{r}_s$ as a Q-vector with elements $r_{sq} = \sum_{d=1}^{s} t_{dq}$. Let $p(\boldsymbol{\theta} \mid \mathbf{r}_s)$ stand for the posterior density of proficiency given $\mathbf{r}_s$. Then the expected losses (21.10), (21.11) and the expected risk (21.12) can be written as $E(L(m, \boldsymbol{\theta}) \mid \mathbf{r}_s)$, $E(L(n, \boldsymbol{\theta}) \mid \mathbf{r}_s)$, and $E(R(\mathbf{r}_{s+1}) \mid \mathbf{r}_s)$, respectively. More specifically, the expected risk is given by

$$E(R(\mathbf{r}_{s+1}) \mid \mathbf{r}_s) = \sum_{\mathbf{r}_{s+1} \mid \mathbf{r}_s} p(\mathbf{r}_{s+1} \mid \mathbf{r}_s) R(\mathbf{r}_{s+1}), \tag{21.18}$$

where the summation is over all scores $\mathbf{r}_{s+1}$ compatible with $\mathbf{r}_s$.

Defining $\mathbf{z}_{s+1} = \mathbf{r}_{s+1} - \mathbf{r}_s$, the posterior predictive distribution defined in Equation (21.13) specializes to

$$\begin{aligned} p(\mathbf{r}_{s+1} \mid \mathbf{r}_s) &= \int, \ldots, \int p(\mathbf{r}_{s+1} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{r}_s) \partial\boldsymbol{\theta} \\ &= \int, \ldots, \int \Gamma_{\mathbf{z}_s+1} \exp(\mathbf{z}'_{s+1}\boldsymbol{\theta}) P_{0(s+1)}(\boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{r}_s) \partial\boldsymbol{\theta}, \end{aligned} \tag{21.19}$$

where $\Gamma_{\mathbf{z}_{s+1}}$ is a shorthand notation for the compound elementary symmetric function of the item parameters of occasions $s + 1$ and $P_{0(s+1)}(\boldsymbol{\theta})$ is equal to Equation (21.17) evaluated using the item parameters of test $s + 1$. That is, $P_{0(s+1)}(\boldsymbol{\theta})$ is equal to the probability of a zero-response pattern on test $s + 1$, given $\boldsymbol{\theta}$.

21.7 Simulation Studies

Simulation studies were designed to investigate the following research questions: (1) What is the performance, in terms of average loss, of MIRT-based sequential mastery testing as a function of the number of items administered per stage? (2) What are the effects on average loss when turning the sequential procedure into an adaptive sequential procedure? (3) In the case of a compensatory loss function, how is average loss influenced when the multidimensional structure is ignored and a unidimensional IRT model is used? The latter research question has two motivations, one that is quickly losing importance and one that will remain important. First, in the case of a multidimensional model, backward induction is a computationally intensive process, while the computing times in the unidimensional case of SMT and in traditional AMT are by now negligible. In the simulation studies below, using a 1700Hz Pentium III, the average time to select an item was about 3 seconds. The maximum was approximately 7 seconds. However, as said, with the emergence of faster and faster computers, these considerations rapidly lose their significance. More important is that the definition of the loss function in the unidimensional case is conceptually much simpler than in the multidimensional case, so, depending on the application and the impact of the decisions made using the test, considering a unidimensional approximation seems worthwhile.

21.7.1 Compensatory Loss Functions

For all simulations pertaining to compensatory loss functions, a three-dimensional compound Rasch model was used. The parameters of the loss function were $(A_1, A_2, A_3) = (-1, -1, -1)$ and $(B_1, B_2, B_3) = (1, 1, 1)$, while the cost of administering one item was set equal to 0.02. The vector of cut-off points was set equal to $\boldsymbol{\theta}_c = \mathbf{0}$. In the studies, the following aspects were varied:

1. The correlation between the latent dimensions. The three-dimensional compound Rasch model was simulated in two conditions: a high-homoge-neity condition where the correlation among all three dimensions was $\rho = 0.80$ and a low-homogeneity condition where this correlation was $\rho = 0.40$. The variance was equal to one for all three dimensions.
2. The test administration design. In the test procedure 27 items could be delivered. These items could be delivered as a fixed test of 27 items, or in a sequential design with 3 stages with 9 items per stage, 9 stages of 3 items, and 27 stages of one item.
3. The test administration mode, which was either sequential or adaptive sequential.

The motivation for the choice of a correlation $\rho = 0.80$ was that it might be expected that a unidimensional approximation might work quite well for this high-homogeneity condition. The correlation $\rho = 0.40$ was chosen because the real data example reported below shows that such a low correlation between subtests is realistic.

For the sequential procedure, the item difficulties b_i were drawn from a standard normal distribution. Further, the items were evenly distributed over the three proficiency dimensions; that is, a third of the items loaded on the first dimension, a third on the second, and a third on the third dimension. Finally, also within a stage, the items were evenly distributed over the three dimensions, with the exception of the one-item stages, where items alternately loaded on a dimension. To avoid capitalization on chance, the item parameters were redrawn in every replication.

For the adaptive sequential mode, a testlet bank was generated in such a way that it could be expected that it supported selection of testlets with differential optimal measurement properties. For the procedures with 3 and 9 stages, the following procedure was adopted:

1. Define the grid $\{\mathbf{h}\} = \{h_1, h_2, h_3\} = \{h(i), h(j), h(k) | i, j, k = 1, \ldots, 5, h(n) = -1.0 + 0.5(n-1)\}$. Notice that this grid has 5^3, that is, 125, points.
2. For each point $\mathbf{h} \in \{\mathbf{h}\}$, draw 3 item difficulties from the multivariate normal distribution with a mean $\mathbf{h}$ and a covariance matrix equal to $0.2\mathbf{I}$. Each item is assumed to load on a different dimension. This is repeated 3 times for each point $\mathbf{h} \in \{\mathbf{h}\}$, so the total number of item parameters is equal to $125 \times 3 \times 3$, which is 1,125. For the procedure with 3 stages, the 9 items form one testlet, for the procedure with 9 stages, three testlets of 3 items are formed.

For the design of 27 stages of one item each, 375 item difficulties were drawn from the standard normal distribution for each proficiency dimension, so also in this case, the item bank consisted of 1,125 items. Also, for the adaptive mode, the item difficulties were redrawn in every replication to avoid capitalization on chance.

The choice of a criterion for adaptive testlet selection in a multidimensional framework is more complicated that in a unidimensional framework. In a undimensional framework (previous chapter), maximum information at the cut-off point and at the expected a posteriori estimate (EAP estimate) of proficiency were used as testlet selecton criteria. In the multidimensional framework, these two criteria are less plausible. In one dimension, both the running estimate of proficiency and the cut-off point are on the same continuum, and any test with high information between these two points will be informative for the decision that has to be made. In a multidimensional framework, the test taker's proficiency is a point in Q-dimensional space and the boundary between masters and nonmasters becomes a line in two dimensions, or a linear manifold in more than two dimensions. Therefore, in this case the relationship between the position of the test taker in the support of the loss function and the optimal testlet will be much more complicated, and it remains a point of further study.

As an alternative, another criterion also studied in the previous chapter will be used. This is motivated by the fact that one is primarily interested in minimizing possible losses due to misclassifications. In the sequential procedure, the decision is based on comparing $L(m, \boldsymbol{\theta})$ and $L(n, \boldsymbol{\theta})$. If, for every possible follow-up testlet $s+1$, the observation $\mathbf{w}_{s+1}$ is available, a natural choice for the follow-up test is the testlet where the posterior variance of the difference between $L(m, \boldsymbol{\theta})$ and $L(n, \boldsymbol{\theta})$, say var $(L(m, \boldsymbol{\theta}) - L(n, \boldsymbol{\theta}) \mid \mathbf{w}_{s+1})$, was minimal. However, the observation $\mathbf{w}_{s+1}$

is not yet available, so a prediction must be made of the likelihood of $\mathbf{w}_{s+1}$. This likelihood is obtained via the predictive distribution $p(\mathbf{w}_{s+1} \mid \mathbf{w}_s)$. So if $\{\mathbf{w}_{s+1}|\mathbf{w}_s\}$ is the set of all possible values $\mathbf{w}_{s+1}$ given $\mathbf{w}_s$, the criterion for selection of the next testlet becomes

$$\sum_{\{\mathbf{w}_{s+1}|\mathbf{w}_s\}} \mathrm{var}(L(m,\boldsymbol{\theta}) - L(n,\boldsymbol{\theta}) \mid \mathbf{w}_{s+1}) p(\mathbf{w}_{s+1} \mid \mathbf{w}_s); \qquad (21.20)$$

that is, a testlet is chosen such that the expected variance of the difference between the losses of the mastery and nonmastery decision is minimal. In the study on the unidimensional case (previous chapter) the performance of the three selection criteria was comparable, with a slight advantage for the procedure based on maximum information at the cut-off point.

The results of the simulation studies for the compensatory model are reported in Table 21.1. Thousand replications were made for each condition. For every replication a true proficiency $\boldsymbol{\theta}$ was drawn from the multivariate normal distribution. At the end of every replication, loss was computed using the true proficiency value. In Table 21.1, it can be seen that the mean loss decreased with the number of items in a testlet. This decrease can be attributed to a decrease in the number of items given. The proportion of correct decisions slightly decreased. It can be seen that using an adaptive testlet selection procedure further decreased mean loss, but this decrease was far less important than the decrease attributable to decrease in the testlet size. These findings are analogous to the findings for the unidimensional case (previous chapter). The losses in the condition with $\rho = 0.40$ are systematically larger than in the condition where $\rho = 0.80$. This is explained by the fact that in the case of a homogeneous item pool, item responses are informative with respect to all proficiency

Table 21.1 Relationship between selection method and mean loss in the compensatory model

ρ	Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
0.80	1	27	Fixed Test	0.87	1.00	0.6347
	3	9	Sequential	0.82	0.41	0.4078
	3	9	Adaptive	0.82	0.41	0.4077
	9	3	Sequential	0.82	0.29	0.3227
	9	3	Adaptive	0.82	0.25	0.3208
	27	1	Sequential	0.79	0.25	0.3446
	27	1	Adaptive	0.80	0.22	0.3060
0.40	1	27	Fixed Test	0.85	1.00	0.6501
	3	9	Sequential	0.80	0.42	0.4055
	3	9	Adaptive	0.80	0.36	0.3969
	9	3	Sequential	0.77	0.31	0.3938
	9	3	Adaptive	0.78	0.22	0.3699
	27	1	Sequential	0.81	0.29	0.3470
	27	1	Adaptive	0.80	0.21	0.3288

Table 21.2 Relationship between selection method and mean loss when multidimensionality is ignored in the compensatory model

ρ	Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
0.80	1	27	Fixed Test	0.81	1.00	0.6985
	3	9	Sequential	0.81	0.41	0.4248
	3	9	Adaptive	0.81	0.43	0.4138
	9	3	Sequential	0.77	0.28	0.4074
	9	3	Adaptive	0.80	0.27	0.3457
	27	1	Sequential	0.80	0.27	0.3721
	27	1	Adaptive	0.80	0.24	0.3295
0.40	1	27	Fixed Test	0.76	1.00	0.8200
	3	9	Sequential	0.73	0.40	0.5781
	3	9	Adaptive	0.73	0.43	0.5017
	9	3	Sequential	0.70	0.29	0.4838
	9	3	Adaptive	0.75	0.27	0.4484
	27	1	Sequential	0.76	0.27	0.4023
	27	1	Adaptive	0.71	0.23	0.4429

dimensions, while in the heterogeneous case, item responses are mainly informative with respect to the proficiency on which they load.

In Table 21.2, the results are given for the conditions where the multidimensional proficiency structure is ignored in the computations supporting the sequential and adaptive sequential procedures. In these conditions, response behavior was simulated and the final mean losses were computed using the "true" item and "true" multidimensional proficiency parameters, while the computations supporting the sequential and adaptive sequential procedure were made using a standard unidimensional Rasch model with the "true" item difficulties b_i and unidimensional standard normally distributed proficiency parameters. So this unidimensional approximation of multidimensional response behavior is based on the assumption that the correlation among the latent abilities is equal to one, i.e., $\rho = 1.0$. Therefore, in the unidimensional case, the losses defined in equations (21.6) and (21.7) were computed using $\theta_1 = \theta_2 = \theta_3 = \theta$, where θ has a standard normal prior, and $\theta_{c1} = \theta_{c2} = \theta_{c3} = \theta_c = 0$.

It can be seen that, in general, the mean losses were higher than the analogous losses in Table 21.2, but the increasing loss remained limited: the maximum increase in loss was found in the condition with a fixed test and $\rho = 0.40$. This decrease was due to a 9% decrease in correct decisions. Overall, the approximation in the case where $\rho = 0.40$ is slightly worse than in the case where $\rho = 0.80$. It can be concluded that the unidimensional approximation based on the assumption $\rho = 1.0$ worked reasonably well. An important exception was the case of adaptive testlet selection with 27 testlets of one item each. In that case, the average loss for the adaptive sequential procedure became higher than the average loss in the nonadaptive sequential testlet selection procedure. So there the combination of a unidimensional approximation of proficiency with the circumstance that the testlets only loaded on one proficiency dimension resulted in a relatively poor performance.

21.7.2 Conjunctive Loss Functions

For all simulations pertaining to conjunctive loss functions, a two-dimensional compound Rasch model was used. The parameters of the loss function were $A_1 = A_2 = -0.5$, $A_3 = A_4 = 0.1$, and $B_1 = B_2 = 1.0$. The cost of administering one item was set equal to 0.01 and the vector of cut-off points was set equal to $\boldsymbol{\theta}_c = \mathbf{0}$.

In the studies, the following aspects were varied:

1. The correlation between the latent dimensions, which was either $\rho = 0.80$ or $\rho = 0.40$. The variance on both dimensions was equal to one.
2. The test administration design. In the test procedure, 32 items could be delivered. These items could be delivered as a fixed test of 32 items, or in a sequential design with 4 stages with 8 items per stage, 8 stages of 4 items, or 32 stages of one item.
3. The test administration mode, which was either sequential or adaptive sequential.

For the sequential procedure, the item difficulties b_i were drawn from a standard normal distribution. Further, the items were evenly distributed over the two proficiency dimensions; that is, half of the items loaded on the first dimension and half loaded on the second dimension. Finally, also within a stage, the items were evenly distributed over the two dimensions, with the exception of the one-item stages, where items alternately loaded on a dimension. The item parameters were redrawn in every replication. For the adaptive sequential mode, a testlet bank was generated in such a way that it could be expected to support the selection of testlets with differential optimal measurement properties. For the design of 32 stages of one item each, this was simply translated into drawing 200 item difficulties for each proficiency dimension from the standard normal distribution. For the procedures with 32, 8, and 4 stages, the following procedure was adopted:

1. Define the grid $\{\mathbf{h}\} = \{h_1, h_2\} = \{h(i), h(j) | i, j = 1, \ldots, 5, h(n) = -1.0 + 0.5(n-1)\}$. Notice that this grid has 5^2, that is, 25 points.
2. For each point $\mathbf{h} \in \{\mathbf{h}\}$, draw 2 item difficulties from the multivariate normal distribution with a mean $\mathbf{h}$ and a covariance matrix equal to $0.2\mathbf{I}$. Each item is assumed to load on a different dimension. This is repeated 8 times for every point $\mathbf{h} \in \{\mathbf{h}\}$, so the total number of item parameters equals 25 × 2 × 8, which is 400. For the procedure with 4 stages, the 16 items form two testlets, for the procedure with 8 stages, four testlets of 4 items are formed.

The testlet selection criterion was the same as in the compensatory case; that is, at each stage the testlet was selected that minimized the criterion defined by expression (21.20). Also, for the adaptive mode, the item difficulties were redrawn in every replication.

Table 21.3 Relationship between selection method and mean loss in the conjunctive model

ρ	Number of Testlets	Items per Testlet	Selection Method	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
0.80	1	32	Fixed Test	0.85	1.00	0.3549
	4	8	Sequential	0.82	0.30	0.1475
	4	8	Adaptive	0.80	0.26	0.1396
	8	4	Sequential	0.78	0.22	0.1306
	8	4	Adaptive	0.80	0.21	0.1302
	32	1	Sequential	0.79	0.20	0.1277
	32	1	Adaptive	0.80	0.20	0.1270
0.40	1	32	Fixed Test	0.80	1.00	0.3999
	4	8	Sequential	0.81	0.30	0.1765
	4	8	Adaptive	0.81	0.24	0.1588
	8	4	Sequential	0.81	0.23	0.1570
	8	4	Adaptive	0.82	0.20	0.1377
	32	1	Sequential	0.80	0.19	0.1375
	32	1	Adaptive	0.81	0.19	0.1373

In Table 21.3, it can be seen that the main effects of the correlation between the dimensions and the number of testlets are analogous to the results for the compensatory model. That is, average loss goes up as the correlation goes down, average loss decreases with the length of the testlets, and there are (relatively small) positive effects on average loss of adaptive testlet selection.

21.8 An Empirical Example

The purpose of this empirical example is to demonstrate the feasibility of the procedure in a real situation and to show the effects of varying the cost parameter. At the end of secondary education in the Netherlands, students participate in central examinations. The grade level they achieve is an important component of the grade level of their certificate. After the examinations are administered, the items belong to the public domain and can be used as practice material or for diagnostic tests to support the educational process. Testing can involve both financial costs (cost of test delivery) and educational costs (time lost testing). The latter cost factor is related to the objective not to overburden the educational process with testing. This example concerns the English Language Comprehension Examination 2000 (at HAVO-level). The test consisted of 45 forced response items and the calibration sample consisted of 1,801 examinees.

The items were calibrated with the multidimensional Rasch model using a method proposed by Béguin and Glas (2001). The top-down procedure starts with deleting items from the complete item set until the remaining set of items forms a unidimensional scale according to some criterion. Then the process is reiterated with the set of deleted items,and so forth, until all, or at least most of the items are scaled. Finally, these subscales are combined using the multidimensional Rasch model. In the present application, the criterion for definition of the subscales consisted of two test statistics: the R_{1c}-statistic by Glas (1988) and the S_i-statistic by Glas and Verhelst (1995). The first statistic is item-oriented and can be used for item selection, and the second is a global test statistic and can be used as a criterion to stop the selection process. Both test statistics have an asymptotic χ^2-distribution.

For the present example, the fit of the unidimensional Rasch model was quite poor ($R1c = 796.58$; df $= 132$; p $= 0.00$). However, 40 of the 45 items proved scalable using three dimensions. The first scale consisted of 19 items ($R1c = 56.18$; df $= 54$; p $= 0.39$), the second of 13 items ($R1c = 31.57$; df $= 36$; p $= 0.68$) and the third of 8 items ($R1c = 29.46$; df $= 21$; p $= 0.10$). The correlations between the dimensions were 0.509, 0.588, and 0.475, respectively. (These results are not atypical; experience with other examination topics and other examination years generally shows that the unidimensional Rasch model is rejected and that most of the items can be modeled into 3 to 5 dimensions). Table 21.4 shows the distribution of the items over the dimensions, the item parameter estimates and their standard errors, and the values of the S_i-statistics. All estimates and tests were computed using the OPLM computer program (Verhelst, Glas & Verstralen, 1995).

Using these parameter estimates, an SMT procedure was simulated using a compensatory model. As above, the parameters of the loss function were $(A_1, A_2, A_3) = (-1, -1, -1)$ and $(B_1, B_2, B_3) = (1, 1, 1)$. The cut-off point was set equal to $\boldsymbol{\theta}_c = \mathbf{0}$. Three values of the cost of administering an item were used: $C = 0.01$, 0.02 and 0.10. These settings were crossed with four testlet administration designs: one test of all 40 items; 5 testlets of 8 items; 8 testlets of 5 items; and 40 testlets of one item each. The person parameters were drawn from a multivariate normal distribution with a covariance matrix as estimated in the calibration phase and 200 replications were thus made. The operating characteristics of the SMT procedure are shown in Table 21.5. As expected, it can be seen that the proportion of testlets administered decreases as the costs go up. Consequently, the proportions of correct classifications go down. Further, also in the present case, the mean loss decreases with the number of testlets in the procedure.

21.9 Conclusions and Further Research

A general theoretical framework for nonadaptive and adaptive sequential testing based on a combination of Bayesian sequential decision theory and multidimensional IRT was presented. This framework was applied to the compound Rasch model. In this model, it is assumed that the test items can be split up into a number of subsets related to specific proficiency dimensions and the relationship between the dimensions is modeled by a covariance structure. Using this model, a number of simulation studies were performed that showed that augmentation of the number of stages in a sequential mastery procedure resulted in a marked decrease in average

Table 21.4 Item calibration for the Examination of English Language Comprehension

Item	Dimension	b_i	$Se(b_i)$	S_i	df	p
1	1	−0.419	0.062	5.67	5	0.34
2	3	0.473	0.056	1.98	3	0.57
5	3	−1.045	0.077	8.05	3	0.04
7	2	0.072	0.047	5.19	6	0.51
8	1	−0.219	0.059	10.89	5	0.05
9	1	0.061	0.056	1.48	5	0.91
10	1	0.768	0.051	6.77	5	0.23
11	1	0.048	0.056	3.52	5	0.62
12	1	1.061	0.050	4.73	5	0.44
13	3	−0.129	0.061	5.79	3	0.12
14	2	−0.826	0.051	6.03	6	0.41
15	2	−0.881	0.052	1.58	6	0.95
16	1	−0.106	0.058	2.50	5	0.77
18	1	1.927	0.053	11.63	5	0.04
19	2	0.527	0.048	3.74	6	0.71
20	1	1.265	0.050	5.57	5	0.35
21	2	−0.225	0.048	3.39	6	0.75
22	1	−0.037	0.057	5.88	5	0.31
23	2	1.446	0.055	3.84	6	0.69
24	3	−0.415	0.065	1.09	3	0.77
25	2	0.094	0.047	2.20	6	0.90
26	1	−2.578	0.139	1.60	3	0.65
27	1	−0.510	0.064	6.06	5	0.30
28	2	−0.179	0.047	5.37	6	0.49
29	2	0.918	0.050	2.71	6	0.84
30	2	−0.237	0.048	8.94	6	0.17
32	3	0.863	0.054	4.14	3	0.24
33	1	−0.694	0.067	2.71	5	0.74
34	1	0.159	0.055	6.17	5	0.29
35	1	−0.395	0.062	4.78	5	0.44
36	3	0.767	0.054	6.08	3	0.10
37	3	−0.367	0.064	3.18	3	0.36
38	1	0.944	0.050	6.31	5	0.27
39	3	−0.146	0.061	2.95	3	0.39
40	1	−0.219	0.059	6.24	5	0.28
41	2	−0.852	0.052	8.42	6	0.20
42	1	0.481	0.052	8.91	5	0.11
43	2	0.091	0.047	2.81	6	0.83
44	1	−1.538	0.090	3.37	4	0.49
45	2	0.049	0.047	6.63	6	0.35

loss. Moving to adaptive sequential mastery testing further reduced average loss, but the effect was far less important than the effect of a nonadaptive sequential procedure. For the compensatory model, the results of the simulation studies showed

Table 21.5 Relationship between cost and operating characteristics

Cost per Item	Number of Testlets	Items per Testlet	Proportion Correct Decisions	Proportion Testlets Given	Mean Loss
0.02	1	40	0.85	1.00	0.8892
	5	8	0.82	0.24	0.4693
	8	5	0.82	0.23	0.3245
	40	1	0.80	0.20	0.3274
0.04	1	40	0.85	1.00	1.6736
	5	8	0.79	0.22	0.5886
	8	5	0.77	0.16	0.4489
	40	1	0.76	0.11	0.4410
0.10	1	40	0.85	1.00	4.0927
	5	8	0.76	0.20	1.0941
	8	5	0.69	0.12	0.8664
	40	1	0.64	0.06	0.6468

that ignoring the multidimensional structure and using a unidimensional approximation resulted in an increase in average losses, but the increase was not dramatic. An exception was adaptive sequential testing with only one item per testlet and a low correlation of the proficiency dimensions. In that case, the average loss was higher than in the analogous case without adaptive item selection.

For the sake of simplicity, it was assumed above that the cost parameter C is constant. However, exposure control and content balancing in ASMT can be handled by the introduction of differential cost parameters for testlets. Further, the cost may also depend on θ, or on the stage s. The generalization to these cases is straightforward. Define $C_{ts'}(\boldsymbol{\theta})$ as the cost of delivering a testlet t at stage s' to a student with proficiency $\boldsymbol{\theta}$. Then the loss function generalizes to $L(d, s, t, \boldsymbol{\theta}) = \max\{\sum_{s'}^{s} C_{ts'}(\boldsymbol{\theta}), \sum_{s'}^{s} C_{ts'}(\boldsymbol{\theta}) + f(d, \boldsymbol{\theta})\}$, where d can assume the values m and n, and $f(d, \boldsymbol{\theta})$ is as defined for the decision d in either the compensatory or the conjunctive model. Since the definition of expected loss and risk [formulas (21.10) – (21.19)] does not depend on the actual form of the loss functions, introduction of this generalization does not present any problems. In fact, also the assumption of linear loss is not essential for the derivations; the procedure easily generalizes to other frequently used loss functions, such as threshold, exponential, quadratic, and normal-ogive loss (refer to Novick & Lindley, 1979).

Several issues may lead to further research. First, the computation of the multiple integrals is done using Gauss–Hermite quadrature, which becomes very time-consuming when more than three dimensions are involved (see, for instance, Glas, 1992). Therefore, problems of higher dimensionality will need other methods, such as simulation methods, for the evaluation of the multiple integrals.

Second, many IRT models, such as, for instance, the "Full Information Factor Analysis" model by Bock, Gibbons, and Muraki (1988), have no sufficient statistics for $\boldsymbol{\theta}$ and will need alternative choices for $\mathbf{w}_s = f(\mathbf{u}_1, \ldots, \mathbf{u}_s)$. In the previous chapter, for the unidimensional 3PLM, it was shown that using unweighted sum scores results in a feasible procedure that produces acceptable results. A generalization to a multidimensional framework would probably be based on a Q-dimensional vector of partial sum scores, but this remains a point of further study.

Another direction for possible further research has to do with the following. A distinction was made here between compensatory and conjunctive-disjunctive mastery models. This distinction is also relevant for IRT models. For instance, the compound Rasch model can be viewed as a compensatory model because the probability of a correct response is based on the sum of the proficiencies on the different dimensions. Therefore, a higher proficiency on one of the dimensions compensates for a lower proficiency on one of the other dimensions. Compensatory IRT models (McDonald, 1967; Lord & Novick, 1968; Reckase, 1985; Glas, 1992; Ackerman, 1996a, 1996b) are by far the most commonly used, but noncompensatory alternatives have been developed (Sympson, 1978; Embretson, 1980, 1984; Ackerman, 1987; Spray, Davey, Reckase, Ackerman & Carlson, 1990; Maris, 1993, 1995). In these models, the probability of a correct response is based on a product of the proficiencies on the different dimensions. Consequently, a low proficiency on one of the dimensions cannot be compensated for by a high proficiency on one of the other dimensions. The status of the loss function and the IRT model in the theory presented above is quite different. The choice of a compensatory or a noncompensatory IRT model is an empirical matter and determined by which IRT model fits the data best. The choice of a compensatory or noncompensatory loss function is a value judgment determined by the opinion of who can be qualified as a master and the judgment of the relative losses due to incorrect classification decisions. Therefore, a combination of the loss functions considered above and noncompensatory IRT models might also be a point for further research.

References

Ackerman, T. A. (1987). *A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data* (ACT Research Report Series 87-12). Iowa City, IA: ACT Inc.

Ackerman, T. A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20,* 309–310.

Ackerman, T. A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–329.

Adams, R. J., Wilson, M. R. & Wang, W. C. (1997). The random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–25.

Adams, R. J., Wilson, M. R. & Wu, M. (1997). Multilevel item response theory models: an approach to errors in variables of regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50,* 3–16.

Béguin, A. A. & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541–562.

Bellman, R. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R. D. & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bradlow, E. T., Wainer, H. & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.

Coombs, C. H. (1960). *A theory of data.* Ann Arbor, MI: Mathesis Press.

Coombs, C. H. & Kao, R. C. (1955). *Nonmetric factor analysis.* [Engng. Res. Bull., No.38]. Ann Arbor, MI: University of Michigan Press.

DeGroot, M. H. (1970). *Optimal statistical decisions.* New York: McGraw-Hill.

Embretson, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479–494.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175–186.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach.* New York: Academic Press.

Fraser, C & McDonald, R. P. (1988). NOHARM: Least Squares item factor analysis. *Multivariate Behavioral Research, 23,* 267–269.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of proficiency. In M. R. Wilson (Ed.), *Objective measurement: Theory into practice, Vol. 1* (pp.236–258). Norwood, NJ: Ablex Publishing Corporation.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. In G. H.Fischer & I. W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications* (pp.69–96). New York, NJ: Springer-Verlag.

Glas, C. A. W., Wainer, H. & Bradlow, E. T. (2000). MML and EAP estimates for the testlet response model. In W. J. van der Linden & C. A. W.Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp.271–287). Boston: Kluwer-Nijhoff Publishing.

Lewis, C. & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement, 14,* 367–386.

Lindgren, B. W. (1976). *Statistical theory* (3rd ed.). New York: Macmillan.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389–404.

Luecht, R. M. & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35*, 239–249.

Luecht, R. M. & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden & C. A. W.Glas (eds.). *Computerized adaptive testing: Theory and practice.* (pp. 117–128). Boston: Kluwer-Nijhof Publishing.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika, 58,* 445–470.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–548.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric monographs, No. 15.*

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden and R. K. Hambleton (Eds.). *Handbook of Modern Item Response Theory* (pp.257–269). New York: Springer-Verlag.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12,* 281–296.

Novick, M. R. & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York, NJ: McGraw-Hill.

Novick, M. R. & Lindley, D. V. (1979). The use of more realistic utility functions in educational applications. *Journal of Educational Measurement, 15,* 81–191.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 237–257). New York: Academic Press.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (eds.). *Handbook of modern item response theory* (pp.271–286). New York: Springer-Verlag.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics, 35,* 1–20.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331–354.

Spray, J. A., Davey, D. C., Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (ACT Research Report Series ONR90-8). Iowa City, IA: ACT Inc.

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference* (pp. 82–98). Minneapolis: University of Minnesota.

van der Linden, W. J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement, 23,* 21–29.

van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

Verhelst, N. D., Glas, C. A. W. & Verstralen, H. H. F. M. (1995). *OPLM: Computer program and manual.* Arnhem: Cito.

Wainer, H. (Ed.). (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H, Bradlow, E. T. & Du, Z. (2000). Testlet response theory: An analogue for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W.Glas (Eds.). *Computer adaptive testing: Theory and practice* (pp. 245–269). Boston: Kluwer-Nijhoff Publishing.

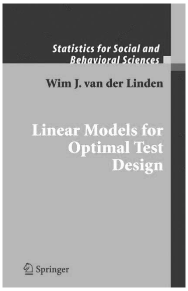
Weiss, D. J. (Ed.). (1983). *New horizons in testing.* New York: Academic Press.

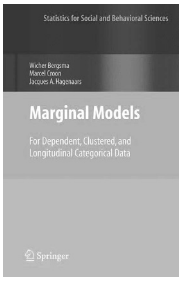
Wilson, D. T., Wood, R. & Gibbons, R. (1991) *TESTFACT: Test scoring, item statistics, and item factor analysis.* (computer software). Chicago: Scientific Software International Inc.

Winston, W. L. (1994). *Operations research: Applications and algorithms.* Belmont, CA: Wadsworth.

Wu, M. L., Adams, R. J. & Wilson, M. R. (1997). *ConQuest: Generalized item response modelling software.* Australian Counsil for Educational Research.

Index


Statistics for Social and Behavioral Sciences
Wim J. van der Linden
Linear Models for Optimal Test Design
Springer


Statistics for Social and Behavioral Sciences
Wicher Bergsma
Marcel Croon
Jacques A. Hagenaars
Marginal Models
For Dependent, Clustered, and Longitudinal Categorical Data
Springer

Statistics for Social and Behavioral Sciences
Mark D. Reckase
Multidimensional Item Response Theory
Springer