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
University of Twente

Wim J. van der Linden
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

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

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

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:

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

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

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

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

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

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:

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

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

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

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:

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

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:

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

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 $\theta = \widehat{\theta}_{u_{i_1},\ldots,u_{i_{k-1}}}$. Formally, it can be presented as

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

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

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

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.

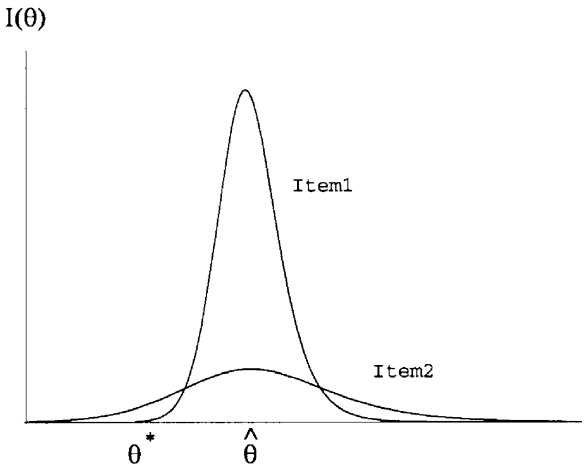

Fig. 1.1 Attenuation paradox in item selection in CAT

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

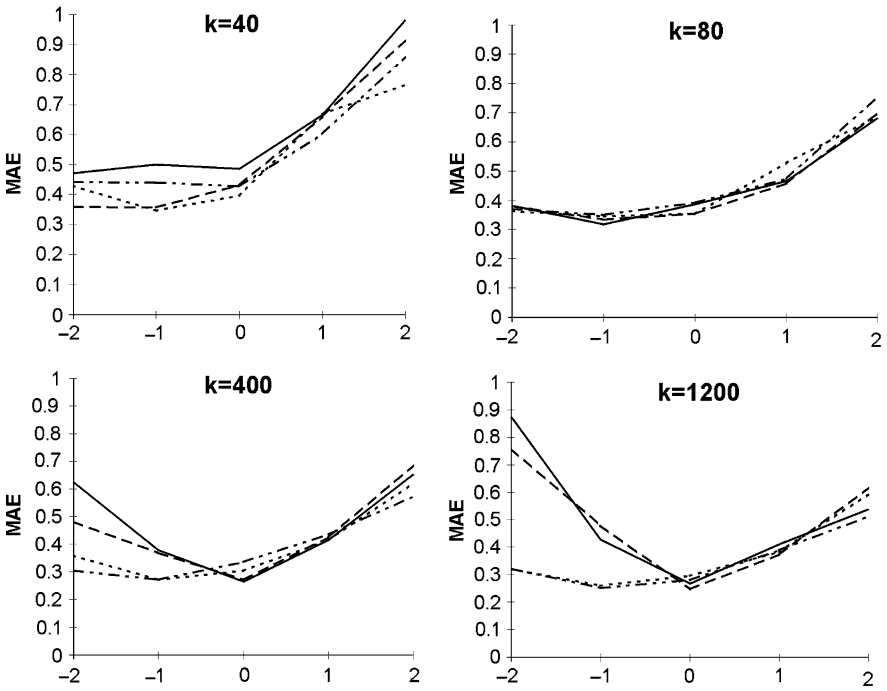

Fig. 1.2 Mean absolute error (MAE) in ability estimation from item pools with $k = 40$, 80, 400, and 1200 items (size of calibration samples: 250: solid; 500: dashed; 1200: dotted; 2500: dashed-dotted)

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

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

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 $p_i(0 \mid u_{i_1}, \ldots, u_{i_{k-1}})$ 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

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

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

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

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

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

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