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.

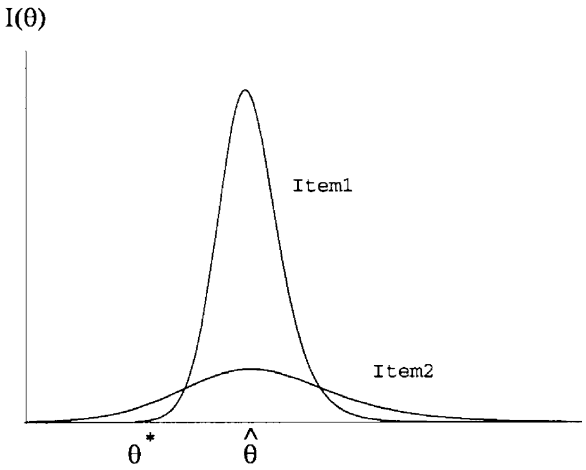

Fig. 1.1 Attenuation paradox in item selection in CAT

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

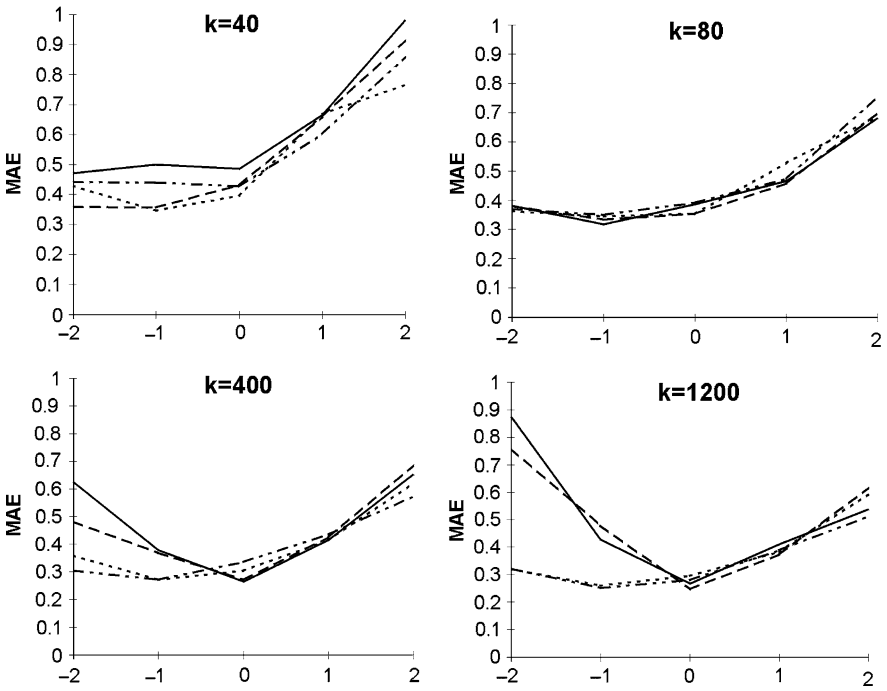

Fig. 1.2 Mean absolute error (MAE) in ability estimation from item pools with $k = 40$, 80, 400, and 1200 items (size of calibration samples: 250: solid; 500: dashed; 1200: dotted; 2500: dashed-dotted)

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

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,

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

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.

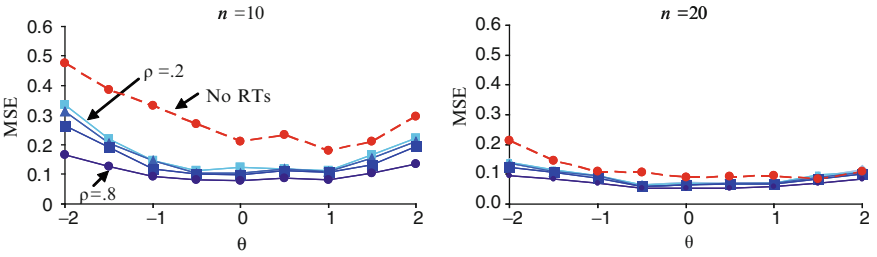

Fig. 1.3 MSE functions of EAP estimator of θ for item selection without RTs (dashed line) and with RTs with $\rho_{\theta\tau} = 0.2, 0.4, 0.6$, and 0.8 (solid lines; the darker the line, the higher the correlation) for tests of $n = 10$ and 20 items. [Reproduced with permission from W. J. van der Linden (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 $k \rightarrow \infty$, 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

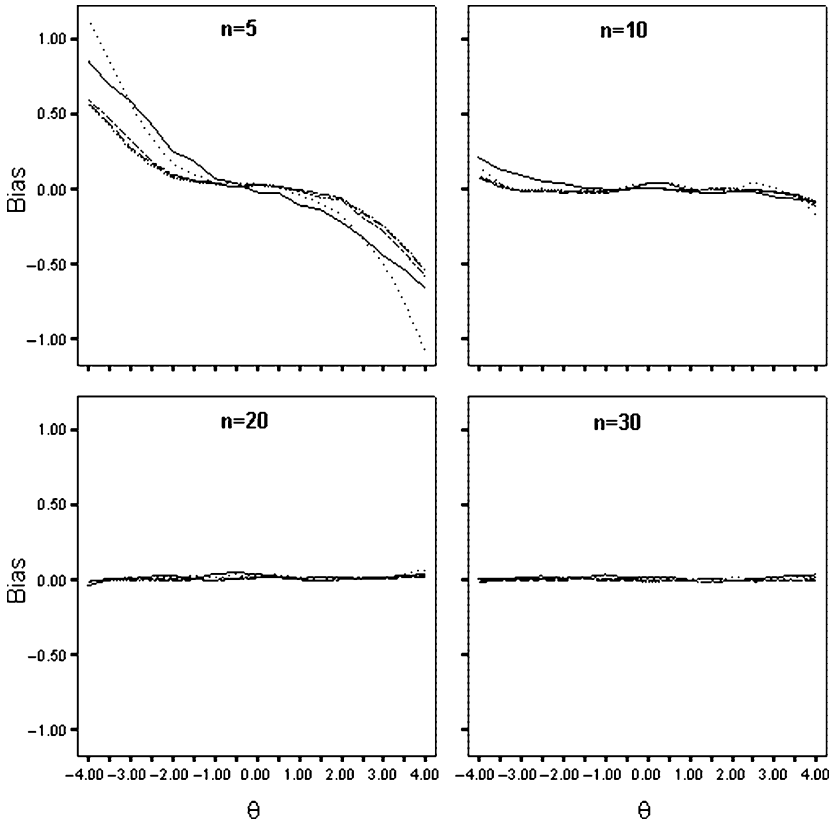

Fig. 1.4 Bias functions for five item-selection criteria after $n = 5, 10, 20, 30$ items (maximum-information with MLE: solid; maximum-posterior weighted Information: dotted; maximum expected information: dashed-dotted; maximum expected posterior variance: dashed; maximum expected posterior weighted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

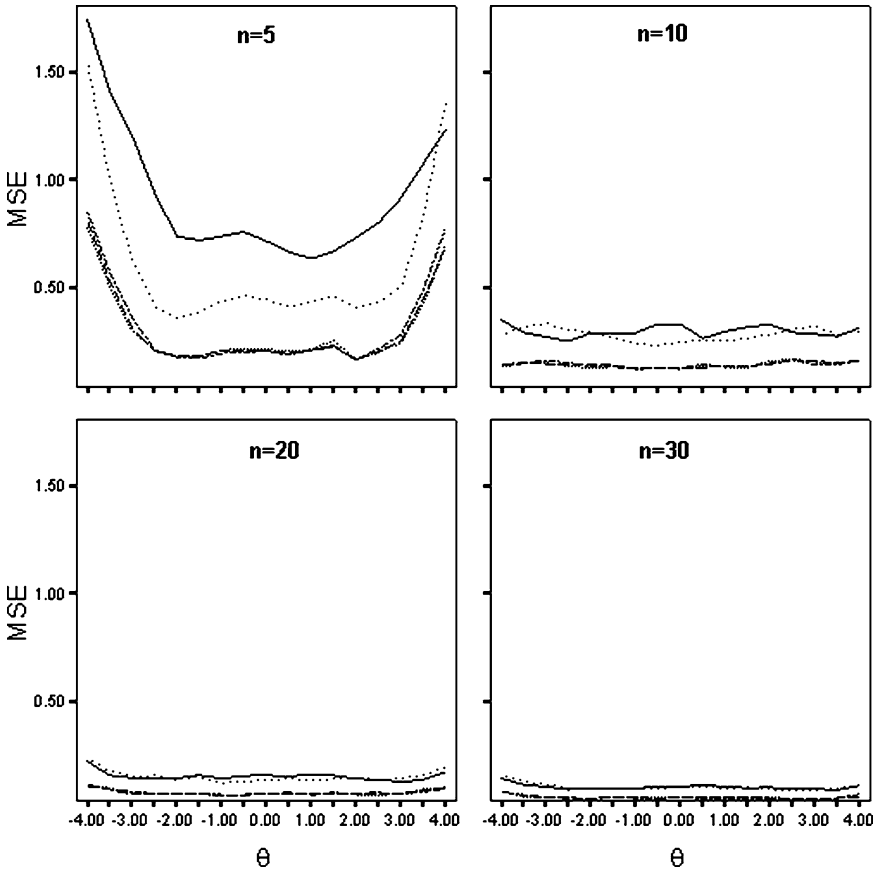

Fig. 1.5 MSE functions for five item-selection criteria after $n = 5$, 10, 20, 30 items (maximum-information with MLE: solid; maximum-posterior weighted information: dotted; maximum expected information: dashed-dotted; maximum expected posterior variance: dashed; maximum expected posterior weighted information: 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.