Statistical Aspects of Adaptive Testing

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

1. Introduction

In psychometrics, it is customary to view the responses of subjects $n = 1, \ldots, N$ to the items $i = 1, \ldots, I$ in an educational or psychological test as realizations $u_{ni}$ of random response variables $U_{ni}$. The distribution of these variables is usually represented by a response model. A large variety of alternative response models has been developed in item response theory (van der Linden and Hambleton, 1997).

Since the responses reveal as much about the properties of the items as the abilities of the subjects that take them, we have to treat the variables $U_{ni}$ as nonidentically distributed across items and subjects. It usually makes sense to assume independence for the variables for the same subject (within-subject or “local” independence). Provided there is no cooperation between them, the variables can also be assumed to be independent between subjects. Statistically, the analysis of response data on tests therefore deals with an instance of nonidentical but independently distributed variables $U_{ni}$.

Suppose the ability of subject $n$ is represented by a (possibly vector-valued) parameter $\theta_n$ and the properties of item $i$ by a parameter vector $\xi_i$. Typically, the interest is in estimating the parameters $\theta_n$ and less in $\xi_i$. The usual approach is to estimate the item parameters with a likelihood marginalized over the subject parameters or in a Bayesian fashion. This step is known as item calibration. Once the item parameters can be treated as known, the response model can be used as a model for measuring future subjects, i.e., estimating their ability parameters $\theta_n$ from realizations $(u_{n1}, \ldots, u_{nI})$.

But even if the parameters $\xi_i$ can be treated as known, the fact that they are item dependent requires special attention when assembling a test. As shown below, collectively the item parameters determine the accuracy of the estimator of $\theta_n$. Inasmuch as the response models are typically nonlinear, the accuracy of these estimators also depends on the unknown parameters $\theta_n$ themselves. The combination of these features creates a problem of optimal design in test theory that is much more complicated than the standard optimal design problem in linear regression with known predictors (van der Linden, 2005, Section 4.7).

For a fixed test, a practical way out of this problem is to specify a target function for the accuracy of the estimator $\hat{\theta}$ of the ability parameter $\theta$ over its space and use the known item parameters $\xi_i, i = 1, \ldots, I$, to select a test from the inventory of calibrated
items that approximates the target best in some sense. A more efficient way, however, is an adaptive test design. In adaptive testing, the test is administered by a computer that has a well-chosen pool of calibrated items stored in its memory. The estimator of $\theta_n$ is updated after an item is completed and the next item is selected to be most efficient at the last update. The test is stopped if a fixed number of items has been taken or a predetermined level of accuracy for $\hat{\theta}$ is realized. Extensive computer simulations have shown that, typically, adaptive testing realizes the same accuracy for $\hat{\theta}$ as a well-designed fixed test with some 50% of the items.

The intuitive idea underlying the first attempts at adaptive testing was that a test had better measurement properties if the difficulties of the items matched the ability of the examinee. Items that are too easy or difficult have predictable responses and cannot provide much information about the ability of the examinee. In fact, this intuition was the same as that of a good examiner in an oral exam who avoids asking questions for which he knows in advance that they are answered correctly or incorrectly by the examinee.

For two reasons it took a long time before adaptive testing could be implemented in large-scale testing programs. First, realistic response models for the distribution of the variables $U_{ni}$ had to be developed before the notion of adaptation could be formalized. Second, estimation of $\theta_n$ is generally computationally intensive, and we had to wait for computers to become more powerful and affordable before estimation in real time was feasible. The first large-scale applications of adaptive testing were introduced in the early 1990s in testing programs for personnel selection, college admission, and licensure and certification testing. Adaptive testing is now also common in such applications as course placement in education and diagnostics in psychology.

The introduction of adaptive testing has led to a host of practical problems that all have serious statistical aspects. In this chapter, we first introduce a few response models that are used in adaptive testing and discuss estimation of their item and subject parameters. We then discuss adaptive estimation of the subject parameters and treat the statistical aspects of several of these more practical problems.

2. Response models

A basic unidimensional model for dichotomously scored responses is the three-parameter logistic (3PL) model (Birnbaum, 1968). In this model, the probability of a correct response $U_{ni} = 1$ of subject $n$ on item $i$ is given by

$$p_i(\theta_n) \equiv \Pr(U_{ni} = 1 \mid \theta_n) \equiv c_i + (1 - c_i)\Psi \left[ a_i(\theta_n - b_i) \right]$$

$$\equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta_n - b_i)]}{1 + \exp[a_i(\theta_n - b_i)]},$$

(1)

where $\theta_n$ is a scalar parameter for the ability of subject $n$, and $\xi_i = (a_i, b_i, c_i)$ contains the parameters for the discriminating power, difficulty, and height of the lower asymptote of the response probability of item $i$, respectively. The last item parameter has been
introduced to represent instances of random guessing on test items. In some applications, which will be discussed below, it proves practical to replace the logistic function $\Psi(\cdot)$ by the normal cumulative distribution function $\Phi(\cdot)$. Then, the probability of a correct response becomes

$$p_i(\theta_n) \equiv c_i + (1 - c_i)\Phi\left[a_i(\theta_n - b_i)\right]. \quad (2)$$

This model is known as the three-parameter normal ogive (3PNO) model.

In practice, it is usually assumed that there is one (dominant) ability $\theta$ that explains the test performances. But empirical model fit analyses may show that a unidimensional response model does not adequately describe the data, for example, when the test items consist of worded mathematical problems whose solution require a verbal in addition to a mathematical skill (Reckase, 1985). In such cases, a response model with a multidimensional ability parameter is more appropriate (McDonald, 1967; Andersen, 1985; Adams et al., 1997).

In a multidimensional response model, the logistic function in (1) and the normal ogive function in (2) are redefined as $\Psi(a_i^t \theta_n - b_i)$ and $\Phi(a_i^t \theta_n - b_i)$, respectively, where $a_i$ is a $Q$-dimensional vector of discrimination parameters $(a_{i1}, \ldots, a_{iq}, \ldots, a_{iQ})$, and $\theta_n$ is a vector of ability parameters $(\theta_{n1}, \ldots, \theta_{nq}, \ldots, \theta_{nQ})$ of the same dimension.

Under the additional assumption that the subjects’ ability parameters can be taken to be a simple random sample from a $Q$-variate normal distribution, that is,

$$\theta_n \sim \text{MVN}(\mu_P, \Sigma_P), \quad (3)$$

the model is equivalent with a full-information factor analysis model (Takane and de Leeuw, 1987). For this reason, the discrimination parameters are sometimes referred to as factor loadings and the ability parameters as factor scores.

In multiple-group IRT, different ability distributions are assumed for different populations.

Note that (3) can be used as a common prior distribution for the ability parameters of the subjects. In addition, it often is useful to introduce a distribution for the items parameters,

$$\xi_i = (a_i, b_i, c_i) \sim \text{MVN}(\mu_I, \Sigma_I), \quad (4)$$

as well. This distribution can also be used in two different ways. First, as a (common) prior distribution to support parameter estimation when the data matrix is ill-conditioned, i.e., contains too little information and can be appropriately described by different sets of parameter values. Mislevy (1986) considers a number of Bayesian approaches that involve prior distributions for the item parameters. Second, the distribution can be used as an empirical distribution for the domain of possible items when modeling hierarchical relations between them. An example is the application to item-cloning techniques, in which a computer algorithm is used to generate families of items from so-called parent items.

Glas and van der Linden (2003; tentatively accepted) and Sinharay et al. (2003) present a multilevel response model for this application in which it is assumed that the parameters of each item are randomly sampled from a multivariate normal distribution associated with a parent item. Analogous to multiple-group IRT, this model allows for normal distributions with different means and covariance matrices for each family.
3. Item calibration

Adaptive testing is based on the availability of a large pool of calibrated test items. The calibration process can have two different designs:

1. A pretesting design in which fixed subsets of items are administered to fixed subsets of subjects in a series of pretest sessions, and the response model is fitted to the data to obtain item-parameter estimates.

2. An adaptive design in which new items are seeded into an operational adaptive test and the responses are used to estimate their item parameters along with the subjects’ ability parameters.

Of course, it is also possible to use combinations of these sampling designs, for example, in a setup with initial item parameter estimation through a fixed design and further improvement of the estimates during operational adaptive testing. In both types of designs, the subjects never respond to all available items. Either type of design can be captured by a design vector $d_n$ with elements $d_{ni}, i = 1, \ldots, I$, that are equal to one if $n$ responded to item $i$ and zero otherwise.

Two different estimation procedures for the item parameters are in use. The first is a marginal maximum likelihood (MML) procedure that builds on the fact that in item calibration only the item parameters are of interest. Consequently, the ability parameters are treated as nuisance parameters, which are integrated out of the likelihood function assuming a population distribution. The resulting marginal likelihood is maximized with respect to the item and population parameters. The second procedure is Bayesian with prior distributions for all parameters. This approach has become feasible through the use of Markov chain Monte Carlo (MCMC) computational methods. These methods circumvent the cumbersome direct evaluation of high-dimensional integrals in MML estimation for more complicated models. The two approaches to item parameter estimation are further explained in the next sections.

3.1. MML estimation

Currently, MML estimation (Bock and Aitkin, 1981) is the most frequently used technique for item calibration. One reason for this is its presence in widely disseminated software packages, for example, Bilog-MG (Zimowski et al., 2002) for unidimensional models for dichotomously scored items and Multilog (Thissen et al., 2003) or Parscale (Muraki and Bock, 2002) for polytomously scored items. MML estimation for multidimensional IRT models (Bock et al., 1988) has been implemented in the software packages Testfact (Wood et al., 2002) and ConQuest (Wu et al., 1997). All these packages compute MML estimates of the structural parameters in the model in a simultaneous fashion; this approach will be outlined below. But MML estimates can also be obtained through a two-step procedure, for example, that implemented in the Mplus software (Muthén and Muthén, 2003).

To obtain generality, the MML estimation equations will be derived for the multidimensional version of the 3PL model. The extension of MML estimation to a Bayesian version of it with prior distributions of the item parameters will be addressed at the end of this section.
Statistical aspects of adaptive testing

Let \( U_n \) be the response vector of subject \( n \) and \( \xi = (\xi_1, \ldots, \xi_I) \). Using the assumption of local independence, the probability of responses \( u_n \) given \( \theta_n \) is

\[
\Pr(u_n | \theta_n, \xi) = \prod_{i=1}^{I} \Pr(U_{ni} = u_{ni} | \theta_n, \xi_i).
\]

(5)

It is assumed that \( u_{ni} \) takes a dummy value if \( d_{ni} = 0 \). We will use the convention \( u_{ni} = 9 \) and, for mathematical treatment, define \( \Pr(U_{ni} = 9 | \theta_n, \xi_i) = 1 \). Since \( u_n \) determines the values of \( d_{ni} \), the latter do not appear on the left side of (5).

Direct maximization of the likelihood function in (5) with respect to \( \theta \) does generally not produce consistent estimators because the number of subject parameters goes to infinity with the sample size. Therefore, Bock and Aitkin (1981) suggested to treat the \( \theta \)s as unobserved realizations of stochastic variables that have the population distribution in (3). Assume that the response vectors of \( N \) subjects have been collected in an \( N \times I \) matrix, \( U \). The marginal likelihood function is defined as

\[
L(\xi, \mu_P, \Sigma_P; U) = \prod_{n=1}^{N} \Pr(u_n | \xi, \mu_P, \Sigma_P),
\]

(6)

with

\[
\Pr(u_n | \xi, \mu_P, \Sigma_P) = \int \cdots \int \Pr(u_n | \theta_n, \xi) g(\theta_n | \mu_P, \Sigma_P) d\theta_n.
\]

(7)

Glas (1992, 1999) showed how to derive the likelihood equations for (6) from Fisher’s identity (Efron, 1977). Let \( \eta \) denote all structural parameters in the model, that is, \( (\xi, \mu_P, \Sigma_P) \), where we choose the diagonal and lower-diagonal elements of \( \Sigma_P \) as free parameters. The first-order derivatives with respect to \( \eta \) can be written as

\[
h(\eta) = \frac{\partial}{\partial \eta} \log L(\eta; U) = \sum_n \mathcal{E}(\omega_n(\eta) | u_n, \eta),
\]

(8)

with

\[
\omega_n(\eta) = \frac{\partial}{\partial \eta} \log p(u_n, \theta_n | \eta),
\]

(9)

and where the expectation is with respect to the posterior distribution

\[
p(\theta_n | u_n, \eta) = \frac{\Pr(u_n | \theta_n, \xi) p(\theta_n | \mu_P, \Sigma_P)}{\int \cdots \int \Pr(u_n | \theta_n, \xi) p(\theta_n | \mu_P, \Sigma_P) d\theta_n}.
\]

(10)

The identity in (8) is closely related to the EM algorithm (Dempster et al., 1977) for finding the maximum of a likelihood. The EM algorithm is applicable in situations where direct inference based on the marginal likelihood is complicated but the complete-data likelihood equations, i.e., equations based on \( \omega_n(\eta) \), are easier to solve. The EM algorithm does so by iteratively improving on estimates \( \eta^{(t-1)} \), solving \( \sum_n \mathcal{E}(\omega_n(\eta^{(t)}) | u_n, \eta^{(t-1)}) = 0 \) with respect to \( \eta^{(t)} \) until satisfactory convergence is reached.
We illustrate the application of Fisher’s identity for the estimation of the population parameters for the $q$th component of $\theta$. If the subject parameters $\theta_n$ had been observed, the mean $\mu_q$ would be equal to

$$\mu_q = \frac{1}{N} \sum_{n=1}^{N} \theta_{nq}. \quad (11)$$

Since they are not observed, we take the posterior expectation of (11) with respect to $\theta_n$ to obtain

$$\mu_q = \frac{1}{N} \sum_{n=1}^{N} \mathcal{E}(\theta_{nq} \mid u_n, \xi, \mu_P, \Sigma_P). \quad (12)$$

Observe that (12) is a function of the response vector $u_n$, the item parameters $\xi$, as well as the population parameters $\mu_P$ and $\Sigma_P$.

Analogously, the equations for the elements of $\Sigma_P$ are

$$\sigma_{qq}^2 = \frac{1}{N} \left[ \sum_{n=1}^{N} \mathcal{E}(\theta_{nq}^2 \mid u_n, \xi, \mu_P, \Sigma_P) - N\mu_q^2 \right], \quad (13)$$

and

$$\sigma_{qq'} = \frac{1}{N} \left[ \sum_{n=1}^{N} \mathcal{E}(\theta_{nq}\theta_{nq'} \mid u_n, \xi, \mu_P, \Sigma_P) - N\mu_q\mu_{q'} \right]. \quad (14)$$

Thus, the population mean of the subjects’ abilities is estimated as the mean of the posterior expected values of the subjects’ parameters. Likewise, the population variances and covariances of these abilities are estimated as the posterior expected variances and covariances between the subjects’ parameters. The estimation equations for the item parameters are found in a similar way; for their equations as well as the standard errors of all estimates, refer to Glas (2000). Further, for the model to become identifiable, a number of restrictions on the parameters is required (see Béguin and Glas, 2001).

The multiple integrals above can be evaluated using Gauss–Hermite quadrature. A critical factor in using Gauss–Hermite quadrature is the number of variables that can be integrated simultaneously. Wood et al. (2002) indicate that the maximum number of variables is 10 with adaptive quadrature, 5 with nonadaptive quadrature, and 15 with Monte Carlo integration.

The consequences of the adding of prior distributions for the item parameters to the MML procedure are not too complicated. If (4) is used as a common prior distribution for the item parameters, the likelihood defined by (6) should be multiplied by the prior density to obtain

$$L(\xi, \mu_P, \Sigma_P; \mu_I, \Sigma_I \mid U) = L(\xi, \mu_P, \Sigma_P; U)p(\xi \mid \mu_I, \Sigma_I). \quad (15)$$

Mislevy (1986) uses the marginal posterior distributions for this likelihood to obtain Bayes modal estimates of the parameters, that is, he estimates these parameters as the mode of their posterior distributions. In one of his approaches, $p(\xi \mid \mu_I, \Sigma_I)$ is treated
as a subjective prior with known hyperparameters $\mu_I$ and $\Sigma_I$. In another approach, labeled empirical Bayes approach, the parameters of $p(\xi | \mu_I, \Sigma_I)$ are estimated along with the ability parameters. In that case, $\mu_I$ and $\Sigma_I$ have their own (known) prior distributions, say, $p(\mu_I, \Sigma_I | \mu_{I0}, \Sigma_{I0})$.

Glas and van der Linden (2003) argue that for some applications (e.g., item cloning), the item parameters should be treated as nuisance parameters as well and consider a likelihood that is marginalized not only with respect to $\theta_n$ but also with respect to $\xi_i$.

### 3.2. Evaluation of model fit in an MML framework

The reliability and validity of adaptive testing is based on the assumption that the response model properly fits the data. Orlando and Thissen (2000) and Glas and Suárez-Falcon (2003) review the available statistical tests of the various assumptions underlying response models. Most statistics are based on the difference between the observed and expected response frequencies.

As an example, consider a statistic of differential item functioning, i.e., differences between the response probabilities of subgroups of subjects that are not related to the abilities that are tested. The particular concern is that these differences disadvantage one of the subgroups. A comparable problem in adaptive testing is item-parameter drift, that is, a change in an item parameter between pretesting and operational testing or during operational testing.

Suppose we have two subgroups, which we label as the focal group and the reference group. A test based on the differences between the observed and expected correct scores would proceed as follows: First, MML estimates of the item parameters are obtained using the data of the two groups simultaneously, typically with the assumption that their subjects are sampled from different ability distributions. Then, the statistic

$$\sum_{n=1}^{N} y_n d_{ni} u_{ni} - \sum_{n=1}^{N} y_n d_{ni} \mathcal{E}(P_i(\theta_n) | u_n, \xi, \mu_P, \Sigma_P)$$

(16)

is calculated, where $y_n = 1$ if subject $n$ belongs to the focal group and zero otherwise. Observe that $\mathcal{E}(P_i(\theta_n) | u_n, \xi, \mu_P, \Sigma_P)$ is the posterior expectation of a correct response of subject $n$ on item $i$. Glas (1998) shows that if the square of (16) is weighted by the estimated variance of $\sum_{n=1}^{N} y_n d_{ni} u_{ni}$, the resulting statistic has an asymptotic $\chi^2$-distribution with one degree of freedom.

This test of differential item functioning belongs to a general class of so-called Rao score tests or Lagrange multiplier tests for item response models (Glas, 1999; Glas and Suárez-Falcon, 2003), which also contains tests of the shape the response function $P_i(\theta_n)$ and the assumption of local independence between item responses.

### 3.3. Bayesian estimation

In Bayesian inference, all parameters are considered as random variables. Samples from their posterior distribution can be generated using an MCMC procedure. These samples can be used to approximate the functionals over the posterior distributions that are of interest. Albert (1992) shows how to apply Gibbs sampling (Gelfand and Smith, 1990)
for the 2PNO model (that is, the model in (2) with \( c_i = 0 \)), and Johnson and Albert (1999) have generalized the procedure to the 3PNO model. We work with the 3PNO instead of the 3PL model because the Gibbs sampler is easier to implement for it. For all practical purposes, the results from the two models are indistinguishable.

To implement a Gibbs sampler, the parameter vector is divided into a number of components. Iterative sampling from the conditional posterior distributions of each component given the previously sampled values for all other components generates a Markov chain. The chain is continued after it has stabilized to generate a sample from the joint posterior distribution of the parameters.

Johnson and Albert (1999, also see Béguin and Glas, 2001) introduced a data-augmentation scheme for the 3PNO that results in tractable conditional posterior distributions. The scheme is based on the following interpretation of the 3PNO. Suppose that the examinee either knows the correct answer with probability \( \Phi(\lambda_{ni}) \), \( \lambda_{ni} = a_i^\prime \theta_n - b_i \), or does not know the correct answer with probability \( 1 - \Phi(\lambda_{ni}) \). In the first case, a correct response is given with probability one, and in the second case the examinee guesses the correct response with probability \( c_i \). Thus, the marginal probability of a correct response is equal to \( \Phi(\lambda_{ni}) + c_i(1 - \Phi(\lambda_{ni})) \).

It proves to be convenient to introduce a vector of binary variables \( W_{ni} \) such that

\[
W_{ni} = \begin{cases} 
1 & \text{if subject } n \text{ knows the correct answer to item } i, \\
0 & \text{otherwise.}
\end{cases}
\] (17)

So, if \( W_{ni} = 0 \), subject \( i \) guesses the response to item \( j \), and if \( W_{ni} = 1 \), subject \( i \) knows the right answer and gives a correct response. Consequently, the conditional probabilities of \( W_{ni} = w_{ni} \) given \( U_{ni} = u_{ni} \) are

\[
P(W_{ni} = 1 | U_{ni} = 1, \lambda_{ni}, c_i) \propto \Phi(\lambda_{ni}),
\]

\[
P(W_{ni} = 0 | U_{ni} = 1, \lambda_{ni}, c_i) \propto c_i(1 - \Phi(\lambda_{ni})),
\]

\[
P(W_{ni} = 1 | U_{ni} = 0, \lambda_{ni}, c_i) = 0,
\]

\[
P(W_{ni} = 0 | U_{ni} = 0, \lambda_{ni}, c_i) = 1.
\] (18)

The observed responses \( u \) of subjects are augmented with latent data \( w = (w_{11}, \ldots, w_{NI})^T \). In addition, following Albert (1992), the data are also augmented with latent data \( z = (z_{11}, \ldots, z_{NI})^T \), which are realizations of independent normally distributed random variables \( Z_{ni} \) with mean \( \lambda_{ni} \) and standard deviation equal to one. For the variables \( w \) and \( z \), it holds that \( z_{ni} > 0 \) if \( w_{ni} = 1 \) and \( z_{ni} \leq 0 \) if \( w_{ni} = 0 \). Hence,

\[
p(z_{ni} | w_{ni}, \lambda_{ni}) \propto \phi(z_{ni}; \lambda_{ni}, 1)
\]

\[
\times \left[ I(z_{ni} > 0)I(w_{ni} = 1) + I(z_{ni} \leq 0)I(w_{ni} = 0) \right],
\] (19)

where \( \phi(\cdot; \lambda_{ni}, 1) \) is the normal density with mean \( \lambda_{ni} \) and standard deviation equal to one and \( I(\cdot) \) is an indicator function that takes the value one if its argument is true and zero otherwise.

We introduce normal prior distributions for \( a \) and \( b \), that is, \( a \sim N(\mu_a, \sigma_a) \) and \( b \sim N(\mu_b, \sigma_b) \), whereas the conjugate distribution Beta\((\alpha, \beta)\) is chosen as the prior
distribution for the guessing probability, $c$. A noninformative prior is obtained by taking $Beta(1, 1)$. As an alternative, we could use (4) as the prior distribution of the item parameters. Finally, the joint prior distribution of $(\mu_P, \Sigma_P)$ is specified as the product of the normal and inverse Wishart densities $p(\mu_P \mid \Sigma_P)$ and $p(\Sigma_P)$, respectively (see, for instance, Box and Tiao, 1973).

The joint posterior of $\eta = (\xi, \theta, z, w, \mu_P, \Sigma_P)$ has density

$$p(\eta \mid U) = p(z, w \mid U; \xi, \theta, \mu_P, \Sigma_P)p(\theta \mid \mu_P, \Sigma_P)p(\mu_P \mid \Sigma_P)p(\Sigma_P)p(\xi)$$

$$= \prod_{n=1}^{N} \prod_{i=1}^{I} p(z_{ni} \mid w_{ni}; \lambda_{ni})p(w_{ni} \mid u_{ni}; \lambda_{ni}, c_i)p(\theta_n \mid \mu_P, \Sigma_P)$$

$$\times p(\mu_P \mid \Sigma_P)p(\Sigma_P)p(a, b)p(c),$$

(20)

where $p(w_{ni} \mid u_{ni}, \lambda_{ni}, c_i)$ and $p(z_{ni} \mid w_{ni}, \lambda_{ni})$ follow from (18) and (19), respectively. Although the distribution has an intractable form, the conditional distributions of the components $\xi, \theta, z, w, \mu, \Sigma_P$ given all others are tractable and easy to sample from.

Therefore, the Gibbs sampler consists of five steps:

1. Draw $\mu_P$ and $\Sigma_P$ conditionally on $\theta$,
2. Draw $w$ and $z$ conditional on $\theta, \xi, \Sigma_P$, and $U$,
3. Draw $\theta$ conditional on $z, a, b, \Sigma_P$, and $\mu$,
4. Draw $a, b$ conditional on $z$ and $\theta$,
5. Draw $c$ conditional on $w$ and $U$.

Details on these steps are given in Béguin and Glas (2001). Steps 3–5 are straightforward: Steps 3 and 4 boil down to sampling from the normal posterior distribution that is associated with a normal regression model and a normal prior distribution. Step 5 involves the well-known case of a binomial model with a beta as prior distribution, which leads to sampling from a beta posterior distribution.

The first two steps are as follows.

**Step 1.** The prior distribution for $(\mu_P, \Sigma_P)$ we have chosen is a normal–inverse–Wishart. Their families are conjugate with the multivariate normal (3). Let $A_0$ be the scale matrix and $v_0$ the degrees of freedom for the prior of $\Sigma_P$. Further, let $\mu_0$ be the prior mean and $k_0$ a weight for the prior of $\Sigma_P$, which can be interpreted as the number of prior measurements on the $\Sigma_P$ scale (Gelman et al., 1995). Then, the posterior distribution is normal–inverse–Wishart with parameters

$$\mu_N = \frac{k_0 N}{k_0 + N} \mu_0 + \frac{N}{k_0 + N} \bar{\theta},$$

$$v_N = v_0 + N,$$

$$\kappa_N = k_0 + N,$$

$$A_N = A_0 + S + \frac{k_0 N}{k_0 + N} (\bar{\theta} - \mu_0)(\bar{\theta} - \mu_0)^T,$$

where $S$ is a matrix with entries $S_{pq} = \sum_{n=1}^{N} (\theta_{np} - \bar{\theta}_p)(\theta_{nq} - \bar{\theta}_q)$ and $\bar{\theta}_p$ and $\bar{\theta}_q$ are the means of the proficiency parameters of dimension $p$ and $q$, respectively.
Samples from the joint posterior distribution of \( \mu_P \) and \( \Sigma_P \) are obtained through the following procedure:

1. Draw \( \Sigma_P \mid \Theta \) from the Inverse–Wishart \( vN(\Lambda^{-1} N) \).
2. Draw \( \mu_P \mid \Sigma_P \) from \( N(\mu_N, \Sigma_P / \kappa_N) \).

The hyperparameters \( \kappa_0, v_0, \mu_0 \) and \( \Lambda_0 \) lead to a noninformative prior distribution if \( \kappa_0 \to 0, v_0 \to -1 \) and \( |\Lambda_0| \to 0 \). This limiting case, which is a multivariate version of Jeffrey’s prior density, leads to the posterior distribution

\[
\Sigma_P \mid \Theta \sim \text{Inverse–Wishart}_{N-1}(S^{-1}),
\]

\[
\mu_P \mid \Sigma_P \sim N(\bar{\Theta}, \Sigma_P / N). \tag{21}
\]

**Step 2.** The posterior \( p(z, w \mid U; \xi, \Theta) \) is factored as \( p(z \mid U; w, \xi, \Theta) p(w \mid U; \xi, \Theta) \), and values of \( w \) and \( z \) are obtained through the following procedure:

1. Draw \( w_{ni} \) from the distribution of \( W_{ni} \) given the data \( U \) and \( \xi, \Theta \) in (18).
2. Draw \( z_{ni} \) from the conditional distribution of \( Z_{ni} \) given \( w, \Theta \) and \( \xi \), which follows from (18) as

\[
Z_{ni} \mid w, \Theta, \xi \sim \begin{cases} 
N(\eta_{ni}, 1) \text{ truncated at the left at 0 if } w_{ni} = 1, \\
N(\eta_{ni}, 1) \text{ truncated at the right at 0 if } w_{ni} = 0.
\end{cases}
\]

As start values for the parameters, we suggest using the MML estimates by one of the computer programs listed earlier. The number of burn-in iterations for the Markov chain to stabilize should be found empirically. One helpful tool is to start multiple MCMC chains from different points and evaluate their convergence by comparing the between- and within-chain variances (see, for instance, Robert and Casella, 1999, p. 366).

### 3.4. Evaluation of model fit in a Bayesian framework

Posterior predictive checks (Meng, 1994) are one of the standard tools for the evaluation of model fit in a Bayesian framework with MCMC computation. They can readily be used to evaluate the fit of IRT models (Hoijting, 2001; Glas and Meijer, 2003; Sinharay, 2005).

Bayesian analogues of the likelihood-based test statistics above can be calculated as a by-product of the Markov chain. Once it has stabilized, at every iteration the sampled values for the parameters \( \eta \) and \( \Theta \) are used to generate replicate data \( U^\text{rep} \) under the response model. The check on the fit of the model consists of a comparison between the same test statistic computed from the original data \( U \) and the generated data \( U^\text{rep} \) for the sampled parameter values. Particularly, we compute the proportion of times the former has a smaller value than the latter, i.e., the Bayesian \( p \)-value.

As an example, a Bayesian analogue of the statistic for differential item function in (16) is given. For the original data, the statistic is calculated as

\[
T(U, \eta) = \sum_{n=1}^{N} y_n d_{ni} u_{ni} - \sum_{n=1}^{N} y_n d_{ni} P_i(\theta_n), \tag{22}
\]
Statistical aspects of adaptive testing

where \( P_i(\theta_n) \) is evaluated at the current draws of the item and subjects’ parameters and \( y_n \) is still the indicator variable used to denote membership of the focal and reference group. The calculation is repeated to obtain the version of the same statistic for the generated data, \( T(U^{rep}, \eta) \). Formally, the Bayesian \( p \)-value is defined as

\[
Pr(T(U, \eta) \mid U) \leq T(U^{rep}, \eta). \tag{23}
\]

Possible other model violations, for instance, with respect to shape of the response functions \( P_i(\theta_n) \), the assumption of local independence, and the behavior of individual subjects, can be detected by replacing (22) by another appropriate Pearson-type statistic.

4. Ability estimation

If the item parameters have been estimated with sufficient precision, they can be treated as the true parameters when estimating the subject parameter, \( \theta \). This practice simplifies real-time ability estimation in operational adaptive testing considerably. Again, likelihood-based and Bayesian methods can be used for estimating \( \theta \). As we will show in the next sections, both classes entail different rules for adapting the item selection to the estimates of \( \theta \).

In the first approach, the likelihood associated with (5),

\[
L(\theta; u_n) = \prod_{i=1}^{l} P(Un_i = u_{ni} \mid \theta, \xi_i)^{d_{ni}}, \tag{24}
\]

is maximized with respect to \( \theta \). The first-order and second-order partial derivatives needed to compute the MLEs and their standard errors can be found in Segall (1996); see also Segall (2000).

For a unidimensional \( \theta \), an attractive alternative to the regular MLE is Warm’s (1989) weighted likelihood estimator (WLE). The estimator is the maximizer of the likelihood in (24) weighted by a function \( w(\theta) \). That is, it maximizes

\[
L(\theta; u_n) = w(\theta)L(\theta \mid u_n), \tag{25}
\]

where the weight function is defined to satisfy

\[
\frac{\partial w(\theta)}{\partial \theta^2} = \frac{H(\theta)}{2I(\theta)}, \tag{26}
\]

with

\[
H(\theta) = \sum_{i=1}^{l} d_{ni} \left[ \frac{p_i'(\theta)[p_i''(\theta)]}{p_i(\theta)[1 - p_i(\theta)]} \right], \tag{27}
\]

\[
I(\theta) = \sum_{i=1}^{l} d_{ni} \left[ \frac{(p_i'(\theta))^2}{p_i(\theta)[1 - p_i(\theta)]} \right] \tag{28}
\]

and \( p_i'(\theta) \) and \( p_i''(\theta) \) are the first- and second-order derivatives of \( p_i(\theta) \). For a linear test, the WLE is attractive because of its unbiasedness to the order \( I^{-1} \).
In a Bayesian framework, estimation is based on the posterior distribution of $\theta$ given $u_n$ in (10). We either use the full posterior distribution or a point estimate derived from it. Examples of point estimates are the posterior mode (Bayes Modal or BM estimation; also known as the Maximum A Posteriori or MAP estimation, Lord, 1980) and its expected value (Expected A Posteriori or EAP estimate, Mislevy, 1986). Use of the EAP estimator in adaptive testing is discussed extensively in Bock and Mislevy (1988).

The choice between likelihood-based and Bayesian estimators in adaptive testing has many aspects. First of all, as is well known, for a noninformative prior, the ML and BM estimators are identical. For informative priors, the choice between an ML and BM estimator is generally one between a favorable bias or mean-squared error (MSE). However, the WLE in (25) has been shown to have smaller bias without sacrificing its MSE. On the other hand, the use of empirical prior distributions estimated from the individual subject’s response times on the items tends to lead to a much more favorable MSE for the Bayesian estimates (see below). Second, in the beginning of the test, when the posterior distribution tends to be considerably skewed, the EAP estimator behaves differently from the BM estimator. Interestingly, the impact of these differences have not yet been researched. Third, the item-selection rule in use for the test interacts with the estimation errors. It is therefore generally impossible to derive the behavior of estimators analytically. The only feasible way to study them is using computer simulation. In fact, more substantial remaining errors in the item parameters may also interact with the selection rule and, for instance, lead to a capitalization on chance (van der Linden and Glas, 2000a). Fourth, a fully Bayesian approach has the advantage of allowing for estimation error in $\theta$ when selecting the items. This feature leads to much more robust item selection in the beginning of the test when point estimates have a large likelihood of being off target and, hence, easily lead to suboptimal item selection. The best choice of estimator might be an eclectic one in which we combine different estimators for different stages in the test (van der Linden and Pashley, 2000, Section 2.3).

5. Empirical examples

Much research has been done on the properties of MML and Bayesian estimators for fixed tests. For adaptive test, however, the properties of the estimators for the item parameters depend on such factors as the calibration design and the distribution of the latent abilities, whereas the estimators of the subject parameters depend on additional factors such as the composition of the item pool as well as the presence of content constraints on the test, collateral information, and item–exposure control. Several of these factors are explained in Sections 5 and 6. Without any claim of generality, we present the results from two small studies, one on item calibration and the other on ability estimation, to illustrate some of the differences between MML and Bayesian estimation.

Table 1 shows the results from a comparison between MML item calibration and calibration using a fully Bayesian approach with MCMC computation. We drew 2,000 ability parameters from a standard normal distribution. For these parameters and the 15 sets of item parameters shown in the first columns of Table 1, item responses vectors were generated according to the 3PNO model. For the Bayesian method, guessing
parameters $c_i$ had a common beta (5, 17) prior distribution; for the two other item parameters, we chose noninformative prior distributions. The Gibbs sampler was run for 12,000 iterations; the first 2,000 were burn-in iterations. As shown in Table 1, the parameter estimates were generally within their estimated confidence band of two standard deviations. Further, we found no systematic differences between the MML and the Bayesian estimates.

To illustrate the difference between EAP and WML estimation of the ability parameter, we generated a pool of 200 items with responses following the 2-parameter logistic model. The item difficulties were drawn from a standard normal distribution and the item discrimination parameter from the distribution of $Y = .5 + .25X$, with for $X$ a $\chi^2$-distribution with one degree of freedom. We then simulated adaptive tests of 10, 20 or 40 items using the maximum-information rule in (34) below. For each test length,
Note that the EAP estimates had a slight outward bias, while the WML estimates were virtually unbiased. Further, the mean absolute errors of the WML estimates were nearly smaller than those of the EAP estimates. So, in the present study, the WML estimator performed better.

6. Rules for adaptive item selection

We follow the same distinction between a likelihood-based and a Bayesian approach as above, and discuss item selection from both perspectives. The following notation will be used. The items in the pool are still denoted as $i = 1, \ldots, I$, but, in addition, we use $k = 1, \ldots, m$ to denote the items in the adaptive test. Thus, $i_k$ is the index of the item in the pool administered as the $k$th item in the test. For convenience, we suppress subject index $n$ in the rest of this chapter.

Assume the $k$th item has to be selected. The preceding $k - 1$ items form the set $S_k = \{i_1, \ldots, i_{k-1}\}$. The set of items in the pool remaining after $k - 1$ items have been selected is $R_k = \{1, \ldots, I\} \setminus S_{k-1}$. The subjects have responses $U_{k-1} = (U_{i_1}, \ldots, U_{i_{k-1}})$ with realizations $u_{k-1} = (u_{i_1}, \ldots, u_{i_{k-1}})$ on the items in $S_k$. The point estimate of $\theta$ based on
these responses is denoted as $\hat{\theta}_{k-1} = \hat{\theta}_k$. The posterior distribution of $\theta$ has density $p(\theta | u_{k-1})$. Item $k$ is selected from set $R_k$ to be optimal at $\hat{\theta}_{k-1}$ or (some functional on) $p(\theta | u_{k-1})$.

The item pool is assumed to be calibrated by the three-parameter logistic model in (1). Item-selection rules for adaptive testing with polytomous items are discussed, for instance, in Dodd et al. (1995) and Koch and Dodd (1989). For rules for adaptive testing with multidimensional items, see Segall (1996, 2000) or Veldkamp and van der Linden (2002).

6.1. Likelihood-based selection

The second-order derivative in (24) reflects the curvature of the observed likelihood function of $\theta$ relative to its scale. Its negative is the observed information measure

$$J_{uk-1}(\theta) = -\frac{\partial}{\partial \theta^2} \ln L(\theta; u_{k-1}).$$

(29)

The expected valued of this measure over the response variables is the Fisher information:

$$I_{uk-1}(\theta) = E[J_{uk-1}(\theta)].$$

(30)

The 3PL model with known item parameters belongs to an exponential family and the observed and expected information measures are identical. For this model, (30) was already given in (28); using our current notation and ignoring the missing data indicator $d_{ni}$, it can be written as

$$I_{uk-1}(\theta) = \sum_{j=1}^{k-1} \frac{[p'_{ij}(\theta)]^2}{p_{ij}(\theta)[1 - p_{ij}(\theta)]},$$

(31)

with

$$p'_{ij}(\theta) \equiv \frac{\partial}{\partial \theta} p_{ij}(\theta).$$

(32)

For several other response models, (29) and (30) are different.

For a fixed test, with increasing test length, the variance of the MLE of $\theta$ is known to converge to the reciprocal of (31). In addition, (31) is easy to calculate and additive in the items. For this reason, Birnbaum (1968) introduced Fisher’s information as the main criterion for fixed-test assembly, where it is typically treated as function of $\theta$ and is more commonly known as the test information function.

A popular choice in adaptive testing is to select the $k$th item to maximize (31) at $\theta = \hat{\theta}_{k-1}$. Formally, the rule can be presented as

$$i_k \equiv \arg \max_j \{ I_{uk-1j}(\hat{\theta}_{k-1}); j \in R_k \}.$$  

(33)

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

$$i_k \equiv \arg \max_j \{ I_{uj}(\hat{\theta}_{k-1}); j \in R_k \}.$$  

(34)

This rule is known as the maximum-information rule in adaptive testing.
For an asymptotic motivation of maximum-information selection, see Chang and Ying (2006). They show that the standard result of the MLE of $\theta$ converging to its true value with sampling variance equal to the reciprocal of (31) still holds, provided the values for the discrimination and guessing parameters in the item pool are bounded away from 0 and $\infty$ and 1, respectively, but item selection is otherwise unconstrained. They also present a slight modification of the likelihood equation for the 3PL model that is sufficient to prevent multiple roots.

An alternative to (31) is the Kullback–Leibler information (Lehmann and Casella, 1996, Section 1.7) defined as the distance between the distribution of response $U_{ij}$ on candidate item $j$ at the true ability $\theta_0$ and the current ability estimate $\hat{\theta}_{k-1}$:

$$K_{ij} (\hat{\theta}_{k-1}, \theta_0) \equiv \mathbb{E} \left[ \ln \frac{L(\theta_0; U_{ij})}{L(\hat{\theta}_{k-1}; U_{ij})} \right],$$

where the expectation is over $U_{ij}$ at $\theta_0$. For the 3PL model,

$$K_{ij} (\hat{\theta}_{k-1}, \theta_0) = p_{ij}(\theta_0) \ln \frac{p_{ij}(\theta_0)}{p_{ij}(\hat{\theta}_{k-1})} + \left[ 1 - p_{ij}(\theta_0) \right] \ln \frac{1 - p_{ij}(\theta_0)}{1 - p_{ij}(\hat{\theta}_{k-1})}.$$  

(35)

Because of local independence, (35) is additive in the items and we can focus on the information by the candidate items. To deal with the fact that $\theta_0$ is unknown, Chang and Ying (1996) propose to integrate (35) over an interval about the current ability estimate, $[\hat{\theta}_{k-1} - \delta_k, \hat{\theta}_{k-1} + \delta_k]$, where $\delta_k$ is an arbitrary constant decreasing with $k$. This suggestion leads to the item-selection rule

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

(37)

If the constants $\delta_k$ are well chosen, (37) can be expected to be robust with respect to the estimation errors in $\hat{\theta}_{k-1}$ in the beginning of the test. As already observed, these errors can be extremely large and it may take some time for the test to recover from initial suboptimal item selection due to large errors, particularly if the subject happens to produce a few unlikely response at this stage.

6.1.1. Likelihood–weighted selection

The same wish of more robust item selection underlies the rule proposed by Veerkamp and Berger (1997). They take the likelihood function associated with the responses $u_{k-1}$ to express the plausibility of the various values of $\theta$ given the data. Hence, their suggestion of a likelihood–weighted information criterion:

$$i_k \equiv \arg \max_j \left\{ \int_{-\infty}^{\infty} L(\theta; u_{k-1}) I_{ik} (\theta) \, d\theta; \ j \in R_k \right\}.$$  

(38)

In the beginning of the test, the likelihood function is flat and values away from the MLE receive substantial weight. Toward the end of the test the likelihood function

becomes more peaked, and nearly all of the weight will go to the values close to the MLE.

Although the idea was presented for Fisher’s information measure, it can easily be extended to include the Kullback–Leibler information and then becomes an alternative to (37). Also, it is interesting to reflect on the differences between (38) and the weighted MLE in (25), where, the other way around, the likelihood function is enhanced with an information-related weight to produce a better point estimate of $\theta$.

6.2. Bayesian selection

Adaptive testing is a natural area for the application of sequential Bayesian methods, with the posterior distribution of $\theta$ after the preceding item serving as the prior distribution for the selection of the next item. If exchangeability of the subjects is assumed and the prior distribution for the first item is taken to be an estimate of the population distribution of $\theta$, such as that in (3), the method becomes fully empirical Bayes. Another motivation for the use of a Bayesian method is further improved robustness of the item selection during the first stage of the test relative to (37) and (38).

The posterior distribution after $k - 1$ items has density $p(\theta | u_{k-1})$. The basic idea is to select the $k$th item such that the posterior distribution (39)

\[
p(\theta | u_{k-1}, U_{ik}) \propto p(\theta | u_{k-1}) p(U_{ik} = u_{ik} | \theta),
\]

or a functional defined on it, becomes optimal in some sense. This general idea is known as preposterior analysis.

6.2.1. Owen’s approximate procedure

Owen (1969; see also 1975) pioneered the application of Bayesian methods to adaptive testing. Due to the computational restrictions of his time, he had to find an approximation to the full preposterior analysis that he actually wanted. To derive his approximation, he choose to work with the 3PNO model in (2).

Owen’s procedure started with a normal prior distribution for the selection of the first item. Next, the rule for choosing items $k = 2, 3, \ldots$ was

\[
|b_{ik} - \mathcal{E}(\theta | u_{k-1})| < \delta
\]

for a small value of $\delta \geq 0$. His suggestion was to stop when the posterior variance became smaller than a predetermined tolerance. Observe that (40) selects an item that is closer than $\delta$ to the current EAP estimate of $\theta$.

Since the normal prior is not conjugate with the binomial distribution of the response variables, the result is an intractable posterior distribution. Much of Owen’s work was devoted to finding closed-form approximations to the true posterior variance and mean. The approximation for the mean that he suggested was motivated by its convergence to the true value of $\theta$ in mean square for $k \to \infty$ (Owen, 1975, Theorem 2).

6.2.2. Preposterior selection

The rules in this section illustrate full preposterior item selection. We show a few rules that predict the responses on the items $i \in R_k$ and select the $k$th item with an optimal expected update of the posterior quantity we are interested in. The prediction of
the response on item $i$ is provided by the predictive posterior distribution, which has probability function

$$p(u_i \mid u_{k-1}) = \int p(u_i \mid \theta) p(\theta \mid u_{k-1}) d\theta. \tag{41}$$

An obvious application is to select the item with the minimum expected posterior variance,

$$i_k \equiv \arg \min \left\{ \frac{1}{u_j = 0} \sum \text{Var}(\theta \mid u_{k-1}, U_j = u_j) p(u_j \mid u_{k-1}) \mid j \in R_k \right\}. \tag{42}$$

The expression for the variance in this criterion follows immediately from (39). The criterion is equivalent to minimization of the preposterior risk under a quadratic loss function for the EAP estimator.

From a Bayesian perspective, the rule in (38) can be generalized by replacing the likelihood function by the posterior density of $\theta$ as weight:

$$i_k \equiv \arg \max \left\{ \int I_{U_j = u_j} (\theta) p(\theta \mid u_{k-1}) d\theta \mid j \in R_k \right\}. \tag{43}$$

A further improvement on (43) can be based on a preposterior argument. We would then predict the posterior weighted information in the responses on the next item and calculate its expectation over the predictions:

$$i_k \equiv \arg \max \left\{ \int J_{u_{k-1}, U_j = u_j} g(\theta \mid u_{k-1}, U_j = u_j) d\theta \times p(u_j \mid u_{k-1}) \mid j \in R_k \right\}. \tag{44}$$

The last two rules can also be formulated with the Kullback–Leibler information and are then alternatives to (37). Also, we could predict more than one item ahead. For a review of these and other Bayesian item-selection rules for the 3PL model as well as empirical comparisons between some of them, see van der Linden (1998) or van der Linden and Pashley (2000). Applications to polytomous models are presented in Penfield (2006).

### 6.2.3. Collateral Information

Item selection can be further improved by using collateral information on the subjects, such as earlier test scores, biographical data, etc. If the information is used in a Bayesian framework, it should be incorporated in the prior distribution of $\theta$. As a result, it allows us to start the test from an individual empirical prior instead of a common (subjective) prior distribution.

Procedures for adaptive testing with a prior distribution regressed on background variables $X_p$, $p = 0, \ldots, P$, are described in van der Linden (1999). They are based on linear regression of $\theta$ on the predictor variables:

$$\theta = \beta_0 + \beta_1 X_1 + \cdots + \beta_P X_P + \varepsilon. \tag{45}$$
with
\[ \varepsilon \sim N(0, \sigma^2). \]  

If the items have been calibrated, substitution of (46) into the response model leads to a logistic regression model with unknown parameters \((\beta_0, \beta_1, \ldots, \beta_P, \sigma)\) and unknown errors \(\varepsilon_n\). It is a standard application in logistic regression to estimate the parameters using an EM algorithm with the \(\varepsilon_n\) as missing data. This should be done along with the regular item calibration. Once the parameters have been estimated, an obvious prior distribution for subject \(n\) is
\[ p(\theta_n) = N(\hat{\theta}_0n, \sigma) \]  
with
\[ \hat{\theta}_0n = \beta_0 + \beta_1 x_1n + \cdots + \beta_P x_Pn. \]

In adaptive testing, the response times on the items form a unique source of collateral information. They are automatically recorded. When a new item has to be selected, the time spent on the previous item can be used to retrofit the previous prior distribution. This retrofitting is in addition to the regular update of the likelihood using the response on the previous item. If the speed and ability at which the population of subjects works are dependent, the retrofitted prior quickly becomes an informative individual empirical prior for the estimation of ability parameter.

To explore the benefits of response times as item-level collateral information, van der Linden (2005) (see also van der Linden, 2006a) used a lognormal model for the response time \(T_i = t_i\) of a subject on item \(i\) operating at speed \(\tau \in (-\infty, \infty)\) during the test,
\[ p(t_i; \tau, \alpha_i, \beta_i) = \frac{\alpha_i}{t_i \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left[ \alpha_i (\ln t_i - (\beta_i - \tau)) \right]^2 \right\}, \]
where \(\beta_i \in (-\infty, \infty)\) and \(\alpha_i \in (0, \infty)\) are parameters for the time intensity and discriminating power of item \(i\). To import the information in \(t_i\) on \(\theta\), the model was extended with a second-level distribution
\[ (\theta, \tau) \sim MVN(\mu_P, \Sigma_P), \]
which relates \(\theta\) to the speed \(\tau\) for the population of subjects. It is assumed that the free parameters in \(\mu_P\) and \(\Sigma_P\) have been estimated during item calibration. (A few of the parameters have to be fixed to make the model identifiable.)

After the response times on \(k - 1\) items have been recorded, the posterior density of the subject’s speed parameter given these times,
\[ p(\tau | t_{k-1}) \propto p(\tau) \prod_{j=1}^{k-1} p(t_{ij}; \tau), \]
is calculated from (49), adopting a prior distribution \(p(\tau)\). Along with the conditional density of the ability given speed, \(p(\theta | \tau)\), which follows from the known distribution...
in (50), the posterior distribution allows us to calculate the posterior predictive density of \( \theta \) given the response times \( t_{k-1} \) as

\[
p(\theta \mid t_{k-1}) = \int p(\theta \mid \tau)p(\tau \mid t_{k-1}) \, d\tau.
\] (52)

When selecting the \( k \)th item, this predictive density is used to retrofit the prior in the posterior distribution of \( \theta \) after the previous \( k - 1 \) items. The selection of the \( k \)th item is then based on the following prior distribution:

\[
p(\theta \mid u_{k-1}, t_{k-1}) \propto p(\theta \mid t_{k-1}) \prod_{j=1}^{k-1} P(u_{ij} \mid \theta).
\] (53)

If the times are measured on a logarithmic scale, both factors in the integrand in (51) are normal, and \( p(\theta \mid t_{k-1}) \) is normal with known mean and variance (for their expressions, see van der Linden, 2005). The procedure is therefore easy to implement.

Figure 1 shows the results from a computer simulation study of adaptive tests of 10 and 20 items. It appears that, for a moderate correlation of \( \rho_{\theta \tau} = .20 \), a 10-item adaptive test with the use of the response times is approximately equally accurate as a 20-item test without it.

Fig. 1. MSE functions for an adaptive test without and with the use of response times for items selection \((n = 10 \text{ and } 20)\).
6.2.4. Item cloning  
Bayesian hierarchical modeling is also convenient for item selection from a pool with families of cloned items. As indicated earlier, an appropriate model for such a pool is a response model extended with a separate second-level distribution of the item parameters for each family.  
Let \( f = 1, \ldots, F \) denote the families in the pool. For notational convenience, we assume that the test consists of one random item from each of a selected subset of families and use \( f_k \) to denote the family in the pool that provides the \( k \)th item in the test. For the 3PL model, an obvious generalization of (4) is to choose

\[
\xi_f = (a_f, b_f, c_f) \sim \text{MVN}(\mu_f, \Sigma_f), \quad f = 1, \ldots, F, \tag{54}
\]

as the second-level distributions. The means \( \mu_f \) and covariance matrices \( \Sigma_f \) are assumed to be estimated from samples of items from the families in the pool during a previous calibration study.  
Item selection now has to follow a two-stage process: first, a family is selected and then an item is picked from the family. The selection of the family can still be adapted to the current estimate of \( \theta \) but the selection of the item has to be random.  
Suppose the \( k \)th family is to be selected. The responses on the items from the \( k - 1 \) previously selected families are denoted as \( u_{k-1} = (u_{f_1}, \ldots, u_{f_{k-1}}) \). The update of the posterior distribution of \( \theta \) after these responses is

\[
p(\theta | u_{k-1}) \propto p(\theta) \prod_{j=1}^{k-1} \int_{\cdot} \int_{\cdot} \int_{\cdot} \int_{\cdot} p(u_{f_j} | \theta, \xi_j) p(\xi_j | \mu_{k-1}, \Sigma_j) d\xi_j. \tag{55}
\]

The selection has to be optimal with respect to this posterior distribution.  
Analogous to (42), the family could be selected as

\[
f_k \equiv \arg \min_j \left\{ \sum_{u_j=0}^1 \text{Var}(\theta | u_{k-1}, U_j = u_j) p(u_j | u_{k-1}); \ j \in R_k \right\}, \tag{56}
\]

where \( R_k \) now denotes the set of candidate families available for selection and \( u_j = 0, 1 \) are the possible responses to a random item from candidate \( j \in R_k \).  
If the interest is in a criterion based on an information measure, we could define the expected information in a random item from family \( f \) as

\[
I_f(\theta) = \int \cdots \int \left[ -\frac{\partial^2}{\partial \theta \partial \theta} \ln L(\theta; u_f, \xi_f) \right] p(\xi_f | \mu_f, \Sigma_f) d\xi_f, \tag{57}
\]

and, for example, use posterior weighting to select the \( k \)th family as

\[
f_k = \arg \max_j \left\{ \int I_j(\theta) p(\theta | u_{k-1}) d\theta; \ j \in R_k \right\}. \tag{58}
\]

The loss of information due to the sampling of the items from their families depends on the relation between the within-family and between-family item variation. For an empirical study of these effects for the rule in (56), see Glas and van der Linden (2003).
7. Other statistical issues in adaptive testing

When the first large-scale testing programs moved to an adaptive format in the early 1990s, it quickly became clear that calibrating an item pool and adopting an item-selection algorithm were not enough. In fact, several new problems with interesting statistical aspects presented themselves. In this section, we discuss a few of them.

First, just like paper-and-pencil tests, to maintain their content validity (and, therefore, their fit to the response model), adaptive tests have to meet an extensive set of content specifications. This set has to be imposed while maintaining the adaptive nature of the item selection. As we will see below, the consequence is a complicated constrained sequential optimization problem, typically with hundreds of constraints, which has to be solved in real time.

Second, an equally urgent new problem was item-pool security. By its very nature, an item-selection rule in adaptive testing picks the best items in the areas on the ability scale where the subjects’ estimates are. Typically, a small set of highly discriminating items tends to dominate the whole scale. For an unconstrained test, this set need not be larger than 5% of the entire item pool or so. Since the pool is in continuous use for some time, it is relatively easy for test takers to find out what the dominant items are and share them with future test takers. As explained below, an effective reaction by the testing agency is to build a random experiment into the test. The experiment has to create more evenly distributed exposure rates for the item without sacrificing too much of the other desired features of the test.

Third, fixed tests are usually assembled with an eye on the time limit in force for them. In order to schedule the subjects efficiently, most real-world adaptive tests also have a fixed length and time limit. But different subjects get different sets of items and the items in the pool typically vary considerably in their time-intensity. As a result, some of the subjects may run out of time whereas others have more than enough time to finish their test. This phenomenon is known as differential speededness of the test. To make adaptive tests fair, they should not suffer from it. As we will show, a solution to this problem entails the necessity to calibrate the items with respect to their time properties as well and use the parameter estimates in the item-selection rule.

Finally, for security reasons, it is not feasible to provide test takers with copies of the test they have taken. In order to give them the opportunity to interpret their scores relative to the content of typical test items, it is standard practice to release a fixed test assembled to the same specifications and equate the score on the adaptive test to a number-correct score on this reference test. The problem of how to equate an adaptive test to a fixed test is not trivial, though.

In the next section, we discuss a solution to adaptive item selection that can not only be used to solve the problem of imposing identical content specifications on the test for all subjects but offers a more general framework that resolves the other problems as well.

7.1. Content specifications

The necessity to impose a set of content specifications on an adaptive test introduces a fundamental dilemma in item selection: To optimize the accuracy of the estimator...
of $\theta$, the items have to be selected sequentially, with an update of the estimator after each item. However, to realize a set of content specifications, we have to select them simultaneously; otherwise, during a test, we might soon observe that some types of items tend to be under- or overrepresented but that attempts to correct for this would lead to the violation of some of the other specifications or the selection of much less favorable items for the accuracy of the ability estimator.

A more general solution of this problem is possible if we treat it as an instance of constrained combinatorial optimization, that is, the choice of combination of items from the pool that satisfies a set of constraints representing the content specifications and is optimal with respect to the ability estimator. This approach would also allow us to deal with any requirement for the test that can be formulated as an explicit constraint on the selection of its items.

To implement the idea, we use a selection process with the following steps:

1. Begin the adaptive test with the selection of a full-size fixed test that is optimal at the initial estimate of $\theta$ and meets all constraints.
2. Administer one item from the test that is optimal at the estimate.
3. Update the estimate of $\theta$.
4. Reassemble the full-size test to be optimal at the updated estimate of $\theta$ while still meeting all constraints and fixing the items that have already been administered.
5. Repeat Steps 2–4 until the adaptive test is completed.

The full-size tests that are assembled, which the subject does not see, are known as shadow tests. Because each of the shadow tests meets all constraints, the adaptive test does. Similarly, since all shadow tests are assembled to be optimal at the current estimate of $\theta$ and the best item at this estimate is administered, the adaptive test is our best choice given all the constraints (van der Linden and Reese, 2003; van der Linden, 2000).

Instead of as a sequence of $m$ constrained combinatorial optimization problems, we can also view the shadow-test approach as an application of a projection method: Because the item-selection algorithm is not allowed to look back and undo any earlier choices of items that appear to lead to constraint violation later in the test, the only option left is to look forward and project the rest of the test. The shadow-test approach does so by calculating a projection that is both feasible with respect to the constraints and optimal with respect to the ability estimator, and then picks the best item from the projection.

The only problem left is that of how to calculate the shadow tests in real time. To do so, we formalize the test-assembly problem more explicitly as a combinatorial optimization problem with binary variables $x_i, i = 1, \ldots, I$, that take the value one if the item is selected for the shadow test and zero otherwise. The variables are used to model the content specifications as a set of constraints, which typically fall into three different classes: (i) constraints on categorical item attributes, such as content categories, item format (e.g., open-ended or multiple-choice format), and problem type (e.g., knowledge, analysis, or application), (ii) constraints on quantitative attributes, such as statistical parameters, word counts, and expected response time, and (iii) logical
constraints to deal with conditional item selection, such as the requirement that one item cannot be selected if any from another set is.

Suppose the test is of size of $m$. As an example of a categorical constraints, we use the set $V_c$ for the indices of the items in the pool that belong to category $c$. Likewise, $q_i$ is the value of item $i$ for a quantitative attribute that we want to constrain. To illustrate the use of logical constraints, we assume that the items with the indices in set $V_e$ cannot be selected for the same test. Finally, the items are selected using the maximum-information rule in (34).

The shadow-test model for the selection of the $k$th item in the adaptive test is

$$\text{maximize } \sum_{i=1}^{l} I_i(\hat{\theta}_{k-1})x_i \quad \text{(test information)}$$

subject to

$$\sum_{i=1}^{l} x_i = m \quad \text{(test length)},$$

$$\sum_{i \in V_c} x_i \geq m_c, \text{ for all } c \quad \text{(categorical attributes)},$$

$$\sum_{i \in V_e} x_i \leq 1, \text{ for all } e \quad \text{(item exclusion)},$$

$$x_i \in \{0, 1\}, \text{ for all } i \quad \text{(range of variables)}.$$
next shadow test can be found using a “hot start” of the solver. For a modern solver with optimized settings, the running times for a real-world adaptive test are negligible.

The model in (59)–(65) can easily be used with any of the other item-selection rule in Section 5. It can also be extended with constraints at other levels in the tests (e.g., item sets with common stimuli and subtests), constraints for more complicated logical selection, etc. For these and other implementation details, see van der Linden (2005).

Figure 2 shows the bias and MSE functions for a 50-item adaptive version of the Law School Admission Test (LSAT) from a pool of 753 items with 100 simulated subjects at $\theta = -2.0, -1.5, \ldots, 2.0$. The test had over 400 constraints that were automatically satisfied for each simulated subject. Except for the lower end of the scale (the item pool
7.2. Item–exposure control

The exposure rates of the items in the pool can be controlled by running a probability experiment before or after a subject begins the test. The first experiments introduced in the literature use the former option. The most popular one is the Symposion–Hetter experiment (1985) (see also Hetter and Sympson, 1997). The experiment is conducted before a newly selected item is administered; the outcome is the decision either to administer the item or to pass and select the next best item at the current estimate of $\theta$. If an item is passed, it is removed from the pool for the subject.

The probability of selecting an item is fixed by the setup of the test (item-selection rule; composition of the item pool; ability estimator; etc.). The only way to control the exposure rates ($=$ probabilities of item administration) for this type of experiment is through manipulation of the conditional probabilities of administering an item given its selection. Because their joint effect on the distribution of the exposure rates is the result of a complicated interaction between them, these conditional probabilities have to be set through an iterative process of simulated adaptive test administrations. The process is generally cumbersome; for a realistic adaptive test it is not unusual to have to run some 150 simulations before a set of probabilities that guarantee a maximum exposure rate below say, .30, at selected ability values is found. For an evaluation and alternative implementations of Symposion–Hetter item–exposure control, see van der Linden (2003).

The other class of methods of random control is based on a probability experiment conducted before a subject begins the test. The experiment is to determine which items are eligible for the subject and which are not. If an item is eligible, it remains in the pool for the subject; otherwise it is removed.

To model the experiment, we consider two different kinds of events: the events of an item being eligible ($E_i$) and an item being administered ($A_i$). It holds that $A_i \subset E_i$ (66) for all $i$. We want to control the conditional exposure rates given $\theta$, $\Pr(A_i \mid \theta)$, because otherwise subjects of comparable ability levels might collude and discover items with dangerously high local exposure rates (Stocking and Lewis, 1998). From (66),

$$\Pr(A_i \mid \theta) = \Pr(A_i, E_i \mid \theta) = \Pr(A_i \mid E_i, \theta) \Pr(E_i \mid \theta)$$

(67)

for all possible values of $\theta$. Imposing a maximum exposure rate, $r_{\text{max}}$,

$$\Pr(A_i \mid \theta) \leq r_{\text{max}},$$

(68)

or

$$\Pr(E_i \mid \theta) \leq \frac{r_{\text{max}}}{\Pr(A_i \mid E_i, \theta)}, \quad \Pr(A_i \mid E_i, \theta) > 0.$$  

(69)
From (67), it follows that
\[ \Pr(E_i \mid \theta) \leq \frac{r^{\max} \Pr(E_i \mid \theta)}{\Pr(A_i \mid \theta)}, \quad \Pr(A_i \mid \theta) > 0. \] (70)

The result in (70) is used to form a recurrence relation. Suppose \( n \) subjects have already taken the test, and we want to know the probabilities of eligibility for subject \( n + 1 \) at selected points \( \theta_g, g = 1, \ldots, G \). The probabilities are
\[ \Pr^{(n+1)}(E_i \mid \theta_g) = \min \left\{ \frac{r^{\max} \Pr^{(n)}(E_i \mid \theta_g)}{\Pr^{(n)}(A_i \mid \theta_g)}, 1 \right\}. \] (71)

By substitution, it is easy to show that the probabilities remain equal, increase, or decrease if they are equal, smaller, or greater than \( r^{\max} \) for the previous subject. This feature of self-adjustment enables us to just begin the test, without the necessity of any prior study to find optimal setting for the control parameters.

To implement the method, we conceive of \( G \) different versions of the item pool, one at each \( \theta_g, g = 1, \ldots, G \), and record the following information: (i) the number of earlier subjects who visited item pool \( g \) and took item \( i \) and (ii) the number of earlier subjects who visited item pool \( g \) when item \( i \) was eligible. Let \( \alpha_{nig} \) and \( \varepsilon_{nig} \) denote these numbers. The probabilities for \( n + 1 \) are estimated as
\[ \Pr(E_i \mid \theta_g) = \min \left\{ \frac{r^{\max} \varepsilon_{nig}}{\alpha_{nig}}, 1 \right\}. \] (72)

During the test, we ignore the differences between the subject's true and estimated ability and use the probabilities at the point \( \theta_g \) closest to the estimated ability. For a choice of \( G = 10 \) or so, the impact of the estimation error on the actual exposure rates has shown to be negligible for all practical purposes.

In order to get smooth behavior of the probabilities of eligibility, it is recommended to update the numbers \( \alpha_{nig} \) and \( \varepsilon_{nig} \) using the technique of fading, which was developed for the update of posterior probabilities in Bayesian networks (e.g., Jensen, 2001, Section 3.3.2). This type of update involves a factor \( w \) slightly smaller than one, which is used to weigh the earlier events relative to current event. For example, \( \alpha_{nig} \) is then updated as \( \alpha^*_{(n+1)i|g} = w\alpha^*_{nig} + 1 \) when item \( i \) is administered to subject \( n \) and as \( \alpha^*_{(n+1)i|g} = w\alpha^*_{nig} \) when it is not. Updating with a fading factor \( w \) produces estimates of the probabilities of eligibility based on an effective sample of the last \( 1/(1-w) \) subjects, with a weight equal to one for subject \( n \) and weights approaching zero for the subjects \( n - 1, n - 2, \ldots \). A practical value is \( w = .999 \), which amounts to an effective sample size of 1,000.

In practice, we should use this method of item–exposure control in combination with the shadow-test approach in the preceding section to allow for all other specifications that have to be imposed on the test. The only thing that has to be done is to introduce the constraints
\[ x_i = 0 \] (73)
for the items that are ineligible in the model for the shadow test in (59)–(65). The update of the probabilities in (71) then requires a slight modification to account for the possibility of overconstraining. (If infeasibility should happen, all items remain in the pool;
the algorithm automatically adjusts the future probabilities of eligibility for an extra exposure.) For these and other details of item–exposure control using random ineligibility constraints, see van der Linden and Veldkamp (2004, 2005).

An example of the results for the method for a 25-item section of the LSAT is given in Figure 3. The results are based on 10,000 simulated subjects. The exposure rates were controlled at the nine values $\theta_g = -2.0, -1.5, \ldots, 2.0$. We show the resulting rates for the cases of no exposure control and control with $r_{\text{max}}$ equal to .15 and .10. The rates are shown at five different values $\theta_g$; the rates at for the other values were entirely comparable.

7.3. Differential speededness

Effective constraints for making a test equally speeded for all subjects can only be formulated when the item pool is calibrated with respect to the time properties of the items as well. Suppose that the lognormal model for response times in (49) has been used. The properties of the items are then captured by (known) parameters $\beta_i$ and $\alpha_i$ for their time intensity and discriminating power, respectively. In addition, we assume that the response times of the subject are used to update the estimate of speed parameter $\tau$ during the test.

From a measurement point of view, the item parameters are nuisance parameters for which we have to control. We will select the items such that their impact on the total amount of time that is required for the test does not vary across the subjects. A moment’s reflection reveals that, provided their values are not unduly large, the parameters $\alpha_i$ do not have much impact on the distribution of the total time. The total logtime has variance $\sum a_i^{-2}/m^2$. Since the time on an item does usually not exceed two minutes or so, the variability of the total time on a 25-item test is already ignorable for all practical purposes. Therefore, for real-world tests, it is typically unnecessary to control for the impact of $\alpha_i$.

The answer to the question of whether we should treat the speed parameter as a nuisance or an intentional parameter depends on the goal of the test. If the test is a pure power test, $\tau$ should be treated as a nuisance parameter. In this case, we should constrain the selection of the items to give each subject enough time to relative to the time limit. If it is intended to have a speed aspect as well, $\tau$ should be treated as an intentional parameter, and the selection of the items should not be constrained with respect to $\tau$. (In fact, it would then be correct to score the test using the vector $(\hat{\theta}, \hat{\tau})$ but this is not common yet.)

We first treat both the item parameters and the speed parameter as nuisance parameters and constrain the time spent on the test with respect to their impact. Since $\tau$ is unknown, it has to be estimated during the test. In order to account properly for the estimation error, we adopt the posterior distribution in (51) as the estimate of $\tau$. This posterior distribution can used be to calculate the predictive distributions on all remaining items in the pool for the subject. If $k-1$ items have been selected, the predictive densities are

$$p(t_i | t_{k-1}) = \int p(t_i | \tau) p(\tau | t_{k-1}) \, d\tau,$$

for all $i \in R_k$. 

\[1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 17\ 18\ 19\ 20\ 21\ 22\ 23\ 24\ 25\ 26\ 27\ 28\ 29\ 30\ 31\ 32\ 33\ 34\ 35\ 36\ 37\ 38\ 39\ 40\ 41\ 42\ 43\ 44\ 45\]
Fig. 3. Conditional item-exposure rates at $\theta = -2.0$ (1) without control (dark curves), control with $r_{\text{max}} = .15$ (gray curves), and control with $r_{\text{max}} = .10$ (light gray curves).
Let $t_{i}^{\pi}$ be the $\pi$th quantile in the predictive distribution for item $i$:

$$
\int_{t_{i}^{\pi}} p(t_{i} \mid t_{k-1}) \, dt_{i} = \pi. \tag{75}
$$

If $t_{\max}$ is the time limit for the test, we impose the following constraint on the shadow test for item $k$:

$$
\sum_{i \in S_{k-1}} t_{i} + \sum_{i \in R_{k}} t_{i}^{\pi} x_{i} \leq t_{\max} \tag{76}
$$

(van der Linden et al., 1999). The first term in (76) is the actual total time recorded for the first $k-1$ items in the test; the second term is the sum of the $\pi$th quantiles in the predicted distributions on the remaining $n-k$ items. The constraint requires the items to be selected such that the projected time on the entire test not be larger than $t_{\max}$.

Observe that (76) is linear in the variables $x_{i}$. Also, if logtimes are used, the densities in (74) are normal, and so is $p(t_{j} \mid t_{k-1})$. The quantiles in (75) are then easy to calculate.
To make the procedure conservative, it makes sense to begin the test with $\pi = .50$ but let $\pi$ increase toward the end of it.

If $\tau$ is an intentional parameter, we control the selection of the items only for the impact of the item parameters. As already indicated, we ignore $\alpha_i$ and focus on $\beta_i$. A convenient way to identify the model in (49) is to set $\mu_{\tau}$ in (50) equal to zero. Under this assumption, for a subject taking the test at the average speed in the population, the expected time on item $i$ on a logarithmic scale is equal to $\beta_i$. Thus, on the natural scale, the expected time is equal to $\exp(\beta_i)$.

The expected total time on the test can now be controlled by including the following constraint in the model for the shadow test:

$$\sum_{i=1}^{I} \exp(\beta_i) x_i \leq t_{\text{max}} - \delta,$$

(77)

where $\delta$ is a speededness parameter to be selected by the testing agency (the larger $\delta$, the more speeded the test). No matter the value of $\delta$ chosen, (77) makes the test equally speeded for all subjects.

Observe that (77) is independent of $k$; the constraint on the previous items in (63) guarantees that its left-hand sum automatically contains the values of $\beta_i$ for the $k-1$ items already administered. Also, the constraints in (76) and (77) are formulated at the level of the total test. If stricter control, in the sense of more homogeneous speededness throughout the test, is required, we can divide the total testing time into equally long periods and impose the constraints at the level of subsequent blocks of items for the periods.

The following example is derived from van der Linden (2005, Section 9.5). For one of the adaptive tests from the Armed Services Vocational Aptitude Battery (ASVAB), we calibrated the items using a constrained version of the lognormal model in (49) and simulated adaptive test administrations for subjects in the range of values for the speed parameter $\tau$ we found in the sample of subjects. The first plot in Figure 4 shows the total time used by the simulated subjects in two of the ability groups. At the lower speed levels, the subjects in these two groups ran out of time. The second plot is for the replication of the same simulation but now with the constraint in (76) in the shadow-test model. The two ability groups no longer ran out of time.

### 7.4. Observed-score reporting

Classical methods of score equating that have been used in adaptive testing are equipercentile equating with randomly equivalent groups of subjects and true-score equating using the test characteristic function (for a description of these methods, see Kolen and Brennan, 2004). The first method requires an equating study in which the population distributions of the scores on the two tests are to be estimated. The equating transformation is defined as the transformation that equates the quantiles in the two distributions. These equating studies are time consuming and expensive. They are particularly cumbersome for the equating of an adaptive test to a fixed reference test, since they have to be repeated each time an item pool is changed. The second method is more efficient...
Fig. 4. Mean time used in adaptive test by subjects in ability groups $\theta = 2.0$ (solid curves) and $\theta = 1.5$ (dashed curves) without and with control for speededness (time limit: 2340 secs = 39 mins).

but confounds the estimated true scores on the reference test with its observed number-correct scores. More importantly, both methods have been shown to be seriously biased because the single equating transformation that they use has to compromise between the observed-score distributions of the subjects at widely different ability levels (van der Linden, 2006b).

A more efficient way of equating the scores on an adaptive test to a linear test is to select the items to yield a number-correct scale for the adaptive test that is identical to that of the reference test for each subject. This can be done using a simple set of linear constraints on the items (van der Linden, 2001). From a practical point of view, the alternative is attractive because it avoids the necessity of an equating study prior to the operational use of the adaptive test. From a statistical point, it is attractive because it equates the scores on the two tests locally at the estimate of $\theta$ and thereby avoids much of the bias inherent in the classical equating methods.

The set of linear constraints follows from a condition on two tests to yield identical observed number-correct scores conditional on $\theta$ derived in van der Linden and Luecht (1998). Suppose we have two tests of $n$, items with number-correct scores $X$ and $Y$. For a subject with ability $\theta$, the probabilities of a correct response on the two tests are
Statistical aspects of adaptive testing

1 denoted as \( p_i(\theta) \) and \( p_j(\theta) \), \( i, j = 1, \ldots, n \). The \( r \)th power of the probability of item \( i \) is denoted as \( p_i^r(\theta) \). The two tests have identical conditional distributions of \( X \) and \( Y \) given \( \theta \) if and only if

\[
\sum_{i=1}^{n} p_i^r(\theta) = \sum_{j=1}^{n} p_j^r(\theta), \quad r = 1, \ldots, n, \tag{78}
\]

that is, if the powers of their success probabilities up to the \( n \)th order are equated. In addition, the importance of these conditions vanishes with the order of the power. In fact, if the test length increases, the conditions for \( r \geq 2 \) become superfluous. For the typical test lengths in educational and psychological testing, the conditions have to be satisfied only for the first 2 or 3 powers.

Observe that the conditions in (78) are linear in the items. It is thus a simple step to incorporate them in the shadow-test model. The model then automatically produces a test for with the same observed-score distribution as the reference test for each of the subjects.

Let \( j = 1, \ldots, n \) denote the items in the reference test and \( T_r(k-1) = T_r(\hat{\theta}_{k-1}) \) the right-hand sum of (78) for the \( r \)th power. These target values can be calculated easily from the response model. The following constraints impose the target values on the shadow tests for item \( k \):

\[
\sum_{i=1}^{l} p_i^r(\hat{\theta}_{k-1}) x_i \leq T_r(k-1) + \varepsilon, \quad \text{for all } r \leq R, \tag{79}
\]

\[
\sum_{i=1}^{l} p_i^r(\hat{\theta}_{k-1}) x_i \geq T_r(k-1) - \varepsilon, \quad \text{for all } r \leq R, \tag{80}
\]

where \( R = 2, 3 \) and \( \varepsilon \) is a small tolerance needed to avoid infeasibility due to an equality constraint. In the beginning of the test, when \( \hat{\theta}_{k-1} \) has a tendency to larger estimation errors, it may be useful to add similar constraints at the values \( \hat{\theta}_{k-1} + \gamma_k \) and \( \hat{\theta}_{k-1} - \gamma_k \), where \( \gamma_k \) should be chosen to decrease with \( k \). Alternatively, we could constraint the sums of powers of the response probabilities at fixed set of values \( \theta_g, g = 1, 2, 3 \).

The best choice of constraints for adaptive testing still has to be researched. We therefore show an example of the use of (78) for the equating of the conditional number-correct-score distributions on two fixed versions of the LSAT. One version served a 101-item reference test; the other was assembled from a pool of over 600 items subject to the constraints in (79) and (80) at the fixed values \( \theta_g = -1.2, 0, 1.2 \) for \( R = 2 \). Their conditional observed-score distributions given \( \theta = -1.5, -0.5, 0.5, 1.5 \) are compared in Figure 5. The distributions appear to be identical for all practical purposes.

8. Concluding comment

The use of adaptive tests is a relatively new phenomenon in educational and psychological testing but over the last decade its technology has quickly matured. In this chapter,
Fig. 5. Distributions of number-correct score on adaptive test (dashed curves) and fixed reference test (solid curves) for subject at $\theta = -1.5$. (Note: differences between the curves are indiscernible.)
we reviewed some of its statistical aspects but had to leave others untouched. For example, we did not discuss any of the aspects of adaptive testing from an item pool with polytomously scored items or items that require a response model with a multidimensional ability parameter. The latter has additional complexity because we have to adapt the item selection to the individual components of $\theta$ or a combination of them. Another area that we did not discuss is the use of statistical procedures for checking response vectors for possible aberrances due to design errors in the test or deviant behavior by subjects. A review of Bayesian procedures for fixed tests is given in Glas and Meijer (2003). But these procedures may not have much power for adaptive tests, where the response probabilities for the model in (1) quickly approximate a value close to .50 if the adaptation is successful and wild response patterns become likely. Additional information, however, can be found in the pattern of the response times on the items. Finally, we have not treated such important areas as item-pool design for adaptive testing, testing with adaptation at a higher level in the test than the individual items (e.g., multistage testing), and newer developments in multidimensional adaptive testing. For these and other aspects, see Chang (2004), van der Linden and Glas (2000b, 2006), or Wainer (2000).

Uncited references

(Chang and Ying, 1999) (Glas and van der Linden, 2006)

References


Statistical aspects of adaptive testing


