Chapter 14
Item Parameter Estimation and Item Fit Analysis

Cees A.W. Glas

14.1 Introduction

Computer-based testing (CBT), as computerized adaptive testing (CAT), is based on the availability of a large pool of calibrated test items. Usually, the calibration process consists of two stages.

(1) A pretesting stage: In this stage, subsets of items are administered to subsets of respondents in a series of pretest sessions, and an item response theory (IRT) model is fit to the data to obtain item parameter estimates to support computerized test administration.
(2) An online stage: In this stage, data are gathered in a computerized assessment environment, proficiency parameters for examinees are estimated, and the incoming data may also be used for further item parameter estimation.

The topic of this chapter is the estimation of the item parameters and the evaluation of item fit, both in the pretest phase and in the online phase. Especially differences in item parameter values in the pretest and online stages are of interest. Such differences are often named *parameter drift*. Evaluation of parameter drift boils down to checking whether the pretest and online data comply with the same IRT model. Parameter drift may have different sources. Security is one major problem in adaptive testing. If adaptive testing items are administered to examinees on an almost daily basis, after a while some items may become known to new examinees. In an attempt to reduce the risk of overexposure, several exposure control methods have been developed. All these procedures prevent items from being administered more often than desired. Typically, this goal is reached by modifying the item selection criterion so that "psychometrically optimal" items are not always selected. Examples of methods of exposure control are the random-from-best-n method (see, e.g., Kingsbury & Zara, 1989, pp. 369–370), the count-down random method (see, e.g., Stocking & Swanson, 1993, pp. 285–286), and the method of Sympson and Hetter (1985; see also Stocking, 1993). With relatively low exposure

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

W.J. van der Linden and C.A.W. Glas (eds.), *Elements of Adaptive Testing*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-0-387-85461-8_14,

rates, items will probably become known later than with high exposure rates. Still, sooner or later some items may become known to some future examinees.

Differences between the pretest and the online stages may also result in other forms of parameter drift. One might, for instance, think of differences in item difficulty resulting from the different modes of presentation (computerized or paper-and-pencil administration) or resulting from a changing curriculum. Also, differences in motivation of the examinees between the pretest and online stages might result in subtle shifts of the proficiency that is measured by the test. Response behavior in these stages might not be properly modeled by the same set of IRT parameters when examinees in the pretest stage are significantly less motivated than those in the high-stakes online stage.

In this chapter, two methods for the evaluation of parameter drift are proposed. The first method is based on a global item-oriented test for parameter drift using a Lagrange multiplier statistic. The method can be viewed as a generalization to adaptive testing of the modification indices for the 2PL model and the nominal response model introduced by Glas (1998, 1999; also see, Glas & Suarez-Falcon, 2003). The second method is targeted at parameter drift due to item disclosure. It addresses the one-sided hypothesis that the item is becoming easier and is losing its discriminative power. The test for this hypothesis is based on a so-called cumulative sum (CUSUM) statistic. Adoption of this approach in the framework of IRT-based adaptive testing was first suggested by Veerkamp (1996) for use with the Rasch model. The present method is a straightforward generalization of this work.

This chapter is organized as follows. First, the most common method of item calibration, marginal maximum likelihood, will be explained. Then the Lagrange multiplier test and the CUSUM test for parameter drift will be explained. Finally, the power of the two classes of tests will be examined in a number of simulation studies.

14.2 Item Parameter Estimation

14.2.1 MML Estimation

Marginal maximum likelihood (MML) estimation is probably the most used technique for item calibration. For the 1PL, 2PL, and 3PL models, the theory was developed by such authors as Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984, 1986), and computations can be made using the software package Bilog-MG (Zimowski, Muraki, Mislevy & Bock, 1996). MML estimation procedures are also available for IRT models with a multidimensional ability structure (see, for instance, Segall, this volume, chap. 3). Under the label "Full Information Factor Analysis", a multidimensional version of the 2PL and 3PL normal-ogive models was developed by Bock, Gibbons, and Muraki (1988) and implemented in TESTFACT (Wilson, Wood & Gibbons, 1991). A comparable model using a logistic rather than a normal-ogive representation was studied by

Reckase (1985, 1997) and Ackerman (1996a and 1996b). In this section, a general MML framework will be sketched, and then illustrated by its application to the 3PL model.

Let $\mathbf{u}_n$ be the response pattern of respondent n, $n = 1, \ldots, N$, and let $\mathbf{U}$ be the data matrix. In the MML approach, it is assumed that the possibly multidimensional ability parameters $\boldsymbol{\theta}_n$ are independent and identically distributed with density $g(\boldsymbol{\theta}; \boldsymbol{\lambda})$. Usually, it is assumed that ability is normally distributed with population parameters $\boldsymbol{\lambda}$ (which are the mean μ and the variance σ^2 for the unidimensional case, or the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Phi}$ for the multidimensional case). Item parameters $\boldsymbol{\beta}$ consist of discrimination parameters (a_i or $\mathbf{a}_i$ for the unidimensional and the multidimensional cases, respectively), item difficulties b_i, and guessing parameters c_i.

In applications of IRT to CAT, students seldom respond to all available items. In the calibration stage, a calibration design is used where samples of students respond to subsets of items, which are often called *booklets* . In the online stage, every student is administered a virtually unique test by the very nature of the item selection mechanism of CAT. Both of these test administration designs are captured by introducing a test administration vector $\mathbf{d}_n$, which has elements d_{in}, $i = 1, \ldots, I$, where I is the number of items in the item pool. The item administration variable d_{in} is equal to one if student n responded to item i, and zero otherwise. The design for all students is represented by an $m \times I$ design matrix $\mathbf{D}$. The definition of the response variable is extended: the vector $\mathbf{u}_n$ has I elements, which are equal to one if a correct response is observed, equal to zero if an incorrect response is observed, and equal to an arbitrary constant if no response is observed. In this context, it is an interesting question whether estimates can be calculated treating the design as fixed and maximizing the likelihood of the parameters conditional on $\mathbf{D}$. If so, the design is called *ignorable* (Rubin, 1976). Using Rubin's theory on ignorability of designs, this question is extensively studied by Mislevy and Wu (1996). They conclude that for the estimation of θ, in adaptive testing the administration design is ignorable. The consequences for item calibration using MML will be returned to in the next section.

MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to $\boldsymbol{\theta}$, rather than maximizing the joint log-likelihood of all person parameters $\boldsymbol{\theta}$ and item parameters $\boldsymbol{\beta}$. Let $\boldsymbol{\eta}$ be a vector of all item and population parameters. Then the marginal likelihood of $\boldsymbol{\eta}$ is given by

$$\log L(\boldsymbol{\eta}; \mathbf{U}, \mathbf{D}) = \sum_n \log \int \ldots \int p(\mathbf{u}_n \mid \mathbf{d}_n, \boldsymbol{\theta}_n, \boldsymbol{\beta}_i) g(\boldsymbol{\theta}_n; \boldsymbol{\lambda}) d\boldsymbol{\theta}_n. \qquad (14.1)$$

The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of person parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur

in IRT models. Kiefer and Wolfowitz (1956) have shown that marginal maximum likelihood estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

The marginal likelihood equations for $\boldsymbol{\eta}$ can be easily derived using Fisher's identity (Efron, 1977; Louis 1982; also see, Glas, 1992, 1998). The first-order derivatives with respect to $\boldsymbol{\eta}$ can be written as

$$\mathbf{h}(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}; \mathbf{U}, \mathbf{D}) = \sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) \,, \tag{14.2}$$

with

$$\boldsymbol{\omega}_n(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{u}_n, \boldsymbol{\theta}_n \mid \mathbf{d}_n, \boldsymbol{\eta}), \tag{14.3}$$

where the expectation is with respect to the posterior distribution $p(\boldsymbol{\theta}_n \mid \mathbf{u}_n, \mathbf{d}_n; \boldsymbol{\eta})$. The identity in (14.2) is closely related to the EM algorithm (Dempster, Laird & Rubin, 1977), which is an algorithm for finding the maximum of a likelihood marginalized over unobserved data. The present application fits this framework when the response patterns are viewed as observed data and the ability parameters as unobserved data. Together they are referred to as the complete data. The EM algorithm is applicable in situations where direct inference based on the marginal likelihood is complicated, and the complete data likelihood equations, i.e., equations based on $\boldsymbol{\omega}_n(\boldsymbol{\eta})$, are easily solved. Given some estimate of $\boldsymbol{\eta}$, say $\boldsymbol{\eta}^*$, the estimate can be improved by solving $\sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}^*) = 0$ with respect to $\boldsymbol{\eta}$. Then this new estimate becomes $\boldsymbol{\eta}^*$ and the process is iterated until convergence.

Application of this framework to deriving the likelihood equations of the structural parameters of the 3PL model proceeds as follows. The likelihood equations are obtained upon equating (14.2) to zero, so explicit expressions are needed for (14.3). Given the design vector $\mathbf{d}_n$, the ability parameter θ_n, and the item parameters of the 3PL model, the probability of response pattern $\mathbf{u}_n$ is given by

$$p(\mathbf{u}_n \mid \mathbf{d}_n, \theta_n, a_i, b_i, c_i) = \prod_i P_i(\theta_n)^{d_{in}u_{in}} (1 - P_i(\theta_n))^{d_{in}(1-u_{in})} \,,$$

where $P_i(\theta_n)$ is the probability of a correct response to item i, as defined in van der Linden and Pashley (this volume, chap. 2, formula 1.1). Define P_{in} and S_{in} by $P_{in} = c_i + (1 - c_i)S_{in}$, so S_{in} is the logistic part of the probability P_{in}. By taking first-order derivatives of the logarithm of this expression, the expressions for (14.3) are found as

$$\omega_n(a_i) = \frac{(u_{in} - P_{in})(1 - c_i)S_{in}(1 - S_{in})(\theta_n - b_i)}{P_{in}(1 - P_{in})}, \tag{14.4}$$

$$\omega_n(b_i) = \frac{(P_{in} - u_{in})(1 - c_i)S_{in}(1 - S_{in})a_i}{P_{in}(1 - P_{in})}, \tag{14.5}$$

and

$$\omega_n(c_i) = \frac{(u_{in} - P_{in})(1 - S_{in})}{P_{in}(1 - P_{in})}. \tag{14.6}$$

The likelihood equations for the item parameters are found upon inserting these expressions into (14.2) and equating the resulting expressions to zero. To derive the likelihood equations for the population parameters, the first-order derivatives of the log of the density of the ability parameters $g(\theta; \mu, \sigma)$ are needed. In the present case, $g(\theta; \mu, \sigma)$ is the well-known expression for the normal distribution with mean μ and standard deviation σ, so it is easily verified that these derivatives are given by

$$\omega_n(\mu) = \frac{(\theta_n - \mu)}{\sigma^2}$$

and

$$\omega_n(\sigma) = \frac{(\theta_n - \mu)^2 - \sigma^2}{\sigma^3}.$$

The likelihood equations are again found upon inserting these expressions in (14.2) and equating the resulting expressions to zero.

Also, the standard errors are easily derived in this framework: Mislevy (1986) points out that the information matrix can be approximated as

$$\mathbf{H}(\boldsymbol{\eta}, \boldsymbol{\eta}) \approx \sum_n E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) E(\boldsymbol{\omega}_n(\boldsymbol{\eta}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})', \tag{14.7}$$

and the standard errors are the diagonal elements of the inverse of this matrix.

The basic approach presented so far can be generalized in two ways. First, the assumption that all respondents are drawn from one population can be replaced by the assumption that there are multiple populations of respondents. Usually, it is assumed that each population has a normal ability distribution indexed by a unique mean and variance parameter. Bock and Zimowski (1997) point out that this generalization together with the possibility of analyzing incomplete item administration designs provides a unified approach to such problems as differential item functioning, item parameter drift, nonequivalent groups equating, vertical equating, and matrix-sampled educational assessment. Item calibration for CAT also fits within this framework.

A second extension of this basic approach is Bayes modal estimation (the term "modal" refers to the mode of the posterior distribution). This approach is motivated by the fact that item parameter estimates in the 3PL model are sometimes hard to obtain because the parameters are poorly determined by the available data. In these instances, item-characteristic curves can be appropriately described by a large number of different item parameter values over the ability scale region where the respondents are located. As a result, the estimates of the three item parameters in the 3PL model are often highly correlated. To obtain "reasonable" and finite estimates, Mislevy (1986) considers a number of Bayesian approaches. Each of

them entails the introduction of prior distributions on item parameters. Parameter estimates are then computed by maximizing the log-posterior density of $\boldsymbol{\eta}$, which is proportional to $\log L(\boldsymbol{\eta};\mathbf{U}) + \log p(\boldsymbol{\eta} \mid \boldsymbol{\zeta}) + \log p(\boldsymbol{\zeta})$, where $p(\boldsymbol{\eta} \mid \boldsymbol{\zeta})$ is the prior density of the $\boldsymbol{\eta}$, characterized by parameters $\boldsymbol{\zeta}$, which in turn follow a density $p(\boldsymbol{\zeta})$. In one approach, the prior distribution is fixed; in another approach, often labeled empirical Bayes, the parameters of the prior distribution are estimated along with the other parameters. In the first case, the likelihood equations in (14.1) change to $\partial \log L(\boldsymbol{\eta};\mathbf{U})/\partial\boldsymbol{\eta} + \partial \log p(\boldsymbol{\eta} \mid \boldsymbol{\zeta})/\partial\boldsymbol{\eta} = \mathbf{0}$. In the second case, in addition to these modified likelihood equations, the additional equations $\partial \log p(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta} = \mathbf{0}$ must also be solved. For details refer to Mislevy (1986). In the following sections, two methods for parameter drift in the framework of the 3PL model and MML estimation will be presented.

14.2.2 Impact of Violations of Ignorability on Item Parameter Estimation

In applications of IRT to CAT, students seldom respond to all available items. Every student is administered a virtually unique test by the very nature of the item selection mechanism of CAT. In the context of CAT, it is an interesting question whether estimates of item parameters can be calculated treating the design matrix $\mathbf{D}$ as fixed by maximizing the likelihood of the parameters conditional on $\mathbf{D}$. If so, the design is called ignorable (Rubin, 1976). In the present section, we assess a number of situations where ignorability is violated. Therefore, first the ignorability principle will be outlined in some detail. Let the potential responses be partitioned into the actually observed responses $\mathbf{u}_{obs}$ and the unobserved responses $\mathbf{u}_{mis}$. As above, the parameter of interest is denoted by $\boldsymbol{\eta}$, and it is assumed that the probability model for $\mathbf{u}_{mis}$ depends on parameters $\boldsymbol{\phi}$. The key concept in the theory of ignorability is "missing at random" (MAR). MAR holds if

$$p(\mathbf{D}|\mathbf{u}_{obs}, \mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X}) = p(\mathbf{D}|\mathbf{u}_{obs}, \boldsymbol{\phi}, \mathbf{X}),$$

where $\mathbf{X}$ are covariates that might play a role. So MAR holds, if the missing data indicators $\mathbf{D}$ do not depend on the missing data $\mathbf{u}_{mis}$, in fact, they only depend on the observed data $\mathbf{u}_{obs}$, and possibly on covariates $\mathbf{X}$. Then, there is a technical condition. In a frequentist framework, the condition is that $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are distinct; that is, the space of $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ factorizes into a $\boldsymbol{\phi}$-space and a $\boldsymbol{\eta}$-space and the two sets of parameters have no mutual functional restrictions. In a Bayesian framework $\boldsymbol{\phi}$ and θ are distinct if $p(\boldsymbol{\phi}|\boldsymbol{\eta}, \mathbf{X}) = p(\boldsymbol{\phi}|\mathbf{X})$, that is, if they have independent priors. Rubin (1976) proved the following:

Theorem

If $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are distinct, and MAR holds,
then
in a frequentist framework $p(\mathbf{u}_{obs}, \mathbf{D} \mid \boldsymbol{\eta}, \boldsymbol{\phi}, \mathbf{X}) \propto p(\mathbf{u}_{obs}, \mid \boldsymbol{\eta}, \mathbf{X})$,
and in a Bayesian framework $p(\boldsymbol{\eta} \mid \mathbf{u}_{obs}, \mathbf{D}, \mathbf{X}) \propto p(\theta \mid \mathbf{u}_{obs}, \mathbf{X})$.

The frequentist version implies that inferences such as maximum likelihood estimation can be based on the likelihood of the observed data, $p(\mathbf{u}_{obs}, \mid \boldsymbol{\eta}, \mathbf{X})$, and the process causing the missingness does not have to be taken into account. In the same manner, the Bayesian version implies that inferences can be based on a posterior $p(\boldsymbol{\eta}|\mathbf{u}_{obs}, \mathbf{X})$ that ignores the probability model for $\mathbf{D}$. It should be noted that conditioning on $\mathbf{D}$ may produce an overestimate of the sample variability of the data, and consequently an underestimate of the standard error of the estimate of θ. Unbiased inferences on standard errors might be obtained if the data are also "observed at random", that is, if $p(\mathbf{D}|\mathbf{u}_{obs}, \mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X}) = p(\mathbf{D}|\mathbf{u}_{mis}, \boldsymbol{\phi}, \mathbf{X})$, so $\mathbf{u}_{obs}$ does not depend on $\mathbf{D}$.

Ignorability in CAT directly follows from the theorem: In CAT the item selection process completely depends on the observed responses and is completely independent of the unobserved responses. Further, ignorability also holds when CAT data are used to calibrate the item and population parameters using maximum marginal likelihood (MML; see Bock and Aitkin, 1981, the impact of targeted designs on MML estimation was studied by Glas, 1988, and Mislevy and Chang, 2000).

In the present chapter, two cases are investigated where the observed data no longer determine the design $\mathbf{D}$: the case where auxiliary information on the students' proficiency is used to select items and the case of item review where the original responses are no longer available. The impact of these violations on the estimates of the item parameters using CAT data in the calibration phase will be assessed using a simulation study.

Consider the response pattern of one student; the index i is dropped for convenience. In a situation of item review, the contribution to the log-likelihood given the original data $\mathbf{u}_{obs}$ and the reviewed data $\mathbf{u}_{mis}$ can be written as

$$
\begin{aligned}
&\log p(\mathbf{u}_{obs}, \mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\eta}) \\
&\quad = \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\mathbf{u}_{mis}, \mathbf{D}; \theta, \boldsymbol{\beta}) g(\theta; \boldsymbol{\lambda}) d\theta \\
&\quad = \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) d\theta \\
&\quad = \log p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) \\
&\qquad + \log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda}) d\theta.
\end{aligned}
$$

Note that this contribution now consists of a term $\log p(\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$ and a term $\log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$. The former gives rise to a log-likelihood associated with a CAT design and if $\mathbf{u}_{mis}$ were observed, these data could be used to obtain consistent estimates of $\boldsymbol{\eta}$. The latter term is the expectation of the probability of $\mathbf{u}_{obs}$ with respect to the posterior distribution $p(\theta|\mathbf{u}_{mis}, \mathbf{D}; \boldsymbol{\beta}, \boldsymbol{\lambda})$. However, if the missing data process is ignored, the expectation of $p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta})$ is considered with respect to $g(\theta; \boldsymbol{\lambda})$; that is, the log-likelihood then becomes a sum of terms

$$\log \int p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta}) g(\theta; \boldsymbol{\lambda}) d\theta. \tag{14.8}$$

The effect is that $p(\mathbf{u}_{obs} \mid \mathbf{D}, \theta, \boldsymbol{\beta})$ is averaged over the wrong proficiency distribution, that is, a distribution with a wrong location parameter and a wrong scale parameter. To assess the effect, consider two students, one with a high θ-value and one with a low θ-value. The first student is administered difficult items, and the second student is administered easy items. However, in (14.8) both their θ-values are assumed to be drawn from the same distribution, and as a result, the easy items are overestimated and the difficult items are underestimated. The effect is due to ignoring the covariates $\mathbf{u}_{mis}$ and $\mathbf{D}$. When the design is governed by auxiliary information about θ, say θ_0, the situation is essentially the same: when the covariate θ_0 is ignored, the proper posterior $p(\theta|\theta_0; \boldsymbol{\beta}, \boldsymbol{\lambda})$ is replaced with $g(\theta; \boldsymbol{\lambda})$, and the result is bias in the estimates of $\boldsymbol{\eta}$.

14.2.3 Simulated Examples

To assess the magnitude of the bias caused by ignoring covariates, simulation studies were conducted. A number of simulation studies were conducted to elucidate the two cases discussed above. The following eight conditions were introduced.

1. Random item selection. In this condition, for every simulee a new set of item parameters was randomly drawn from the standard normal distribution and responses to this randomly assembled test were generated. So this condition did not entail CAT; it was used as a baseline for reference.
2. Computerized adaptive testing.
3. Computerized adaptive testing with item review. In this condition, new responses were generated for all the selected items. So the condition is far more extreme than what can be expected in real-life testing situations.
4. Computerized adaptive testing with item review only for proficiency levels $\theta > 0.0$. In this first set of simulations, the results will just be a combination of the two previous conditions; the purpose of this condition will become apparent in the simulation studies pertaining to item calibration.
5. Computerized adaptive testing where the first half of the test items were chosen to be optimal at the true proficiency value.
6. Computerized adaptive testing where all items were chosen to be optimal at the true proficiency value.
7. Computerized adaptive testing where the first half of the test items were chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and a standard deviation equal to 1.0.
8. Computerized adaptive testing where the first half of the test items were chosen to be optimal at θ_0, where θ_0 was drawn from a normal distribution with a mean equal to the true proficiency parameter, and a standard deviation equal to 2.0.

Adaptive test data were generated for 1,000 simulees with parameters drawn from the standard normal distribution. The item bank consisted of 200 items equally spaced between −2.0 and 2.0, and the test length was 20 items. The one-parameter logistic model (1PLM) was used to avoid contamination of the results by the possibly poor identification of the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM). Unless indicated otherwise, the starting value of the proficiency estimate was equal to zero. The proficiency parameter was estimated by maximum likelihood and maximum information was used as a selection criterium. Using these adaptive test data, MML estimates of the item parameters were computed under the assumption that θ had a standard normal distribution.

In every condition reported below, 100 replications were made. In the condition of random item selection, the test of 20 items was resampled from the item bank for every simulee.

The results are shown in Table 14.1. For five items from the item bank, the last three columns give the bias, standard error, and mean of the estimates over the replications, respectively. The following conclusions can be drawn.

1. Comparing random item selection and CAT, it can be seen that the latter greatly reduced the standard error. In both cases, the bias was relatively small.
2. In all other conditions, the bias was substantial.
3. In CAT with item review, there is inward bias; that is, easy items are overestimated and difficult items are underestimated.
4. If only simulees with $\theta > 0$ review the items, the bias in the easy items vanishes.
5. Choosing the complete test to be optimal at the true θ completely contaminates the calibration in the sense that all item parameters shrink to zero.

14.3 Item Fit Analysis

14.3.1 Lagrange Multiplier Tests

The idea behind the Lagrange multiplier (LM) test (Aitchison & Silvey, (1958), and the equivalent efficient score test (Rao, 1947), can be summarized as follows. Consider some general parameterized model and a special case of the general model, the so-called restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by setting one or more parameters of the general model to constants. The LM test is based on evaluating a quadratic function of the partial derivatives of the log-likelihood function of the general model evaluated at the ML estimates of the restricted model. The LM test is evaluated using the ML estimates of the parameters of the restricted model. The unrestricted elements of the vector of the first-order derivatives are equal to zero because their values originate from solving

Table 14.1 Squared bias and standard errors for calibration of β

Item Selection Mode	β	Bias	S.E.	Mean
Random selection	−2.0	0.01	0.32	−1.96
	−1.0	0.05	0.20	−0.94
	0.0	0.01	0.23	−0.01
	1.0	0.01	0.29	1.01
	2.0	0.08	0.29	2.08
CAT	−2.0	0.02	0.19	−2.00
	−1.0	0.03	0.26	−1.03
	0.0	0.01	0.08	−0.01
	1.0	0.00	0.21	0.99
	2.0	0.00	0.19	2.00
CAT with item review	−2.0	0.64	0.22	−1.33
	−1.0	0.34	0.29	−0.65
	0.0	0.01	0.07	−0.01
	1.0	0.28	0.22	0.71
	2.0	0.60	0.18	1.39
CAT with item review	−2.0	0.07	0.21	−1.90
if $\theta > 0.0$	−1.0	0.15	0.29	−0.84
	0.0	0.01	0.07	0.01
	1.0	0.20	0.19	0.79
	2.0	0.52	0.22	1.47
50% optimal	−2.0	0.43	0.23	−1.54
at true θ	−1.0	0.38	0.25	−0.61
	0.0	0.00	0.17	−0.00
	1.0	0.33	0.22	0.66
	2.0	0.40	0.22	1.59
100% optimal	−2.0	1.92	0.39	−0.05
at true θ	−1.0	0.92	0.21	−0.07
	0.0	0.04	0.18	0.04
	1.0	0.93	0.22	0.06
	2.0	1.84	0.38	0.15
50% initial responses at	−2.0	0.21	0.20	−1.76
$\widehat{\theta}$ with s.d.$(\widehat{\theta}) = 1.0$	−1.0	0.17	0.19	−0.82
	0.0	0.00	0.17	0.00
	1.0	0.08	0.21	0.91
	2.0	0.24	0.20	1.75

the likelihood equations. The magnitude of the elements of the vector of first-order derivatives corresponding with restricted parameters determines the value of the statistic: the closer they are to zero, the better the model fits.

More formally, the principle can be described as follows. Consider a null hypothesis about a model with parameters $\boldsymbol{\eta}_0$. This model is a special case of a general model with parameters $\boldsymbol{\eta}$. In the case discussed here, the special model is derived from the general model by setting one or more parameters to zero. So if the parameter vector $\boldsymbol{\eta}_0$ is partitioned as $\boldsymbol{\eta}_0 = (\boldsymbol{\eta}_{01}, \boldsymbol{\eta}_{02})$, the null hypothesis entails $\boldsymbol{\eta}_{02} = 0$. Let $\mathbf{h}(\boldsymbol{\eta})$ be the partial derivatives of the log-likelihood of the general model, so

$\mathbf{h}(\boldsymbol{\eta}) = \partial \log L(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in $\boldsymbol{\eta}$. The test will be based on the statistic

$$LM = \mathbf{h}(\boldsymbol{\eta}_{\mathbf{02}})^t \boldsymbol{\Sigma}^{-1} \mathbf{h}(\boldsymbol{\eta}_{\mathbf{02}}), \tag{14.9}$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{10} \boldsymbol{\Sigma}_{00}^{-1} \boldsymbol{\Sigma}_{01}$$

and

$$\boldsymbol{\Sigma}_{pq} = \sum_n \mathbf{h}_n(\boldsymbol{\eta}_{0p}) \mathbf{h}_n(\boldsymbol{\eta}_{0q})^t.$$

The statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_{02}$ (Aitchison & Silvey, 1958; Rao, 1947).

Recently, the LM principle has been applied in the framework of IRT for evaluating differential item functioning (Glas, 1998) and the axioms of unidimensionality and local stochastic independence (Glas, 1999). Though originally presented in the framework of a fixed item administration design, these tests can also be applied in the framework of the stochastic design characteristics for CAT. However, the result with respect to the asymptotic distribution of the statistics does not automatically apply to the case of a stochastic design. The ignorability principle ensures consistency of estimators in a CAT design, but it does not apply to sample inferences, such as confidence intervals and the distributions of statistics for evaluation of model fit (Mislevy & Chang, 1998). Therefore, for the applications presented here, a power study under the null model will be part of the example to be presented. The results will show that the asymptotic distribution of the LM statistics is hardly affected by CAT.

14.3.2 An LM Test for the Fit of Item-Characteristic Curves

The idea of the LM test and modification index presented here will be to partition the latent ability continuum into a number of segments, and to evaluate whether an item's ICC conforms to the form predicted by the null model in each of these segments. However, the actual partitioning will take place on the observed number-correct scale rather than on the θ scale. Usually, the unweighted sum score and the associated estimate of θ will highly correlate. Let the item of interest be labeled k, and let the other items be labeled $i = 1, 2, \ldots, k-1, k+1, \ldots, K$. Let $r_n^{(k)}$ be the unweighted sum score on the response pattern of student n without item k. The possible scores $r_n^{(k)}$ will be partitioned into S disjoint subsets using boundary scores $r_0 < r_1 < r_2 \ldots < r_s < \ldots < r_S$, with $r_0 = 0$ and $r_S = K - 1$. Further, define

$$w_s\left(r_n^{(k)}\right) = \begin{cases} 1 & \text{if } r_{s-1} < r_n^{(k)} < r_s, \\ 0 & \text{otherwise,} \end{cases} \tag{14.10}$$

so $w_s\left(r_n^{(k)}\right)$ is an indicator function that assumes a value equal to one if the number-correct score without item k is in score range s. As an alternative model to the 2PLM and 3PLM, consider a model where the item discrimination and difficulty parameters are redefined as $a_n + \sum_s w_s\left(r_n^{(k)}\right)\delta_{1s}$ and $b_n + \sum_s w_s\left(r_n^{(k)}\right)\delta_{2s}$. The simultaneous hypothesis $\delta_{1s} = 0$ and $\delta_{2s} = 0$ ($s = 2, \ldots, S$; that is, $s = 1$ is used as a baseline) can be evaluated using an LM test. For respondents with a number-correct score in category s, it holds that

$$\omega_n(\delta_{1s}) = \theta_n(u_{in} - P_i(\theta_n)) \tag{14.11}$$

and

$$\omega_n(\delta_{2s}) = P_i(\theta_n) - u_{in}, \tag{14.12}$$

where $\omega_n(\delta_{1s})$ and $\omega_n(\delta_{2s})$ are defined as in (14.3). Using (14.2) it can be inferred that the elements of the vectors of first-order derivatives $h(\delta_1)$ and $h(\delta_1)$ are given by

$$\begin{aligned}&\sum_n w_s\left(r_n^{(k)}\right) E(\omega_n(\delta_{1s}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\\ &\quad = \sum_n w_s\left(r_n^{(k)}\right) E(\theta_i(u_{in} - P_i(\theta_n)) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\end{aligned} \tag{14.13}$$

and

$$\begin{aligned}&\sum_n w_s\left(r_n^{(k)}\right) E(\omega_n(\delta_{2s}) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta})\\ &\quad = \sum_n w_s\left(r_n^{(k)}\right) E(P_i(\theta_n) \mid \mathbf{u}_n, \mathbf{d}_n, \boldsymbol{\eta}) - \sum_n w_s\left(r_n^{(k)}\right) u_{in}.\end{aligned} \tag{14.14}$$

Notice that (14.14) is the difference between the observed number of persons of sub-sample s with a correct score on item i, and its expected value. So (14.14) can be seen as a residual. A test for the simultaneous hypothesis $\delta_{1s} = 0$ and $\delta_{2s} = 0$, for $s = 1, \ldots, S-1$, can be based on a statistic with an asymptotic χ^2 distribution with $2(S-1)$ degrees of freedom, where the statistic defined by (14.9) is evaluated using MML estimates of the null model, that is, the 2PL model or 3PL model. It is also possible to define separate tests for $\delta_{1s} = 0$ or $\delta_{2s} = 0$ ($s = 1, \ldots, S-1$). These tests are based on LM statistics with $S-1$ degrees of freedom.

14.3.3 An LM Test for Parameter Drift

We noted earlier that parameter drift can be evaluated by checking whether pretest and online data can be properly described by the same IRT model. Consider G

groups labeled $g = 1, \ldots, G$. It is assumed that the first group partakes in the pretesting stage, and the following groups partake in the online stage. The application of the LM tests to monitoring parameter drift is derived from the LM test for differential item functioning proposed by Glas (1998) for the 2PL model. This is a test of the hypothesis that the item parameters are constant over groups, that is, the hypothesis $a_{ig} = a_i$ and $b_{ig} = b_i$, for all g. To see the relation with the LM framework, consider two groups, and define a variable y_n that is equal to one if n belongs to the first group and zero if n belongs to the second group. Defining $a_{iy} = a_i + y_n\delta_1$ and $b_{iy} = b_i + y_n\delta_2$, the hypothesis given by $\delta_1 = 0$ and $\delta_2 = 0$ can be evaluated using the LM test. For more than two groups, more dummy variables y_n are needed to code group membership. This approach can of course also be used to monitor parameter drift in CAT. Further, generalization to the 3PL model entails adding $\delta_3 = 0$, with $c_{iy} = c_i + y_n\delta_3$, to the null hypothesis.

For actual implementation of this approach using adaptive testing data, the high correlation of estimates of the three item parameters discussed in the previous section must be taken into account. Another parameter estimation problem arises specifically in the context of adaptive testing. Guessing (which may be prominent in the calibration stage) may rarely occur in the online stage because items are tailored to the ability level of the respondents. Therefore, a test focused on all three parameters simultaneously often proves computationally unstable. In the present chapter, three approaches will be studied. In the first, the LM test will be focused on simultaneous parameter drift in a_i and b_i; in the second approach, the LM test will be focused on parameter drift in c_i. These two tests will be labeled $LM(a_i, b_i)$ and $LM(c_i)$, respectively. In the third approach, the guessing parameter will be fixed at some plausible constant, say, the reciprocal of the number of response alternatives of the items, and the LM statistic will be used to test whether this fixed guessing parameter is appropriate in the initial stage and remains so when the adaptive testing data are introduced. So the hypothesis considered is that $c_{ig} = c_i$ for all g. Using simulation studies, it will be shown that the outcomes of these three approaches are quite comparable.

14.3.4 A CUSUM Test for Parameter Drift

The CUSUM chart is an instrument of statistical quality control used for detecting small changes in product features during the production process (see, for instance, Wetherill, 1977). The CUSUM chart is used in a sequential statistical test, where the null hypothesis of no change is never accepted. In the present application, loss of production quality means that the item is becoming easier and less discriminating.

Contrary to the case of the LM test, the CUSUM test needs estimation of the item parameters for every group of students $g = 1, \ldots, G$. As above, the first group partakes in the pretesting stage, and the following groups take an adaptive test. However, estimation of the guessing parameter is problematic in a CAT situation because, as already mentioned, guessing may be prominent in the calibration stage,

while it may rarely occur in the online stage, where the items are tailored to the ability level of the respondents. Two possible solutions include fixing the guessing parameter to some plausible constant such as the reciprocal of the number of response options, or concurrent estimation of the item guessing parameter using all available data. In either approach, the null hypothesis is $a_{ig} - a_{i1} \geq 0$ and $b_{ig} - b_{i1} \geq 0$, for the respondent groups $g = 1, \ldots, G$. Therefore, a one-sided CUSUM chart will be based on the quantity

$$S_i(g) = \max \left\{ S_i(g-1) + \frac{a_{i1} - a_{ig}}{Se(a_{ig} - a_{i1})} + \frac{b_{i1} - b_{ig}}{Se(b_{i1} - b_{ig} \mid a_{i1} - a_{ig})} - k, 0 \right\}, \tag{14.15}$$

where $Se(a_{ig} - a_{i1}) = \sigma_a$ and $Se(b_{i1} - b_{ig} \mid a_{i1} - a_{ig}) = \sqrt{\sigma_b^2 - \sigma_{ab}^2/\sigma_a^2}$, with σ_a^2, σ_b^2, and σ_{ab} the appropriate elements of the covariance matrix of the parameter estimates given by (14.7). Further, k is a reference value determining the size of the effects one aims to detect. The CUSUM chart starts with $S_i(1) = 0$ and the null hypothesis is rejected as soon as $S_i(g) > h$, where h is some constant threshold value. The choice of the constants k and h determines the power of the procedure. In the case of the Rasch model, where the null hypothesis is $b_{ig} - b_{i1} \geq 0$, and the term involving the discrimination indices is lacking from (14.15), Veerkamp (1996) successfully uses $k = 1/2$ and $h = 5$. This choice was motivated by the consideration that the resulting test has good power against the alternative hypothesis of a normalized shift in item difficulty of approximately half a standard deviation. In the present case, one extra normalized decision variable is employed, namely, the variable involving the discrimination indices. So, for instance, a value $k = 1$ can be used to have power against a shift of one standard deviation of both normalized decision variables in the direction of the alternative hypothesis. However, there are no compelling reasons for this choice; the attractive feature of the CUSUM procedure is that the practitioner can choose the effect size k to meet the specific characteristics of the problem. Also, the choice of a value for h is determined by the targeted detection rate, especially by the trade-off between Type I and II errors. In practice, the values of h and k can be set using simulation studies. Examples will be given below.

14.4 Examples

In this section, the power of the procedures suggested above will be investigated using a number of simulation studies. Since all statistics involve an approximation of the standard error of the parameter estimates using (14.7), first the precision of the approximation will be studied by assessing the power of the statistics under the null model, that is, by studying the Type I error rate. Then the power of the tests will be studied under various model violations. These two topics will first be studied for the LM tests, then for the CUSUM test.

Table 14.2 Type I error rate of LM test

K	L	N_g	Percentage at 10% $LM(c_i)$	Significant $LM(a_i, b_i)$
50	20	500	8	9
		1000	10	10
	40	500	9	10
		1000	11	8
100	20	500	12	10
		1000	8	9
	40	500	10	12
		1000	10	10

In all simulations, the ability parameters θ were drawn from a standard normal distribution. The item difficulties b_i were uniformly distributed on $[-1.5, 1.5]$, the discrimination indices a_i were drawn from a log-normal distribution with a zero mean and a standard deviation equal to 0.25, and the guessing parameters were fixed at 0.20, unless indicated otherwise. In the online stage, item selection was done using the maximum information principle. The ability parameter was estimated by its expected a posteriori value (EAP); the initial prior was standard normal.

The results of eight simulation studies with respect to the Type I error rate of the LM test are shown in Table 14.2. The design of the study can be inferred from the first three columns of the table. It can be seen that the number of items K in the item bank was fixed at 50 for the first four studies and at 100 for the next four studies. In both the pretest stage and the online stages, test lengths L of 20 and 40 were chosen. Finally, as can be seen in the third column, the number of respondents per stage, N_g, was fixed at 500 and 1000 respondents. So summed over the pretest and online stage, the sample sizes were 1000 and 2000 respondents, respectively. For the pretest stage, a spiraled test administration design was used. For instance, for the $K = 50$ studies, for the pretest stage, five subgroups were used; the first subgroup was administered items 1 – 20, the second items 11 – 30, the third items 21 – 40 the fourth items 31 – 50, and the last group received the items 1 – 10 and 41 – 50. In this manner, all items drew the same number of responses in the pretest stage. For the $K = 100$ studies, for the pretest stage four subgroups administered 50 items were formed, so here the design was 1 – 50, 26 – 75, 51 – 100 and 1 – 25 and 76 – 100. For each study, 100 replications were run. The results of the study are shown in the last two columns of Table 14.2. These columns contain the percentages of $LM(c_i)$ and $LM(a_i, b_i)$ tests that were significant at the 10% level. It can be seen that the Type I error rates of the tests conform to the nominal value of 10%. These results support the adequacy of the standard error approximations for providing accurate Type I error rates.

The second series of simulations pertained to the power of the LM statistics under various model violations. The setup was the same as in the above study with $K = 100$ items in the item bank, a test length $L = 50$, $N_1 = 1000$ simulees in the pretest stage and $N_2 = 1000$ simulees in the online stages. Two model violations were simulated. In the first, the guessing parameter c_i went up in the online stage; in

Table 14.3 Power of LM test

Model Violation		Percentage at 10% $LM(a_i, b_i)$	Significant $LM(c_i)$
$c_i = 0.25$	Hits	25	15
	False alarm	08	10
$c_i = 0.30$	Hits	45	35
	False alarm	13	11
$c_i = 0.40$	Hits	95	85
	False alarm	17	20
$b_i = -0.20$	Hits	25	30
	False alarm	13	12
$b_i = -0.40$	Hits	55	70
	False alarm	15	20
$b_i = -0.60$	Hits	80	95
	False alarm	13	27

the second, the item difficulty b_i went down in the online stage. Six conditions were investigated: c_i rose from 0.20 to 0.25, 0.30, and 0.40, respectively, and b_i changed from the initial value by -0.20, -0.40 and -0.60, respectively. These model violations were imposed on the items 5, 10, 15, etc. So 20 out of the 100 items were affected by this form of parameter drift. 100 replications were made for each condition. Both the $LM(c_i)$ and $LM(a_i, b_i)$ tests were used. The results are shown in Table 14.3. This table displays both the percentage of "hits" (correctly identified items with parameter drift) and "false alarms" (items without parameter drift erroneously identified as drifting). Three conclusions can be drawn. Firstly, it can be seen that the power of the tests increases as the magnitude of the model violation grows. Secondly, the power of the test specifically aimed at a model violation is always a little larger than the power of the other test, but the differences are quite small. For instance, in the case $b_i = -0.60$, the power of $LM(a_i, b_i)$ is 0.95, while the power of $LM(c_i)$ is 0.85. The third conclusion that can be drawn from the table is that the percentage of "false alarms" is clearly higher than the nominal 10% error rate. A plausible explanation might be that the improper parameter estimates of the 20% items with parameter drift influence the estimates of the 80% non-affected items. Finally, it can be noted that the agreement between the two tests with respect to the flagged items was high; agreement between the two tests was always higher than 0.84.

As mentioned above, the power of the CUSUM procedure is governed by choosing an effect size k and a critical value h. A good way to proceed in a practical situation is to calibrate the procedure when the pretest data have become available. First, the practitioner must set an effect size k of interest. Then, assuming no parameter drift, online data can be simulated using the parameter estimates of the pretest stage. Finally, CUSUM statistics can be computed to find a value for h such that an acceptable Type I error rate is obtained. An example will be given using the same set-up as above: there were $K = 100$ items in the item bank, test length was $L = 50$, and the pretest data consisted of the responses of $N_1 = 1000$ simulees.

Table 14.4 Type I error rate of CUSUM test

Effect Size	$h = 2.5$	$h = 5.0$	$h = 7.5$	$h = 10.0$
$k = 0.50$	17	04	01	00
$k = 1.00$	09	06	01	00
$k = 2.00$	01	00	00	00

Then, four batches of responses of $N_g = 1000$ ($g = 2, \ldots, 5$) simulees were generated as online data, and CUSUM statistics $S_i(g)$ were computed for the iterations $g = 2, \ldots, 5$. This procedure was carried out for three effect sizes k and four thresholds h; the values are shown in Table 14.4.

In the table, the percentages items flagged in the fifth iteration ($g = 5$) of the procedure are shown for the various combinations of k and h. Since no parameter drift was induced, the percentages shown can be interpreted as Type I error rates. For an effect size $k = 0.50$, it can be seen that a value $h = 2.5$ results in 17% flagged items, which is too high. A value $h = 5.0$ results in 4% flagged items, which might be considered an acceptable Type I error rate. Also, for an effect size $k = 1.00$ a critical value $h = 5.0$ seems a good candidate. Finally, for $k = 2.00$, all four values of h produce low Type I error rates. So it must be concluded that, given the design and the sample size, detection of parameter drift with an effect size of two standard deviations may be quite difficult.

This result was further studied in a set of simulations where model violations were introduced. These studies used the setup $K = 100$, $L = 50$, and $N_g = 1000$, for $g = 1, \ldots, 5$. The model violations were similar to the ones imposed above. So in six conditions, the guessing parameter c_i rose from 0.20 to 0.25, 0.30, and 0.40, respectively, and b_i changed from the initial value by -0.20, -0.40, and -0.60, respectively. Again, for each condition, 20 of the 100 items were affected by the model violation. The results are shown in Table 14.5. For the simulation studies with effect sizes $k = 0.50$ and $k = 1.00$, a critical value $h = 5.0$ was chosen; for the studies with effect size $k = 2.00$, the critical value was $h = 2.5$.

For every combination of effect size and model violation, 20 replications were made. The last four columns of Table 14.5 give the percentages of "hits" (flagged items with parameter drift) and "false alarms" (erroneously flagged items per condition) for the iterations $g = 2, \ldots, 5$. The percentages are aggregated over the 20 replications per condition. As expected, the highest percentages of "hits" were obtained for the smaller effect sizes $k = 0.50$ and $k = 1.00$, and the larger model violations. The top is the combination $k = 1.00$ and $b_i = -0.60$, which, for $g = 5$, has an almost perfect record of 99% "hits". In this condition, the percentage of "false alarms" remained at a 10% level. The worst performances were obtained for combinations of $k = 0.50$ and $k = 2.00$ with small violations as $c_i = 0.25$, $c_i = 0.30$, and $b_i = -0.20$. These conditions both show a low "hit" rate and a "false alarm" rate of approximately the same magnitude, which is relatively high for a "false alarm" rate.

Table 14.5 Power of CUSUM test

Effect Size	Model Violation		Iteration $g = 2$	$g = 3$	$g = 4$	$g = 5$
$k = 0.50$	$c_i = 0.25$	Hits	00	00	05	15
		False alarm	00	04	05	13
	$c_i = 0.30$	Hits	00	05	10	20
		False alarm	00	03	05	06
	$c_i = 0.40$	Hits	00	30	75	85
		False alarm	00	00	01	03
$k = 1.00$	$c_i = 0.25$	Hits	15	25	30	45
		False alarm	05	13	17	21
	$c_i = 0.30$	Hits	15	35	55	50
		False alarm	03	03	03	06
	$c_i = 0.40$	Hits	30	75	90	85
		False alarm	03	04	06	09
$k = 2.00$	$c_i = 0.25$	Hits	00	05	15	15
		False alarm	00	00	03	00
	$c_i = 0.30$	Hits	05	15	15	20
		False alarm	03	01	04	04
	$c_i = 0.40$	Hits	15	30	55	60
		False alarm	00	01	01	01
$k = 0.50$	$b_i = -0.20$	Hits	00	00	10	15
		False alarm	00	00	06	05
	$b_i = -0.40$	Hits	00	15	45	60
		False alarm	01	06	09	15
	$b_i = -0.60$	Hits	05	35	65	80
		False alarm	00	00	04	04
$k = 1.00$	$b_i = -0.20$	Hits	00	20	40	35
		False alarm	00	01	03	05
	$b_i = -0.40$	Hits	25	50	55	65
		False alarm	01	04	06	09
	$b_i = -0.60$	Hits	20	75	95	99
		False alarm	03	06	10	10
$k = 2.00$	$b_i = -0.20$	Hits	00	00	05	05
		False alarm	00	01	03	03
	$b_i = -0.40$	Hits	05	10	30	35
		False alarm	00	00	03	01
	$b_i = -0.60$	Hits	00	25	75	75
		False alarm	01	03	04	03

14.5 Discussion

This chapter showed how to evaluate whether the IRT model of the pretest stage also fits the online stage. Two approaches were presented. The first was based on LM statistics. It was shown that the approach supports the detection of specific model violations and has the advantage of known asymptotic distributions of the

statistics on which it is based. Two specific model violations were considered here, but the approach also applies to other model violations, such as violation of local independence and multidimensionality (see Glas, 1999). The second approach is based on CUSUM statistics. The distribution of these statistics is not known, but an appropriate critical value h can be found via simulation. An advantage, however, is that the practitioner can tune the procedure to the needs of the specific situation. When choosing h, the subjective importance of making "hits" and avoiding "false alarms" can be taken into account, and the effect size k can be chosen to reflect the magnitude of parameter drift judged relevant in a particular situation. Summing up, both approaches provide practical tools for monitoring parameter drift.

References

Ackerman, T. A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20,* 309–310.

Ackerman, T. A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement, 20,* 311–329.

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29,* 813–828.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., Gibbons, R. D. & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement, 12,* 261–280.

Bock, R. D. & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York: Springer-Verlag.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B. 39,* 1–38.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Laird, and D. Rubin). *J. R. Statist. Soc. B., 39,* 1–38.

Fischer, G. H. & Scheiblechner, H. H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12,* 23–51.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1) (pp. 236–258). Norwood, NJ: Ablex Publishing Corporation.

Glas, C. A. W. (1988). The Rasch Model and multi-stage testing. *Journal of Educational Statistics, 13,* 45–52.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.

Glas, C. A. W. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64,* 273–294.

Glas, C. A. W. & Suarez-Falcon, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27,* 87–106.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–903.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359–375.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226–233.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.
Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.
Mislevy, R. J. & Chang, H.-H. (2000). Does adaptive testing violate local independence? *Psychometrika, 65,* 149–156.
Mislevy, R. J. & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.
Neyman, J. & Scott, E. L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16*, 1–32.
Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.
Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9,* 401–412.
Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden and R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.
Rigdon S. E. & Tsutakawa, R. K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567–574.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.
Stocking, M. L. (1993). *Controlling exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton, NJ: Educational Testing Service.
Stocking, M. L. & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.
Sympson, J. B. & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual meeting of the Military Testing Association* (pp. 973–977). San Diego: Navy Personnel Research and Development Center.
Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.
Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing.* Unpublished doctoral thesis, Twente University, the Netherlands.
Wetherill, G. B. (1977). *Sampling inspection and statistical quality control* (2nd ed.). London: Chapman and Hall.
Wilson, D. T., Wood, R. & Gibbons, R. D. (1991) *TESTFACT: Test scoring, item statistics, and item factor analysis* (computer software). Chicago: Scientific Software International, Inc.
Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.
Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International, Inc.