DOCUMENT RESUME

ED 480 898 TM 035 251

AUTHOR	Glas, C.A.W.
TITLE	Quality Control of Online Calibration in Computerized Assessment. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-CTR-97-15
PUB DATE	2003-09-00
NOTE	15p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Item Response Theory; *Quality Control; Simulation
IDENTIFIERS	*Calibration; *Item Parameters

ABSTRACT

In computerized adaptive testing, updating item parameter estimates using adaptive testing data is often called online calibration. This study investigated how to evaluate whether the adaptive testing data used for online calibration sufficiently fit the item response model used. Three approaches were investigated, based on a Lagrange multiplier (LM) statistic, a Wald statistic, and a cumulative sum (CUSUM) statistic. The power of the tests was evaluated with a number of simulation studies. It was found that the tests had moderate to good power to detect shifts in the values of the guessing and difficulty parameters, and all tests were equally sensitive to all shifts in the values of all parameters. The practical conclusion is that all of these statistics can be used very well to detect if something has happened to the item parameters but that it may be difficult to attribute the problems to specific parameters. (Contains 3 tables and 17 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

Quality Control of Online Calibration in Computerized Assessment

C. A. W. Glas

Department of Educational Measurement and Data Analysis
Faculty of Educational Science and Technology
University of Twente, Enschede, the Netherlands

Law School Admission Council
Computerized Testing Report 97-15
September 2003

A Publication of the Law School Admission Council

LSAC RESEARCH REPORT SERIES

■ Quality Control of Online Calibration in Computerized Assessment

C. A. W. Glas

Department of Educational Measurement and Data Analysis
Faculty of Educational Science and Technology
University of Twente, Enschede, the Netherlands

■ Law School Admission Council
Computerized Testing Report 97-15
September 2003

A Publication of the Law School Admission Council

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 201 law schools in the United States and Canada.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The introduction of computerized adaptive testing (CAT) has made it necessary to build large pools of test items with the item statistics (commonly called parameters) needed to describe the characteristics of the items. The process of obtaining item parameters usually consists of the following two stages:

1. *Pretest stage.* In a series of sessions, sets of items are administered to groups of test takers, and a mathematical model called item response theory (IRT) is used to obtain estimates of item parameters representing such features as item difficulty, discriminating power (the ability of the item to distinguish between more and less able test takers), or susceptibility to guessing.

2. *Online stage.* The test is operational and administered online but the responses are also used for parameter estimation, for example, to keep improving the precision of previous estimates or to obtain estimates for new items added to the pool.

In this paper it is proposed that methods of quality control be used in the calibration process, for example, to check if the values of the item parameters have not drifted between the pretest and the online stage. If parameter drift is found, the response data cannot be pooled to increase the precision of the parameter estimates. Methods of quality control can also be used to detect security breaches in an online stage. Three different statistics for quality control are proposed: (1) a Lagrange multiplier (LM) statistic; (2) a Wald statistic; and (3) a cumulative sum (CUSUM) statistic. The power of the tests based on these statistics, that is, their ability to detect shifts in the parameter values, was evaluated.

It was found that the tests had moderate to good power to detect shifts in the values of the guessing and difficulty parameters. In addition, all tests were equally sensitive to shifts in the values of all parameters, even if the null hypothesis of no shift was formulated for only one of them. This result is not surprising because estimates of the parameters in the model evaluated are usually highly correlated. The practical conclusion from the study is that all of these statistics can be used very well to detect if something has happened to the item parameters but that it may be difficult to attribute the problems to specific parameters.

Abstract

In computerized adaptive testing, updating item parameter estimates using adaptive testing data is often called online calibration. This paper investigates how to evaluate whether the adaptive testing data used for online calibration sufficiently fit the item response model used. Three approaches are investigated, based on a Lagrange multiplier (LM) statistic, a Wald statistic, and a cumulative sum (CUSUM) statistic. The power of the tests is evaluated with a number of simulation studies.

Introduction

Computerized assessment, such as CBT (computer based testing) and CAT (computer adaptive testing), is based on the availability of a large pool of calibrated test items. Usually, the calibration process consists of two stages.

1. *The pretesting stage.* In this stage, subsets of items are administered to subsets of respondents in a series of pretest sessions, and an item response (IRT) model is fitted to the data to obtain item parameter estimates to support computerized test administration.

2. *The online stage.* In this stage, data are gathered in a computerized assessment environment. There may be several motives for using these data for further parameter estimation. The interest may be to continuously update estimates to attain the greatest possible precision. Or new, previously uncalibrated items may be entered into the bank and can only be calibrated using incoming responses.

Closely related to the motives for online calibration, but also an aim in itself, is quality control, that is, checking whether pretest and online results comply with the same IRT model. In the present paper, three methods of quality control are proposed. The first method is based on the Lagrange multiplier statistic (LM). The method can be viewed as a generalization to adaptive testing of the modification indices for the 2-PL model and the nominal response model introduced by Glas (1997, 1998). The second method is based on a Wald statistic. The third method is based on a so-called cumulative sum (CUSUM) statistic. This last approach stems from the field of statistical quality control (see, for instance, Wetherill, 1977). Using this method in the framework of IRT-based adaptive testing was first suggested by Veerkamp (1996) in the framework of the Rasch model. In this paper, the procedure will be generalized to the 3-PL model.

This paper is organized as follows. The section that follows outlines a framework for estimation of the 2-PL model. This model will subsequently be used for a general introduction of the LM statistic in the next section. The LM and the Wald and CUSUM statistics will then be applied to quality control in adaptive testing. Section 6 evaluates the performance of the proposed methods with a number of simulation studies. Finally, in Section 7 some conclusions and suggestions for further research will be formulated.

Before proceeding, a remark, with respect to the scope of this paper. Strictly speaking, the methods proposed here also apply to a situation where there is no pretest stage and the item bank is bootstrapped during the online stage. However, without a pretest stage, in the initial stages of online calibration, the data on some of the items may be prohibitively scarce or even ill-conditioned, in the sense that there is too little information in the data to estimate all relevant parameters. Below, it will be assumed that the data are such that parameter estimates can be obtained. Generalization of the methods to be proposed to ill-conditioned data, probably by introducing prior distributions on the item parameters, is beyond the scope of the present paper and will be treated later. Further, it will be assumed that the number of items in the bank is such that standard errors of estimates can be computed using the complete information matrix. Also application of the procedures to very large item banks, where other approximations to the standard errors have to be made, are points of future research.

Preliminaries

Consider dichotomous items where responses of persons labeled n to items labeled i are coded $x_{ni} = 0$, and $x_{ni} = 1$. The, probability of a correct response is given by

$$\begin{aligned}\phi_i(\theta_n) &= Pr(X_{ni} = 1|\theta_n, \alpha_i, \beta_i, \gamma_i)\\ &= \gamma_i + (1-\gamma_i)\psi_i(\theta_n)\\ &= \gamma_i\,(1-\gamma_i)\frac{\exp(\alpha_i\theta_n - \beta_i)}{1+\exp(\alpha_i\theta_n - \beta_i)},\end{aligned} \tag{1}$$

where θ_n is the ability parameter of person n and α_i, β_i, and γ_i are the discrimination, difficulty, and guessing parameter of item i, respectively. Since simultaneous ML estimates of all item parameters are hard to obtain (see, for instance, Swaminathan & Gifford, 1986), in the present paper it will be assumed that γ_i is fixed to some plausible constant, say, to the guessing probability. Using priors on γ_i to facilitate its estimation is a topic for future study. Below, the well-known theory of MML estimation for IRT models will be reiterated. In this presentation the formalism of Glas (1992, 1997, 1998) will be used, which, as will become apparent in the sequel, is especially suited for the introduction of the procedures below. The choice of a distribution of ability is not essential to the theory presented here; it can be the parametric MML framework (see Bock & Aitkin, 1981) or the nonparametric MML framework (see DeLeeuw & Verhelst, 1986; Follmann, 1988). However, to make the presentation explicit, it is assumed that the ability distribution is normal with parameters μ and σ. Further, for reasons of simplicity, it is assumed that all respondents belong to the same population. Modern software for the 2- and 3-PL model, such as Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), does not have this restriction, but this generalization is straightforward. So, let $g(\theta n; \mu, \sigma)$ be the density of θ. Further, let the item administration variable d_{ni} take the value one if the item was administered to n, and zero if this was not the case. If $d_{ni} = 0$ it will be assumed that $x_{ni} = c$, where c is some arbitrary constant.

Let $\boldsymbol{x}_n$ and $\boldsymbol{d}_n$ be the response pattern and the item administration vector of respondent n, respectively. With a reference to the ignorability principle by Rubin (1976), Mislevy (1986) asserts that in adaptive testing consistent ML estimates of the model parameters can be obtained maximizing the likelihood of responses $\boldsymbol{x}_n$ conditionally on the design $\boldsymbol{d}_n$, that is, the design can be ignored. So, if $\xi' = (\alpha', \beta', \mu, \sigma)$ is the vector of all item and population parameters, the log-likelihood to be maximized can be written as

$$\ln L(\xi; \boldsymbol{X}, \boldsymbol{D}) = \Sigma_n \ln Pr(\boldsymbol{x}_n | \boldsymbol{d}_n; \xi), \tag{2}$$

where $\boldsymbol{X}$ stands for the data matrix and $\boldsymbol{D}$ stands for the design matrix.

To derive the MML estimation equations, it proves convenient to introduce the vector of derivatives

$$\boldsymbol{b}_n(\xi) = \frac{\partial}{\partial\xi}\ln Pr(\boldsymbol{x}_n\theta_n | \boldsymbol{d}_n; \xi) = \frac{\partial}{\partial\xi}[\ln Pr(\boldsymbol{x}_n | \boldsymbol{d}_n, \theta_n, \alpha, \beta, \gamma) + \ln g(\theta_n | \mu, \sigma)], \tag{3}$$

with

$$Pr(\boldsymbol{x}_n|\boldsymbol{d}_n,\theta_n,\alpha,\beta,\gamma)=\Pi_i\ \phi_i(\theta_n)^{d_{ni}x_{ni}}(1-\phi_i(\theta_n))^{d_{ni}(1-x_{ni})}. \tag{4}$$

Glas (1992, 1997, 1998) adopts an identity due to Louis (1982) to write the first order derivatives of Equation 2 with respect to ξ as

$$h(\xi)=\frac{\partial}{\partial\xi}\ln L(\xi;X,D)=\Sigma_n\ E(\boldsymbol{b}_n(\xi)|\boldsymbol{x}_n,\boldsymbol{d}_n,\xi). \tag{5}$$

This identity greatly simplifies the derivation of the likelihood equations. For instance, using the shorthand notation $\psi_{ni}=\psi_i(\theta_n)$ and $\phi_{ni}=\phi_i(\theta_n)$, from Equations 3 and 4 it can be easily verified that

$$\boldsymbol{b}_n(\alpha_i)=\boldsymbol{d}_{ni}\frac{(\boldsymbol{x}_{ni}-\phi_{ni})(1-\gamma_i)\theta\,\psi_{ni}(1-\psi_{ni})}{\phi_{ni}(1-\phi_{ni})} \tag{6}$$

and

$$\boldsymbol{b}_n(\beta_i)=\boldsymbol{d}_{ni}\frac{(\phi_{ni}-\boldsymbol{x}_{ni})(1-\gamma_i)\psi_{ni}(1-\psi_{ni})}{\phi_{ni}(1-\phi_{ni})}. \tag{7}$$

The likelihood equations for the item parameters are found upon inserting these expressions into Equation 5 and equating these expressions to zero. To derive the likelihood equations for the population parameters, using Equation 3 results in

$$\boldsymbol{b}_n(\mu)=(\theta_n-\mu)\sigma^{-2} \tag{8}$$

and

$$\boldsymbol{b}_n(\sigma)=-\sigma^{-1}+(\theta_n-\mu)^2\sigma^{-3}. \tag{9}$$

The likelihood equations are again found inserting these expressions in Equation 5 and equating these expressions to zero.

For computing estimation errors, and the LM, Wald, and CUSUM statistics, also the second order derivatives of the log-likelihood function are needed. As with the derivation of the estimation equations, also for the derivation of the matrix of second order derivatives, the theory by Louis (1982) can be used. Using Glas (1992), it follows that the observed information matrix, which is the opposite of the matrix of second order derivatives, that is,

$$H(\xi,\xi)=-\frac{\partial^2\ln L(\xi;X,D)}{\partial\xi\ \partial\xi'} \tag{10}$$

evaluated using MML estimates, is given by

$$H(\xi,\xi)\ =\ -\Sigma_n[E(B_n(\xi,\xi)|\boldsymbol{x}_n,\boldsymbol{d}_n,\xi)-E(\boldsymbol{b}_n(\xi)\boldsymbol{b}_n(\xi)|\boldsymbol{x}_n,\boldsymbol{d}_n,\xi)],, \tag{11}$$

where

$$B_n(\xi,\xi) = \frac{\partial^2 \ln Pr(x_n,\theta_n|d_n;\xi)}{\partial\xi\ \partial\xi'}. \qquad (12)$$

Unfortunately, for the 3-PL model, the exact expressions for the second order derivatives become prohibitively complicated. However, Mislevy (1986) points out that the observed information matrix can be approximated as

$$H(\xi,\xi) \approx \Sigma_n\ E(b_n(\xi)b_n(\xi)|x_n,d_n,\xi). \qquad (13)$$

Simulation studies by Glas (1997) in the framework of the 2-PL model and the nominal response model (Bock, 1972) show that this approximation is quite good, in the sense that statistics based on this approximation attain their theoretical distribution. In the sequel, it will become apparent that this must also hold for the 3-PL model.

Lagrange Multiplier Tests

Earlier applications of LM tests to the framework of IRT have been described by Glas and Verhelst (1995) and Glas (1997, 1998). The principle of the LM test (Aitchison & Silvey, 1958), and the equivalent efficient-score test (Rao, 1948) can be summarized as follows. Consider a null-hypothesis about a model with parameters ϕ_0. This model is a special case of a general model with parameters ϕ. In the present case the special model is derived from the general model by fixing one or more parameters to known constants. Let ϕ_0 be partitioned as $\phi_0' = (\phi_{01}', \phi_{02}') = (\phi_{01}', c')$, where c is the vector of the postulated constants and ϕ_{01} is the vector of free parameters of the special model. Let $h(\phi)$ be the partial derivatives of the log-likelihood of the general model, so $h(\phi) = (\partial / \partial\phi)\ln L(\phi)$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in ϕ. Let $H(\phi,\phi)$ be defined as $-(\partial^2 / \partial\phi\partial\phi')\ln L(\phi)$. Then the LM statistic is given by

$$LM = h(\phi_0)'\ H(\phi_0,\phi_0)^{-1}\ h(\phi_0). \qquad (14)$$

If Equation 14 is evaluated using the ML estimate of ϕ_{01} and the postulated values of c, it has an asymptotic χ^2 distribution with degrees of freedom equal to the number of parameters fixed (Aitchison & Silvey, 1958).

An important computational aspect of the procedure is that at the point of the ML estimates $\hat{\phi}_{01}$ the free parameters have a partial derivative equal to zero. Therefore, Equation 14 can be computed as

$$LM(c) = h(c)' W^{-1} h(c) \qquad (15)$$

with

$$W = H_{22}(c,c) - H_{21}(c,\hat{\phi}_{01})\, H_{11}(\hat{\phi}_{01},\hat{\phi}_{01})^{-1}\ H_{12}(\hat{\phi}_{01},c), \qquad (16)$$

where the partitioning of $H(\phi_0,\phi_0)$ into $H_{22}(c,c)$, $H_{21}(c,\hat{\phi}_{01})$, $H_{11}(\hat{\phi}_{01},\hat{\phi}_{01})$, and $H_{12}(\hat{\phi}_{01},c)$ is according to the partition $\phi_0' = (\phi_{01}', \phi_{02}') = (\phi_{01}', c')$.

Notice that $H(\hat{\phi}_{01},\hat{\phi}_{01})$ also plays a role in the Newton-Raphson procedure for solving the estimation equations and in computation of the observed information matrix. So its inverse will usually by available at the end of the estimation procedure. Further, if the validity of the model of the null-hypothesis is tested against various alternative models, the computational task is relieved because the inverse of $H(\hat{\phi}_{01},\hat{\phi}_{01})$ is already available and the order of W is equal to the number of parameters fixed, which must be small to keep the interpretation of the outcome tractable.

The interpretation of the outcome of the test is supported by observing that the value of Equation 15 depends on the magnitude of $h(c)$, that is, on the first order derivatives with respect to the parameters ϕ_{02} evaluated in c. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free, and the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free, that is, the values at which these parameters are fixed in the special model are adequate and the test is not significant, therefore, the special model is not rejected.

Lagrange Multiplier Statistics for Quality Control

In the introduction section, it was noted that simultaneous ML estimates of all item parameters in the 3-PL model are hard to obtain (see, for instance, Swaminathan & Gifford, 1986). Therefore, in the present paper it will be assumed that the guessing parameter γ_i is fixed to some plausible constant, say, to the guessing probability. In this section, it will be shown how an LM statistic can be used for testing whether this fixed guessing parameter is appropriate and remains appropriate when confronted with the adaptive testing data.

Consider G groups labeled $g = 1,\ldots, G$ and $y_{ng} = 1$ if person n belongs to group g, $y_{ng} = 0$ otherwise. In this paper, the first group partakes in the pretesting stage, and the following groups partake in the online stage. Given this partition, several hypotheses can be tested. For instance, Glas (1998) suggests evaluating DIF by testing the hypothesis that item parameters are constant over groups, i.e., testing the hypothesis that $\alpha_{ig} = \alpha_i$ and $\beta_{ig} = \beta_i$, for $g = 1, \ldots, G$. This can, of course, also be applied in an adaptive testing situation for monitoring parameter drift. However, in the present paper, a test for the hypothesis that $\gamma_{ig} = \gamma_i$, for $g = 1, \ldots, G$ will be given as an example of applying the LM approach to quality control of adaptive testing. The LM statistic for testing this hypothesis is based on the first order derivatives with respect to γ_{ig}. For using Equation 3, the first order derivatives of Equation 4 with respect to γ_{ig}, $b_n(\gamma_{ig})$, need to be computed. It is easily verified that

$$b_n(\gamma_{ig}) = y_{ng} d_{ni} \frac{(x_{ni} - \phi_{ni})(1 - \psi_{ni})}{\phi_{ni}(1 - \phi_{ni})}. \tag{17}$$

Let Γ_i be a vector of the elements, γ_{ig}, $g = 1, \ldots, G$. A test for the null-hypothesis $\gamma_{ig} = \gamma_i$ can be based on

$$LM(\Gamma_i) = h(\Gamma_i)' W^{-1} h(\Gamma_i) \tag{18}$$

with

$$W = H_{22}(\Gamma_i, \Gamma_i) - H_{21}(\Gamma_i, \xi)\, H_{11}(\xi, \xi)^{-1}\; H_{12}(\xi, \Gamma_i), \tag{19}$$

where ξ is the vector of the parameters of the null-model. Therefore, $H_{11}(\xi, \xi)$ is the matrix of second order derivatives with respect to these parameters, that is, it is equivalent to the matrix defined by Equation 10. If $h(\Gamma_i)$ and W are evaluated using MML estimates of the null-model, that is, the estimates of ξ, the LM(Γ_i) statistic has an asymptotic χ^2-distribution with G degrees of freedom.

A Wald Test and a CUSUM Chart for Quality Control

The CUSUM chart is an instrument of statistical quality control used for detecting small changes in product features during the production process. The CUSUM chart is used in a sequential statistical test, where the null-hypothesis of no change is never accepted (Veerkamp, 1996). In the present case, the alternative hypothesis is that the item is becoming easier and is loosing its discriminating power. Therefore, the null-hypothesis is $\alpha_{ig} - \alpha_{i1} \geq 0$ and, $\beta_{ig} - \beta_{i1} \geq 0$, for groups of respondents labeled $g = 1, \ldots, G$. As above, the first group partakes in the pretesting stage, and the following groups are taking an adaptive test.

Before turning to the one-sided hypothesis $\alpha_{ig} - \alpha_{i1} \geq 0$ and $\beta_{ig} - \beta_{i1} \geq 0$, first consider the two-sided null-hypothesis that $\alpha_{ig} - \alpha_{i1} = 0$ and, $\beta_{ig} - \beta_{i1} = 0$. Let d_{ig} be a vector defined by $d_{ig} = (\alpha_{ig} - \alpha_{i1}, \beta_{ig} - \beta_{i1})'$. This two-sided hypothesis can be evaluated with the Wald statistic

$$Q_i = d_{ig}' W_{ig}^{-1} d_{ig}, \tag{20}$$

where W_{ig} is the covariance matrix of d_{ig}. Since the statistic is computed using independent estimates of the item parameters in two groups, it holds that $W_{ig} = \Sigma_{ig} + \Sigma_{i1}$ where Σ_{ig} and Σ_{i1} can be approximated using the relevant elements of the inverse of the opposite of Equation 13, computed with the MML estimates obtained in group g and group 1, respectively. This statistic defined in Equation 20 has an asymptotic χ^2 distribution with two degrees of freedom. However, the interest is in a one-sided test, so also the signs of the elements of d_{ig} are needed. Since Equation 20 is a quadratic form, its signed square root is of interest. Further, it may be interesting to test the hypothesis iteratively. Therefore, a one-sided cumulative sum chart will be based on the quantity

$$S_i(g)=\max\left\{S_i(g-1)+\frac{\alpha_{il}-\alpha_{ig}}{Se(\alpha_{ig}-\alpha_{il})}+\frac{\beta_{il}-\beta_{ig}}{Se(\beta_{il}-\beta_{ig}|\alpha_{il}-\alpha_{ig})}-k_i,0\right\}, \tag{21}$$

where $Se(\alpha_{ig}-\alpha_{il})=\sigma_\alpha$ and $Se(\beta_{il}-\beta_{ig}|\alpha_{il}-\alpha_{ig})=\sqrt{\sigma^2_\beta-\sigma^2_{\alpha,\beta}/\sigma^2_\alpha}$, with $\sigma^2_\alpha,\sigma^2_\beta$ and $\sigma_{\alpha,\beta}$

the appropriate elements of the covariance matrix W_{ig}, which is also used in Equation 20. Further, k_i is a reference value. The CUSUM chart starts with

$$S_i(0)=0, \tag{22}$$

and the null-hypothesis is rejected as soon as

$$S_i(j)>h_i, \tag{23}$$

where h_i is some constant threshold value. The choice of the constants k_i and h_i determines the power of the procedure. In the case of the Rasch model, where the null-hypothesis is $\beta_{ig}-\beta_{il}\geq 0$, and the term involving the discrimination indices is lacking from Equation 21, Veerkamp (1996) successfully uses $k = 1/2$ and $h_i = 5$. This choice was motivated by the consideration that this set up has good power against the alternative hypothesis of a normalized shift in item difficulty of approximately one standard deviation. In the present case one extra normalized decision variable is employed, i.e., the variable involving the discrimination indices. To have power against a shift of one standard deviation of both normalized decision variables in the direction of the alternative hypothesis, a value $k_i = 1$ will be tried out below. The value $h_i = 5$ will not be changed.

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 estimate of the standard error of the parameter estimates, and this standard error is approximated using Equation 13, the precision of this approximation will be studied first by assessing the power of the statistics under the null-model. Then the power of the tests will be studied under various model violations.

For all simulations reported below, the ability parameters θ were drawn from a standard normal distribution. The item difficulties β_i were uniformly distributed on [-1.0, 1.0], the discrimination indices α_i were drawn from a log-normal distribution with a zero mean and a standard deviation equal to 0.10, and the guessing parameter γ_i was generally fixed at 0.20. In the online phase, item selection was done using the maximum information principle. The ability parameter θ 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 power of the statistics under the null-model are shown in Table 1. 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. Both in the pretest phase and the online phase, test lengths L of 20 and 40 were chosen, the exact setup is shown in the first two columns of Table 1. Finally, in the third column it can be seen that the number of respondents per phase was fixed at 500 and 1,000 respondents. So summed over the pretest and online phase, the sample sizes were 1,000 and 2,000 respondents, respectively. For the pretest phase, a spiraled test administration design was used. For instance, for the $K = 50$ studies, for the pretest phase, five subgroups were used, the first subgroup was administered items 1 to 20, the second, items 11 to 30, the third, items 21 to 40, the fourth, items 31 to 50, and the fifth group received items 1 to 10 and 41 to 50. In this manner, all items drew the same number of responses in the pretest phase. For the $K = 100$ studies, the pretest phase consisted of four subgroups administered 50 items. Here the design was 1-50, 26-75, 51-100 and 1-25 and 76-100. One hundred replications were run for each study.

TABLE 1

Power of LM and Wald test under the null-model (100 replications)

			Percentage Significant at 10%	
K	*L*	N_g	LM Test	Wald Test
50	20	500	8	9
		1,000	10	10
	40	500	9	10
		1,000	11	8
100	20	500	12	10
		1,000	8	9
	40	500	10	12
		1,000	10	10

Note. K = size of the item pool; L = test length; N_g = number of persons in calibration and adaptive testing batches.

The results of the study are shown in the last two columns of Table 1. These columns contain the percentages of LM and Wald tests that were significant at the 10% level. It can be seen that the power of the tests conforms to its theoretical value of 10%. Therefore, it can be concluded that the approximations of the standard errors were quite close.

A second series of simulations focused on the power in the case that the online responses were given using a value for the guessing parameter γ_i that was different from the value of the pretest phase. Results are shown in Table 2. The first panel of the table pertains to a situation where, for the items 5, 10, 15, etc., γ_i changes from 0.00 in the pretest phase to 0.25 in the online phase. So 20% of the items do not fit the null-model of the pretest phase. In the fourth and fifth column, the rejection rate of aberrant items using a 10% significance level is shown for the LM and Wald test, respectively. The number of replications was 100. It can be seen that the power of both tests is quite large. Then, for 20 replications, 9 more batches of size N_g of respondents were generated and for each new batch, the CUSUM statistic defined by Equation 21 was computed. In the last six columns the percentage of the detected aberrant items is shown. Non-aberrant items were detected at chance level, in this case 5%. It can be seen that approximately 100% of the aberrant items are detected after 4 iterations, which can be considered quite good.

TABLE 2
Detection of aberrant items: changes in γ_i (per row: 100 replications for LM/Wald and 20 replications for CUSUM)

			Significant at 10%		CUSUM Detected After Iteration					
K	L	N_t	LM Test	Wald Test	2	3	4	6	8	10
from $\gamma_i = .00$ to $\gamma_i = .25$										
50	20	500	95	69	72	77	88	100	100	100
		1,000	100	70	85	90	100	100	100	100
	40	500	100	100	77	83	99	100	100	100
		1,000	100	100	93	98	100	100	100	100
100	20	500	92	93	69	75	92	100	100	100
		1,000	98	92	81	95	100	100	100	100
	40	500	100	100	73	87	100	100	100	100
		1,000	100	100	88	99	100	100	100	100
from $\gamma_i = .20$ to $\gamma_i = .30$										
50	20	500	10	25	2	3	4	12	33	45
		1,000	40	60	2	4	4	35	58	66
	40	500	31	22	3	3	4	22	44	65
		1,000	55	73	4	6	7	45	56	78
100	20	500	18	21	1	2	10	11	45	50
		1,000	58	47	4	5	5	13	54	67
	40	500	42	44	3	4	7	32	45	75
		1,000	49	77	2	6	7	22	50	76
from $\gamma_i = .20$ to $\gamma_i = .40$										
50	20	500	50	44	10	15	19	40	66	70
		1,000	90	60	12	18	22	50	81	82
	40	500	89	97	18	26	33	76	89	100
		1,000	100	99	17	24	38	73	100	100
100	20	500	52	44	9	12	18	34	75	86
		1,000	88	73	11	22	25	68	79	100
	40	500	90	82	19	24	31	57	83	100
		1,000	100	100	18	29	30	77	100	100

The positive picture of the power of the LM, Wald, and CUSUM changes dramatically if $\gamma_i = .20$ changes from 0.20 in the pretest phase to 0.30 in the online phase. From the second panel of Table 2, it can be seen that in this case the power of the LM and Wald test is quite low, while even after 10 iterations the CUSUM procedure has only detected about half of the aberrant items. In the last panel of Table 2, γ_i changes from 0.20 to 0.40, and the power becomes better, although for the $L = 20$ studies, the power is still quite low.

Note that in the above simulations, only the LM test is strictly aimed at the alternative that γ_i has changed. However, the estimates of the three parameters of the 3-PL model are highly correlated. This implies that changes in parameters are often confounded and it is very difficult to identify the actual parameter that is changing. For instance, if an item becomes known, this can both be translated into an augmentation of γ_i, that is, in an augmentation of a correct response unassociated with θ, in a loss of discriminating power, and in a lowering of item difficulty. As a consequence, a test that should be sensitive to changes in γ_i may also have power against changes in α_i and β_i. The latter case was investigated using the same simulation setup as above. The results are displayed in Table 3, the first panel pertains to a change of –0.50 in the difficulty of the items 5, 10, 15, 20, and so on; the second panel pertains to a change –1.00 in the difficulty of these items. It can be seen that all tests are indeed sensitive to these changes, especially the power for the change –1.00 is very high.

TABLE 3
Detection of aberrant items: changes in β_i (per row: 100 replications for LM/Wald and 20 replications for CUSUM)

			Significant at 10%		CUSUM Detected After Iteration					
K	L	N_g	LM Test	Wald Test	2	3	4	6	8	10
change –0.50										
50	20	500	25	24	12	17	33	65	96	100
		1,000	27	28	10	17	28	80	91	100
	40	500	24	22	12	26	38	88	99	100
		1,000	30	20	15	22	41	83	100	100
100	20	500	23	21	12	18	31	94	95	100
		1,000	44	33	5	20	32	78	89	100
	40	500	50	42	19	24	54	87	100	100
		1,000	53	44	17	23	55	87	100	100
change –1.00										
50	20	500	99	89	80	99	100	100	100	100
		1,000	90	90	85	90	100	100	100	100
	40	500	89	96	87	83	100	100	100	100
		1,000	94	96	89	98	100	100	100	100
100	20	500	96	98	87	95	98	100	100	100
		1,000	99	92	83	95	100	100	100	100
	40	500	89	94	93	97	100	100	100	100
		1,000	99	99	98	99	100	100	100	100

Discussion

This paper explored how to evaluate whether the adaptive testing data used for online calibration sufficiently fit the item response model used. Three approaches were studied, one based on a Lagrange multiplier (LM) statistic, the others on a Wald and a cumulative sum (CUSUM) statistic, respectively. The theoretical advantage of the latter procedure is that it is based on a directional hypothesis and can be used iteratively. The power of the tests was evaluated with a number of simulation studies. It was found that the power of the procedures ranged from rather moderate for a change from $\gamma_i = 0.20$ to $\gamma_i = 0.30$, to good for a change from $\gamma_i = 0.00$ to $\gamma_i = 0.25$. Further, it was found that the tests are equally sensitive to changes in item difficulty and the guessing parameter. So the bottom line here is that all these statistics detect that something has happened to the parameters, but it will be very difficult to attribute misfit to specific parameters.

References

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

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

DeLeeuw, J., & Verhelst, N. D. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11,* 183–196.

Follmann, D. (1988). Consistent estimation in the Rasch model based on non-parametric margins. *Psychometrika, 53,* 553–562.

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). Westport, CT: Ablex Publishing Corporation.

Glas, C. A. W. (1997). *Some modification indices for the 2-PL model and the nominal response model.* Enschede, The Netherlands: Twente University, Department of Educational Measurement and Data Analysis.

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

Glas, C. A. W., & Verhelst, N. D. (1995). Tests of fit for polytomous Rasch models. In G. H. Fischer & I. W. Molar (Eds.), *Rasch models. Their foundation, recent developments and applications.* New York: Springer.

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

Rao, C. R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Swaminathan, H. & Gifford, J. A. (1986). Bayesian estimation for the one-parameter logistic model, *Psychometrika, 47,* 349–364.

Veerkamp, W.J.J. (1996). *Statistical methods for computerized adaptive testing.* Unpublished doctoral thesis, Twente University, Enschede, The Netherlands.

Wetherill, G. B. (1977). *Sampling inspection and statistical quality control, (2nd ed.).* London: Chapman and Hall.

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.