Journal of Educational and Behavioral Statistics
Winter 2000, Vol. 25, No. 4, pp. 373–389

Detection of Known Items in Adaptive Testing with a Statistical Quality Control Method

Wim J. J. Veerkamp
Cees A. W. Glas
University of Twente

Keywords: *computer adaptive testing, CUSUM charts, differential item functioning, item response theory, parameter drift*

Due to previous exposure of items in adaptive testing, items may become known to a substantial portion of examinees. A disclosed item is bound to show drift in the item parameter values. In this paper, it is suggested to use a statistical quality control method for the detection of known items. The method is worked out in detail for the 1-PL and 3-PL models. Adaptive test data are used to re-estimate the item parameters, and these estimates are used in a test of parameter drift. The method is illustrated in a number of simulation studies, including a power study.

One of the major problems in computer adaptive testing (CAT) is security. 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. These procedures have in common that they prevent the items from being administered more often than desirable. Typically, this goal is reached through modifying the item selection criterion such that the 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; also see, Stocking, 1993). With relatively low exposure rates, items will probably become known later than with high exposure rates. Still, sooner or later some items may become known to a part of the future examinees.

In this paper, it is suggested to use a method of statistical quality control, the so-called CUSUM charts, to detect possibly known items. After detection such an item can be removed from the item bank. The methods proposed in this paper are primarily focused on parameter drift due to item disclosure. However, parameter drift may also occur as a result of differences between the pretest and the operational phase. One might think of a change in the mode of presentation

The authors would like to thank Wim van der Linden, Martijn Berger, and Norman Verhelst for their valuable suggestions and comments.

(computerized or paper-and-pencil administration) as well as of a change in the motivation of the respondents. However, the method proposed below entails a one-sided test that the item is becoming easier and is losing its discrimination. The method will have power for parameter drift fitting this description, and not for parameter drift in the opposite direction.

Two IRT models will be considered, the 1-PL and the 3-PL model. In practical situations, the choice of a model is often related to the number of respondents available for parameter estimation. Lord (1983) has shown that for small sample sizes (under about 500) the parameter estimates in the 2- and 3-PL models become inaccurate, and in these cases the 1-PL model is to be preferred. For larger sample sizes, however, the 2- and 3-PL models are preferred for their greater flexibility, which generally results in better model fit.

This paper is organized as follows: In the next section, the effect of item disclosure on the item parameters will be described as parameter drift. It will be shown that tests for parameter drift are special cases of tests for item bias or differential item functioning (DIF). Then it will be explained how the CUSUM chart can be used for the detection of disclosed items. Finally, the method is illustrated in a number of simulations, and a power study is presented.

Item Parameter Drift

If an item is known by a fair number of respondents, the item follows a different item response curve than it is supposed to. Assume that an item known in advance is answered correctly with probability one. Then the following consequences for parameter estimates can be derived.

The 1-PL Model

Assume that item i is known by a proportion of c_i of the respondents and that the original item response curve describing the probability of a correct response as a function of the proficiency parameter θ is $P_i^0(\theta)$. Then the probability of a random examinee with proficiency level θ giving a correct response to item i is

$$c_i + (1 - c_i)\, P_i^0(\theta), \tag{1}$$

with

$$P_i^0(\theta) = \frac{\exp(\theta - \beta_i^0)}{1 + \exp(\theta - \beta_i^0)}, \tag{2}$$

and β_i^0 the original difficulty parameter of item i. Note that (1) is equivalent to a 3-PL item response curve with guessing parameter c_i and a discrimination parameter equal to one. Under these assumptions, the fact that respondents know an item in advance leads to an apparent shift of its guessing parameter away from zero. Also, the item p-value will increase due to the examinee's advance knowledge of the item. In the 1-PL model, an increase of an item p-value implies a decrease of the value of the item's difficulty parameter. Before

describing how to detect parameter drift, the effect of respondents knowing items in advance on the difficulty parameter is derived.

Samejima (1984) gives a 1-PL item response curve that is close to the 3-PL item response curve given in (1). The difficulty parameter β_i^1 of this 1-PL item response curve,

$$P_i^1(\theta) = \frac{\exp(\theta - \beta_i^1)}{1 + \exp(\theta - \beta_i^1)}, \tag{3}$$

is derived equating (3) and (1) for $\theta = \beta_i^0$. So from $P_i^1(\beta_i^0) = c_i + (1 - c_i) P_i^0(\beta_i^0)$, it follows that $\beta_i^0 - \beta_i^1 = \text{logit}\,(\frac{1}{2} + \frac{1}{2}c_i)$, which results in $\beta_i^1 = \beta_i^0 - \log\left(\frac{1 + c_i}{1 - c_i}\right)$ and

$$\beta_i^1 \approx \beta_i^0 - 2c_i.$$

Hence, if 1-PL curves are used to describe item responses, and percentage c_i of the respondents knows item i in advance, the difficulty parameter seems to have decreased by approximately $2c_i$.

The 3-PL Model

To derive the impact of item disclosure in the 3-PL model, a probabilistic interpretation of the 3-PL model as a response model must be given first. The probability of a correct response of a random respondent with proficiency parameter θ is given by

$$P_i(\theta) = \gamma_i + (1 - \gamma_i)\psi_i(\theta), \tag{4}$$

with

$$\psi_i(\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{1 + \exp(\alpha_i(\theta - \beta_i))},$$

where α_i, β_i, and γ_i are the discrimination, difficulty and guessing parameter, respectively. To derive an interpretation of the model, rewrite (4) as

$$P_i(\theta) = \gamma_i(1 - \psi_i(\theta)) + \psi_i(\theta). \tag{5}$$

This can be interpreted as a model where there is a probability $\psi_i(\theta)$ that the respondent can give a correct response using the relevant proficiency, a probability $1 - \psi_i(\theta)$ that the respondent cannot give a correct response using the relevant proficiency, and guesses with γ_i as the probability of a correct response. Note that it is assumed that conditionally on the case where the respondent can give a correct response using the relevant proficiency, the correct response is given with probability equal to one. So the total probability of a correct response is a sum of a term $\psi_i(\theta)$ and a term $\gamma_i(1 - \psi_i(\theta))$. Now suppose that the item

has become known to a proportion of c_i of the respondents, and all these respondents give a correct response with probability equal to one. Then using (5), it follows that the probability of a correct response becomes

$$c_i + (1 - c_i)[\gamma_i(1 - \psi_i(\theta)) + \psi_i(\theta)],$$

which can be rewritten as

$$(c_i + \gamma_i + c_i\gamma_i) + (1 - (c_i + \gamma_i + c_i\gamma_i))\,\psi_i(\theta). \tag{6}$$

Notice that (6) entails a 3-PL model with guessing parameter $c_i + \gamma_i + c_i\gamma_i$, and discrimination and difficulty parameters equal to the original parameters α_i and β_i, respectively. So the disclosure of the item has translated itself in an augmentation of the lower asymptote of the item response curve.

Detection of Parameter Drift With the CUSUM Charts

Parameter drift has much in common with DIF. In both situations, one distinguishes between two or more groups of respondents. In DIF studies, one group serves as a reference group, and whether or not the response behavior of focal groups differs from that of the reference group is evaluated. In studies of parameter drift, one may distinguish a calibration phase from a CAT phase and evaluate whether or not response behavior differs. Therefore, it may come as no surprise that the statistical tools for the two kinds of studies are related. Lord (1980, Chap. 14) suggests a test for DIF to determine whether or not item parameters differ for two groups of respondents. The same test can be used to test whether the item has become easier, that is, to test the null hypothesis $\beta_i^1 - \beta_i^0 \geq 0$ against the alternative $\beta_i^1 - \beta_i^0 < 0$, where β_i^0 is the parameter value in the calibration phase and β_i^1 is the parameter value during the administration of the adaptive test. Assuming that β_i^0 is calibrated (perhaps recalibrated after the first use of the item in the adaptive test) as $\hat{\beta}_i^0$, with standard error $\sigma(\hat{\beta}_i^0)$, the test statistic suggested by Lord becomes

$$\frac{\hat{\beta}_i^0 - \hat{\beta}_i^1}{\sqrt{\sigma^2(\hat{\beta}_i^0) + \sigma^2(\hat{\beta}_i^1)}}, \tag{7}$$

where $\hat{\beta}_i^1$ and $\sigma(\hat{\beta}_i^1)$ are the difficulty parameter estimate and its standard error based on data collected during the administration of the adaptive tests. Since β_i^0 and β_i^1 are estimated using independent samples, the estimates do not covary and the standard error of the difference $\hat{\beta}_i^0 - \hat{\beta}_i^1$ can be computed as the denominator of (7).

The test statistic given in (7) fits in the general framework of Wald-type tests (Glas & Verhelst, 1995, pp. 89–92) and has an asymptotic standard normal distribution. In the framework of DIF studies, this test is not much used; most practitioners prefer using the Mantel-Haenszel procedure (Holland & Thayer,

1988), though Fischer (1995) points out that this approach has some serious shortcomings in situations where the 1-PL method does not hold. However, it is beyond the scope of this paper to delve deeper into this matter; for more information on methods for detection of DIF, see, for instance, Camilli and Shephard (1994).

In this paper, it is suggested that the cumulative sum (CUSUM) chart be used for testing parameter drift. The procedure can be viewed as a sequential series of Wald tests. The CUSUM chart is an instrument used in statistical quality control (see, e.g., Wetherill, 1977) for detecting changes in product features during the production process. It is used in a sequential statistical test, where the null hypothesis of no change is never accepted. The test always continues until the null hypothesis of no change is rejected. It is only a matter of how many samples it takes until either a certain change is detected or the null hypothesis is erroneously rejected. In other words, the power of the test is a function of the number of samples. For quality control of items in an adaptive testing item pool, this method can be based on cumulated deviations of difficulty parameter estimates from the value found in the calibration study.

Description of the CUSUM Procedure

Because the procedure is conceptually simpler for the 1-PL than for the 3-PL, the CUSUM procedure will be described for the former model first. Consider a CAT program where the item parameters are re-estimated every once in a while, for example, each time the test has been taken by N respondents. It is assumed that in this period item i has been administered n_i times. Each new difficulty estimate is based on the item responses collected after the previous re-estimation. The estimate found at the j-th re-estimation, $\hat{\beta}_i^j$, is compared with the value of the estimate found in the intial calibration, that is, $\hat{\beta}_i^0$. For the 1-PL model, a one-sided cumulative sum chart can be based on the quantity

$$S_i(j) = \max\left\{ S_i(j-1) + \frac{\hat{\beta}_i^0 - \hat{\beta}_i^j}{\sqrt{\sigma^2(\hat{\beta}_i^0) + \sigma^2(\hat{\beta}_i^j)}} - k, 0 \right\}, \qquad (8)$$

where k is a constant reference value indicating the size of the standardized shift considered worth charting. In the sequel, k will be referred to as the effect size. The cumulative sum chart starts with

$$S_i(0) = 0,$$

and the null hypothesis is rejected as soon as

$$S_i(j) > h, \qquad (9)$$

where h is some constant threshold value.

Generalization of the procedure to the 3-PL model is complicated by a problem specific to CAT. The problem is that guessing may be prominent in the calibration phase, while it may occur less frequently in the CAT phase, because there the items are tailored to the proficiency level of the respondents. Therefore, it will be assumed that the guessing parameter is fixed to some plausible constant, say, to the reciprocal of the number of response alternatives available. Since the guessing parameter is fixed, augmentation of the lower asymptote of the item characteristic curve, as modeled in (9), will now be translated into lowering of the item discrimination or difficulty parameter, or both. For the 3-PL model with fixed guessing parameter, the procedure starts with the parameter estimates α_i^0 and β_i^0 originally obtained. Then, for new batches of respondents $j = 1, \ldots, J$ taking the adaptive test, the alternative hypothesis entails that the item is becoming easier and is losing its discriminating power. Therefore, the simultaneous null-hypothesis is $\alpha_i^j - \alpha_i^0 \geq 0$ and $\beta_i^j - \beta_i^0 \geq 0$ for $j = 1, \ldots, J$. A cumulative sum chart will be based on the quantity

$$S_i(g) = \max\left\{ S_i(g-1) + \frac{\hat{\alpha}_i^0 - \hat{\alpha}_i^j}{Se(\hat{\alpha}_i^0 - \hat{\alpha}_i^j)} + \frac{\hat{\beta}_i^0 - \hat{\beta}_i^j}{Se(\hat{\beta}_i^0 - \hat{\beta}_i^j \mid \hat{\alpha}_i^0 - \hat{\alpha}_i^j)} - k_i, 0 \right\}. \quad (10)$$

Since α_i^0 and α_i^1 are estimated using independent samples, the standard error $Se(\hat{\alpha}_i^0 - \hat{\alpha}_i^j)$ can be computed as the square root of $\sigma^2(\hat{\alpha}_i^0) + \sigma^2(\hat{\alpha}_i^j)$. Further, within a sample the estimates of discrimination and difficulty parameters are highly correlated. Therefore, in (10) the difference $\hat{\beta}_i^0 - \hat{\beta}_i^j$ is weighted with the conditional standard error $Se(\hat{\beta}_i^0 - \hat{\beta}_i^j \mid \hat{\alpha}_i^0 - \hat{\alpha}_i^j)$, which is computed as the square root of $\sigma^2(\hat{\beta}_i^0) + \sigma^2(\hat{\beta}_i^j) - (\sigma^2(\hat{\alpha}_i^0,\hat{\beta}_i^0) + \sigma^2(\hat{\alpha}_i^j,\hat{\beta}_i^j))/(\sigma^2(\hat{\alpha}_i^0) + \sigma^2(\hat{\alpha}_i^j))$, where $\sigma(\hat{\alpha}_i^0,\hat{\beta}_i^0)$ and $\sigma(\hat{\alpha}_i^j,\hat{\beta}_i^j)$ are the covariance of the estimates of the discrimination and difficulty parameters obtained in the calibration sample and the j-th CAT sample, respectively. More details on the computation of these standard errors will be given below.

Parameters of the CUSUM Procedure

The values of k, h, and N determine the success of the procedure. The choices of these values are often based on the run length distribution or the average run length (ARL) for the in-control state (no parameter drift) and a specific out-of-control state (the amount of drift of the parameter value to be detected). The run length is defined as the number of samples, that is, the number of times the difficulty parameter of the item is re-estimated, taken before the chart indicates a lack of control, a drift of the difficulty parameter. Obviously, the ARL of the in-control state should be large, whereas the ARL of the out-of-control state should be small. Montgomery (1991, p. 295) gives recommendations for the values of these parameters. He suggests using $k \approx 1/2$ and h is four or five. These values provide a CUSUM chart with good power (in terms of ARL)

against the alternative hypothesis of a shift in item difficulty level of a size approximately one standard deviation of the sample variable used in the CUSUM chart. In (9) this standard deviation is equal to one, because one normalized variable is used. In other words, the variable $\hat{\beta}_i^0 - \hat{\beta}_i^j$ is divided by its standard deviation. The suggestions should be viewed as rules of thumb that can be adapted if the resulting run length distribution is not considered satisfactory. For the test for the 3-PL model, one extra normalized decision variable is employed: 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 $k = 1$ suggests itself. Depending on the effect size the practitioner is interested in, other values are of course also possible. Below it will be shown how a suitable value of h in a concrete situation can be found by performing a simulation study.

Estimation of Item Parameters and Standard Errors

For the CUSUM chart approach described above new estimates of the difficulty parameter are needed. For the 1-PL model, standard item parameter estimation techniques are conditional maximum likelihood (CML) and marginal maximum likelihood (MML) (see, for instance, Molenaar, 1995). However, Glas (1988) shows that CML is not feasible for multi-stage and adaptive testing designs. Further, it is shown that MML can be used for the estimation of item parameters in these response-contingent designs in a very broad class of IRT models. For the estimation of the item parameters in these designs, the process causing missing responses can be ignored and the MML estimation equations have the same form as in the case of a fixed, non-response-contingent design (Glas, 1988; Mislevy & Wu, 1988, 1996). However, for sampling inferences, such as computation of standard errors, an adaptive test design cannot be ignored (see, for instance, Little & Rubin, 1987, p. 88). This implies that in these cases the asymptotic variance of the MML parameter estimator is not equal to the reciprocal of Fisher information. In the present paper, it is assumed that neglecting this problem introduces only a minor additional bias in the standard error estimator. The simulation studies presented below corroborate that assumption.

Therefore, standard errors are computed as follows. Let the item administration variable d_{ni} take the value one if item i was administered to respondent n and zero if this was not the case. If $d_{ni} = 1$, x_{ni} is the binary response of respondent n to item i, if $d_{ni} = 0$, it will be assumed that x_{ni} is equal to some arbitrary constant. Let $\boldsymbol{x}_n$ and $\boldsymbol{d}_n$ be the response pattern and the item administration vector of respondent n, respectively. It will be assumed that proficiency is normally distributed. Further, the groups of respondents $j = 1, \ldots, J$ may have their own proficiency distribution. So let $g(.;\mu_{j(n)},\sigma_{j(n)})$ be the density of θ_n, $j(n)$ is the index of the j-th proficiency distribution. Further, ξ is a vector of all

item and proficiency distribution parameters in the model. The log-likelihood to be maximized in MML can be written as

$$\begin{aligned} L(\xi;\boldsymbol{X},\boldsymbol{D}) &= \Sigma_n \log p(\boldsymbol{x}_n \mid \boldsymbol{d}_n;\xi) \\ &= \Sigma_n \log \int p(\boldsymbol{x}_n,\theta_n \mid \boldsymbol{d}_n;\xi)\, d\theta_n \\ &= \Sigma_n \log \int p(\boldsymbol{x}_n \mid \boldsymbol{d}_n,\theta_n,\alpha,\beta,\gamma) g(\theta_n;\mu_{j(n)},\sigma_{j(n)})\, d\theta_n, \end{aligned}$$

where $\boldsymbol{X}$ stands for the data matrix and $\boldsymbol{D}$ for the design matrix. Mislevy (1986) suggests that the asymptotic covariance matrix of the parameter estimates be computed by inverting

$$\boldsymbol{H}(\xi,\xi) \approx \Sigma_n \boldsymbol{h}_n(\xi)\boldsymbol{h}_n(\xi)' \tag{11}$$

with

$$\boldsymbol{h}_n(\xi) = \frac{\partial}{\partial \xi} \log p(\boldsymbol{x}_n \mid \boldsymbol{d}_n;\xi) = E(\boldsymbol{b}_n(\xi) \mid \boldsymbol{x}_n,\boldsymbol{d}_n,\xi),$$

and

$$\boldsymbol{b}_n(\xi) = \frac{\partial}{\partial \xi} \log p(\boldsymbol{x}_n,\theta_n \mid \boldsymbol{d}_n;\xi).$$

Glas (1999) shows that (11) is an approximation of the expected Fisher information matrix. In the examples given below, this matrix was computed using the MML parameter estimates issued by Bilog-MG (Zimoski, Muraki, Mislevy, & Bock, 1996).

The CUSUM method entails independent parameter estimates in different groups. The scales obtained in the groups must be identified in such a way that the identification restrictions interfere as little as possible with the conclusions on parameter drift. When the proficiency distributions of the groups differ, which they usually do, fixing the parameters of the proficiency distributions of the groups leads to a shifting of the two latent scales, which will result in erroneous conclusions. Using the parameters of one of the items to identify the scales is even more problematic, especially when the item used happens to be biased itself. Therefore, in the framework of a discussion of a Wald test for DIF in the Rasch model, Glas and Verhelst (1995, p. 91) argue that the best way to identify independent estimates is to rescale the parameter estimates in each group in such a way that the sum of the estimates is zero. For the 3-PL model, an analogous approach entails imposing the restrictions that both the sum of the difficulties and the sum of the logarithms of the discrimination parameters are zero. These parameter transformations also imply a transformation of the covariance matrix of the estimates; for details, refer to Glas and Verhelst (1995, p. 92).

An Illustration

In this section, the method is illustrated for the 1-PL model in a number of simulation studies. In the next section, a more systematic power study will be presented for the 3-PL model. The method suggested above was applied to a simulated item pool, including items previously known by a certain percentage of the respondents. The bank consisted of 1-PL items, with difficulty levels drawn from the standard normal distribution. From the 150 items in the pool, 15 randomly chosen items were assumed to be known. Three items were assumed known by 25% of the respondents, three by 20%, three by 15%, three by 10%, and three items by 5% of the respondents.

Calibration data were generated according to the following design. The items were divided into six booklets. Each booklet consisted of 50 items. The first booklet consisted of items 1 to 50, the second, items 26 to 75, the third, 51–100, the fourth, 76–125, the fifth, 101–150, and the sixth, 126–150 and 1–25. So, each item figured in two booklets. Each respondent was presented with one booklet. Sample sizes of 100 and 500 respondents per booklet were used. So, in total, 600 and 3000 respondents were used for calibration, and as each respondent was given one third of the pool, each item was presented to 200 or 1000 respondents, respectively. The proficiency distributions of the six groups of respondents were assumed to be normal with standard deviation one, and with means varying from -0.25 to $+0.25$, with a difference of 0.1 between successive groups.

Using these data, MML estimates of the difficulty parameters were computed using a subroutine of the OPLM-package (One parameter logistic model; Verhelst, Glas, & Verstralen, 1994), called OPMML. This program computes both the estimates of the item and proficiency distribution parameters and their standard errors. Separate normal distributions were assumed for the proficiency parameters in each group.

Adaptive Test Design

In the adaptive testing phase, respondents' proficiency parameters were drawn from a normal distribution with mean 0.2 and variance 1. The responses were generated according to the 1-PL model (2) for the items assumed to be unknown to the respondents. The responses to the previously known item were generated according to the 3-PL model given in (1). In the adaptive testing phase, the examinee's proficiency was estimated using weighted maximum likelihood (Warm, 1989). The selection of the first item was made with the proficiency estimate equal to zero. The maximum-information criterium was used for item selection. Since a 1-PL item bank was used, an optimal item was an item with difficulty estimate closest to the current proficiency estimate.

Exposure Control

In this study, exposure of the items was controlled through random item selection (see, e.g., Kingsbury & Zara, 1989). Each time the four most informa-

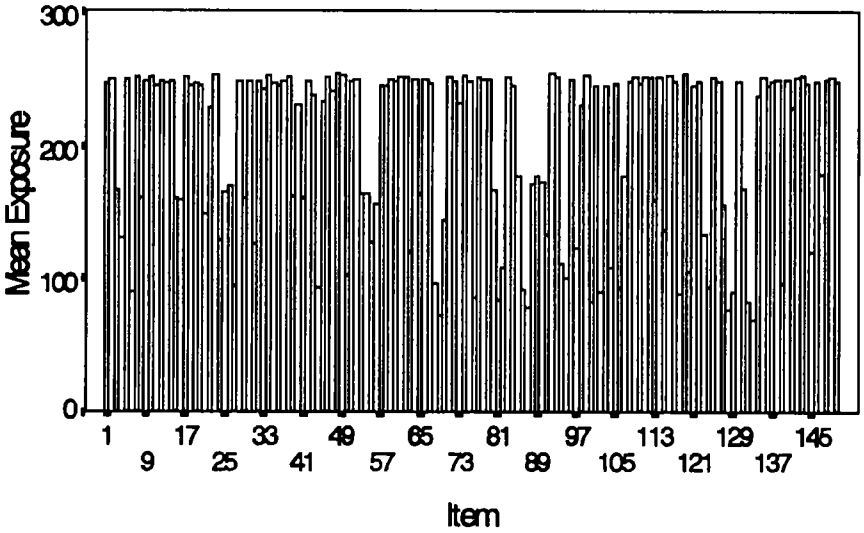


FIGURE 1. *An example of mean item exposure where exposure is the number of times the item is exposed in 1000 tests (The mean is taken over 25 replications.)*

tive items were selected, and from this set an item was randomly chosen to be administered. Selected items were not selected again. As a result, the exposure rates of the individual items were not far from the average exposure rate of 20% (30 items in the test, 150 items in the pool).

In Figure 1, an example of the exposure distribution is given. Items are ordered according to the difficulty values found in the calibration study. The exposure rates are computed as the mean over 25 replications, where every replication consisted of 1000 adaptive tests. About half of the items had an exposure rate of about 25%, and an exposure rate of less than 10% was found for less than 10% of the items.

The CUSUM Procedure

The CUSUM procedure was carried out with parameters $k = 1/2$, $h = 5$, and $N = 1000$. Twenty-five batches of 1000 respondents were generated; in every batch, the parameters were re-estimated using MML and these estimates were compared with the values obtained in the initial calibration. Finally, the normalized differences were added to the cumulative sums (8).

In Table 1, the minimum, mean, and maximum values of the estimates of the standard errors used in the CUSUM chart are shown. As the values of the parameters h and k were chosen to provide a CUSUM chart with power against the alternative of a shift in difficulty of one standard error, the last line in the table indicates the size of the parameter shifts for which the current CUSUM chart can be used.

TABLE 1
Estimates of Standard Errors of Item Parameter Estimators

Sample Size	100			500		
Standard Error	Min	Mean	Max	Min	Mean	Max
$\sigma(\hat{\beta}_i^0)$	0.28	0.29	0.37	0.12	0.13	0.15
$\sigma(\hat{\beta}_i^1)$	0.14	0.18	0.44	0.14	0.18	0.40
$\sqrt{\sigma^2(\hat{\beta}_i^0) - \sigma^2(\hat{\beta}_i^1)}$	0.31	0.34	0.57	0.18	0.22	0.43

In Figure 2, the observed detection rates of the items are shown as a function of run lengths. Items were divided into groups with the same percentage of disclosure: five groups of three items with percentages ranging from 5% to 25%, and one group of 135 items not known by any respondents. As the quality control procedure was repeated three times for each of the two calibration sample sizes, the detection rates were calculated over three replications of three (known) or 135 (unknown) items each.

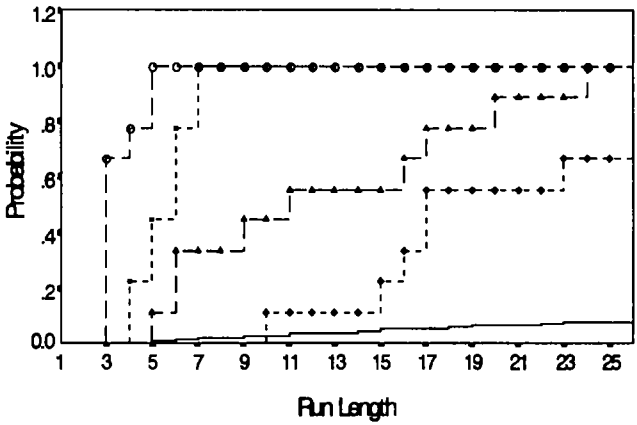
Because the results in Figure 2 are based on a small data set, the conclusions are only tentative. The following observations can be made. The items with $c_i = 0.2$ and 0.25 were usually detected soon. Items with $c_i = 0.1$ and 0.15 were usually detected, but detection was sometimes not possible within 25 runs. The low detection rate for the items with $c_i = 0.15$ in Figure 2b can be explained by the high estimate of the standard error for one of the three items. Items with $c_i = 0.05$ were usually not detected within 25 runs. The type I error after 25 runs was about 7.5%. Finally, the differences between the two item pools were rather small.

A Power Study

In this section, a simulation study of the power of the CUSUM test for the 3-PL model will be presented. As mentioned above, the power of the CUSUM procedure is governed by choosing an effect size k and a critical value h. A practical procedure to set the parameters of the CUSUM procedure may be the following. First, the practitioner must set an effect size of interest. Then, when the pretest data have become available, CAT data can be simulated using the parameter estimates of the pretest stage without assuming parameter drift. Finally, CUSUM statistics can be computed to find a value of h such that an acceptable Type I error rate is obtained.

This is illustrated in a simulation with the following design. The item bank consisted of 100 items, with difficulties equally spaced on the interval -1.00 to 1.00 and discrimination parameters drawn from a log-normal distribution with mean zero and standard deviation equal to 0.10. The guessing parameter was fixed at 0.20. Proficiency parameters were standard normally distributed throughout the study. For the pretesting phase, four groups of 250 testees each were generated. Each group responded to 50 items: the first group responded to

(a)



(b)

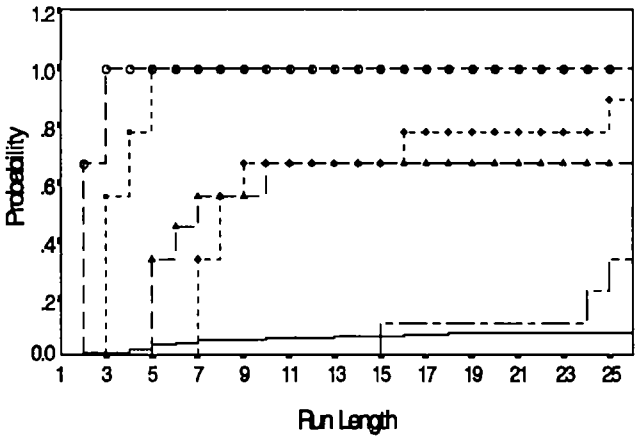


FIGURE 2. *The observed relative frequency of detecting an item as known, as a function of the run length, for the items known by 25% (long slashed line, marked by open circles), 20% (short slashes, squares), 15% (long slashes, triangles), 10% (short slashes, diamonds), 5% (long and short slashes), and 0% (solid line), of the respondents (Items were taken from two item pools that only differed in their calibration design, namely with 100 (a) and 500 (b) respondents per booklet.)*

TABLE 2
Type I error rate of CUSUM test after 5 iterations (20 replications)

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

the items 1 to 50, the second group to 26 to 75, the third group to 51 to 100, and the fourth group to the items 1 to 25 and 76 to 100. So in the pretest design every item was presented to 500 respondents. Next, four batches, $j = 1, \ldots, 4$, of 1000 CAT simulees were generated. Every testee responded to 20 items. In this study, respondent parameters were estimated by their posterior expectation using a standard normal prior, and item selection was by the maximum information criterium. Difficulty, discrimination, and proficiency distribution parameters were estimated by MML; the guessing parameter was fixed at its true value 0.20. Finally, CUSUM statistics $S_i(j), j = 1, \ldots, 4$, were computed.

This procedure was carried out for three effect sizes k and four thresholds h; the values are shown in Table 2. In the table, the percentages of items flagged in the fourth iteration ($j = 4$) 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 $h = 2.5$ resulted in flagging of 17% of the items, which is too high. A value $h = 5.0$ resulted 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 produced 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 will be further studied in the next set of simulations where model violations were introduced.

These studies pertain to a set-up similar to the previous one; however, in the present case parameter drift was imposed on every fifth item, so that 20 of the 100 items were affected. Parameter drift was imposed in six conditions: in the first three, items were known to 5%, 10%, and 20% of the respondents, respectively, and in the next three the difficulty parameter changed from the initial value by -0.20, -0.40 and -0.60, respectively. The results on the detection of these items are shown in Table 3. 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 3 give the percentages of the affected items detected by the CUSUM method. As expected, the highest percentages of detection were obtained for the smaller effect sizes $k = 0.50$ and $k = 1.00$ and the larger model violations. The

TABLE 3
Detection of parameter drift under various model violations (20 replications)

Effect Size	Model Violation	Iteration $j = 1$	$j = 2$	$j = 3$	$j = 4$
$k = 0.50$	$c_i = 0.05$	00	00	05	15
	$c_i = 0.10$	00	05	10	20
	$c_i = 0.20$	00	30	75	85
$k = 1.00$	$c_i = 0.05$	15	25	30	45
	$c_i = 0.10$	15	35	55	50
	$c_i = 0.20$	30	75	90	85
$k = 2.00$	$c_i = 0.05$	00	05	15	15
	$c_i = 0.10$	05	15	15	20
	$c_i = 0.20$	15	30	55	60
$k = 0.50$	$\beta_i - 0.20$	00	00	10	15
	$\beta_i - 0.40$	00	15	45	60
	$\beta_i - 0.60$	05	35	65	80
$k = 1.00$	$\beta_i - 0.20$	00	20	40	35
	$\beta_i - 0.40$	25	50	55	65
	$\beta_i - 0.60$	20	75	95	99
$k = 2.00$	$\beta_i - 0.20$	00	00	05	05
	$\beta_i - 0.40$	05	10	30	35
	$\beta_i - 0.60$	00	25	75	75

best detection record was obtained by the combination $k = 1.00$ and a shift in difficulty of -0.60, which, for $j = 4$, has an almost perfect detection rate of 99%.

In Table 4, the detection rates of the 80 items not affected by parameter drift are shown. The last four columns of this table give the percentage false alarms, that is, items erroneously flagged as drifting. It can be seen that the percentage of false alarms remained relatively low. The worst performances were obtained for combinations of $k = 0.50$ and $k = 2.00$ with small violations, such as item disclosure to 5% or 10% of the respondents, or a shift in difficulty of -0.20. Besides a relatively low hit-rate, these conditions show a false-alarm-rate of approximately 10%, which is relatively high.

Discussion

Recently, exposure control has received much attention in the literature about adaptive testing. The aim of exposure control is to preserve the quality of the item pool, and, consequently, of the test. Test validity decreases when items become known to the respondents in advance. Exposure control decreases the risk of items becoming known, but item exposure can never be ruled out. Therefore, in this paper, a method to detect known items is proposed. By routinely re-estimating item parameters on the basis of adaptive test data and comparing the estimates with the values found in the initial calibration, drift in the parameters of known items can be detected. The method suggested here is

TABLE 4
Type I error rate under various model violations (20 replications)

Effect Size	Model Violation	Iteration $j = 1$	$j = 2$	$j = 3$	$j = 4$
$k = 0.50$	$c_i = 0.05$	00	04	05	13
	$c_i = 0.10$	00	03	05	06
	$c_i = 0.20$	00	00	01	03
$k = 1.00$	$c_i = 0.05$	05	13	17	21
	$c_i = 0.10$	03	03	03	06
	$c_i = 0.20$	03	04	06	09
$k = 2.00$	$c_i = 0.05$	00	00	03	00
	$c_i = 0.10$	03	01	04	04
	$c_i = 0.20$	00	01	01	01
$k = 0.50$	$\beta_i - 0.20$	00	00	06	05
	$\beta_i - 0.40$	01	06	09	15
	$\beta_i - 0.60$	00	00	04	04
$k = 1.00$	$\beta_i - 0.20$	00	01	03	05
	$\beta_i - 0.40$	01	04	06	09
	$\beta_i - 0.60$	03	06	10	10
$k = 2.00$	$\beta_i - 0.20$	00	01	03	03
	$\beta_i - 0.40$	00	00	03	01
	$\beta_i - 0.60$	01	03	04	03

based in the so-called cumulative sum or CUSUM charts (see, e.g., Wetherill, 1977). With this method, any specified change in the value of the difficulty parameter can be detected. The method is illustrated with a simulated example and a power study. Results show that the detection rate of the procedure is quite acceptable and that type I error rate is well under control.

Once parameter drift is detected by the CUSUM chart method, a more thorough analysis should always follow. In the first place, the significance of the parameter change should be checked with the aid of other statistical tests. If the results affirm the first conclusion, a search for the precise cause of the parameter drift should be performed. The origin of parameter change is not necessarily previous knowledge of the respondents. For example, Bock, Muraki, and Pfeiffenberger (1988) suggest that the values of item parameters can drift when the same items are used for a long time, merely because of educational, technical, and cultural changes. Finally, it should be decided whether or not items detected can be used in any future way.

References

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275–285.

Camilli, G., & Shephard, L. A. (1994). Methods for identifying biased test items. Vol. 4 of *Measurement methods for the social sciences series*. Thousand Oaks, CA: Sage Publications.

Fischer, G. H. (1995). Some neglected problems in IRT. *Psychometrika, 60*, 459–487.

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

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

Glas, C.A.W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 69–95). New York: Springer-Verlag.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer and H. I. Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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

Little, R.J.A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M. (1983). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 51–61). New York: Academic Press.

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

Mislevy, R. J., & Wu, P. K. (1988). *Inferring examinee ability when some item responses are missing* (Research Report RR-88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J. & Wu, P. K. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing* (ETS Research Report RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 39–52). New York: Springer-Verlag.

Montgomery, D. C. (1991). *Introduction to statistical quality control* (2nd ed.). New York: John Wiley & Sons.

Samejima, F. (1984). *Results of item parameter estimation using Logist 5 on simulated data* (ONR/RR-84-3). Knoxville, Tennessee: University of Tennessee.

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, CA: Navy personnel Research and Development Center.

Verhelst, N. D., Glas, C.A.W., & Verstralen, H.H.F.M. (1994). *OPLM: Computer program and manual*. Arnhem, the Netherlands: CITO.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wetherill, G. B. (1977). *Sampling inspection and 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.

Authors

WIM J. J. VEERKAMP wrote this article as a Phd.-student at the Department of Educational Measurement and Data Analysis, Faculty of Educational Science and Technology University of Twente, P. O. Box 217, 7500AE, Enschede, the Netherlands. He is currently working as an econometrician at the Zwolse Algemene Bank, Nieuwegein, the Netherlands. He specializes in forecasting and risk-analysis.

CEES A. W. GLAS is an Associate Professor at the Department of Educational Measurement and Data Analysis, Faculty of Educational Science and Technology University of Twente, P. O. Box 217, 7500AE, Enschede, the Netherlands; glas@edte.utwente.nl. He specializes in psychometrics and educational measurement.

Received July 1998
Revision Received June 1999
Accepted January 2000