LSAC RESEARCH REPORT SERIES

The Use of Statistical Process Control-Charts for Person-Fit Analysis in Computerized Adaptive Testing

Rob R. Meijer
Edith M. L. A. van Krimpen-Stoop
University of Twente

Law School Admission Council
Computerized Testing Report 98-12
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

A person's item response pattern may reveal such undesirable behavior as faking responses on personality tests, guessing, or knowledge of the correct answers due to test preview. These undesirable behaviors may result in inappropriate test scores and may have serious consequences for practical test use, for example, they could result in decision errors in testing for job and educational selection.

In analyzing Law School Admission Test (LSAT) data operationally, item response theory (IRT) is applied. IRT is a mathematical model whereby the probability that a test taker will answer an item correctly is related to the ability level of the test taker and the characteristics of the item. To detect item response patterns that do not fit regular test taking behavior under an IRT model several person-fit statistics have been proposed. Nearly all statistics are based on the difference between the observed item scores and those expected under the IRT model. When the distribution of a statistic under the hypothesis of regular response behavior is known, item response patterns can be classified as fitting or nonfitting the model. To date, almost all fit statistics have been proposed for conventionally administered paper-and-pencil (P&P) tests. With the advent of computerized adaptive testing (CAT), research is needed with respect to the application of person-fit statistics in CAT. In earlier research, the empirical distributions under CAT of a frequently used fit statistic, l_Z and an adaptation l_Z* were studied. These authors found that for simulated P&P data the empirical distribution of l_Z* was more in agreement with the standard normal distribution than the distribution of l_Z. For CAT data, however, there was a large discrepancy between the empirical and standard normal distribution for both statistics. Consequently, inadvertent application of these person-fit tests to CAT data may result in inaccurate decisions.

In this paper, several new fit statistics especially designed for CAT were studied. These statistics were based on the theory of statistical process control (SPC). In industrial applications, a (production) process is in a state of statistical control if the variable being measured has a stable distribution. One technique from SPC is based on Shewhart control charts. These charts are used to determine if a process is in statistical control by examining past data. An example of a Shewhart control chart is the X-chart, where the observed averages of the variable being measured in a sample of size *(n)* are calculated over time. X-charts are very effective in detecting large shifts in the quality of a production process. Another technique from statistical process control is the cumulative sum procedure. This paper shows that this procedure could be successfully applied to detect test takers in a CAT whose response behavior is not in accordance with the IRT model. For example, suppose an examine becomes more and more careless during the test because of fatigue. As a result, in the first part of the CAT responses will tend to alternate between correct and incorrect reponses, whereas in the second part of the test more and more items are answered incorrectly due to carelessness. The statistics studied in this paper were designed to be sensitive to such changes in response behavior. The power of these statistics with respect to a large variety of nonfitting response behaviors was determined. Recommendations on the appropriate choice from these statistics for CAT programs were made.

Abstract

In this study a cumulative-sum or CUSUM-procedure, from the theory of Statistical Process Control was modified and applied in the context of person-fit analysis in a computerized adaptive testing (CAT) environment. Six person-fit statistics were proposed using the CUSUM-procedure; three of which can be used to investigate the CAT in online test taking. The usefulness of these statistics was explored in a small simulation study. In most CUSUM-procedures standard normally distributed statistics are used; based on this normality, boundaries can be determined to decide when a process is out-of-control. The statistics proposed in this study, however, are not standard normally distributed and therefore, in this study, boundaries were determined using simulated data. This study investigated whether the numerical values of the upper and lower threshold were dependent on the latent trait, θ. Results showed that the upper and lower bound were rather stable across θ-levels. However, the height of the boundaries was dependent on the particular statistic chosen.

Introduction

The development of group ability tests has enabled the comparison of an individual's total test score with the scores of a population norm group, thus allowing for more meaningful interpretation of the ability level. However, a person's test score may not reveal the operation of other undesirable influences of test taking behavior such as faking on biodata questionnaires and personality tests, guessing, or knowledge of the correct answers due to test preview These and other influences may result in inappropriate test scores which may have serious consequences for practical test use, for example, in job and educational selection, where classification errors may result.

To detect item response patterns that do not fit an item response theory (IRT) model several person-fit statistics have been proposed (Drasgow; Levine, & Williams, 1985; Molenaar & Hoijtink, 1990; Kianer & Rettig, 1990). In almost all statistics, for an individual person with a latent trait value θ the difference between the observed and expected item score on the basis of an IRT model is compared across items. When the distribution of a statistic is known under a null model, item score patterns can be classified as fitting or nonfitting.

To date, almost all fit statistics have been proposed in the context of conventionally administered tests or paper-and-pencil (P&P) tests. With the increasing use of computerized adaptive testing (CAT), additional research is needed with respect to the application of person-fit statistics in CAT. Recently, Meijer and van Krimpen-Stoop (1998) investigated the empirical distribution of an often used fit statistic, l_z (Drasgow et al., 1985) and an adaptation l_z* (Snijders, 1998) that correct for the use of the estimated latent trait value, $\hat{\theta}$, instead of θ in l_z. Both statistics were assumed to be standard normally distributed. Meijer and van Krimpen-Stoop (1998) found that for simulated P&P data when $\hat{\theta}$ was used instead of θ the empirical distribution of l_z* was more in agreement with the standard normal distribution than the distribution of l_z. For CAT data, however, there was a large discrepancy between the empirical and theoretical distribution for both statistics. Consequently, decisions about the fit of a score pattern on the basis of theoretical critical values will be inaccurate. As an alternative, Meijer and van Krimpen-Stoop (1998) proposed to simulate the distribution for a given $\hat{\theta}$ through parametric bootstrapping. Given a fixed θ value, P&P and CAT response vectors were generated and the distribution of the significance probabilities was determined on the basis of $\hat{\theta}$. For P&P tests the results were promising in the sense that the significance probabilities were in agreement with the expected percentages. However, for CAT and $\hat{\theta}$ the probabilities were too low, which hamper the use of these statistics in a CAT environment.

This study proposes new fit statistics especially designed for a CAT. These statistics are based on the theory of statistical process control as explained below. The use of these statistics will be explored by means of two small simulation studies.

Statistical Process Control

A (production) process is in a state of statistical control if the variable being measured has a stable distribution. The analysis of determining whether a process is in statistical control is often called statistical process control (SPC). One technique from SPC is using a Shewart control chart, originally proposed by Shewart (1931); these charts are used to determine if a process is in statistical control by examining past data. An example of a Shewart control chart is the $\overline{X}$-chart, where the observed averages ($\overline{X}$) of the variable being measured in a sample of size n are measured over time. An $\overline{X}$-chart is very effective in detecting large mean shifts in the production process. However, a disadvantage of Shewart control charts is that the chart only uses the information about the process contained in the last sample taken; the information of the entire sequence of samples is ignored. As a result, Shewart charts are rather ineffective in detecting small shifts in the process. A technique from SPC that is more effective in detecting smaller shifts in the mean is the cumulated sum (CUSUM) procedure, originally proposed by Page (1954). In a CUSUM-procedure, sums are accumulated, but a value of the statistic, obtained from a sample of size n, is only accumulated if it exceeds "the goal value" by more than k units. Suppose that Z_i is the value of statistic Z obtained from a sample at time i, k is the reference value, and h is some threshold. Then, the two-sided CUSUM procedure can be written in terms of S_i^H and S_i^L, where

$$
\begin{aligned}
S_1^H &= \max [0, Z_1 - k] \\
S_2^H &= \max [0, (Z_1 - k) + (Z_2 - k)] \\
&= \max [0, (Z_2 - k) + S_1^H] \\
S_3^H &= \max [0, (Z_3 - k) + S_2^H] \\
&\ldots\ldots \\
S_i^H &= \max [0, (Z_i - k) + S_{i-1}^H], \qquad (1)
\end{aligned}
$$

and analogously

$$S_i^L = \min[0, (Z_i + k) + S_{i-1}^L \quad (2)$$

with starting values $S_0^H = S_0^L = 0$. Note that the cumulations can be running on both sides concurrently. The sum of *consecutive positive values* of $Z_i - k$ is reflected by S_i^H and the sum of *consecutive negative values* of $Z_i + k$ by S_i^L. Thus, as soon as $|Z_i| > k$ the CUSUM-chart starts. The process is in an "out-of-control" state when $S^H > h$ or $S^L < -h$ and "in-control" otherwise. This means that, after a number of consecutive positive or negative values of the statistic, the process can become out-of-control. One assumption underlying the CUSUM procedure is that the Z_i-values computed are approximately standard normally distributed; the values of k and h are based on this assumption. The value of k is usually selected as one half of the mean shift (in Z_i-units) one wishes to detect; for example, $k = 0.5$ is the appropriate choice for detecting a shift of one times the standard deviation of Z_i.

In practice, CUSUM charts with $k = 0.5$ and $h = 4$ or $h = 5$ are often used (see for example, Montgomery, 1991, p. 295).

Computerized Adaptive Testing, Person Fit, and Statistical Process Control

IRT models describe the probability of a correct response to an item, as a function of the item and person parameters. In this study we use the two-parameter logistic IRT model (2-PLM) because it is less restrictive with respect to empirical data than the one-parameter logistic model and it does not have the estimation problems of the guessing parameter in the three-parameter logistic model (e.g., Baker, l992, pp. 109–112) The 2PLM has been shown to have a reasonable fit to achievement and personality data (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996).

Let X_{vj} be the binary (0,1) response of test taker v to item j ($j = 1, ..., J$), where 1 denotes a correct or keyed response, and 0 denotes an incorrect or not keyed response. Further, let a_j denote the item discrimination parameter, b_j the item difficulty parameter, and θ_v the latent trait value of test taker v. Then the probability of correctly answering an item according to the 2-PLM can be written as

$$P_{vj} = P_j(\theta_v) = \frac{\exp\left[a_j(\theta_v - b_j)\right]}{1 + \exp\left[a_j(\theta_v - b_j)\right]}, \quad (3)$$

For a P&P test, the administered test consists of the same items for all test takers; in a CAT, items are selected based on the responses to previous items, and an item is selected from a large pool of items to adapt to the ability of the test taker. As a result, test takers with high θ-values respond to more difficult items than test takers with low θ-values and, especially at the end of the test, the probability of a correct answer to the selected item is close to 0.5 for all test takers. An important implication of this selection process is that the response patterns of *normal* responding test takers consists of an alternation of 1 and 0 scores.

To investigate a test taker's fit to an IRT model in almost all person-fit statistics the difference between the observed and expected item score is compared across items. A general form in which most person-fit statistics in the context can be expressed is

$$W(\theta) = \sum_{j=1}^{J}\left(X_{vj} - P_j(\theta)\right)w_j(\theta) \quad (4)$$

where the statistic is of the centered form, that is, the expected value of the statistic equals 0 and $w_j(\theta)$ denotes a suitable weight: often the variance of an item score is taken into account to obtain a standardized version of the statistic. A person-fit statistic is said to be standardized when the distribution is the same across θ values.

For example, Wright and Stone (1979) proposed a person-fit statistic based on standardized residuals, where the weight

$$w_j(\theta) = \frac{1}{JP_j(\theta)\left(1 - P_j(\theta)\right)}$$

was taken resulting in

$$U(\theta) = \sum_{j=1}^{J} \frac{\left(X_{vj} - P_j(\theta)\right)^2}{JP_j(\theta)\left(1 - P_j(\theta)\right)} \tag{5}$$

U can be interpreted as the mean of the squared standardized residuals based on J items.

One of the drawbacks of U is that negative and positive residuals cannot be distinguished. In a CAT this is interesting because a string of negative or positive residuals may indicate aberrant behavior. For example, suppose a test taker with an average θ-value responds to a test and during the test the test taker becomes more and more careless because he/she becomes tired. As a result, in the first part of the test the responses will be an alternation of zeros and ones, whereas in the second part of the test more and more items are answered incorrectly due to carelessness; thus, in the second part of the test, consecutive negative residuals will occur.

The sum of *consecutive* negative or positive residuals can be investigated by slightly changing the CUSUM-procedure described above. A CAT can be viewed as a multistage test, where each item is a stage. Let i denote the stage of the CAT. Further, let the statistic T_i be a function of the residuals at stage i, m_v the final test length for test taker v, and j the item selected from an item pool Ω_n, containing n items; that is, $i = 1, ..., m_v$ and $j \in \{1, 2, ..., n\}$. Then, for test taker v per stage i of a CAT

$$S_i^H = \max\left[0, (T_i - k) + S_{i-1}^H\right], \tag{6}$$

$$S_i^L = \min\left[0, (T_i + k) + S_{i-1}^L\right], \text{ and} \tag{7}$$

$$S_0^H = S_0^L = 0, \tag{8}$$

where S^H and S^L reflect the sum of *consecutive* positive and negative residuals, respectively. Let UB and LB be some appropriate upper and lower bound, respectively, then, when $S^H > UB$ or $S^L < LB$ the response pattern can be classified as nonfitting the model.

New Person-Fit Statistics

In this section several person-fit statistics are proposed that are especially developed for a CAT and are based on the CUSUM-procedure discussed above.

In a CAT, θ_v is estimated in each stage i based on the response given up to that stage. Let $\hat{\theta}_{vi}$ denote the estimated θ_v at stage i, then on the basis of $\hat{\theta}_{vi}$, the item for the next stage, $i + 1$, is selected. Let $\hat{\theta}_v = \hat{\theta}_{v,m_v}$ denote the final estimate of θ_v, X_{vij} the binary (0, 1) response of test taker v to item j at stage i, and Δ_{vi} the set of items administered to test taker v up to and including stage i. Then, the probability of answering item j correctly, evaluated at $\hat{\theta}_{vi}$ can be written as

$$P_{vij} = P\left(X_j = 1 \mid \hat{\theta}_{vi}\right) = \frac{\exp\left[a_j\left(\hat{\theta}_{vi} - b_j\right)\right]}{1 + \exp\left[a_j\left(\hat{\theta}_{vi} - b_j\right)\right]}. \tag{9}$$

For notational convenience the index v will be dropped below.

To take the sequential nature of a CAT into account, the probability of a correct answer to the administered item j at stage i can be evaluated at $\hat{\theta}_{i-1}$; that is, $P_{i-1,j}$ where $\hat{\theta}$ from the previous stage is used. Define

$$T_{ij}^1 = X_{ij} - P_{i-1,j}, \tag{10}$$

$$T_{ij}^2 = T_{ij}^1 \times \left[P_{i-1,j}\left(1 - P_{i-1,j}\right)\right]^{-\frac{1}{2}}, \text{ and} \tag{11}$$

$$T_{ij}^3 = T_{ij}^1 \times \left[I_v^{i-1}\right]^{-\frac{1}{2}}, \tag{12}$$

where

$$I^{i} = \sum_{k \in \Delta_{vi}} I_k\left(\hat{\theta}_i, a_k, b_k\right) = \sum_{k \in \Delta_{vi}} a_k^2 P_{ik}\left(1 - P_{ik}\right). \tag{13}$$

That is, I^i is the test information function according to the 2-PLM, of a test containing the items administered up to and including stage i, evaluated at the estimated θ at stage i. Thus, T_{ij}^1 is the residual of the response of test taker v at item j at stage i and the probability of a correct response to item j at stage i, evaluated at the estimated ability at the previous stage and T_{ij}^2 and T_{ij}^3 are functions of the residuals. The statistics T^1, T^2, and T^3 are proposed to investigate the sum of consecutive positive or negative residuals, if desired in an on-line situation.

Define

$$T_{ij}^4 = X_{ij} - P_{m_v, j}, \tag{14}$$

$$T_{ij}^5 = T_{ij}^4 \times \left[P_{m_v, j}\left(1 - P_{m_v, j}\right)\right]^{-\frac{1}{2}}, \text{ and} \tag{15}$$

$$T_{ij}^6 = T_{ij}^4 \times \left[\tilde{I}^{\,i-1}\right]^{-\frac{1}{2}}, \tag{16}$$

where

$$\tilde{I}^{\,i} = \sum_{k \in \Delta_i} I_k\left(\hat{\theta}, a_k, b_k\right) = \sum_{k \in \Delta_i} a_k^2 P_{m_v, j}\left(1 - P_{m_v, j}\right). \tag{17}$$

$\tilde{I}^{\,i}$ is the test information function according to the 2-PLM, of a test containing the items administered up to and including stage i, evaluated at the final estimated θ. The statistics T^4, T^5, and T^6 are proposed to investigate the sum of consecutive negative or positive residuals, evaluated at the final estimate of θ_v. As a result of using $\hat{\theta}$ instead of $\hat{\theta}_{i-1}$, the development of the accumulated residuals can no longer be investigated in an online situation.

To determine upper and lower bounds in a CUSUM-procedure it is assumed that the statistic computed at each stage is approximately standard normally distributed. However, the distributions of T^1 through T^6 are far from standard normal; T^1 and T^4 follow a binomial distribution with only *one* observation and the other statistics are standardized versions of T^1 and T^4, and also based on only *one* observation. Therefore, setting $k = 0.5$ and the upper and lower bound to 5 and –5, respectively, might not be appropriate. As has been stated in Ryan (1989, p.140), the value of h could be adjusted to compensate for this (i.e., lack of normality) if there were guidelines for adjusting h for various nonnormal situations. Due to the absence of guidelines for correction of h for nonnormally distributed statistics, a simulation study was performed to determine upper and lower bounds for the CUSUM using T^1 through T^6.

Two Simulation Studies

Two simulation studies were conducted to get a first impression of the usefulness of the CUSUM-procedure in the context of person-fit analysis. First, in Study 1, the CUSUM procedure was applied to different types of nonfitting score patterns; second, in Study 2, it was investigated whether the numerical values of the upper and lower threshold of the CUSUM-procedure depend on the θ-level.

Study 1

Method

In this study the CUSUM-procedure described in Equations 6 through 8 was performed using three different types of simulated adaptive response vectors: one normal response vector and two nonfitting response vectors. To get an idea of the usefulness of the CUSUM-procedure the statistics T^1 and T^4 were used; these statistics are illustrative for the other statistics which are functions of T^1 and T^4. Two types of

aberrant response behavior were considered. First, guessing on all items in the test was simulated. This type of aberrant behavior is realistic when an test taker only takes a CAT to become familiar with the item content and randomly guesses the correct answer with a probability of 0.2 (assuming a test with five alternatives per item). Second, a response vector was simulated using two different θ-values to respond to the test. Examples of this aberrant behavior are sleeping, carelessness, and preknowledge about some items in the test. A response vector was simulated assuming that the simulated test taker responds to the first half of the test with θ_1 and to the second half of the test with θ_2, where the correlation between the two abilities is characterized by a parameter ρ. Here the value $\rho = 0$ was taken. A pool of 400 items fitting the 2-PLM with $a_i \sim N(1; 0.2)$ and $b_i \sim U(-3; 3)$ was used to simulate the three adaptive response vectors.

The normal response vector was simulated as follows. First, the true θ of a simulated test taker was set to a fixed θ-level, here $\theta = 2$. Then, the first item of the CAT selected was the item with maximum information given $\theta = 0$. For this item, $P(\theta)$, according to Equation 3 was determined. To simulate the answer (1 or 0), a random number y from the uniform distribution on the interval [0,1] was drawn; when $y < P(\theta)$ the response to item i was set to 1 (correct response), 0 otherwise. The first four items of the CAT were selected with maximum information for $\theta = 0$, and based on the responses to these four items, $\hat{\theta}$ was obtained using weighted maximum likelihood estimation (Warm, 1989). The next item selected was the item with maximum information given $\hat{\theta}$ at that stage. For this item, P (θ) was computed, a response was simulated, θ was estimated and another item was selected based on maximum information given $\hat{\theta}$ at that stage. This procedure was repeated until the test attained the length of 20 items.

The response vector generated by a simulated test taker guessing on all items was simulated as follows. The first item of the CAT selected was the item with maximum information given $\theta = 0$. To simulate the answer (1 or 0), a random number y from the uniform distribution on the interval [0, 1] was drawn; when $y < P(\theta) = 0.2$ the response to item i was set to 1 (correct response), 0 otherwise. The first four items of the CAT were selected with maximum information for $\theta = 0$, and based on the responses to these four items, $\hat{\theta}$ was obtained using weighted maximum likelihood estimation (Warm 1989). The next item selected was the item with maximum information given $\hat{\theta}$ at that stage. For this item, a response was simulated with $P(\theta) = 0.2$, θ was estimated and another item was selected based on maximum information given $\hat{\theta}$ at that stage. This procedure was repeated until the test attained the length of 20 items.

The simulated test taker with a two-dimensional ability parameter, with $\rho = 0$, was simulated by first drawing two values θ_1 and θ_2 from the standard normal distribution, with correlation $\rho = 0$. Then, the first item of the CAT selected was the item with maximum information given $\theta = 0$. For the responses to the first 10 items $P(\theta_1)$ was used, and for the last 10 items $P(\theta_2)$ was used.

Results

In Figures 1 and 2 the CUSUM-procedure was plotted for the normal response vector with $\theta = 2$ using statistics T^1 and T^4, respectively. The difference between T^1 and T^4 was that in T^1, $\hat{\theta}_{i-1}$ was used; that is, the estimated θ obtained from the previous stage in the CAT, and in T^4, $\hat{\theta}$ was used, that is, the final estimate. Figure 1 showed an increase in S^H during the first seven items. This can be explained by the fact that this simulated test taker had a high θ-value; in the beginning of the CAT, the simulated test taker responded to relatively easy items given his/her θ-value and as a result a sequence of positive residuals, that is, correct responses, appeared. Later on in the CAT, the items became more and more difficult; the alternation of zeros and ones was reflected by increasing S^H and decreasing S^L at the same time. Figure 2 reflected the same process, but now the increase in the first seven items was much smaller. This was because, a simulated test taker with a high θ-value will, in general, also have a high final estimate ($\hat{\theta}$); due to the use of this final estimate, the probability of a correct response to the first seven (easy) items was close to 1, and therefore, the residual became close to zero. In contrast, using $\hat{\theta}_{i-1}$ (Figure 1), resulted in a probability of a correct response close to 0.5, because here the starting point of estimating θ was set to 0.

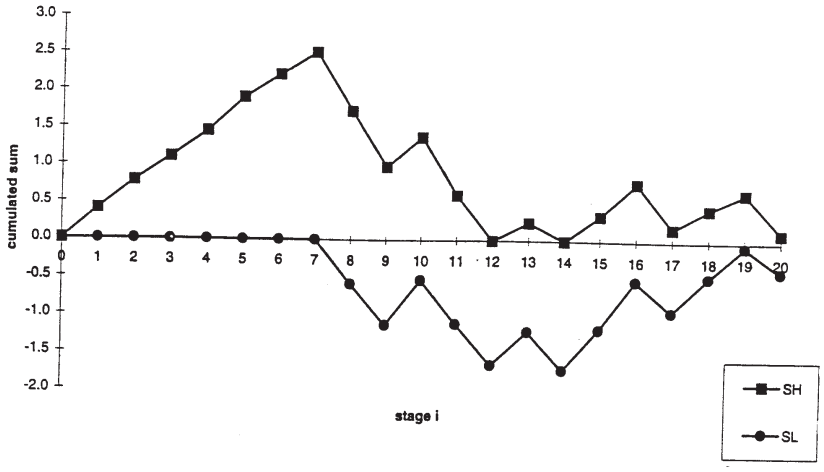

FIGURE 1. *CUSUM of a normal response vector, $\theta = 2$ using statistic T1*

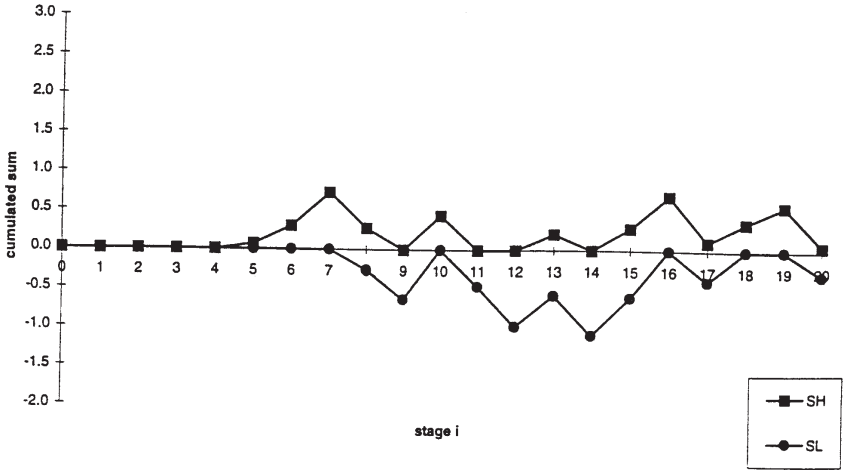

FIGURE 2. *CUSUM of a normal response vector, $\theta = 2$ using statistic T4*

In Figures 3 and 4 the CUSUM-procedure was plotted for the guessing response vector using statistic T^1 and T^4, respectively. In Figure 3 the sequence of consecutive incorrect answers was reflected by decreasing S^L. In Figure 4 the decrease of S^L was smaller, again, due to the use of $\hat{\theta}$. A guessing simulated test taker will obtain a relatively low $\hat{\theta}$, and almost all responses will be incorrect. As a result $P(\hat{\theta})$ will be close to 0 for the first items, and the residuals of the incorrect answers will become negative, but close to 0. This also explained the large increase in Figure 4 for S^H: the probability of a correct response was close to 0 for the first items and correct responses then resulted in residuals close to 1.

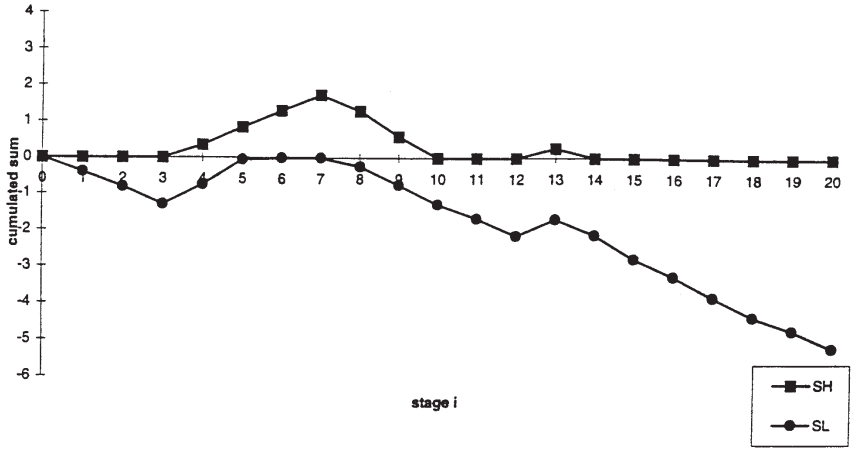

FIGURE 3. *CUSUM of a "guessing" response vector using statistic T1*

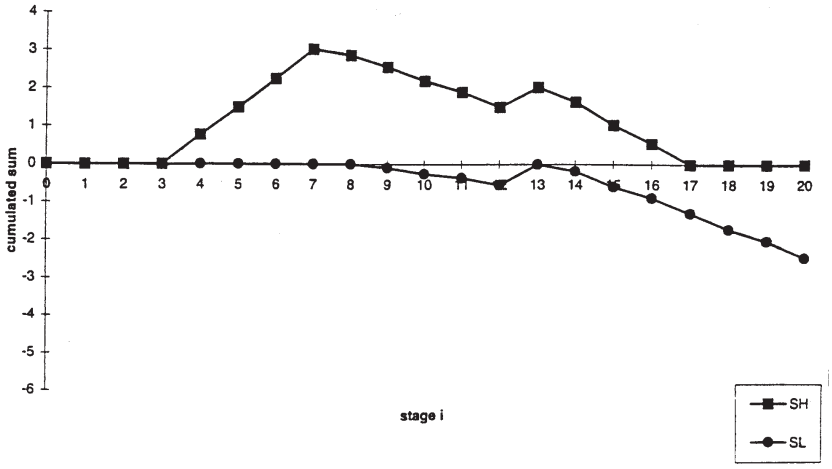

FIGURE 4. *CUSUM of a "guessing" response vector using statistic T4*

In Figures 5 and 6 the CUSUM-procedure was plotted for the response vector with a two-dimensional θ, using statistics T^1 and T^4, respectively. Figure 5 showed first an increase of S^H followed by a decrease of S^L. This phenomenon suggested a higher θ-value during the first half of the CAT than during the second half of the test. In Figure 6 the increase during the first half of the test was smaller, but still present, again due to the use of $\hat{\theta}$.

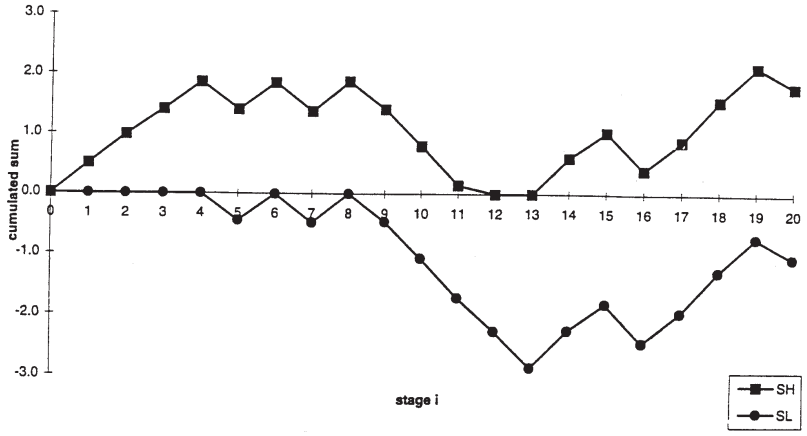

FIGURE 5. *CUSUM of a response vector with $\rho = 0$ using statistic T1.*

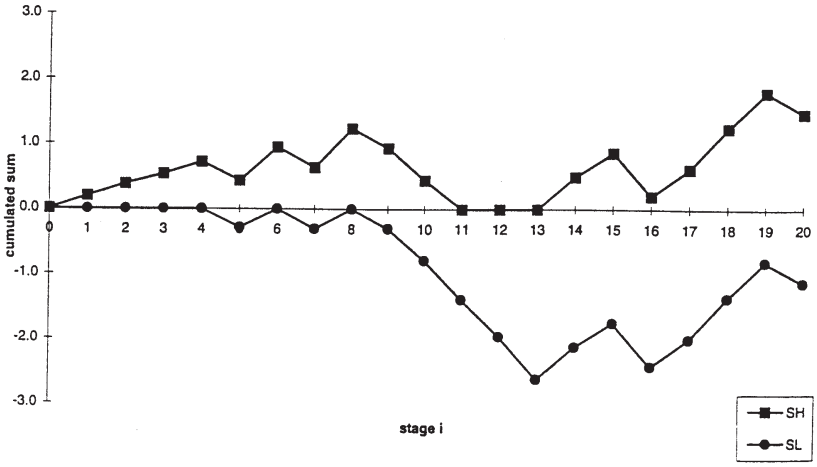

FIGURE 6. *CUSUM of a response vector with $\rho = 0$ using statistic T4*

Study 2

Method

In this simulation study it was investigated whether the numerical values of the upper and lower threshold of the CUSUM-procedure depend on the θ-level. When these boundaries are independent of θ, the results of the CUSUM-procedure are comparable across simulated test takers.

Ten datasets consisting of normal adaptive response vectors were constructed. Five datasets of 200 response vectors were constructed at five different θ-levels, and five datasets of 1,000 response vectors were constructed at five different θ-levels, where θ = -2, -1, 0, 1, and 2. Two different sizes of the datasets were used to investigate the influence of sample size on the determination of the boundaries, UB and LB, of the CUSUM-procedures. Different θ-levels were used to investigate the numerical values of UB and LB across θ-levels.

A pool of 400 items fitting the 2-PLM with $a_i \sim N(1, 0.2)$ and $b_i \sim U(-3, 3)$ was used to simulate the adaptive response vectors. An adaptive response pattern was simulated as described in Study 1, where the true θ of a simulated test taker was set to a fixed θ-level, dependent on the dataset constructed. For each simulated test taker, six different statistics, T^1 through T^6, were used in the CUSUM-procedure described in Equations 6, 7, and 8. Then, for each simulated test taker and for each statistic,

$$\max S^H = \max_i \left(S_i^H\right) \text{ and} \tag{18}$$

$$\min S^L = \min_i \left(S_i^L\right) \tag{19}$$

were determined resulting in 200 or 1,000 values of S^H and S^L for each dataset and for each statistic. Then, for each dataset, the upper threshold, UB, was determined as the value of max S^H for which 5% of the simulated test takers had higher max S^H-values and the lower threshold, LB, was determined as the value of min S^L for which 5% of the simulated test takers had lower min S^L-values. That is, a two-sided test at α = 0.10 was conducted, where P (max $S^H \geq UB$) = P (min $S^L \leq L_B$) = 0.05. In other words, for each statistic two boundaries (upper and lower bound) per dataset were determined, where 10% of the simulated test takers attained values outside these boundaries.

Results

In Figures 7 and 8 the numerical values of the upper and lower bound, UB and LB, of the CUSUM-procedure using statistics T^1 and T^4, respectively, were plotted against θ. Thus, the influence of the use of $\hat{\theta}_{i-1}$ in T^1 compared with $\hat{\theta}$ in T^4 was reflected in Figure 7 and Figure 8. Figure 7 showed that UB of the CUSUM using T^1 was approximately the same across θ-levels and across sample size. The LB of the CUSUM using T^1 was less stable across θ-levels and across sample sizes; the differences across sample sizes were largest for θ = 1 and θ = 2. For T^4 (Figure 8) the effect of the sample size on the use of statistic T^4 on both UB and LB was negligible; however, for LB, there was a slight difference across θ-levels between sample sizes. The values of UB and LB were both nearly equivalent across θ-levels. Figures 7 and 8 also showed that the UB and LB of T^1 were more stable than T^4. These figures also showed that, when $\hat{\theta}$ was used in T^4, the UB was slightly lower than when $\hat{\theta}_{i-1}$ was used in T^1, whereas LB was approximately the same.

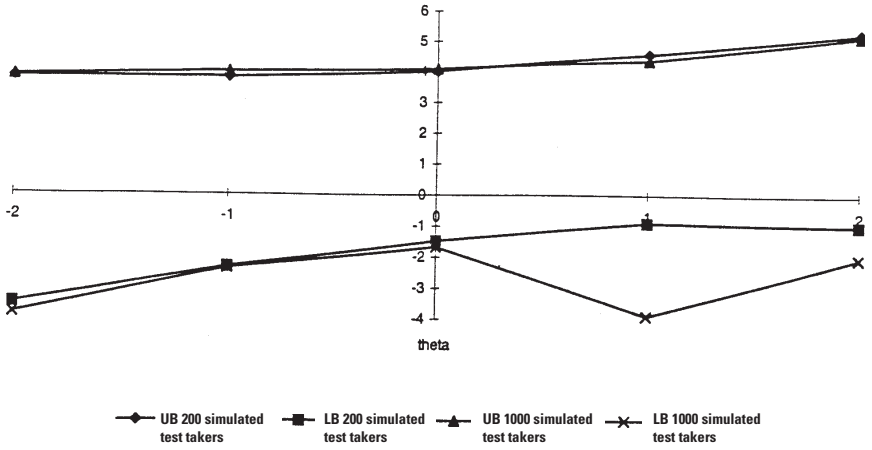

FIGURE 7. *Upper and lower bound of CUSUM for T1, where* $\alpha = 0.10$ *(two-sided)*

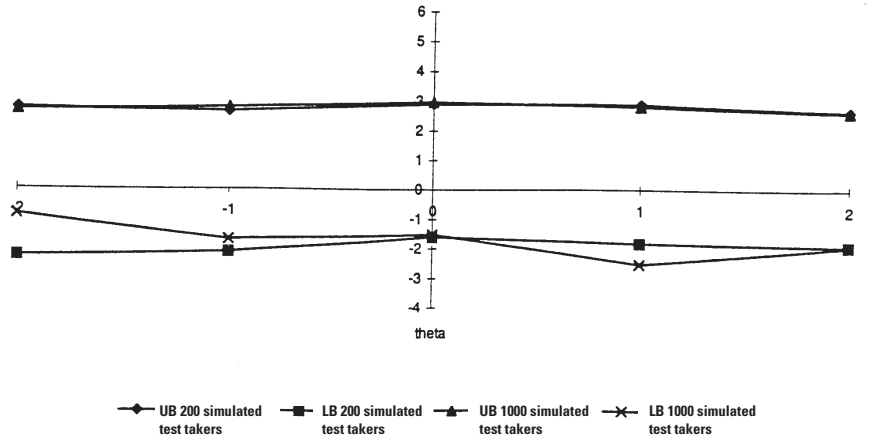

FIGURE 8. *Upper and lower bound of CUSUM for T4, where* $\alpha = 0.10$ *(two-sided)*

In Figures 9 and 10 the numerical values of the upper and lower bound, *UB* and *LB* of the CUSUM-procedure using statistics T^2 and T^5, respectively, were plotted against θ. Figure 9 showed that the values of *UB* were almost equivalent across θ-levels for both sample sizes, and that the values of *LB* were slightly different across θ-levels and across sample sizes; again, for $\theta = 1$ and $\theta = 2$ there were differences across sample sizes. Figure 10 showed that when using $\hat{\theta}$ in T^4 instead of $\hat{\theta}_{i-1}$ in T^1 the differences across sample size became smaller.

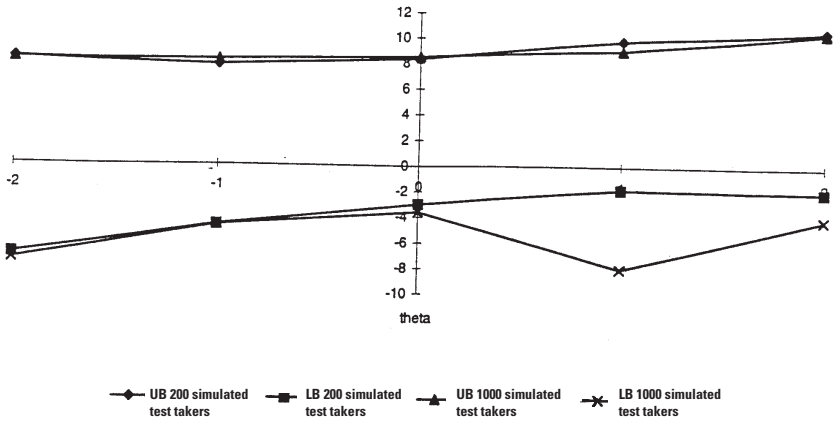

FIGURE 9. *Upper and lower bound of CUSUM for T2, where $\alpha = 0.10$ (two-sided)*

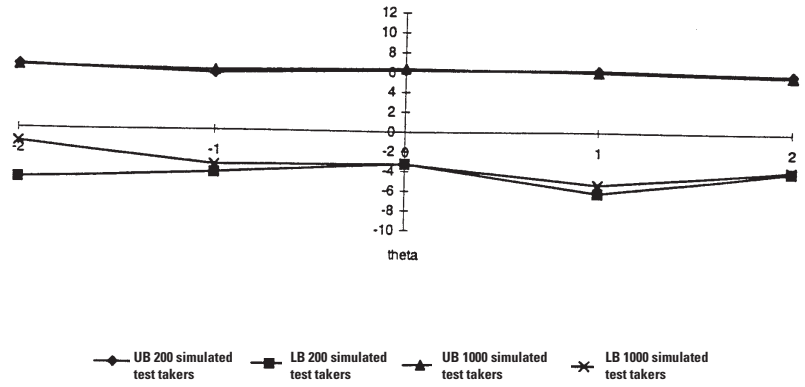

FIGURE 10. *Upper and lower bound of CUSUM for T5, where $\alpha = 0.10$ (two-sided)*

Finally, in Figures 11 and 12 the numerical values of the upper and lower bound, *UB* and *LB*, of the CUSUM-procedure using statistics T^3 and T^6, respectively, were plotted against θ. Here the use of $\hat{\theta}$ (T^6)resulted in a larger fluctuation in the numerical values of *UB* and *LB* than the use of $\hat{\theta}_{i-1}$ (T^3). Figure 11 showed that both *UB* and *LB* were nearly equivalent across different θ-levels and across sample size, whereas the *UB* and *LB* of T^6 were quite different across θ-levels. Figure 11 also showed that the influence of sample size on the distribution of *UB* was negligible. However, the distribution of *LB* was slightly different across sample sizes.

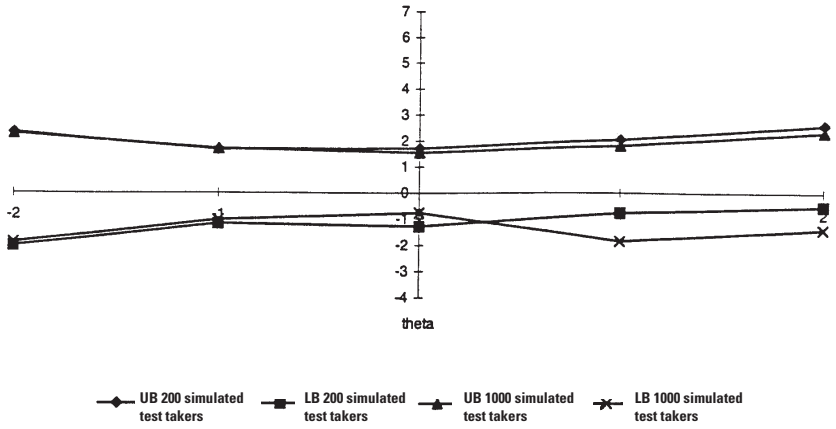

FIGURE 11. *Upper and lower bound of CUSUM for T3, where $\alpha = 0.10$ (two-sided)*

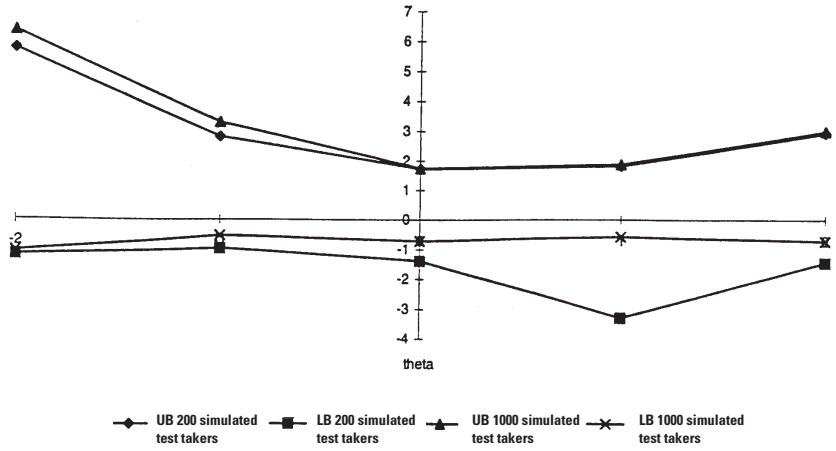

FIGURE 12. *Upper and lower bound of CUSUM for T6, where $\alpha = 0.10$ (two-sided)*

Discussion

In this study, we explored the usefulness of CUSUM-procedures in the context of person-fit analysis. Several statistics were proposed and their behavior was investigated in a small simulation study (Study 1). In Study 1 the effects of particular aberrant behavior was investigated. Results show that a CUSUM-procedure is sensitive to nonfitting response vectors and that to distinguish normal from aberrant response vectors it may be advisable to use $\hat{\theta}$ instead of $\hat{\theta}_{i-1}$ because the statistics using $\hat{\theta}_{i-1}$ are instable in the first part of the CAT. To classify a pattern as aberrant an upper and lower threshold should be determined. Therefore, in Study 2 it was investigated whether the numerical values of the upper and lower bound were dependent on θ. It was shown that, for all statistics, the upper and lower bound were rather stable across θ-levels. Also, the use of $\hat{\theta}$ and $\hat{\theta}_{i-1}$ was examined. Results showed differences in the numerical values of the upper and lower bound between statistics using $\hat{\theta}$ and $\hat{\theta}_{i-1}$.

In future research the detection rates for several different types of aberrant response behavior when CUSUM-procedures were used to classify response patterns as nonfitting will be compared with other person-fit statistics.

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Dekker.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

KIaner, K. C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*, 193–206.

Meijer, R. R. & van Krimpen-Stoop, E. M. L. A. (1998). *Simulating the null distribution of person-fit statistics for conventional and adaptive tests.* Unpublished report, University of Twente, Ensehede, The Netherlands.

Molenaar, I. W, & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55,* 75–106.

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

Page, E. S. (1954). Continuous inspection schemes. *Biometrika, 41,* 100–115.

Reise, S. P. & Waller (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45–58.

Ryan, T. P. (1989). *Statistical methods for quality improvement.* New York: Wiley.

Shewart, W. A. (1931). *Economic control of quality of manufactured product.* New York: Van Nostrand.

Snijders, T. (1998) *Asympotic distribution of person-fit statistics with estimated person parameter*. Unpublished report, University of Groningen, The Netherlands.

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

Wright, B. D., & Stone, M. H. (1997). *Best test design: Rasch Measurement*. Chicago: Mesa Press.

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 71–87.