DOCUMENT RESUME

ED 428 099 TM 029 500

AUTHOR van Krimpen-Stoop, Edith M. L. A.; Meijer, Rob R.
TITLE Person Fit Based on Statistical Process Control in an Adaptive Testing Environment. Research Report 98-13.
INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE 1998-10-00
NOTE 28p.
AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adaptive Testing; *Computer Assisted Testing; *Goodness of Fit; Simulation
IDENTIFIERS Paper and Pencil Tests; *Person Fit Measures; *Statistical Process Control

ABSTRACT

Person-fit research in the context of paper-and-pencil tests is reviewed, and some specific problems regarding person fit in the context of computerized adaptive testing (CAT) are discussed. Some new methods are proposed to investigate person fit in a CAT environment. These statistics are based on Statistical Process Control (SPC) theory. A technique from SPC that is effective in detecting small shifts in the mean of the variable being measured is the cumulative sum (CUSUM) procedure. How CUSUM is applied to CATs is outlined, and eight statistics are proposed to investigate the sum of consecutive negative or positive residuals. Two simulation studies evaluated the use of these eight statistics. Results show that the detection rates of these statistics are dependent on the type of aberrance simulated. Conditions under which the statistics can be used or should not be used are discussed. (Contains 2 tables and 26 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 099

TM

Person Fit based on Statistical Process Control in an Adaptive Testing Environment

Research Report 98-13

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Helissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

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

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

TM029500

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

ERIC

Person Fit Based on Statistical Process Control in an Adaptive Testing Environment

Edith M. L. A. van Krimpen-Stoop

Rob R. Meijer

University of Twente

ERIC Full Text Provided by ERIC

Person Fit Based on Statistical Process Control in an Adaptive Testing Environment

Introduction

The aim of computerized adaptive testing (CAT) is to construct an optimal test for an individual examinee. To achieve this, the ability of the examinee is estimated during test administration and items are selected that match the current ability estimate. This is done using an item response theory (IRT) model that is assumed to describe an examinee's response behavior. It is questionable however, whether the assumed IRT model gives a good description for each individual's test behavior. For those individuals for whom this is not the case, as a measure of the ability level the current ability estimate may be inadequate and as a result the construction of an optimal test may be difficult.

There are all sorts of causes that may invalidate an ability estimate. For example, examinees may take a CAT to familiarize themselves with the questions to be asked and randomly guess the correct answers on almost all items in the test, or examinees may have preknowledge of some of the items in the item pool and correctly answer these items independent of their trait level and the item characteristics. These types of aberrant behavior may invalidate the ability estimate and it seems therefore useful to investigate the fit of a score pattern to the test model. Research with respect to methods that provide information about the fit of a score pattern to a test model is usually referred to as appropriateness measurement or person-fit measurement. Most studies in this area are, however, in the context of paper-and-pencil (P&P) tests. As will be argued below, the application of person-fit theory presented in the context of P&P tests cannot simply be generalized to CAT.

The aim of this article is to first give an introduction to existing person-fit research in the context of P&P tests, then to discuss some specific problems regarding person fit in the context of CAT, and finally to explore some new methods to investigate person-fit in a CAT environment.

Person Fit and Paper and Pencil Tests

Three methods have been proposed to investigate the fit of an examinee to an IRT model: person-fit statistics, person-fit tests, and the person response function.

In IRT, the probability of obtaining a correct answer on item i ($i = 1, ..., n$) is explained by an examinee's latent trait value (θ) and the characteristics of the item (Hambleton

& Swaminathan, 1985). Let U_i denote the binary $(0, 1)$ response to item i, a_i the item discrimination parameter, b_i the item difficulty parameter, and c_i the item guessing parameter. The probability of correctly answering an item according to the three-parameter logistic IRT model (3PLM) is defined by

$$P(U_i = 1 \mid \theta) = P_i(\theta) = c_i + (1 - c_i) \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}.$$

When $c_i = 0$, the 3PLM becomes the two-parameter logistic IRT model (2PLM):

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \tag{1}$$

Person-Fit Statistics

Most person-fit research has been conducted using fit statistics. A general form in which most person-fit statistics can be expressed is

$$\sum_{i=1}^{n} [U_i - P_i(\theta)] w_i(\theta), \tag{2}$$

where $w_i(\theta)$ is a suitable weight (see, e.g., Snijders, 1998). Note that, as a result of dichotomous items U_i^2 equals U_i; thus, statistics of the form

$$\sum_{i=1}^{n} [U_i - P_i(\theta)]^2 v_i(\theta),$$

can be re-expressed as statistics of the form in Equation 2. The expected value of the statistic equals 0 and often the variance is taken into account to obtain a standardized version of the statistic. For example, Wright and Stone (1979) proposed a person-fit statistic based on standardized residuals, where the weight

$$w_i(\theta) = \frac{1}{nP_i(\theta)[1 - P_i(\theta)]}$$

was taken resulting in

$$V = \sum_{i=1}^{n} \frac{[U_i - P_i(\theta)]^2}{nP_i(\theta)[1 - P_i(\theta)]}. \quad (3)$$

V can be interpreted as the corrected mean of squared standardized residuals based on n items; relatively large values of V point to a deviant item score pattern.

Most studies in the literature have been conducted using some suitable function of the log likelihood function

$$l = \sum_{i=1}^{n} \{U_i \ln P_i(\theta) + [1 - U_i] \ln [1 - P_i(\theta)]\}.$$

This statistic, first proposed by Levine & Rubin (1979), was further developed by Drasgow, Levine, and Williams (1985). Two problems exist when using l as a fit statistic. The first problem is that the numerical values of l depend on the trait level. As a result, examinees may or may not be classified as aberrant depending on the trait level. A second problem is that for classifying a response pattern as aberrant, a distribution of the statistic is needed. Using a distribution, the probability of exceedance or significance probability can be determined. Because large negative values of l indicate aberrance, the significance probabilities in the left tail of the distribution are of interest.

To overcome both problems Drasgow, Levine, and Williams (1985) proposed a standardized version of l, l_z, that was less confounded with the trait level and was assumed to be standard normally distributed for long tests due to the central limit theorem

$$l_z = \frac{l - E(l)}{[\text{var}(l)]^{\frac{1}{2}}},$$

where $E(l)$ and var (l) denote the expectation and variance of l, respectively. These quantities are given by

$$E(l) = \sum_{i=1}^{n} \{P_i(\theta) \ln P_i(\theta) + [1 - P_i(\theta)] \ln [1 - P_i(\theta)]\},$$

and

$$\text{var}(l) = \sum_{i=1}^{n} P_i(\theta)\left[1 - P_i(\theta)\right]\left[\ln \frac{P_i(\theta)}{1-P_i(\theta)}\right]^2.$$

However, as argued by Molenaar and Hoijtink (1990) l_z is standard normally distributed for long tests when true θ is used, but in practice θ is replaced by its estimate $\hat{\theta}$, which effects the distribution of l_z. This was shown in Molenaar and Hoijtink (1990) and also in van Krimpen-Stoop and Meijer (in press). It both studies it was found that when maximum likelihood estimation was used to estimate θ, the variance of l_z was smaller than expected under the standard normal distribution. In particular for short tests and tests of moderate length (tests of 50 items or less) the variance was found to be seriously reduced. As a result, the statistic was very conservative in classifying score patterns as aberrant. In the context of the Rasch (1960) model, Molenaar & Hoijtink (1990) proposed three approximations to the distribution of l conditional on the total score: using (1) complete enumeration, (2) Monte Carlo simulation, and (3) a chi-square distribution, where the mean, standard deviation, and skewness of l were taken into account. These approximations are conditional on the total score which in the Rasch model is a sufficient statistic for θ. Recently, Snijders (1998) derived the asymptotic distribution for a family of person-fit statistics that are linear in the item responses and in which θ is replaced by $\hat{\theta}$.

For a review of person-fit statistics, see Meijer and Sijtsma (1995).

Person-Fit Tests

A second way to investigate the fit of a score pattern is to test the null model of normal response behavior against an alternative model of aberrant response behavior. Levine and Drasgow (1988) proposed a method for the identification of aberrant persons which was statistical optimal; that is, no other method can achieve a higher rate of detection at the same Type I error rate. They calculated a likelihood ratio statistic that provides the most powerful test for the null hypothesis that an item score pattern is normal versus the alternative hypothesis that it is aberrant. Here, the researcher in advance has to specify a model for normal behavior and a model that specifies a particular type of aberrant behavior. Klauer (1991, 1995) followed the same strategy and used uniformly most powerful tests in the context of the Rasch model to test against person-specific item discrimination, violations against local stochastic independence, and unidimensionality.

Person Response Function

A third alternative to investigate person fit is to use a person response function (PRF). A PRF gives the probability of a correct response for a person with a fixed θ as a function of b. In IRT, the item response function is assumed to be a nondecreasing function of θ, whereas the PRF is assumed to be nonincreasing in b. The fit of an examinees score pattern can be investigated by comparing the observed and expected PRF. Large differences between observed and expected PRFs may indicate nonfitting response behavior. Let n items be ordered by their b-values and let item rank numbers be assigned accordingly, such that

$$b_1 < b_2 < ... < b_n$$

Furthermore, let $\hat{\theta}$ be the maximum likelihood estimate of θ under the 3PLM. Choose n such that A_r $(r = 1, .., R)$ ordered classes can be formed, each containing m items; thus $A_1 = \{1, ..., m\}$, $A_2 = \{m+1, ..., 2m\}$, ..., $A_R = \{k-m+1, ..., k\}$. Within each class, for a particular θ value the difference of observed and expected correct scores is taken and this difference is divided by the number of items in the subtest. Using $E\left(U_i|\hat{\theta}\right) = P_i\left(\hat{\theta}\right)$, this yields

$$D_s\left(\hat{\theta}\right) = m^{-1} \sum_{i \in A_s} \left[U_i - P_i\left(\hat{\theta}\right)\right], \; all \; s = 1, ..., S.$$

Next, the Ds are added across subtests, which yields

$$D\left(\hat{\theta}\right) = \sum_{s=1}^{S} D_s\left(\hat{\theta}\right).$$

$D(\hat{\theta})$ was taken as a measure of an individual's fit to the model by Trabin and Weiss (1983). For example, if an examinee copies the answers to the most difficult items, the scores on the most difficult subtests are likely to be substantially higher than predicted by the expected PRF. Nering and Meijer (1998) compared the PRF approach with the l_z statistic using simulated data and found that the detection rate of l_z in most cases was higher than using the PRF. They suggested that the PRF and l_z can be used in a complementary way: misfitting score patterns can be detected using l_z, and differences between expected and observed PRFs can be used to retrieve more information at a subtest level. Klauer and Rettig (1990) statistically refined the methodology of Trabin and Weiss (1983).

Person Fit in Computerized Adaptive Testing

To investigate the fit of a score pattern in a CAT one obvious option is to use one of the methods proposed for P&P tests. However, research conducted with these methods in a CAT showed that this is not straightforward. Nering (1997) evaluated the first four moments of the distribution of l_z for a CAT. His results were in concordance with the results using P&P tests: the variance and the mean were smaller than expected and the distributions were negatively skewed. As a result the normal approximation in the tails of the distribution was inaccurate. Van Krimpen-Stoop and Meijer (in press) simulated the distributions of l_z and l_z^* , an adapted version of l_z in which the variance was corrected for $\hat{\theta}$ according to the theory presented in Snijders (1998). They simulated item scores with a fixed set of administered items and item scores generated according to a stochastic design, where the choice of the administered items depended on the responses to the previous items administered. Results indicated that the distribution of l_z and l_z^* differed substantially from the theoretical distribution although the item characteristics and the test length determined the severeness of the difference. Glas, Meijer, and van Krimpen (1998) adapted the person tests discussed by Klauer (1995) to the 2PLM and investigated the detection rate of these statistics in a CAT. They found very low detection rates for most simulated types of aberrant response behavior: the detection rates varied between 0.01 and 0.24 at significance level $\alpha = 0.10$ (one-sided).

A possible explanation for these results is that the characteristics of a CAT are unfavorable for the assessment of person fit using existing person-fit statistics. The first problem is that a CAT contains relatively few items compared to a P&P test. Because the detection rate of a person-fit statistic is sensitive to test length - longer tests will result in higher detection rates (e.g., Meijer, Molenaar, & Sijtsma, 1993) - the detection rate for a CAT will, in general, be lower than for P&P tests. A second problem is that almost all person-fit statistics use the spread in the item difficulties: generally speaking, aberrant response behavior consists of many 0 scores to easy items and many 1 scores to difficult items. In a CAT, the spread in the item difficulties is relatively modest: in particular at the end of the test when $\hat{\theta}$ is close to true θ, items with similar item difficulties will be selected and as a result it is difficult to distinguish normal from aberrant score patterns.

An alternative to the use of P&P person-fit statistics is to use statistics that are especially designed for CAT. Few examples of research using such statistics exist. McLeod and Lewis (1998) discussed a Bayesian approach for the detection of a specific kind of aberrant response behavior, namely for examinees with preknowledge of the items. They proposed to use a

posterior log-odds ratio as a method for detecting item preknowledge in a CAT. However, the effectiveness of this ratio to detect aberrant response patterns was unclear because of the absence of a distribution and critical values of the log-odds ratio.

Below we will explore the usefulness of new methods to investigate person-fit in CAT. We will propose new statistics that can be used in a CAT and that are based on theory from Statistical Process Control. Also, the results of simulation studies investigating the critical values and detection rates of the statistics will be presented.

Statistical Process Control

In this section theory from Statistical Process Control (SPC) is introduced.. SPC is often used to control production processes. Consider, for example, the process of producing tea bags, where each tea bag has a certain weight. Too much tea in each bag is undesirable because of the increasing costs of material (tea) for the company. On the other hand, the customers will complain when there is too few tea in the bags. Therefore, the weight of the tea bags need to be controlled during the production process. This can be done using techniques from SPC.

A (production) process is in state of statistical control if the variable being measured has a stable distribution. A technique from SPC that is effective in detecting small shifts in the mean of the variable being measured, is the cumulative sum (CUSUM) procedure, originally proposed by Page (1954). In a CUSUM procedure, sums are accumulated, but only if they exceed 'the goal value' by more than d units. Let Z_t be the value of statistic Z (e.g., the standardized mean weight of tea bags) obtained from a sample of size N at time point t. Furthermore, let d be the reference value, and h some threshold. Then, the two-sided CUSUM procedure can be written in terms of C_t^H and C_t^L, where

$$
\begin{aligned}
C_1^H &= \max\left[0, Z_1 - d\right] \\
C_2^H &= \max\left[0, (Z_1 - d) + (Z_2 - d)\right] \\
&= \max\left[0, (Z_2 - d) + C_1^H\right] \\
C_3^H &= \max\left[0, (Z_3 - d) + C_2^H\right] \\
&\cdots\cdots \\
C_t^H &= \max\left[0, (Z_t - d) + C_{t-1}^H\right],
\end{aligned}
$$

and analogously

$$C_t^L = \min\left[0, (Z_t + d) + C_{t-1}^L\right],$$

with starting values $C_0^H = C_0^L = 0$. Note that the cumulations can be running on both sides concurrently. The sum of consecutive positive values of $Z_t - d$ is reflected by C_t^H and the sum of consecutive negative values of $Z_t + d$ by C_t^L. Thus, as soon as $|Z_t| > d$ the CUSUM procedure starts. The process is 'out-of-control' when $C^H > h$ or $C^L < -h$ 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_t-values are approximately standard normally distributed; the values of d and h are based on this assumption. The value of d is usually selected as one-half of the mean shift (in Z_t-units) one wishes to detect; for example, $d = 0.5$ is the appropriate choice for detecting a shift of one times the standard deviation of Z_t. In practice, CUSUM-charts with $d = 0.5$ and $h = 4$ or $h = 5$ are often used (for a reference of the rationale see Montgomery, 1991, p.295). Setting these values for d and h results in a significance level of approximately $\alpha = 0.0027$ (two-sided). Note, that in person-fit research α is fixed and critical values are derived from the distribution of the statistic. In this study, we will also use a fixed α and derive critical values through simulation.

CAT and Statistical Process Control

CUSUM procedures investigate strings of positive and negative values of a statistic. Person-fit statistics are often defined in terms of the difference between observed and expected scores, see Equation 2. A common used statistic is V, the mean of the squared standardized residuals based on n items, defined in Equation 3. One of the drawbacks of V is that negative and positive residuals can not be distinguished which in a CAT is interesting because a string of negative or positive residuals may indicate aberrant behavior. For example, suppose an examinee with an average θ-value responds to a test and during test administration the examinee 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 incorrectly answered due to carelessness; thus, in the second part of the test, consecutive negative residuals may occur.

Sums of consecutive negative or positive residuals can be investigated using a CUSUM procedure. This can be explained as follows. A CAT can be viewed as a multistage test, where

each item is a stage and each stage can be seen as a timepoint; at each stage a response to one item is given. Let i_k denote the kth item in the CAT; that is, k is the stage of the CAT. Further, let the statistic T_k be a function of the residuals at stage k, n the final test length, and let, without loss of generality, the reference value d be equal to 0. Below, some examples of statistic T are proposed. For each examinee, at each stage k of a CAT, the CUSUM procedure can be determined as

$$C_k^H = \max\left[0, T_k + C_{k-1}^H\right], \tag{4}$$

$$C_t^L = \min\left[0, T_k + C_{k-1}^L\right], \text{ and} \tag{5}$$

$$C_0^H = C_0^L = 0, \tag{6}$$

where C^H and C^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 $C^H > UB$ or $C^L < LB$ the response pattern can be classified as nonfitting the model, otherwise, the response pattern is normal.

Some Person-Fit Statistics

Let S_k denote the set of items administered as the first k items in the CAT and $R_k = \{1, ..., I\} \setminus S_{k-1}$ the set of remaining items in the pool; from R_k the kth item in the CAT is administered. A principle of CAT is that θ is estimated at each stage k based on the responses to the previous administered items; that is, the items in set S_{k-1}. Let $\hat{\theta}_{k-1}$ denote the estimated θ at stage $k-1$ and $\hat{\theta}_n$ the final estimate of θ. Then, based on this value $\hat{\theta}_{k-1}$, the item for the next stage, k, is selected from R_k. The probability of correctly answering item i_k, according to the 2PLM, evaluated at $\hat{\theta}_{k-1}$ can be written as

$$P_{i_k}\left(\hat{\theta}_{k-1}\right) = \frac{\exp\left[a_{i_k}\left(\hat{\theta}_{k-1} - b_{i_k}\right)\right]}{1 + \exp\left[a_{i_k}\left(\hat{\theta}_{k-1} - b_{i_k}\right)\right]}. \tag{7}$$

Two sets of four statistics, all corrected for test length and based on the difference between observed and expected item scores, are proposed. The first four statistics, T^1 through T^4, are proposed to investigate the sum of consecutive positive or negative residuals in an on-line situation, when the test length of the CAT is fixed. These four statistics use as expected score the probability of correctly answering the item, evaluated at the updated ability estimate, defined

in Equation 7. The other four statistics, T^5 through T^8, use as expected score the probability of correctly answering the item, evaluated at the final ability estimate $\hat{\theta}_n$. As a result of using $\hat{\theta}_n$ instead of $\hat{\theta}_k$, the development of the accumulated residuals can no longer be investigated in an on-line situation.

All statistics are based on the general form defined in Equation 2: a particular statistic is defined by choosing a particular weight. In line with the literature on person fit in P&P tests, see for example Equation 3, two statistics are proposed where the residual $U - P(\cdot)$ is weighted by the estimated standard deviation, $[P(\cdot)(1 - P(\cdot))]^{-\frac{1}{2}}$.

In order to construct person-fit statistics that are sensitive for nonfitting behavior in the beginning part of the CAT, two statistics are proposed where the residual is weighted by reciprocal of the square root of the test information function containing the items administered up to and including stage k. This information function is a monotone increasing function of the stage of the CAT and as a result, the residuals in the beginning of the test become a larger weight than the residuals in the last part of the CAT.

Also, two statistics are proposed to detect nonfitting behavior at the end of the CAT. Here, the residuals are multiplied by the square root of the stage of the CAT, $\sqrt{k}$. Due to the increasing function $\sqrt{k}$, the residuals at the beginning of the CAT are less weighted than residuals at the later part of the CAT.

Define

$$\begin{aligned} T_k^1 &= \frac{1}{n}\left\{U_{i_k} - P_{i_k}\left(\hat{\theta}_{k-1}\right)\right\}, \\ T_k^2 &= T_k^1 \times \left\{P_{i_k}\left(\hat{\theta}_{k-1}\right)\left[1 - P_{i_k}\left(\hat{\theta}_{k-1}\right)\right]\right\}^{-\frac{1}{2}}, \\ T_k^3 &= T_k^1 \times \left\{I\left(\hat{\theta}_{k-1}\right)\right\}^{-\frac{1}{2}}, \text{ and} \\ T_k^4 &= \sqrt{k} \times T_k^1, \end{aligned}$$

where $I\left(\hat{\theta}_k\right)$ is the test information function according to the 2-PLM, of a test containing the items administered up to and including stage k, evaluated at $\hat{\theta}_k$; that is,

$$I\left(\hat{\theta}_k\right) = \sum_{i_g \in S_k} I_{i_g}\left(\hat{\theta}_k, a_{i_g}, b_{i_g}\right) = \sum_{i_g \in S_k} a_{i_g}^2 P_{i_g}\left(\hat{\theta}_k\right)\left[1 - P_{i_g}\left(\hat{\theta}_k\right)\right].$$

Thus, T_k^1 is the residual of the response and the probability of a correct response to item i_k, where the probability is evaluated at the estimated ability at the previous stage; T_k^2, T_k^3, and T_k^4

ERIC
BEST COPY AVAILABLE

are functions of these residuals. Due to the use of the updated ability estimate, the sequential nature of the CAT is taken into account.

Define

$$
\begin{aligned}
T_k^5 &= \frac{1}{n}\left\{U_{i_k} - P_{i_k}\left(\hat{\theta}_n\right)\right\}, \\
T_k^6 &= T_k^5 \times \left\{P_{i_k}\left(\hat{\theta}_n\right)\left[1 - P_{i_k}\left(\hat{\theta}_n\right)\right]\right\}^{-\frac{1}{2}}, \\
T_k^7 &= T_k^5 \times \left\{I\left(\hat{\theta}_n\right)\right\}^{-\frac{1}{2}}, \text{ and} \\
T_k^8 &= \sqrt{k} \times T_k^5,
\end{aligned}
$$

where $I\left(\hat{\theta}_n\right)$ is the test information function, of a test containing the items administered up to and including stage i, evaluated at the final estimated ability, $\hat{\theta}_n$. Thus,

$$
I\left(\hat{\theta}_n\right) = \sum_{i_g \in S_k} I_{i_g}\left(\hat{\theta}_n, a_{i_g}, b_{i_g}\right) = \sum_{i_g \in S_k} a_{i_g}^2 P_{i_g}\left(\hat{\theta}_n\right)\left[1 - P_{i_g}\left(\hat{\theta}_n\right)\right].
$$

The statistics T^5 through T^8 are proposed to investigate the sum of consecutive negative or positive residuals, evaluated at the final estimate $\hat{\theta}_n$. Due to the use of $\hat{\theta}_n$ instead of $\hat{\theta}_k$, the development of the accumulated residuals can no longer be investigated in an on-line situation.

These eight statistics can be used in the CUSUM procedure described in Equation 4 through 6. As a result of the use of the CUSUM procedures, the sum of positive and negative residuals is updated after each item response.

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^8 are far from standard normal; T^1 and T^5 follow a binomial distribution with only one observation, the other statistics are standardized versions of T^1 and T^5, also based on only one observation. As a result, setting $d = 0.5$ and the upper and lower bound to 5 and -5, respectively, might not be appropriate in this context. Therefore, in this study, the numerical values of the upper and lower bound are investigated through simulation, with the fixed values $\alpha = 0.05$ and $d = 0$.

A Simulation Study

The parametric CUSUM procedure investigates strings of correct or incorrect responses in a CAT. A drawback of the CUSUM procedure is the absence of guidelines for determination

of the upper and lower bound for nonnormally distributed statistics. Therefore, in Study 1, a simulation study was conducted to investigate the numerical values of the upper and lower threshold of the CUSUM procedures using statistics T^1 through T^8 across θ-levels. When these bounds are independent of θ, a fixed upper and lower bound for each statistic can be used. In Study 2, the detection rate of the CUSUM procedures with the statistics T^1 through T^8 for several types of nonfitting response behavior are investigated.

In this simulation study we use the 2PLM 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, 1992, pp.109-112). The 2PLM has shown to have a reasonable fit to several achievement and personality data (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996). Furthermore, in these two studies true item parameters were used. This is realistic when item parameters are estimated using large samples: Molenaar and Hoijtink (1990) found no serious differences between true and estimated item parameters for samples consisting of $1,000$ examinees or more.

Study 1

Method

Five datasets consisting of 10,000 normal adaptive response vectors each were constructed at five different θ-levels; $\theta = -2, -1, 0, 1,$ and 2. An item pool of 400 items fitting the 2PLM with $a_i \sim N(1; 0.2)$ and $b_i \sim U(-3; 3)$ was used to generate the adaptive response vectors.

The normal response vector was simulated as follows. First, the true θ of a simulee was set to a fixed θ-level. 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 1 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, $\widehat{\theta}$ was obtained using weighted maximum likelihood estimation (Warm, 1989). The next item selected was the item with maximum information given $\widehat{\theta}$ at that stage. For this item, $P(\theta)$ was computed, a response was simulated, θ was estimated and another item was selected based on maximum information given $\widehat{\theta}$ at that stage. This procedure was repeated until the test attained the length of 30 items.

For each simulee, eight different statistics, T^1 through T^8, were used in the CUSUM

procedure described in Equations 4, 5, and 6. Then, for each simulee and for each statistic,

$$\begin{aligned} \max C^H &= \max_k \left(C_k^H\right) \text{ and} \\ \min C^L &= \min_k \left(C_k^L\right) \end{aligned}$$

were determined resulting in 10,000 values of C^H and C^L for each dataset and for each statistic. Then, for each dataset and each statistic, the upper bound, UB, was determined as the value of $\max C^H$ for which 2.5% of the simulees had higher $\max C^H$-values and the lower bound, LB, was determined as the value of $\min C^L$ for which 2.5% of the simulees had lower $\min C^L$-values. That is, a two-sided test at $\alpha \leq 0.05$ was conducted, where $P\left(\max C^H \geq UB\right) = P\left(\min C^L \leq LB\right) = 0.025$. In other words, for each statistic two bounds (upper and lower bound) per dataset were determined, where 5% of the simulees attained values outside these bounds.

When the bounds are stable across θ, it becomes possible to use one fixed upper and one fixed lower bound for each statistic. Therefore, to construct fixed bounds, the weighted average of the upper and lower bound were calculated across θ for each statistic, with different weights for different θ-values; weights 0.05, 0.2, and 0.5, for $\theta = \pm 2$, ± 1, and 0, respectively were used. These weights represent a realistic population distribution of abilities.

Results

In Table 1 the upper and lower bounds, at $\alpha \leq 0.05$ (two-sided), of statistics T^1 through T^8 are tabulated at five different θ-levels. Table 1 shows that, for most statistics the upper and lower bounds were stable across θ-levels. For statistic T^7, the bounds were less stable across θ-values. Table 1 also shows that, for most statistics except T^4 and T^8, the weighted average bounds were approximately symmetrical around 0.

As a result of the stable bounds for almost all statistics across θ, one fixed upper and lower bound can be taken as bounds for the CUSUM procedures.

Study 2

Method

To examine the detection rates for the eight proposed statistics, different types of nonfitting response behavior were simulated. Aberrant item scores were simulated for all items, and for the first or second part of the response pattern. Six datasets containing 1,000 nonfitting

adaptive response patterns were constructed; an item pool of 400 items with $a_i \sim N(1.0; 0.2)$ and $b_i \sim U(-3; 3)$ was used. The detection rate was defined as the proportion of nonfitting response patterns that were classified as aberrant. For each response vector, CUSUM procedures using T^1 through T^8 were performed; a response vector was classified as nonfitting when

$$\begin{aligned} \max C^H(T^q) &> UB(T^q) \text{ or} \\ \min C^L(T^q) &< LB(T^q) \end{aligned}$$

for $q = 1, ...8$. The upper and lower bounds for each statistic were set to the weighted average of the values presented in Table 1. To facilitate comparisons, a dataset containing 1,000 model fitting adaptive response patterns was constructed and the percentage of normal response vectors classified as aberrant was determined.

Types of aberrant response behavior

Random response behavior.

To investigate nonfitting behavior to all items of the test, random response behavior to all items was simulated. This type of response behavior may be the result of guessing the answers to the items of a test and was empirically studied by Van den Brink (1977). He described persons who took a multiple-choice test only to familiarize themselves with the questions that would be asked. Because returning an almost completely blank answering sheet may focus a teacher's attention on the ignorance of the examinee, each examinee randomly guessed the correct answers on almost all items of the test. "Guessing" simulees were assumed to answer the items by randomly guessing the correct answers on each of the 30 items in the test with a probability of 0.2. This probability corresponds to the probability of obtaining the correct answer by guessing in a multiple-choice test with five alternatives per item.

Non-invariant ability.

To investigate nonfitting response behavior in the first or in the second half of the CAT, response vectors with a two-dimensional θ were simulated (Klauer, 1991). It was assumed that during the first half of the test an examinee had another θ value than during the second half. Carelessness, fumbling, or memorization of some items can be the cause of non-invariant abilities. Two datasets containing response vectors with a two-dimensional θ were simulated by drawing two ability values, θ_1 and θ_2, from a bivariate standard normal distribution; the correlation between the two values was modeled by the parameter ρ. Thus, during the first half of the test $P(\theta_1)$ was used and during the second half $P(\theta_2)$ was used to simulate the responses

to the items. The values $\rho = 0$ and $\rho = 0.5$ were used here to simulate the response patterns.

Violations of local stochastic independence.

To examine aberrance in second part of the CAT, response vectors with violations of local stochastic independence between all items in the test were simulated. When previous items provide new insights that are useful for answering the next item, or when the process of answering the items is exhausting, the assumption of local independence between the items may be violated. A generalization of a model proposed by Jannarone (1986) (see Glas, Meijer, and van Krimpen, 1998) was used to simulated response vectors with local independence between all subsequent items

$$P\left(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta\right) \propto \exp\left[\sum_{j=i}^{i+1} x_j a_j \left(\theta - b_j\right) + x_i x_{i+1} \delta_{i,i+1}\right],$$

where $\delta_{i,i+1}$ is a parameter modeling association between items (see Glas et al., 1998, for more details). Using this model, the probability of correctly answering an item is now determined by the item parameters a and b, the person parameter θ and the association parameter δ. When $\delta = 0$ the model equals the 2PLM. Compared to the 2PLM, positive values of δ result in a higher probability of a correct response (e.g., learning-effect), and negative values of δ result in a lower probability of correctly answering an item (e.g., carelessness). The values $\delta = -2$, -1, 1, and 2 were used to simulate nonfitting response patterns.

Results

In Table 2 the detection rates for the eight different CUSUM procedures are given for several types of nonfitting response behavior and normal response behavior. For the dataset of normal response patterns and for most statistics the percentage of simulees classified as nonfitting was around the expected 0.05. For T^3, however, the percentage of simulees classified as nonfitting was 0.13.

Furthermore, the detection rates for guessing simulees were high for most statistics, except for T^5 and T^8; for example, the detection rate was 0.19 for T^5, whereas for T^7 it was 0.97.

For most statistics the detection rates for non-invariant abilities were lower than for guessing. For $\rho = 0$, the detection rates were approximately 0.30 for most statistics except for T^3 (0.13) and T^7 (0.13). For $\rho = 0.5$ the detection rates were approximately 0.20 for most statistics, except for T^3 and T^7 (0.11 for both).

Furthermore, Table 2 shows that, for violations of local independence, highest detection

rates occurred for $\delta = 2$. For $\delta = 1$ the detection rates were approximately 0.10 for all statistics, whereas for $\delta = 2$ the detection rates were approximately 0.30 for most statistics, except for T^1 (0.22) and T^7 (0.18).

Discussion

The results of Study 1 showed, that the bounds of the CUSUM procedures were rather stable across θ-values for all statistics except T^7. As a result, for most statistics, one fixed UB and LB can be used as threshold for the CUSUM procedures. In Study 2 these fixed bounds were used to determine the detection rates for the eight CUSUM procedures. Results showed that using statistic T^3 the percentage of normal response patterns as nonfitting was larger than the expected 0.05; T^3 resulted in a high percentage of normal response patterns classified as nonfitting, thus, the CUSUM procedure using statistic T^3 might result in a liberal classification of nonfitting response behavior.

Because there are few other studies using person fit in a CAT, it is difficult to compare the detection rates with other studies. An alternative is to compare the results with results from studies using P&P data. For example, despite differences in simulating the data, the results of this study were similar to the results of Zickar and Drasgow (1996). In the Zickar and Drasgow (1996) study, real data were used where some examinees were instructed to distort their own responses; detection rates between 0.01 and 0.32 were found for P&P data.

This study showed that the detection rates were dependent on the type of aberrance simulated. It was shown that most statistics were sensitive to random response behavior to all items in the test. It was also shown that most statistics were sensitive to response behavior when simulees used a two-dimensional θ. None of the proposed statistics was shown to be sensitive to local dependence between all items in the test when the probability of a correct response decreased during the test (negative values of δ); an example of this behavior is an examinee who becomes more and more careless along the test because he or she becomes tired. However, most statistics were shown to be sensitive to violations of local independence when the probability of a correct response increased during the test (positive values of δ); for example, when previous items provide new insights that are useful for answering the next item.

Detection rates were low for some types of nonfitting response behavior. Testing the null model of normal response behavior against an alternative model of specific nonfitting response behavior may be a suitable alternative.

The proposed CUSUM-procedure using statistics T^1 through T^4 can be used in an on-line

situation where during administration the statistics can be computed. One application of person fit in an on-line situation is to take action when an examinee is getting 'out-of-control'. Then, items from a different item pool containing new items can be selected to prevent inflated test scores due to item preknowledge.

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.

Glas, C. A. W., Meijer R. R. & van Krimpen, E. M. L. A. (1998). *Statistical tests for person misfit in computerized adaptive testing.* Research report 98-01, University of Twente, Enschede.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 535-547.

Klauer, K. C. (1995). The assessment of person fit. In: G. H. Fischer & I. W. Molenaar. *Rasch models, foundations, recent developments, and applications*, 97-110.

Klauer, 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.

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53*, 161-176.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1993). Item, test, person and group characteristics and their influence on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111-120.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant items score patterns: a review and new developments. *Applied Measurement in Education, 8*, 261-272.

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.

Nering, M. L. (1997). The distribution of indexes of person-fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115-127.

Nering, M. L., & Meijer, R. R. (in press). A comparison of the person response function and the lz statistic to person-fit measurement. *Applied Psychological Measurement.*

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

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests.* Copenhagen: Nielsen & Lydiche.

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

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

Trabin, T. E., & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D.J. Weiss (Ed.), *New horizons in testing.* New York: Academic Press.

van den Brink, W. P. (1977). Het verken-effect [The scouting effect]. *Tijdschrift voor Onderwijsresearch, 2*, 253-261.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (in press). Simulating the null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement.*

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

Wright, B. D. & Stone, M. H. (1979). *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.

Author Note

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the authors and not necessarily reflect the position or policy of LSAC.

Table 1 Bounds of CUSUM using statistics T^1 through T^8.

		T^1		T^2		T^3		T^4	
θ	weights	LB	UB	LB	UB	LB	UB	LB	UB
-2	.05	-.23	.19	-.47	.40	-.12	.09	-.13	1.81
-1	.20	-.20	.19	-.42	.40	-.08	.07	-.13	1.83
0	.50	-.20	.20	-.41	.42	-.07	.07	-.13	1.86
1	.20	-.20	.20	-.41	.43	-.07	.09	-.13	1.86
2	.05	-.18	.23	-.41	.47	-.09	.11	-.13	2.02
	average	-.20	.20	-.41	.42	-.07	.07	-.13	1.86
		T^5		T^6		T^7		T^8	
θ	weights	LB	UB	LB	UB	LB	UB	LB	UB
-2	.05	-.13	.13	-.27	.29	-.11	.30	-.10	1.72
-1	.20	-.13	.13	-.28	.28	-.07	.13	-.10	1.73
0	.50	-.13	.14	-.29	.29	-.07	.06	-.11	1.76
1	.20	-.13	.13	-.28	.28	-.12	.07	-.10	1.73
2	.05	-.13	.13	-.29	.27	-.28	.11	-.10	1.85
	average	-.13	.13	-.28	.28	-.09	.09	-.10	1.75

Table 2 Detection rates of CUSUM procedures.

		T^1	T^2	T^3	T^4	T^5	T^6	T^7	T^8
normal		.05	.06	.13	.04	.04	.04	.07	.04
guessing		.66	.72	.89	.59	.19	.59	.97	.21
$\rho =$	0	.31	.34	.13	.28	.33	.33	.12	.28
	0.5	.20	.23	.11	.16	.18	.20	.11	.16
$\delta =$	-2	.03	.05	.11	.01	.00	.01	.12	.00
	-1	.03	.04	.11	.01	.01	.01	.10	.01
	1	.10	.13	.19	.11	.11	.12	.11	.12
	2	.22	.27	.33	.28	.29	.32	.18	.34

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-98-13	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*
RR-98-12	W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*
RR-98-11	W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*
RR-98-10	W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*
RR-98-09	B.P. Veldkamp, *Multiple Objective Test Assembly Problems*
RR-98-08	B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*
RR-98-07	W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*
RR-98-06	W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*
RR-98-05	W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*
RR-98-04	C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*
RR-98-03	C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*
RR-98-02	R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*
RR-98-01	C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*
RR-97-07	H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*
RR-97-06	H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*
RR-97-05	W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*
RR-97-04	W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

BEST COPY AVAILABLE

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of ~~Educational Science~~ and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM029500

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").