LSAC RESEARCH REPORT SERIES

Some New Methods to Detect Person Fit in CAT

Rob R. Meijer
Edith M. L. A. van Krimpen-Stoop
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 99-03
March 1999

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

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 author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

The purpose of person-fit analysis is to detect persons with response patterns that do not fit the expectations from a reasonable model of response behavior. The analysis may help to reveal the operation of such undesirable influences on test takers' behavior as guessing or knowledge of correct answers due to test preview. The occurrence of misfitting response patterns may result in inappropriate test scores and, thus, involve serious consequences for test use, for example, a high volume of classification errors in educational and job selection.

To detect response patterns that do not fit a test model, several person-fit statistics have been proposed. Nearly all statistics are a mathematical function of the differences between the observed and expected item scores compared across items for a single examinee. If the distribution of the person-fit statistic is known, a statistical test can be used to classify response patterns as fitting or nonfitting.

To date, most fit statistics were proposed for use with conventionally administered paper-and-pencil (P&P) tests. With the increasing use of computerized adaptive testing (CAT), additional research is needed to develop person-fit statistics for use in CAT. In an earlier project, several existing person-fit statistics for P&P tests were studied in a CAT environment. Results showed that the use of these person-fit statistics was problematic because their empirical distributions were not in agreement with the theoretical distributions. The reason for this discrepancy is that CATs are typically much shorter than P&P tests and have items that are selected in an adaptive mode.

In the current project, eight new statistics based on cumulative-sum (CUSUM) procedures from Statistical Process Control theory are proposed. Four of these statistics were developed specifically to analyze person-fit in a CAT environment. The power of these statistics was explored in a large simulation study. With the original CUSUM procedures, normally distributed statistics are assumed. From this assumption, boundaries can be determined to decide when a process is out of control. In the current study, the statistics were not assumed to be normally distributed, but their boundaries were determined using simulated data. As it appeared, the boundaries were stable across the ability levels of the examinees. They can, therefore, be used safely in a large variety of applications. The results also showed that the statistics perform well and have detection rates comparable to those of traditional person-fit statistics for P&P tests.

Abstract

Person fit is concerned with detecting nonfitting item-score patterns. Most person-fit statistics have been proposed in the context of conventionally administered tests or paper-and-pencil (P&P) tests. In this study, we will first review some existing person-fit studies in a computerized adaptive testing (CAT) context and then investigate the usefulness of some new fit statistics that are based on the specific characteristics of a CAT. Both the use of statistical process control and the use of nonparametric tests is explored. The results of a simulation study to detect nonfitting response patterns in a CAT showed that the detection rate of these statistics is comparable to the detection rate of person-fit statistics in P&P tests.

Introduction

In the context of item-response theory (IRT) modeling, several methods have been proposed to detect item-score patterns that are not in agreement with the item score pattern expected based on a particular test model. These item-score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level (θ). This area of research is commonly referred to as person-fit, and the majority of the research on person fit has concentrated on the development of statistics that can be used to identify nonfitting response vectors. These statistics are based on either a likelihood approach or a residual-based approach in which the difference between observed and expected item scores is evaluated. Examples can be found in Drasgow, Levine, and Williams (1985); Molenaar and Hoijtink (1990); and Klauer and Rettig (1990). In almost all of these person-fit statistics, for an individual person with a latent trait value θ, the difference between the observed and expected item scores 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.

Most fit statistics have been proposed in the context of conventionally administered tests or paper-and-pencil (P&P) tests. Meijer and van Krimpen-Stoop (2003) proposed the use of statistical process control to detect nonfitting response patterns in computerized adaptive testing (CAT). In this study, we will investigate the use of some fit statistics based on statistical process control and on nonparametric tests. This study is organized as follows: First, the principles of person fit in an IRT context is discussed. Next, we will review some existing person-fit studies in a CAT context. Then, we will propose some new fit statistics that are based on the specific characteristics of a CAT. Finally, we will present the results of simulation studies that investigate the characteristics of the proposed person-fit statistics.

Item Response Theory and Person-Fit Research in Paper-and-Pencil Tests

IRT models describe the probability of a correct response to an item as a function of the item and person parameters. Let U_{ij} be the binary (0, 1) response of examinee j (j = 1, ..., J) to item i (i = 1, ..., I), where 1 denotes a correct or keyed response, and 0 denotes an incorrect or not keyed response. Further, let a_i denote the item discrimination parameter, b_i the item difficulty parameter, c_i the item guessing parameter, and θ the latent trait value. The probability of correctly answering an item according to the three-parameter logistic IRT model (3PLM) can be written as

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

When parameter c_i = 0, the 3PLM becomes the two-parameter logistic IRT model (2PLM); the probability of a correct response to an item, according to the 2PLM, is defined by

$$P_i(\theta) = \frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}. \tag{2}$$

In this 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 3PLM (e.g., Baker, 1992, pp.109–112). The 2PLM has been shown to have a reasonable fit to several types of achievement and personality data (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996).

A P&P test consists of the same items for all examinees. To investigate an examinee'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 IRT context can be expressed is (see Snijders, 1998)

$$Q(\theta) = \sum_{i=1}^{I} (U_{ij} - P_i(\theta))\, w_i(\theta) \tag{3}$$

where the statistic is of the centered form; that is, the expected value of the statistic equals 0 and $w_i(\theta)$ 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

$$w_i(\theta) = \frac{1}{IP_i(\theta)(1 - P_i(\theta))} \tag{4}$$

was taken resulting in

$$V(\theta) = \sum_{i=1}^{I} \frac{(U_{ij} - P_i(\theta))^2}{IP_i(\theta)(1 - P_i(\theta))}. \tag{5}$$

V can be interpreted as the mean of the squared standardized residuals based on I items.

Person-Fit Research in CAT

In contrast to P&P tests, little research has been done with respect to person fit in CAT. In a CAT, items are selected from a large pool of items to adapt to the ability of the examinee, and items are selected based on the responses to previous items. Often, items are selected for which the probability of correctly answering an item is close to 0.5 because the information contained in that item is higher compared with other items. As a result, examinees with high θ-values respond to more difficult items than examinees 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 examinees. An important implication of this selection process is that the response patterns of normally responding examinees consists of an alternation of 1 and 0 scores.

van Krimpen-Stoop and Meijer (1999) Study

van Krimpen-Stoop and Meijer (1999) investigated the empirical distribution of an often-used fit statistic in the context of P&P tests, l_z (Drasgow et al., 1985) and an adaptation l_z^* (Snijders, 1998) that corrects for the use of the estimated latent trait value $\hat{\theta}$ instead of true θ in l_z. Both statistics were assumed to be standard normally distributed. van Krimpen-Stoop and Meijer 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 were inaccurate. As an alternative, they proposed to simulate the asymptotic sampling 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 in the tails of the distribution were too low, which hamper the use of these statistics in a CAT environment.

Bradlow and Weiss (1997) Study

Bradlow and Weiss (1997) conducted a study in which several classes of statistics were introduced to identify nonfitting response patterns in CAT. One class of statistics was based on the (nonparametric) theory of runs and another class was based on the differences between observed and expected item scores. Bradlow and Weiss (1997) estimated the significance probabilities by simulating the distribution of the statistics using four different methods. Let g_{obs} denote the observed value of statistic g and $v(g)$ denote an appropriate distribution of g. The significance probability was determined as

$$\int_{g_{obs}}^{+\infty} v(g)dg,$$

assuming, without loss of generality, that larger values of g_{obs} are more aberrant. Four different distributions $v_s(g)$, for $s \in \{1, 2, 3, 4\}$ were simulated to estimate the significance probability: (1) empirical distribution computed across a population of examinees, (2) prior predictive distribution, (3) posterior predictive distribution, and (4) asymptotic (sampling) distribution. The empirical significance probability was determined by categorizing the population of examinees into regions of examinees who had taken a test with the same test length, and comparing examinees to those in the same class of test length.

Let $\omega = (\theta, \hat{\phi})$ be the vector of the unknown person parameter θ and the known item parameters ϕ (estimated from large samples); U_{obs} is the observed response pattern, and $f(U \mid \omega)$ is the probability distribution of the response pattern according to an IRT model, conditional on ω. For the prior predictive distribution,

$$v_2(U) = \int f(U \mid \omega)p(\omega)d\omega \tag{6}$$

was used, with a standard normal prior for θ (note that $p(\omega) = p(\theta)$ due to the use of known item parameters). For each response pattern U, a value of the statistic g was computed; these values of g constituted the prior predictive distribution of g.

As posterior predictive distribution,

$$v_3(U) = \int f(U \mid \omega)p(\omega \mid U_{obs})d\omega \tag{7}$$

was used. Again, for each simulated response vector, the value of g was computed; the values of g constituted the posterior predictive distribution of g.

The asymptotic sampling distribution was taken as

$$v_4(U) = f(U \mid \hat{\omega}), \tag{8}$$

where $\hat{\omega}$ was the maximum likelihood estimator of ω based on $f(U \mid \omega)$.

The following simulation algorithm was used to compute the integrals. First, a value of ω was drawn from the appropriate density function; that is, for the prior predictive distribution, the value of ω was drawn from $p(\omega)$, the standard normal prior for θ and known item parameters. For the posterior predictive distribution, θ was drawn from $p(\omega \mid \mathbf{U}_{obs})$. For the asymptotic sampling distribution, θ was set to the maximum likelihood estimate obtained from the original response pattern. Second, based on the value of θ drawn in the first step, a response pattern was replicated according to the assumed IRT model. And third,

for each replicated response vector, the value of statistic g was computed. In total, 99 response vectors were replicated. The 99 values of g constituted the relevant simulated distribution $v_s(g)$. On the basis of this distribution, the significance probability was determined. In replicating the response patterns, the stochastic CAT item selection procedure was not taken into account; as a result, the replicated response vectors contained the same items in the same order as the original response pattern.

The statistics were applied to 100 examinees randomly selected from the 1995 National Council Licensure Examination (NCLEX, 1995) CAT item pool; the Rasch model was used to describe the assumed IRT model on the NCLEX. For each examinee, the values of the statistics were calculated and the significance probabilities were estimated according to the four simulation methods described above. An examinee was classified as nonfitting at significance level $\alpha = 0.05$. The results showed that, for a fixed response pattern, the significance probabilities were different for each statistic and for each simulation method. The results also showed that, in general, using the prior predictive distribution classified more examinees as nonfitting than using the posterior predictive distribution, which classified about the same number of examinees as nonfitting as the sampling distribution. The number of examinees classified as aberrant using the length of the longest run (*LLR*) was low, whereas the number of examinees classified as aberrant for the total number of runs (*TNR*) was higher: 0.01 and 0.19, respectively, using the prior predictive distribution. The number of examinees classified as aberrant using the statistics based on the differences between observed and expected scores varied between 0.15 and 0.34, using the prior predictive distribution.

McLeod and Lewis (1998) Study

McLeod and Lewis (1998) examined whether examinees were successful in attaining higher test scores in a CAT, when they had preknowledge of some of the items that were used. They used a Bayesian approach where the nonfitting behavior of item preknowledge was modeled by a modified 3PLM, where the probability of a correct response was a combination of the probability of obtaining a correct response based on preknowledge and the probability of a correct response based on the ability of the examinee. An assumption of the model was that the probability of a correct response when the item was memorized was equal to one. Let $p(m_i)$ denote the probability that item i was memorized. Furthermore, let δ denote the state that the examinee is using item preknowledge, that is, the examinee memorized at least one of the items in the pool. For modeling item preknowledge, McLeod and Lewis (1998) used the probability of correctly answering an item, given that the examinee is using item preknowledge as

$$P(U_i = 1 \mid \theta, \delta) = p(m_i) + (1 - p(m_i))\left(c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]} \right) \tag{9}$$

$$= c_i^* + (1 - c_i^*)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{10}$$

where

$$c_i^* = p(m_i) + c_i - c_i p(m_i). \tag{11}$$

Let $P_\delta(U_i) = P(U_i = 1 \mid \theta, \delta)^{u_1}\ P(U_i = 0 \mid \theta, \delta)^{1 - u_1}$. The probability that the examinee is using item preknowledge, the prior $p(\delta)$, was updated after each item response. Let $p_0(\delta)$ denote the initial probability that an examinee is using item preknowledge. The posterior probability that an examinee is using item preknowledge after n items was updated after the response to the previous items, and can be written as

$$p_n(\delta) \propto \int P_\delta(U_n) p_{n-1}(\theta, \delta) d\theta. \tag{12}$$

McLeod and Lewis (1998) used the odds ratio index

$$\frac{p_n(\delta)/(1 - p_n(\delta))}{p_0(\delta)/(1 - p_0(\delta))} \tag{13}$$

to identify examinees using item preknowledge.

Results showed that examinees were successful in attaining higher test scores using item preknowledge. McLeod and Lewis (1998) suggest using the final odds ratio as an index to identify examinees who use item

preknowledge. However, criteria to classify a person as having item preknowledge were not yet established at the time of their study. Furthermore, note that their method can only be used to detect examinees with item preknowledge.

Purpose of the Study

The present study was designed to explore several alternative methods of detecting nonfitting response patterns. In this study, statistics are proposed that are updated after each item response, using theory from Statistical Process Control (SPC). Furthermore, statistics are proposed that can be applied in the context of a CAT based on testlets. The Bradlow & Weiss (1997) study was extended by determining the significance probabilities of *TNR* and *LLR* using exact theoretical distributions of runs (Mood, 1940, Mosteller, 1941) instead of simulation methods.

Statistical Process Control

In this section, theory from Statistical Process Control (SPC), often used to control production processes, is introduced. For an earlier discussion of this technique, see also Meijer and van Krimpen-Stoop (2003). Consider, for example, the process of producing bags of candy, in which each bag has a certain weight. Too much candy in each bag is undesirable for financial reasons and customers will complain when too few pieces of candy are in the bags. Therefore, the weight of the bags of candy needs to be controlled during the production process. This can be done using techniques from SPC.

A (production) process is in a state of statistical control if the variable being measured has a stable distribution. One technique from SPC is using a Shewhart control chart, originally proposed by Shewhart (1931); 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 $\overline{X}$-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 Shewhart 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, Shewhart 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 cumulative 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 d units. Suppose that Z_t is the value of statistic Z obtained from a sample at time point t, d is the reference value, and h is some threshold. Then, the two-sided CUSUM procedure can be written in terms of C_t^H and C_t^L, where

$$\begin{aligned} C_1^H &= \max[0,\, Z_1 - d] \\ C_2^H &= \max[0,\, (Z_1 - d) + (Z_2 - d)] \\ &= \max[0,\, (Z_2 - d) + C_1^H] \\ C_3^H &= \max[0,\, (Z_3 - d) + C_2^H] \\ &\cdots\cdots \\ C_t^H &= \max[0,\, (Z_t - d) + C_{t-1}^H], \end{aligned} \tag{14}$$

and analogously

$$C_t^L = \min[0,\, (Z_t + d) + C_{t-1}^L], \tag{15}$$

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 $Z_t + d$ is reflected by C_t^L. Thus, as soon as $|Z_t| > d$, the CUSUM chart starts. The process is in an "out-of-control" state 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 computed 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, 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 3. A commonly used statistic is V, defined in Equation 5, the mean of the squared standardized residuals based on I items. One of the drawbacks of V is that negative and positive residuals cannot 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 the test, 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 will occur.

Sums of consecutive negative or positive residuals can be investigated by using a CUSUM procedure. This can be explained as follows. A CAT can be viewed as a multistage test, in which 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, let n be the final test length, and let, without loss of generality, the reference value be equal to 0. Below, some examples of statistic T are proposed. For examinee j, at each stage k of a CAT, the CUSUM procedure can be determined as

$$C_k^H = \max[0,\ T_k + C_{k-1}^H], \tag{16}$$

$$C_t^L = \min[0,\ T_k + C_{k-1}^L], \text{ and} \tag{17}$$

$$C_0^H = C_0^L = 0, \tag{18}$$

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 classified as nonfitting to the model; otherwise, the response pattern is normal.

Person Fit Statistics

Let S_k denote the set of items administered as the first k items in the CAT and $R_k = \{1, ..., I\} \backslash S_{k-1}$ denote 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 previously 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 $\hat{\theta} = \hat{\theta}_n$ denote 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 answering item i_k correctly, evaluated at $\hat{\theta}_{k-1}$, can be written as

$$P_{i_k}(\hat{\theta}_{k-1}) = \frac{\exp[a_{i_k}(\hat{\theta}_{k-1} - b_{i_k})]}{1 + \exp[a_{i_k}(\hat{\theta}_{k-1} - b_{i_k})]}. \tag{19}$$

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 online situation when the test length of the CAT is fixed. These four statistics use as the expected score the probability of answering the item correctly, evaluated at the updated ability estimate, defined in Equation 19. The other four statistics, T^5 through T^8, use as the expected score the probability of answering the item correctly, evaluated at the final ability estimate $\hat{\theta}$. As a result of using $\hat{\theta}$ instead of $\hat{\theta}_k$, the development of the accumulated residuals can no longer be investigated in an online situation. All statistics are based on the general form defined in Equation 3: A particular statistic is defined by choosing a particular weight. In two statistics, the residual $U - P(\cdot)$ is weighted by the estimated standard deviation,

$$P(\cdot)(1 - P(\cdot))^{-\frac{1}{2}}.$$

In two other statistics, $U - P(\cdot)$ is weighted by the square root of the test information function containing the items administered up to and including stage k, which is a monotone increasing function of the stage of the CAT. As a result, the residuals in the beginning of the test become a larger weight than the residuals in the last part of the test. These two statistics may be sensitive to nonnormal responses in the earlier part of the test. In two other statistics, 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. These two statistics may be sensitive to nonnormal responses in the later part of the CAT.

Define

$$T_k^1 = \frac{1}{n}[U_{i_k} - P_{i_k}(\hat{\theta}_{k-1})], \tag{20}$$

$$T_k^2 = T_k^1 \times [P_{i_k}(\hat{\theta}_{k-1})(1 - P_{i_k}(\hat{\theta}_{k-1}))]^{-\frac{1}{2}}, \tag{21}$$

$$T_k^3 = T_k^1 \times [I(\hat{\theta}_{k-1})]^{-\frac{1}{2}}, \text{ and} \tag{22}$$

$$T_k^4 = \sqrt{k} \times T_k^1, \tag{23}$$

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

$$I(\hat{\theta}_k) = \sum_{i_g \in S_k} I_{i_g}(\hat{\theta}_k, a_{i_g}, b_{i_g}) = \sum_{i_g \in S_k} a_{i_g}^2 P_{i_g}(\hat{\theta}_k)(1 - P_{i_g}(\hat{\theta}_k)). \tag{24}$$

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

$$T_k^5 = \frac{1}{n}[U_{i_k} - P_{i_k}(\hat{\theta})], \tag{25}$$

$$T_k^6 = T_k^5 \times [P_{i_k}(\hat{\theta})(1 - P_{i_k}(\hat{\theta}))]^{-\frac{1}{2}}, \tag{26}$$

$$T_k^7 = T_k^5 \times [I(\hat{\theta})]^{-\frac{1}{2}}, \text{ and} \tag{27}$$

$$T_k^8 = \sqrt{k} \times T_k^5, \tag{28}$$

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

$$I(\hat{\theta}) = \sum_{i_g \in S_k} I_{i_g}(\hat{\theta}, a_{i_g}, b_{i_g}) = \sum_{i_g \in S_k} a_{i_g}^2 P_{i_g}(\hat{\theta})(1 - P_{i_g}(\hat{\theta})). \tag{29}$$

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

These eight statistics can be used in the CUSUM procedure described in Equations 16 through 18. 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, while the other statistics are standardized versions of T^1 and T^5, also based on only one observation. Therefore, 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 $d = 0$. Another alternative is to increase the sample size of the sample obtained at each stage; in a CAT, the sample is of size 1 because one item at each stage is administered. Under certain conditions, increasing the sample size can be done by considering a CAT based on testlets, where the testlets contain N items (with N large enough). Then, the sample obtained at each timepoint becomes of size $N > 1$.

Use of CUSUM-Based Statistics in Testlets

A testlet is defined by Wainer and Kiely (1987) as "a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow"; in other words, testlets are small tests. In a CAT based on testlets, the first testlet selected is based on an initial estimate of θ. After each testlet, θ is estimated and the next testlet selected is the testlet with maximum information at the updated $\hat{\theta}$. The CAT ends when the precision of $\hat{\theta}$ is adequate or when a certain number of testlets is administered. Thus, now the stages, k, of the CAT are testlets instead of items.

The upper and lower bounds of SPC control charts and CUSUM procedures are based on the assumption of measurements from a normal distribution. One of the SPC procedures is a chart for the number of defectives, with a binomial parameter p (i.e., the probability of a defective product in the sample) where a normal approximation to the binomial distribution can be applied when the sample is large enough. Quessenberry (1991) showed that the normal approximation is sufficiently accurate at significance level α when the sample size N at each time point is

$$N > \frac{\ln \alpha}{\ln(1-p)}.$$

In the case of a CAT, the probability of a correct response is close to 0.5. So, setting $p = 0.5$ and the significance level $\alpha = 0.05$, the sample size at each time point needs to be at least 4.3. However, in this present case the sample size at each time point is only one; that is, one item at each stage of the CAT. Therefore, it is difficult to use the normal approximation to a binomial distribution in this present context. However, in case of testlets, with $N \geq 5$ items per testlet, this theory may be applied in a proper way.

Let $X_{s_i k}$ denote the response to item $i = 1, ..., N$ of testlet $s = 1, ..., R$ at stage k of the CAT; that is, the kth testlet administered in the CAT. Define

$$X_{sk} = \sum_{i=1}^{N} X_{s_i k}$$

as the total score of testlet s. Then, according to the 2PLM, the expected value and variance of X_{sk} can be determined as

$$E(X_{sk} \mid \theta) = E\left(\sum_{i=1}^{N} X_{s_i k} \mid \theta\right) = \sum_{i=1}^{N} E(X_{s_i k} \mid \theta) = \sum_{i=1}^{N} P(X_{s_i k} = 1 \mid \theta) \tag{30}$$

and analogously

$$var(X_{sk} \mid \theta) = \sum_{i=1}^{N} P(X_{s_i k} = 1 \mid \theta)[1 - P(X_{s_i k} = 1 \mid \theta)]. \tag{31}$$

Because in practice θ is unknown, $\hat{\theta}_{k-1}$ can be used as an alternative. When N is large enough, statistic

$$Z_{sk} = \frac{X_{sk} - E(X_{sk})}{\sqrt{var(X_{sk})}} \tag{32}$$

is approximately standard normally distributed. When $P(X_{s_i k} = 1 \mid \hat{\theta}_{k-1})$ is used to determine the expected value and the variance of X_{sk}, Z_{sk} can be rewritten in terms of residuals,

$$Z_{sk} = \frac{\sum_{i=1}^{N} [X_{s_i k} - P(X_{s_i k} = 1 \mid \hat{\theta}_{k-1})]}{\sqrt{var(X_{sk})}}.$$

That is, Z_{sk} is the standardized sum of residuals of the items in the testlet. Using statistic Z_{sk} in the CUSUM procedure described in Equations 14 and 15 (with $d = 0.5$, the appropriate choice for detecting a shift of one times the standard deviation of Z_{sk}) results in the following procedure:

$$C_k^H = \max[0, Z_{sk} - 0.5) + C_{k-1}^H]$$

$$C_k^L = \min[0, Z_{sk} + 0.5) + C_{k-1}^L]$$

with starting values $C_0^H = C_0^L = 0$. An examinee can be classified as nonfitting the IRT model when $C_k^H > h$ or $C_k^L < -h$. That is, after a number of consecutive positive or negative standardized residuals Z_{sk}, an examinee may be classified as nonfitting. Setting $d = 0.5$ and $\alpha = 0.05$ will result in a threshold of $h = 3$ (see e.g., Montgomery, 1991, pp. 291–293).

Hulin, Drasgow and Parsons (1983, p. 113) describe the aspect of test preview; because complete memorization of the tests is unlikely due to, for example, test length, it is likely that "cheaters" memorize blocks of items. Especially in the context of CAT based on testlets, it is likely that preknowledge of all items in some testlets occurs and all items in the testlet are answered correctly; consecutive positive residuals Z_{sk} or large positive values of Z_{sk} may be the result of these nonfitting response patterns. The proposed CUSUM procedure might be suitable for detecting these types of nonfitting response patterns.

Total Number of Runs and Length of Longest Run

An alternative is to apply the commonly used (nonparametric) tests for randomness in a sequence of alternatives: tests based on runs, for example, the total number of runs, or the length of the longest run. A run is defined as a succession of similar events followed and preceded by different events; here the different events are 0 for an incorrect response and 1 for a correct response. Due to the CAT item-selection procedure, an alternation of zeros and ones is expected; that is, many runs are expected. Few long runs may indicate aberrant response behavior. For example, an examinee is guessing the answers to all the items of a test. Due to the random response behavior, the probability of a correct response becomes small. Therefore, many incorrect responses and few correct responses will occur during the test, resulting in fewer and longer runs than expected.

Total Number of Runs (TNR)

Suppose an examinee j responds to an adaptive test of length n. Let the vector $U_j = (U_{ji_1}, U_{ji_2}, ..., U_{ji_n})$ be the dichotomous response vector to the administered adaptive test, a vector of zeros and ones. Let n_0 denote the number of incorrect responses, n_1 the number of correct responses, and r the total number of runs in the sequence of responses. The probability distribution of TNR, the total number of runs of $n = n_0 + n_1$ objects, is defined as

$$P(TNR = r) = \begin{cases} \dfrac{\binom{n_0 - 1}{r/2 - 1}\binom{n_1 - 1}{r/2 - 1}}{\binom{n_0 + n_1}{n_0}} & \text{if } r \text{ is even} \\ \dfrac{\binom{n_0 - 1}{(r-1)/2}\binom{n_1 - 1}{(r-3)/2} + \binom{n_0 - 1}{(r-3)/2}\binom{n_1 - 1}{(r-1)/2}}{\binom{n_0 + n_1}{n_0}} & \text{if } r \text{ is odd} \end{cases} \quad (33)$$

for $r = 2, 3, ..., n$ (Gibbons & Chakraborti, 1992, pp. 72–73). Because few runs might indicate aberrance, the significance probability of the observed response vector is the probability of r or less runs and is defined by

$$p^*(TNR) = P(TNR \le r) = \sum_{q=2}^{r} P(TNR = q). \tag{34}$$

Length of the Longest Run (LLR)

Let r_{vw} denote the number of runs of type $v = 0, 1$ which are of length $w = 1, ..., n_v$. Let l denote the length of the longest run observed in the response pattern U. Longer runs are more aberrant, thus the significance probability of the random variable *LLR*, the length of the longest run of $n = n_0^* + n_1$ objects, is the probability of getting at least one run of length l or more of either type 1 or 0. This probability was derived by Mosteller (1941) and can be written as

$$p^*(LLR) = P(r_{1w} \ge 1 \text{ or } r_{2w} \ge 1 \text{ or both}; LLR \ge l) = 1 - \frac{A}{\binom{n_0 + n_1}{n_0}} \tag{35}$$

where

$$A = \sum_{r_0 > \frac{n}{l}} \left\{ \left[\sum_{s=0}^{r_0} (-1)^s \binom{r_0}{s} \binom{n_0 - 1 - s(l-1)}{r_0 - 1} \right] \left[\sum_{r_1 = r_0 - 1}^{r_0 + 1} F(r_0, r_1) \sum_{t=0}^{r_1} (-1)^t \binom{r_1}{t} \binom{n_1 - 1 - t(l-1)}{r_1 - 1} \right] \right\},$$

and

$$F(r_0, r_1) = \begin{cases} 0 \text{ if } |r_0 - r_1| > 1 \\ 1 \text{ if } |r_0 - r_1| = 1 \\ 2 \text{ if } |r_0 - r_1| = 0 \end{cases}.$$

Examinees can be classified as nonfitting when the significance probability of the total number of runs and/or the length of the longest run is smaller than some predefined significance level α, that is, $p^*(TNR) < \alpha$ and/or $p^*(LLR) < \alpha$.

Simulation Studies

The two methods, the parametric CUSUM procedures and the nonparametric runs tests (*TNR* and *LLR*), both investigate strings of correct or incorrect responses in a CAT; with the CUSUM procedure, it can be tested whether the responses fit to an IRT model; the runs tests do not assume any IRT model. An advantage of *TNR* and *LLR* is that the significance probabilities can be exactly determined using the theoretical distributions described in Equations 33 through 35. A drawback of the CUSUM procedure is the absence of guidelines for determination of the upper and lower boundary for non-normally 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 boundaries are independent of θ, a fixed upper and lower boundary 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 were investigated, using fixed and simulated boundaries. Furthermore, in Study 2, the detection rates of *TNR* and *LLT* (based on the exact significance probabilities) for several types of nonfitting response behavior were examined, this in contrast with the Bradlow and Weiss (1997) study in which only empirical data were used and the significance probabilities were estimated instead of exactly determined.

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 2 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, $P(\theta)$ 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 30 items.

For each simulee, eight different statistics, T^1 through T^8, were used in the CUSUM procedure described in Equations 16, 17, and 18. Then, for each simulee and for each statistic,

$$\max C^H = \max_k \left(C_k^H \right) \text{ and} \tag{36}$$

$$\min C^L = \min_k \left(C_k^L \right) \tag{37}$$

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 threshold 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 threshold 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 $\alpha = 0.05$ was conducted, where $P(\max C^H \geq UB) = P(\min C^L \leq LB) = 0.025$. In other words, for each statistic, two boundaries (upper and lower bound) per dataset were determined, where 5% of the simulees attained values outside these boundaries.

Also, for each statistic, the weighted average of the upper and lower boundary was calculated, with different weights for different θ-values: weights 0.05, 0.2, 0.5, 0.2, and 0.05, for $\theta = -2, -1, 0, 1$, and 2, respectively. Weights were used to represent a "realistic" distribution of abilities.

Results

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

As a result of the stable boundaries for almost all statistics across θ, one fixed upper and lower boundary for each statistic might be taken as thresholds for the CUSUM procedures.

TABLE 1
Upper and lower boundaries of CUSUM procedure with T^1 through T^8

		T^1		T^2		T^3		T^4	
	weights	*LB*	*UB*	*LB*	*UB*	*LB*	*UB*	*LB*	*UB*
$\theta = -2$	0.05	–0.23	0.19	–0.47	0.40	–0.12	0.09	–0.13	1.81
–1	0.2	–0.20	0.19	–0.42	0.40	–0.08	0.07	–0.13	1.83
0	0.5	–0.20	0.20	–0.41	0.42	–0.07	0.07	–0.13	1.86
1	0.2	–0.20	0.20	–0.41	0.43	–0.07	0.09	–0.13	1.86
2	0.05	–0.18	0.23	–0.41	0.47	–0.09	0.11	–0.13	2.02
weighted average		–0.20	0.20	–0.41	0.42	–0.07	0.07	–0.13	1.86
		T^5		T^6		T^7		T^8	
	weights	*LB*	*UB*	*LB*	*UB*	*LB*	*UB*	*LB*	*UB*
$\theta = -2$	0.05	–0.13	0.13	–0.27	0.29	–0.11	0.30	–0.10	1.72
–1	0.2	–0.13	0.13	–0.28	0.28	–0.07	0.13	–0.10	1.73
0	0.5	–0.13	0.14	–0.29	0.29	–0.07	0.06	–0.11	1.76
1	0.2	–0.13	0.13	–0.28	0.28	–0.12	0.07	–0.10	1.73
2	0.05	–0.13	0.13	–0.29	0.27	–0.28	0.11	–0.10	1.85
weighted average		–0.13	0.13	–0.28	0.28	–0.09	0.09	–0.10	1.75

Study 2

Method

Six datasets containing 1,000 nonfitting adaptive response patterns were constructed with three types of aberrant response behavior; 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 was classified as aberrant. For each response vector, CUSUM procedures using T^1 through T^8 were performed and the significance probabilities p^* of *TNR* and *LLR* were determined, as in Equations 34 and 35. When the CUSUM procedure with T^1 through T^8 was used, a response vector was classified as nonfitting when max $C^H(T^i) > UB(T^i)$ or max $C^L(T^i) < LB(T^i)$ for $i = 1, ...8$. The upper boundary and lower boundary for each statistic was set to the weighted average of the values presented in Table 1. When *TNR* or *LLR* was used, a simulee was classified as aberrant when $p^* < \alpha$. 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 analogously determined.

Types of Aberrant Response Behavior

Random response behavior. The first type of aberrant response behavior that was simulated was random response behavior to all items of the test. 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. Second, 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. Third, response vectors with violations of local stochastic independence between the items of 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, & van Krimpen-Stoop, 1998) was used to simulate response vectors with local independence between all subsequent items

$$P(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta) \propto \exp\left[\sum_{j=i}^{i+1} x_j a_j(\theta - b_j) + x_i x_{i+1}\delta_{i,\,i+1}\right], \tag{38}$$

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, and negative values of δ result in a lower probability of correctly answering an item. 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 and statistics *TNR* and *LLR* are given for several types of nonfitting response behavior. Table 2 shows that for the dataset of normal response patterns and for most statistics, the percentage of simulees classified as nonfitting was around 0.05; for *LLR* and T^3, the percentage of simulees classified as nonfitting was 0.01 and 0.13, respectively.

Table 2 shows that the detection rates for guessing simulees were quite different for each statistic; for example, the detection rate was 0.04 for *LLR* and 0.97 for T^7. Table 2 also shows that, for $\rho = 0$ and $\rho = 0.5$, the detection rates were rather high for almost all statistics; for $\rho = 0$, the lowest detection rate was 0.12 for T^7, the highest detection rate was 0.34 for T^2. For violations against local independence, significant detection rates occurred only for $\delta = 1$ and 2; for example, for statistic *TNR*, the detection rate was 0.79 and 0.36 for $\delta = 2$ and 1, respectively, compared with 0.00 for both $\delta = -2$ and -1. However, for $\delta = 1$ and 2, the detection rates were rather high: for $\delta = 2$, the lowest detection rate was 0.18 for T^7, whereas the highest detection rate was 0.79 for *TNR*.

TABLE 2
Detection rates

Response Behavior		*TNR*	*LLR*	T^1	T^2	T^3	T^4	T^5	T^6	T^7	T^8
normal		0.07	0.01	0.05	0.06	0.13	0.04	0.04	0.04	0.07	0.04
guessing		0.11	0.04	0.66	0.72	0.89	0.59	0.19	0.59	0.97	0.21
$\rho =$	0	0.18	0.13	0.31	0.34	0.13	0.28	0.33	0.33	0.12	0.28
	0.5	0.11	0.05	0.20	0.23	0.11	0.16	0.18	0.20	0.11	0.16
$\delta =$	–2	0.00	0.00	0.03	0.05	0.11	0.01	0.00	0.01	0.12	0.00
	–1	0.00	0.00	0.03	0.04	0.11	0.01	0.01	0.01	0.10	0.01
	1	0.36	0.05	0.10	0.13	0.19	0.11	0.11	0.12	0.11	0.12
	2	0.79	0.20	0.22	0.27	0.33	0.28	0.29	0.32	0.18	0.34

Discussion

The results of Study 1 showed that the boundaries of the CUSUM procedures were rather stable across θ-values for all statistics except T^3 and T^7. As a result of the stable boundaries across θ-levels, for all statistics T^1 through T^8, one fixed *UB* and *LB* was used as a threshold for the CUSUM procedures. In Study 2, these fixed boundaries were used to determine the detection rates for the eight CUSUM procedures. Also, the exact significance probabilities of *TNR* and *LLR* were computed to determine the detection rates for these statistics. Results showed that, using statistics *LLR* and T^3, the percentage of normal response patterns as nonfitting were deviant from 0.05. *LLR* resulted in a low percentage of normal response patterns classified as nonfitting; as a result, *LLR* might result in a conservative classification of nonfitting response behavior. 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. The high percentage of normal response patterns classified as nonfitting of statistic T^3 might be caused by the use of a fixed *UB* and *LB* for each θ-value, although the boundaries were not stable across θ-levels. Detection rates for the CUSUM procedure using statistic T^3 might, therefore, be improved by simulating boundaries for each simulee instead of using a fixed *UB* and *LB*.

Due to the scarceness of literature, it is difficult to compare the detection rates with other studies using CAT data. 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 was used where some examinees were distorting their own responses; detection rates between 0.01 and 0.32 were found for the P&P data.

From our own experience, the detection rates found in this study were high compared with other person-fit statistics in a CAT environment (e.g., Bradlow & Weiss, 1997).

References

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

Bradlow, E. T., & Weiss, R. E. (1997). *Outlier measures and norming methods for computerized adaptive tests*. Retrieved from http://rem.ph.ucla.edu/~rob/papers/list.html.

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.

Gibbons, J. D., & Chakraborti, S. (1992). *Nonparametric statistical inference*. New York: Marcel Dekker.

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

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory.* Homewood, Il: Dow Jones-Irwin.

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., & 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.

McLeod, L. D., & Lewis, C. (1998, April.) *A Bayesian approach to detection of item preknowledge in a CAT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Meijer, R. R., & van Krimpen-Stoop, E. M. L. A. (2003). *The use of statistical process control-charts for person-fit analysis in computerized testing* (Computerized Testing Report 98-12). Newtown, PA: Law School Admission Council.

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.

Mood, A. M. (1940). The distribution theory of runs. *Annals of Mathematical Statistics, 11*, 367–392.

Mosteller, F. (1941). Note on an application of runs to quality control charts. *Annals of Mathematical Statistics, 12*, 228–232.

NCLEX (1995). National Council Licensure Examination, National Council of State Boards of Nursing, Chicago, Ill.

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

Quessenberry, C. P. (1991). SPC Q charts for a binomial parameter p: short or long runs. *Journal of Quality Technology, 23*, 239–246.

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

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

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

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

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of a person-fit statistic in fixed and computerized adaptive tests. *Applied Psychological Measurement*, *23*, 327–345.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *Journal of Educational Measurement, 24*, 185–210.

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.