LSAC RESEARCH REPORT SERIES

The Use of Person-Fit Statistics in Computerized Adaptive Testing

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

Law School Admission Council
Computerized Testing Report 97-14
September 2005

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

Table of Contents

Executive Summary

In computerized adaptive testing (CAT), it is often assumed that an item response theory (IRT) model adequately describes the response behavior of the test takers. IRT is a mathematical theory that models the likelihood that a test taker of a particular level of ability will answer a particular test item correctly. In reality, test takers may not always respond to test items in ways that are consistent with the IRT model. Some possible reasons for this "nonfitting" behavior include test anxiety, guessing, cheating (achievement tests), or faking responses (personality inventories).

Several person-fit statistics have been proposed to detect test takers with nonfitting response behavior. To evaluate these statistics, their distributions under the null assumption of perfect model fit must be known. For conventional standardized tests, the distributions of several fit statistics have been investigated, and the results are well documented. However, less is known about these distributions for adaptive tests.

In this paper, three different simulation studies are reported. Each study simulated the use of a frequently used person-fit statistic, namely the standardized log-likelihood statistic, l_z. In the first study, the null distribution of the l_z-statistic (i.e., the distribution one would expect if test takers respond in a manner predicted by the mathematical model studied) was estimated both for the true ability level as well as an estimate of it. In the second study, the null distributions for a fixed-format test and a CAT were compared. In the final study, the statistic for detecting several types of aberrant response behavior in CAT was studied.

The results indicated that the empirical null distribution of l_z for CATs differed from those for tests with a fixed format, and that the differences were larger if estimates of the ability parameters were used rather than their true values. As a result, the detection rates for different types of aberrant behavior were generally low. However, the results also indicated that the estimation errors in the ability estimates due to aberrant behavior on some items were not dramatically different from those for the test takers that behaved according to the model on all items. Thus, in practice, ability can be estimated reasonably well if the response behavior of the test takers does not fit the model for some of the items.

Abstract

Several person-fit statistics have been proposed to detect item score patterns that do not fit an item response theory model. To classify response patterns as not fitting a model, a distribution of a person-fit statistic is needed. Recently, the null distributions of several fit statistics have been investigated using conventional administered tests. For computerized adaptive testing (CAT), however, less is known about the distribution of fit statistics. In this study, a three-part simulation study was conducted. First, the theoretical distribution of the often used l_z-statistic across θ levels in a conventional testing CAT environment was investigated, where θ and $\hat{\theta}$ were used to calculate l_z. Second, two procedures for simulating the distribution of l_z were examined: (1) item scores were simulated with a fixed set of administered items, and (2) item scores were generated according to a stochastic design, where the choice of the administered item $i+1$ depended on item i. Finally, a study was performed to evaluate the usefulness of person-fit in the CAT environment to detect aberrant response patterns. Results indicated that the distribution of l_z differed from the theoretical distribution, and that simulating the sampling distribution of l_z was problematic when $\hat{\theta}$ was used to determine the values of l_z. As a result, the detection rates for different types of aberrant behavior were relatively low. However, it was also shown that, for some types of aberrant behavior, the error in $\hat{\theta}$ was not dramatically different from that of normal-responding simulated test takers, and that, in practice, $\hat{\theta}$ can be reasonably well estimated, even for simulated test takers not fitting the model.

Introduction

Item responses that do not fit the assumed item response theory (IRT) model may cause the latent trait value θ to be inaccurately estimated. Possible interpretations for nonfitting test behavior include test anxiety, guessing, cheating on achievement tests, or response distortion as a result of faking the answers on personality inventories (Zickar & Drasgow, 1996). Statistics have been proposed to detect nonfitting test takers (e.g., Drasgow & Levine, 1986; Meijer, 1994; Tatsuoka, 1984), and the effectiveness of these statistics to detect nonfitting response vectors has been investigated (e.g., Drasgow, Levine, & McLaughlin, 1987, 1991). However, most person-fit studies concentrated on conventionally administered tests. With the increasing use of computerized adaptive tests (CAT), additional research is needed with respect to the application of person-fit statistics using these types of tests.

A few studies investigated the usefulness of person-fit analysis in CAT. Candell (1988; cited in Drasgow, Levine, & Zickar, 1996) used optimal person-fit statistics in which the likelihood of a normal-responding person was compared with the likelihood under an aberrant model to study the ability of a likelihood ratio test to identify simulated aberrant test takers. Although this approach is interesting because it has maximum

power against a specified alternative, the drawback is that, for other types of aberrant responding, the power is low.

Nering (1997) examined the distribution of two fit statistics, l_z (Drasgow, Levine, & Williams, 1985), and $ECI4_z$ (Tatsuoka, 1984), in a CAT environment and found that the empirical distribution was dramatically different from the theoretical distribution. As a result, Nering concluded that "when attempting to classify a response vector as model divergent (...) cutscores may have to be based on such factors as item pool size, item pool discrimination, and so forth." An alternative to using a critical value derived from a theoretical distribution is to simulate for each test taker a distribution of a person-fit statistic based on the characteristics of the item bank and the estimated latent trait value of θ ($\hat{\theta}$). Then, using the simulated distribution, it can be determined how likely a response vector is under the IRT model.

The purpose of the present study was to extend the Nering study by examining the distribution of a person-fit statistic in CAT and to investigate two different ways to simulate the distribution of a person-fit statistic. Furthermore, the detection rate of a person-fit statistic to detect nonfitting response vectors in a CAT will be investigated.

Person-Fit Analysis

In person-fit analysis, several fit statistics have been used in the context of the one-, two-, and three-parameter logistic model (Hambleton & Swaminathan, 1985, pp. 35–48). In this study, we use the two-parameter logistic model (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 been shown to have a reasonable fit to several achievement and personality data (e.g., Reise & Waller, 1990; Zickar & Drasgow, 1996).

Let X_i be the binary (0, 1) response to item i, where one denotes a correct or keyed response, and zero denotes an incorrect or not-keyed response. Further, let a_i denote the item-discrimination parameter and b_i the item-difficulty parameter, then the probability of correctly answering an item according to the 2PLM can be written as

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

For simplicity, P_i will be written as P, and X_i as X.

Levine and Rubin (1979) proposed the log-likelihood statistic, denoted as l_0, as a measure of departure from the logistic IRT models. l_0 can be written as

$$l_0 = \ln\left[\prod_i P\left(\hat{\theta}\right)^X \left(1 - P\left(\hat{\theta}\right)\right)^{1-X}\right], \tag{2}$$

where the product is across all items in the test. Because l_0 is confounded with $\hat{\theta}$, Drasgow, Levine, and Williams (1985) proposed to use the standardized version of l_0, denoted as l_z.

This statistic equals

$$l_z = \frac{l_0 - E(l_0)}{\left[\mathrm{var}(l_0)\right]^{\frac{1}{2}}}, \tag{3}$$

where E (l_0) and var (l_0) denote the expectation and variance of l_0, respectively. These quantities are given by

$$E(l_0) = \sum_i \left\{P\left(\hat{\theta}\right)\ln\left[P\left(\hat{\theta}\right)\right] + \left(1 - P\left(\hat{\theta}\right)\right)\ln\left[1 - P\left(\hat{\theta}\right)\right]\right\}, \tag{4}$$

and

$$\mathrm{var}(l_0) = \sum_i P\left(\hat{\theta}\right)\left(1 - P\left(\hat{\theta}\right)\right)\left[\ln\frac{P\left(\hat{\theta}\right)}{1 - P\left(\hat{\theta}\right)}\right]^2. \tag{5}$$

For classifying a response pattern as aberrant, an important tool is the probability of exceedance, also called the significance probability. Let t be the observed value of the person-fit statistic T, then the significance probability is defined as the probability under the sampling distribution that the value of the test statistic is smaller than the observed value of the statistic: $p^* = P(T \leq t)$. Because large negative values of l_z indicate aberrance, the significance probabilities in the left tail of the distribution are of interest. The value

of a statistic with $p^* = \alpha$ will be denoted as the critical value at significance level α. For example, for a standard normally distributed statistic, the critical value at level $\alpha = 0.05$ is –1.65.

Drasgow et al. (1985) purported, in the context of conventional testing, that l_z was distributed standard normal for long tests; that is, tests consisting of 85 items. However, several studies (e.g., Meijer & Nering, 1997; Molenaar & Hoijtink, 1990) showed that l_z was not standard normally distributed for tests of realistic length (20–60 items). It was found that the distribution of l_z was negatively skewed and that the normal approximation was inaccurate, especially in the tails of the distribution. As an alternative, Molenaar and Hoijtink (1990) proposed for the Rasch model three approximations to the distribution of l_0, conditional on the total score: using (1) complete enumeration, (2) Monte Carlo simulation, and (3) a chi-squared distribution, where the mean, standard deviation, and skewness of l_0 were taken into account. Complete enumeration is suitable for very short tests (see Liou and Chang (1992) for a network algorithm for tests of 30 items or less). For very long tests, an accurate calculation of the moments, needed for the chi-squared approximation, is difficult and, as an alternative, Monte Carlo simulation can be applied. In the Rasch model, the total score is a sufficient statistic for θ; for the 2PLM this is not the case; that is, the distribution of a person-fit statistic conditional on the total score is dependent on θ. As an alternative, $\hat{\theta}$ can be used in the case of the 2PLM or 3PLM. However, care should be taken by doing this because, as Molenaar and Hoijtink (1990) noticed, the statistical results on the distribution of a person-fit statistic may change when substituting $\hat{\theta}$ for θ.

Person Fit in CAT

Nering (1997) examined the distribution of l_z within CAT by evaluating the first four moments (mean, standard deviation, skewness, and kurtosis) of the distribution. His results were in concordance with the results using conventional tests: the empirical distribution of l_z was different from the theoretical distribution. An important finding was that the normal approximation in the tails of the distribution was inaccurate: using a critical value of –1.65 resulted in a conservative classification of aberrant response patterns; $p^* = P\,(l_z \leq -1.65) << 0.05$. On the basis of these results, it can be concluded that the theoretical distribution of l_z is not useful to obtain a critical value in CAT. A possible solution for determining significance probabilities in CAT is to approximate the distribution of the statistic by using simulation methods.

Simulating Distributions of Person-Fit Statistics in CAT

The significance probabilities can be determined by simulating the distribution of l_z. This can be realized in at least two ways. One possibility is to determine the distribution of l_z by drawing a large number of θ-values from the standard normal distribution, each θ-value representing a person responding to the test. Then, item scores can be simulated for each θ-value, according to the assumed IRT model, and the l_z-value for each pattern can be calculated; the l_z-values of these patterns (with different θ-values) constitute the simulated distribution based on the characteristics of the item bank. Another possibility is to simulate a distribution of l_z for each θ-value; that is, given θ, a large number of response patterns are simulated and, for each pattern, the l_z-value is calculated. Next, a distribution is simulated, based on the item bank and on θ. Based on this distribution, a critical value at level α can be determined.

The first method results in using one critical value for all test takers, whereas the second method will probably result in using different critical values at different θ-values. When the distribution of l_z is the same across all θ-levels, the first method will result in an appropriate simulated distribution.

Using the second method, the distribution of l_z can be simulated using a fixed sequence of items (test design) in which for each θ-value a large number of response vectors are simulated given the observed test design. Thus, each response vector consists of responses to the same items. However, an important aspect when simulating the distribution of a statistic is the stochastic process of item selection in a CAT (Glas, Meijer, & van Krimpen, 1997). When a test taker responds to a CAT, the test design may be different for each test taker. Therefore, the distribution can be simulated using a stochastic test design in which, for each θ-value, a large number of adaptive response patterns are simulated with, in principle, different test designs.

Let the vector d denote the test design; that is, a vector of the numbers of administered items in CAT, and $T(X)$ a statistic of the observed response vector $X = (X_1, \ldots, X_k)$ of test length k. In CAT, item selection is based on responses to previously administered items that are dependent on the ability of the test taker. Therefore, X is conditional on d and θ, and a function of X, for example the statistic $T(X) = T$, is also conditional on d and θ. Thus, the distribution of T, conditional on d and θ, is defined as

$$f\left(T \mid d, \theta\right). \tag{6}$$

To obtain the unconditional distribution of T, Equation 6 can be multiplied by the probability distribution of the design d, which results in

$$f(T, d, \theta) = f(d, \theta) f(T|d, \theta)$$
$$= f(\theta) f(d|\theta) f(T|d, \theta). \quad (7)$$

where $f(\theta)$ represents the probability distribution of θ.

To determine whether the observed value of the statistic T is unlikely under the null-model, the distribution of T given in Equation 7 should be determined. In a conventional test, in which every test taker responds to the same test with the same items, the probability of the design d is the same for every test taker. Thus, for a conventional test $f(d|\theta) = f(d)$, which results in $f(T, d, \theta) = f(\theta) f(d) f(T|d, \theta)$. However, in an adaptive test, every test taker may respond to a unique set of items. Therefore, in a CAT environment, the distribution $f(T, d, \theta)$ has to be determined for classifying observed response patterns as aberrant.

Equation 7 can be approximated using simulation methods. First, for each simulated test taker, $f(\hat{\theta}|\theta)$ is determined as the normal distribution with mean $\hat{\theta}$ and variance $I^{-1}(\hat{\theta})$; the reciprocal of the test information function which is given by

$$I(\theta) = \sum_j I_j(\theta, a_j, b_j) = \sum_j a_j^2 P_j(\theta)(1 - P_j(\theta)), \quad (8)$$

where $P_j(\theta)$ is defined by Equation 1, $\hat{\theta}$ is substituted for θ, and the sum is across all items of the test. Then, n abilities can be drawn from $N(\hat{\theta}, I^{-1}(\hat{\theta}))$, and for all n values of parameter θ, an adaptive response vector can be simulated. By simulating adaptive response patterns for each simulated test taker, the distribution of d given θ is taken into account, and by computing the value of T for all adaptive response patterns, $f(d|\theta) f(T|d, \theta)$ is obtained.

Purpose of This Study

This study was designed to investigate (1) the distribution of l_z across different θ-levels in CAT and the influence of estimation errors of θ on the distribution of l_z (Study 1); (2) the influence of the stochastic nature of the test design in CAT and the estimation errors of θ on simulating the distribution of l_0 and l_z (Study 2); and (3) the detection rate of l_0 and l_z using different types of aberrant behavior and different simulation methods (Study 3).

Study 1

In this study, the distribution of l_z was investigated in a CAT. Nering (1997) examined the distribution of l_z by first drawing 10,000 θ-values from the standard normal distribution and then simulating an adaptive response vector for each θ-value. For each response vector, $\hat{\theta}$ was used to determine the value of l_z. These 10,000 values constituted the simulated distribution of l_z. In this study, the simulated distribution was determined by (1) drawing true θ from a standard normal distribution, and (2) fixing true θ at different levels. For these θ-values, adaptive response vectors were generated, and θ was estimated by $\hat{\theta}$; $\hat{\theta}$ was used to determine the value of l_z. Doing so, the critical values obtained by Nering can be compared with the critical values obtained when the distribution of l_z is simulated at a fixed θ-level. Finally, l_z was also calculated using true θ, that is $P(\theta)$ was used to determine the value of l_z. This enables us to investigate the influence of estimation errors in $\hat{\theta}$ on the distribution of this statistic.

Method

Ten datasets consisting of 10,000 adaptive response patterns were constructed. Nine datasets were constructed at nine different θ-levels: $\theta = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5$, and 2; one dataset was constructed where 10,000 θ's were drawn from a standard normal distribution. First, for each response vector, the value of l_z was calculated using $\hat{\theta}$ and these 10,000 values of l_z were used to obtain the distribution of l_z for each dataset. For all simulated distributions, the critical values at level α were determined and compared with the critical values at level α of the standard normal distribution, where $\alpha = 0.01, 0.02, 0.03, 0.04$, and 0.05. Furthermore, the first four moments (mean, standard deviation, skewness, and kurtosis) of the simulated distribution of l_z were computed and compared with the moments of the standard normal distribution. Second, l_z was also calculated using true θ for each response vector to constitute the distribution of l_z without presence of estimation errors.

Generating Adaptive Response Vectors

A pool of 1,000 items fitting the 2PLM with $a_i \sim U\left(\frac{1}{2}, 2\right)$ and b~ U(-2, 2) was used. An adaptive response pattern was simulated as follows. First, the true θ of a simulated test taker was drawn from a standard normal distribution or was set to a fixed θ-level, dependent on the dataset constructed. 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 $P(\theta) > y$ the response to item i was set to one (correct response), zero otherwise. The first four items of the adaptive test 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); this estimator is less biased than other maximum likelihood estimators, and also exists for perfect and 0-scores. The next item selected was the item with maximum information given $\hat{\theta}$. 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 asymptotic standard error of θ was 0.25; this is an often-used value; see for example DeAyala (1992). The asymptotic standard error of θ was determined by

$$SE(\theta) = \left[\sum_i a_i^2 P_i(\theta)\left(1 - P_i(\theta)\right) \right]^{-\frac{1}{2}}, \tag{9}$$

where the sum was across all administered items and $P_i(\theta)$ was defined by the 2PLM given in Equation 1; the standard error was estimated by substituting $\hat{\theta}$ for θ.

Results

In Table 1, the first four moments and the critical values at level α of the simulated distributions of l_z using $\hat{\theta}$ are given for different θ-levels. Table 1 shows that the first two moments are different from 0 and 1, respectively, for all θ-levels; both mean and variance fluctuated around 0.5. Skewness and kurtosis were also different from zero. The distribution was found to be positively skewed, and the highest skewness observed was 4.07 for $\theta = 0$. The highest kurtosis was also found for $\theta = 0$ where the kurtosis was 6.69. Thus, the distribution of l_z using $\hat{\theta}$ was rather different from the standard normal distribution. Table 1 also shows that the critical values at $\alpha = 0.05$ of the simulated distributions were much closer to zero than the expected –1.65 under the normal distribution. For example, for $\theta = 0$, only 5% of the simulated test takers obtained a l_z-value below –0.48. Thus, using $l_z \leq -1.65$ will result in too few simulated test takers being classified as aberrant; that is, the decision rule will result in a conservative classification of aberrant response behavior. Table 1 also shows that the critical values were different across θ-levels. For example, for $\theta = 0$, the critical value at $\alpha =$ 0.01 was -0.95, whereas the critical value for $\theta = 2$ was –1.53. When θ was drawn from the standard normal distribution, the critical values were also closer to 0 than expected; for example, the critical value at $\alpha = 0.05$ was -0.70. Table 1 shows that for $\theta \sim N(0, 1)$, the critical values at level α are close to the mean critical value at level α across the nine θ-levels. For example, for $\theta \sim N(0, 1)$, the critical value at $\alpha = 0.03$ was –0.91 and the mean critical value at $\alpha = 0.03$ across θ-values was –0.92 (not tabulated).

TABLE 1
Distributional characteristics of the simulated distribution of l_z using $\hat{\theta}$

					Critical Value				
	Mean	Variance	Skewness	Kurtosis	0.01	0.02	0.03	0.04	0.05
$\theta \sim$ N(0, 1)	0.61	0.56	2.64	3.19	-1.30	-1.07	-0.91	-0.79	-0.70
θ = -2.0	0.30	0.32	0.67	1.78	-1.49	-1.30	-1.19	-1.03	-0.89
-1.5	0.53	0.47	1.71	1.52	-1.48	-1.29	-1.12	-0.99	-0.90
-1.0	0.56	0.58	1.94	1.39	-1.37	-1.19	-1.07	-0.96	-0.87
-0.5	0.65	0.56	2.98	3.72	-1.13	-0.94	-0.81	-0.72	-0.62
0.0	0.74	0.52	4.07	6.69	-0.95	-0.77	-0.64	-0.56	-0.48
0.5	0.63	0.55	2.92	3.61	-1.07	-0.94	-0.82	-0.74	-0.66
1.0	0.59	0.57	2.49	2.53	-1.20	-1.03	-0.91	-0.82	-0.73
1.5	0.49	0.55	1.54	1.09	-1.53	-1.30	-1.15	-1.03	-0.94
2.0	0.26	0.34	0.47	1.20	-1.53	-1.35	-1.19	-1.10	-1.01

In Table 2 the distributional characteristics of l_z using true θ are given. Although the mean was closer to zero than using $\hat{\theta}$, it was larger than expected (approximately 0.10 for each θ-level), whereas the variance was smaller than expected; values around 0.8 were found. Results with respect to the skewness and kurtosis were also more in agreement with the standard normal distribution. However, the distribution now tended to be negatively skewed, with the largest value of –0.81 for θ = –1.5. For θ = –2, the highest kurtosis was obtained (1.44). It can be concluded that the distribution of l_z, using true θ, was more in agreement with the standard normal distribution than when $\hat{\theta}$ was used. Similar conclusions pertain for the critical values. However, note that the critical values were different across θ-levels; for example, the critical value at α = 0.03 for θ = 0.5 was –1.68 whereas for θ = –1.5 the critical value was –1.94.

TABLE 2
Distributional characteristics of the simulated distribution of l_z, using θ

					Critical Value				
	Mean	Variance	Skewness	Kurtosis	0.01	0.02	0.03	0.04	0.05
$\theta \sim N(0,1)$	0.10	0.86	-0.20	0.23	-2.40	-2.06	-1.83	-1.67	-1.56
θ = -2.0	0.10	0.73	-0.68	1.44	-2.45	-2.09	-1.87	-1.68	-1.53
-1.5	0.11	0.79	-0.81	1.17	-2.62	-2.20	-1.94	-1.75	-1.59
-1.0	0.07	0.88	-0.40	-0.08	-2.42	-2.09	-1.87	-1.71	-1.58
-0.5	0.11	0.86	-0.08	0.21	-2.31	-1.99	-1.78	-1.62	-1.50
0.0	0.11	0.87	0.04	0.17	-2.35	-2.01	-1.80	-1.63	-1.48
0.5	0.11	0.80	-0.01	0.09	-2.24	-1.93	-1.68	-1.51	-1.39
1.0	0.08	0.89	-0.29	0.01	-2.40	-2.14	-1.89	-1.75	-1.61
1.5	0.08	0.81	-0.47	0.29	-2.41	-2.06	-1.85	-1.71	-1.57
2.0	0.09	0.78	-0.63	0.82	-2.51	-2.13	-1.90	-1.75	-1.64

Study 2

In Study 1 it was shown that (1) the distribution of l_z does not follow a standard normal distribution in a CAT, in particular when $\hat{\theta}$ was used, and (2) the distribution of l_z across θ-levels; as a result, it is advisable to simulate the distribution conditional on θ or $\hat{\theta}$. This second study was designed to investigate the influence of the stochastic nature of the test design in CAT and the estimation errors of θ on simulating the distribution of l_z.

TABLE 3

Distribution of significance probabilities using a fixed and stochastic design, for true θ, $\hat{\theta}$, and $\tilde{\theta}$, to simulate the null distribution of l_0 and l_z

Test Design		Significant Probabilities in Interval										
		[0, 0.1]	[0.1, 0.2]	[0.2, 0.3]	[0.3, 0.4]	[0.4, 0.5]	[0.5, 0.6]	[0.6, 0.7]	[0.7, 0.8]	[0.8, 0.9]	[0.9, 1.0]	X^2
θ												
Fixed	l_0	0.078	0.103	0.103	0.073	0.110	0.103	0.093	0.115	0.133	0.093	11.1
Stochastic	l_0	0.110	0.120	0.080	0.088	0.110	0.090	0.118	0.088	0.105	0.098	7.2
Fixed	l_z	0.078	0.103	0.103	0.073	0.110	0.103	0.093	0.115	0.133	0.093	11.1
Stochastic	l_z	0.110	0.103	0.103	0.080	0.103	0.095	0.093	0.125	0.095	0.095	5.1
$\hat{\theta}$												
Fixed	l_0	0.003	0.033	0.060	0.070	0.098	0.128	0.110	0.130	0.173	0.198	128.6
Stochastic	l_0	0.035	0.065	0.080	0.058	0.120	0.108	0.123	0.105	0.165	0.143	58.7
Fixed	l_z	0.003	0.033	0.060	0.070	0.098	0.128	0.110	0.130	0.173	0.198	128.6
Stochastic	l_z	0.005	0.058	0.053	0.073	0.110	0.095	0.130	0.125	0.158	0.195	111.3
$\tilde{\theta}$												
Fixed	l_0	0.000	0.018	0.058	0.063	0.098	0.130	0.110	0.133	0.193	0.200	162.6
Stochastic	l_0	0.000	0.040	0.073	0.060	0.078	0.128	0.133	0.118	0.183	0.190	134.0
Fixed	l_z	0.000	0.018	0.058	0.063	0.098	0.130	0.110	0.133	0.193	0.200	162.6
Stochastic	l_z	0.000	0.008	0.050	0.050	0.083	0.123	0.105	0.155	0.205	0.223	213.8
Expected		0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	

Method

A dataset of 400 model fitting adaptive response patterns was constructed using the item pool and procedure described in Study 1, where true θ was drawn from the standard normal distribution. The distributions of l_0 and l_z were simulated in two different ways. First, for each adaptive response vector, these distributions were simulated using a fixed test design; that is, for each $\hat{\theta}$ estimated from the obtained response pattern, 500 response patterns were replicated where the test design was set to the observed test design. Thus, for each simulated test taker, the administered test was viewed as a *conventional test*, and this conventional test was replicated 500 times where the value of parameter θ was set to the value $\hat{\theta}$ obtained from the observed response pattern; the value of $\hat{\theta}$ was assumed to be a true value of parameter θ. The values of l_0 and l_z, given in Equations 2 and 3, were computed for each replicated response pattern to obtain the distribution given the value $\hat{\theta}$ and d. Then, the significance probability $P(T < t)$ was determined by comparing the values of l_0 and l_z of the original response pattern with the simulated distribution, where T is l_0 or l_z and t equals the observed value of the statistic.

Second, the distributions of l_0 and l_z were simulated using the stochastic test design. For each simulated test taker, an *adaptive test* was replicated 500 times where the value of parameter θ was set to the value $\hat{\theta}$ obtained from the observed response pattern. For each simulated test taker, 500 adaptive response patterns given $\hat{\theta}$ were replicated as follows: First, ability was set to the observed value of $\hat{\theta}$; this value was assumed to be a true value of parameter θ. The first item selected was the item with maximum information given $\theta = 0$. For this item, the probability of answering correctly to this item, given the value $\hat{\theta}$, was determined according to the 2PLM, and a response was simulated. The first four items of the CAT were selected with maximum information for $\theta = 0$, and, based on the responses to these items, the weighted maximum likelihood estimator was determined. The next item selected was the item with maximum information given the estimated parameter θ at that stage. The probability of answering this item correctly was computed and a response was simulated, then an item was selected based on the estimated θ at that stage. This procedure was repeated until $SE(\theta) \leq 0.25$. Thus, for each simulated test taker, 500 adaptive response patterns were simulated conditional on the value of $\hat{\theta}$. For each replicated response pattern, the values of l_0 and l_z were determined to obtain the simulated distribution. Then, the values of l_0 and l_z of the original response patterns were compared with the simulated distribution.

The distributions of l_0 and l_z were simulated using three types of ability to determine the values of l_0 and l_z: true θ, $\hat{\theta}$, and $\tilde{\theta}$ where $\tilde{\theta}$ was drawn from $N(\hat{\theta}, I^{-1}(\hat{\theta}))$. That is, for each simulated test taker, 500 response patterns were simulated where the value of true θ, $\hat{\theta}$, or $\tilde{\theta}$ was assumed to be a true value of parameter θ; that is, $P(\theta)$, $P(\hat{\theta})$, or $P(\tilde{\theta})$ was used to generate responses to items. When the replications were replicated using $\hat{\theta}$ or $\tilde{\theta}$, all l_0- and l_z-values were determined by using $\hat{\theta}$ in Equations 2 and 3. When the replications were replicated using θ, all l_0- and l_z-values were computed using θ in order to determine a distribution without the presence of estimation errors. Although, in practice, θ is unknown, determining the distribution based on θ allows us to investigate the influence of $\hat{\theta}$ on the distribution of l_0 and l_z. The two ways of simulating, using a fixed and a stochastic test design, and the use of three types of ability, θ, $\hat{\theta}$, or $\tilde{\theta}$, resulted in 2 x 3 = 6 different methods of simulating the distribution of l_0 and l_z.

Results

In Table 3, the distribution of the significance probabilities of l_0 and l_z using a fixed and a stochastic design are given when θ, $\hat{\theta}$, or $\tilde{\theta}$ were used. To illustrate the distribution of the significance probabilities in Table 3, ten intervals are considered, each of length 0.10. The expected proportion of simulated test takers with a significance probability in a particular interval was 0.10; for example, it was expected that 10% of the simulated test takers have l_z-values with $0.4 < p^* \leq 0.5$. To test whether the distribution of significance probabilities approached the uniform distribution, Pearson's chi-squared tests, X^2, can be calculated, with $E(X^2) = 9$. Table 3 shows that using θ and using a fixed or stochastic test design resulted in an approximately uniform distribution of the significance probabilities for both l_0 and l_z. However, conditioning on $\hat{\theta}$ and using a fixed or stochastic test design resulted in an inappropriate approximation of the distribution of l_0 and l_z. For example, using $\hat{\theta}$ and a stochastic test design with l_z as fit index resulted in $X^2 = 111.3$, which is highly significant. Especially, the probabilities in the left tail were much too small. Using a stochastic design, only 0.5% of the simulated test takers attained a l_z-value with $0 \leq p^* \leq 0.1$. Using $\tilde{\theta}$ results were even worse; for example, using a stochastic design and l_z resulted in $X^2 = 213.8$. Again, the distributions of l_0 and l_z were inappropriately approximated, especially in the left tail of the distribution.

Study 3

Study 2 showed that approximating the distribution of l_0 and l_z for each θ-value using simulation methods was problematic. Study 3 was designed to explore the influence of inappropriately approximated distributions of l_0 and l_z on detecting aberrant response patterns. Furthermore, the relationship between the detection rate and the error in $\hat{\theta}$ was investigated.

Method

Seven datasets containing 200 nonfitting adaptive response patterns were constructed, with three types of aberrant response behavior. The first type was guessing on all the items in a test. This guessing model mimics the type of answering behavior studied empirically by Van den Brink (1977). He described test takers who took a multiple-choice exam without preparation, and the only purpose of taking the exam was to become familiar with the type of questions that would be asked. Because returning an almost completely blank answer sheet may focus attention on a test taker's ignorance, the test taker would randomly guess the correct answer on almost all items in the test. Also in a CAT, test-taking behavior to get familiar with the content of the items is realistic (Wainer, 1990). "Guessing" simulated test takers were simulated by randomly guessing the correct answer on each item in a CAT with a probability of 0.2 (assuming a test with five alternatives per item).

Second, response vectors with a two-dimensional θ parameter were simulated: item responses for a simulated test taker were based on different ability levels during the first and second halves of the test. Carelessness, fumbling, or memorization of some items can be the cause of noninvariant abilities. Two datasets containing response vectors with a two-dimensional ability parameter 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 $\rho = 0.8$ and $\rho = 0.6$ were used here to simulate the response patterns.

The third type of aberrant response vectors simulated were vectors with violations against local stochastic independence between the items of the test. 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. Four datasets were constructed with violations of the local independence assumption. These adaptive response vectors were simulated according to a model proposed by Jannarone (1986). The appendix describes the model in detail. 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 these nonfitting response patterns.

In order to determine the detection rates, the distributions of l_0 and l_z were simulated using a completely crossed 2 x 2 factorial design. The first factor was the test design: fixed or stochastic test designs were considered. The second factor was the type of θ used: θ or $\hat{\theta}$.

For classifying an original response pattern as aberrant, the distributions of l_0 and l_z were simulated for each simulated test taker by replicating 500 response patterns. The detection rate of a statistic is defined here as the proportion of detected nonfitting response patterns. An adaptive response vector was classified as nonfitting the 2PLM when the observed value of l_0 and l_z was below the critical value at level $\alpha = 0.05$, found in the simulated distribution of l_0 and l_z, respectively. It was shown in Study 2 that, conditional on true θ, the left tail probability was reasonably in agreement with the expected 0.10. However, when $\hat{\theta}$ was used, the left tail probability was shown to be inappropriately approximated. The influence on the detection rate is unknown, however. Furthermore, the detection rate of the l_z based on the decision rule $l_z \leq -1.65$ was determined, where $\hat{\theta}$ was used to determine l_z.

For every dataset, the mean absolute bias was determined as

$$MAB = \frac{1}{n}\sum_{i=1}^{n} | \hat{\theta}_i - \theta_i |$$

and the mean bias was calculated as

$$MB = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\theta}_i - \theta_i \right).$$

These variables were determined to investigate the trade-off between bias and detection rate.

Results

In Table 4 the detection rates are given. Table 4 shows that the detection rates for l_0 and l_z were higher for true θ than for $\hat{\theta}$ for all types of aberrant behavior. For example, for $\rho = 0.6$, using a stochastic design and l_z resulted in a detection rate of 0.34 using θ, whereas the detection rate decreased to 0.05 when $\hat{\theta}$ was used. When θ was used, the differences between a fixed and a stochastic test design were negligible. The largest difference occurred for $\delta = 2$, where the detection rate is 0.34 for a fixed design and 0.41 for a stochastic design, for both l_0 and l_z. Conditional on $\hat{\theta}$, the detection rate decreased dramatically. For example, for $\delta = 1$ and l_0, only 1% of the true aberrant simulated test takers were detected when a fixed design was used and 4% when a stochastic test design was used. When $l_z \leq -1.65$ was used, the detection rate was, as expected, also very low.

TABLE 4
Detection rates for several types of aberrant behavior, for a fixed and stochastic test design to simulate the null distribution of l_0 and l_z, and using $l_z \leq -1.65$

			θ				$\hat{\theta}$				
			Test Design				Test Design				$l_z \leq$ -1.65
			Fixed		Stochastic		Fixed		Stochastic		
	MAB	*MB*	l_0	l_z	l_0	l_z	l_0	l_z	l_0	l_z	
Guessing	2.134	-2.100	0.94	0.94	0.96	0.94	0.21	0.21	0.37	0.28	0.23
ρ = 0.6	0.432	0.012	0.31	0.31	0.35	0.34	0.02	0.02	0.09	0.05	0.03
0.8	0.327	-0.003	0.17	0.17	0.23	0.21	0.01	0.01	0.07	0.03	0.02
δ = -2	0.403	-0.382	0.27	0.27	0.27	0.31	0.00	0.00	0.01	0.00	0.00
-1	0.284	-0.245	0.12	0.12	0.11	0.14	0.00	0.00	0.01	0.00	0.00
1	0.300	0.245	0.13	0.13	0.17	0.18	0.01	0.01	0.04	0.01	0.01
2	0.567	0.564	0.34	0.34	0.41	0.41	0.01	0.01	0.04	0.01	0.01

Table 4 shows that the *MAB* varied from 0.284 (for $\delta = -1$) to 2.134 (for "guessing"), respectively. To have an idea of the seriousness of these biases, the *MAB* for the set with normal simulated test takers was also calculated. For the normal simulated test takers, the *MAB* = 0.196 was found. Thus, it can be concluded that for local stochastic dependence and two-dimensional abilities, *MAB* is only a little larger than for the normal simulated test takers, whereas for guessing on all items the *MAB* is dramatically different. Moreover, Table 4 shows that for high values of the *MAB*, the detection rate was also relatively high. For example, for guessing simulated test takers, the *MAB* = 2.134, and using $\hat{\theta}$ and l_z for a fixed and a stochastic design resulted in a detection rate of 0.21 and 0.28, respectively, whereas for $\delta = -1$ with a *MAB* = 0.284, the detection rate was 0.00 under the same conditions.

Discussion

To detect test takers with inappropriate test scores in a CAT, the use of person-fit statistics was investigated in this study. In particular, the distribution and the detection rates using theoretical and simulated distributions, and using θ and $\hat{\theta}$, were explored. Results showed that the distribution of l_0 and l_z differed across θ-levels and that the critical values described from the theoretical distribution differed substantially from the empirical distributions using $\hat{\theta}$. In a conventional testing context, Molenaar and Hoijtink (1990) proposed to condition on the total score. Also, Liou and Chang (1992) proposed an algorithm for determining the person-fit distribution for tests of 30 items or less using the total score as a sufficient statistic for θ. In an adaptive testing environment, however, it is quite problematic to construct such a distribution, because of the absence of a sufficient statistic for estimating $\hat{\theta}$ (even for the Rasch model, see Glas, 1989). An alternative is to simulate the distribution of the statistic for every θ-value. In practice, however, θ is unknown and $\hat{\theta}$ is used. The results of this study indicate that using $\hat{\theta}$ resulted in an inappropriate approximation of the distribution, especially in the tails. As a result of the inappropriate approximation in the left tail of the distribution, the detection rates for several types of aberrant response behavior were low.

However, from a practical point of view, the low detection rates may not be much of a problem, because the bias for local stochastic dependence and two-dimensional ability was only a little larger than for a set of normal simulated test takers. For guessing on all items, the bias was relatively high: *MAB* = 2.134, which also

resulted in a relatively high detection rate: using $\hat{\theta}$ and a stochastic test design, the detection rates for l_0 and l_z were 0.37 and 0.28, respectively.

In this study, the true item parameters were used. In practice, however, these parameters are unknown and have to be estimated. Therefore, further research is needed to investigate the influence of using estimated item parameters on the estimating process of θ and on the empirical distribution of l_z . However, note that using estimated item parameters probably will result in a more inaccurately approximated distribution of person-fit statistics and lower detection rates.

References

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

Candell, G. L. (1988). Application of appropriateness measurement to a problem in adaptive testing. Unpublished doctoral dissertation, University of Illinois, Champaign.

DeAyala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement, 16,* 327–343.

Drasgow, F., Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10,* 59–67.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59–79.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

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.

Drasgow, F., Levine, M. V., & Zickar, M. J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education, 9(1)*, 47–64.

Glas, C. A. W. (1989). *Contributions to estimating and testing Rasch models.* Doctoral Thesis, Enschede: University of Twente, Enchede, The Netherlands.

Glas, C. A. W., Meijer R. R., & van Krimpen, E. M. L. A. (1997). *Statistical tests for person misfit in computerized adaptive testing.* Unpublished manuscript, University of Twente, Enschede, The Netherlands.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (2nd ed.). Boston: Kluwer-Nijhoff Publishing.

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

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

Liou, M., & Chang, C. H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57*, 169–181.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18,* 311–314.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfitting response vectors. *Applied Psychological Measurement, 21,* 321–336.

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

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

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

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95–110.

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

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

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

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

Appendix: Modeling Local Dependence

Let x_i, a realization of X_i, be the response to item i. One of the models Jannarone (1986) presented was a conjunctive Rasch-model; local dependence between two subsequent items can be modeled by

$$P\left(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta\right) = \frac{\exp\left[\sum_{j=1}^{i+1} x_j\left(\theta - b_j\right) + x_i x_{i+1}\left(\theta - \xi_{i,i+1}\right)\right]}{1 + \exp[\theta - b_i] + \exp\left[\theta - b_{i+1}\right] + \exp\left[\sum_{j=1}^{i+1}\left(\theta - b_j\right) + \left(\theta - \xi_{i,i-1}\right)\right]}, \tag{A1}$$

where $\delta_{i,i+1}$ is a parameter modeling association between items i and $i + 1$. This model can be generalized to a conjunctive 2PLM, which can be written as

$$P\left(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta\right) \propto \exp\left[\sum_{j=1}^{i+1} x_j a_j\left(\theta - b_j\right) + x_i x_{i+1}\gamma_{i,i+1}\left(\theta - \xi_{i,i+1}\right)\right], \tag{A2}$$

where γ and ξ are parameters modeling association between items.

In this study the model used to simulate response vectors with local independence between all subsequent items was

$$P\left(X_i = x_i, X_{i+1} = x_{i+1} \mid \theta\right) \propto \exp\left[\sum_{j=1}^{i+1} x_j a_j\left(\theta - b_j\right) + x_i x_{i+1}\delta_{i,i+1}\right], \tag{A3}$$

where $\delta_{i,i+1}$ is a parameter modeling association between items. the four possible realizations of (X_i, X_{i+1}) have the following probabilities

$$P\left(X_i = 0, X_{i+1} = 0\right) \propto 1,$$

$$P\left(X_i = 1, X_{i+1} = 0\right) \propto \exp\left[a_i\left(\theta - b_i\right)\right],$$

$$P\left(X_i = 0, X_{i+1} = 1\right) \propto \exp\left[a_{i+1}\left(\theta - b_{i+1}\right)\right], \text{ and}$$

$$P\left(X_i = 1, X_{i+1} = 1\right) \propto \exp\left[a_i\left(\theta - b_i\right) + a_{i+1}\left(\theta - b_{i+1}\right) + \delta_{i,i+1}\right].$$

The conditional probability of a correct response to item i +1 given a correct response to item i can be written as

$$P\left(X_{i+1} = 1 \mid X_i = 1, \theta\right) = \frac{P\left(X_i = 1, X_{i+1} = 1 \mid \theta\right)}{P\left(X_i = 1, X_{i+1} = 0 \mid \theta\right) + P\left(X_i = 1, X_{i+1} = 1 \mid \theta\right)}$$

$$= \frac{\exp\left[a_i\left(\theta - b_i\right) + a_{i+1}\left(\theta - b_{i+1}\right) + \delta_{i,i+1}\right]}{\exp\left[a_i\left(\theta - b_i\right)\right] + \exp\left[a_i\left(\theta - b_i\right) + a_{i+1}\left(\theta - b_{i+1}\right) + \delta_{i,i+1}\right]}, \tag{A4}$$

and the probability of a correct response to item $i + 1$ given an incorrect response to the previous item can be written as

$$P\left(X_{i+1} = 1 \mid X_i = 0, \theta\right) = \frac{\exp\left[a_{i+1}\left(\theta - b_{i+1}\right)\right]}{1 + \exp\left[a_{i+1}\left(\theta - b_{i+1}\right)\right]}, \tag{A5}$$

which is the 2PLM. The conditional probabilities in Equation A4 and the 2PLM were used to simulate the responses to the items when the items are local-stochastic dependent.