DOCUMENT RESUME

ED 421 547 TM 028 873

AUTHOR	Glas, Cees A. W.; Meijer, Rob R.; van Krimpen-Stoop, Edith M. L. A.
TITLE	Statistical Tests for Person Misfit in Computerized Adaptive Testing. Research Report 98-01.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
PUB DATE	1998-00-00
NOTE	28p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Foreign Countries; *Responses; *Testing Problems
IDENTIFIERS	Local Independence (Tests); *Person Fit Measures; Power (Statistics); Two Parameter Model

ABSTRACT

Recently, several person-fit statistics have been proposed to detect nonfitting response patterns. This study is designed to generalize an approach followed by Klauer (1995) to an adaptive testing system using the two-parameter logistic model (2PL) as a null model. The approach developed by Klauer is described, and some difficulties in generalizing it to a computerized adaptive testing model are explored. Alternative approaches are presented, and the results of a power study and the consequences for person-fit measurement in adaptive testing situations are discussed. The first part of the simulation concerns the ability to ignore the process that causes missing data and the impact of the fact that data in adaptive testing are not observed at random. Then the power of the proposed three tests is studied. Results suggest that tests against local independence in the 2PL have little power, but that the power for testing against nonvariant abilities is relatively larger but still low. (Contains 4 tables and 18 references.) (SLD)

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

ED 421 547

TM

Statistical Tests for Person Misfit in Computerized Adaptive Testing

Research Report
98-01

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

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Melissen

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

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

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

□ Minor changes have been made to improve reproduction quality.

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

TM028873

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

ERIC
Full Text Provided by ERIC

Statistical Tests for Person Misfit in Computerized Adaptive Testing

Cees A. W. Glas

Rob R. Meijer

Statistical Tests for Person Misfit in Computerized Adaptive Testing

Summary

In case both the null and alternative model are exponential family models, an unbiased uniformly most powerful test can be based on the minimal sufficient statistics of the alternative model (Lehmann, 1986). Strictly speaking, this approach is not valid in the framework of the two parameter logistic model (2PL), since this model does not define an exponential family. However, the notion of using statistics related to the parameters of an alternative model as the basis of a test is intuitively appealing, even though a uniformly most powerful test probably does not exist. Below, $\hat{\theta}$ will be the weighted ML estimate (Warm, 1989). In an adaptive testing environment, the following alternatives to the 2PL will be considered.

1 Non-invariant abilities: in the alternative model it is assumed that the 2PL is valid during the whole testing session, but that the respondent's ability parameter changes during test-taking. More specifically, it will be assumed that a person has two person parameters, say θ_1 and θ_2; the first parameter governs the responses on the first half of the test, the second parameter governs the second part. The test will be based on the statistic $T=(T_1,T_2)$, where T_1 and T_2 are the unweighted sum scores on the first and second part, respectively. Of course, splitting the test into two halves is arbitrary, the practitioner can use any non-trivial, suitable partition of the test.

2 Lack of local stochastic independence: it is assumed that the probability of a correct response on an item is augmented by a previous correct response. This is modeled by introducing a transfer parameter δ, such that the probability of two consecutive responses x_i and x_{i+1} is given by

$$P(X_i=x_i, X_{i+1}=x_{i+1}|\theta,\alpha,\beta,\delta) \propto \exp\left[x_i(\alpha_i\theta-\beta_i)+x_{i+1}(\alpha_{i+1}\theta-\beta_{i+1})+x_i x_{i+1}\delta\right]$$

The test will be based on the statistic $L=\sum_{i=1}^{k-1} X_i X_{i+1}$.

3 Person-specific item parameters. The test will be based on classifying the probability of the observed response pattern $\mathbf{x}=(x_1,\ldots,x_i,\ldots,x_k)$, $P\left(\mathbf{X}=\mathbf{x}\middle|\hat{\theta},\alpha,\beta\right)$, among all possible values $P\left(\mathbf{Y}=\mathbf{y}\middle|\hat{\theta},\alpha,\beta\right)$, where **Y** is a possible response pattern. The statistic will be denoted by X.

Significance probabilities for the tests based on *T*, *L* and *X* are calculated by simulating their respective distributions under the null-hypothesis. Simulation studies were carried out to assess the power of the proposed tests. One important point here is that the item administration design is contingent on the responses given. It is well known (see, for instance, Mislevy, 1986, and Glas, 1988) that the stochastic nature of the design does not threaten the consistency of the estimates. This is due to the fact that in adaptive testing unobserved responses are missing at random (MAR, Rubin, 1976). Therefore, for the estimation of model parameters the process that causes missing data can be ignored. However, in adaptive testing the data are not observed at random (OAR, Rubin, 1976). As a result, the set of possible response patterns given the design is only a subset of the possible response patterns in a fixed design. The impact of these restrictions was one of the focuses of the simulation studies performed.

Results

1 For proper simulation of the distributions of *T*, *L* and *X* the stochastic nature of the design has to be taken into account.

2 The accuracy with which the distributions of *T*, *L* and *X* can be simulated seems to depend on the model considered (RM versus 2PL) and the distribution of item parameters in the item bank.

3 The power of the tests under the alternative model is very low. Only the power of the test based on *T* under the non-invariant abilities model reaches a non-trivial value.

4 The error in the estimation of θ is not much increased by the model violations introduced, so the estimates are quite robust.

It must be stressed that the second and third result are preliminary: especially the simulation algorithm for the distribution of *T* is not yet completely satisfactory.

Introduction

Person-fit analysis (or appropriateness measurement) is concerned with the detection of response patterns that are unusual (nonfitting) given what is expected under an item response theory (IRT) model. When a person's response pattern is nonfitting according to the model, it is questionable whether the test score is an adequate description of the trait being measured. Typical forms of aberrant response behavior are guessing and cheating (e.g., as a result of item or test preview) which may result in spuriously high or spuriously low test scores.

Recently, several person-fit statistics have been proposed to detect nonfitting response patterns (e.g., Drasgow, Levine, & Williams, 1985, Molenaar & Hoijtink, 1990, Meijer, 1994). A popular approach is to determine the (log)likelihood of a response pattern and to classify response patterns with an extreme likelihood statistic as nonfitting. Note that following this approach, it is only tested whether a response pattern is unlikely given the model: the type of nonfitting response behavior is not identified. In the context of the Rasch (1960) model, Klauer (1995) addressed this problem by specifying alternative models for response behavior and testing the null hypothesis of model conform behavior against the specified alternative model.

Klauer (1995) considered three types of nonfitting response behavior: violations of local stochastic independence between the items, violations of invariant ability across subtests of the total test, and person-specific item discrimination. For these three model violations, person tests were constructed and a power analysis was performed. A limitation of the Klauer (1995) study is that it was restricted to fixed-format tests and to the Rasch model. The present study was designed to generalize the approach followed by Klauer (1995) to an adaptive testing setting using the two-parameter logistic model (2PL) as a null model.

This report is organized as follows. First, the approach by Klauer is presented in Section 2. Second, some problems with respect to the generalization of his approach to an adaptive testing situation are discussed. An alternative approach is presented in Section 3. Finally, the results of a power study and the consequences for person-fit measurement in adaptive testing situations are discussed in Section 4.

The Rasch model and Uniformly Most Powerful Tests

To test normal response behavior against a specified alternative model, Klauer (1995) constructed uniformly most powerful (UMP) tests using the Rasch model. Because UMP tests are central in this approach the principle of these tests will be characterized first.

The Rasch model is based on the assumptions of local stochastic independence, unidimensionality of the latent trait θ, sufficiency of a respondent's unweighted sum score and some technical assumptions beyond the scope of the present paper. Let β_i denote the difficulty for item i, where i=1, ...,k. Then the probability of a correct response on item i given θ, according to the Rasch model (RM) can be written as

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

It is well-known that the RM belongs to the exponential family of distributions.

The general derivation of a UMP test in an exponential family model proceeds as follows. Let $\mathbf{x}=(x_1,\ldots,x_k)$ be a realization of $\mathbf{X}=(X_1,\ldots,X_k)$ and let ξ and η be two parameters. The likelihood of the two-parameter exponential family can then be written as,

$$P(\mathbf{X} = \mathbf{x}|\xi,\eta) = \mu(\xi,\eta)h(\mathbf{x})\exp\{\eta T(\mathbf{x}) + \xi R(\mathbf{x})\}, \tag{2}$$

where $\mu(\xi,\eta)$ and $h(\mathbf{x})$ are normalizing functions and $T(\mathbf{x})$ and $R(\mathbf{x})$ sufficient statistics. Considering the factorization criterion (e.g., Lindgren, 1993, p.231), the statistics $T(\mathbf{x})$ and $R(\mathbf{x})$ are minimal sufficient statistics for the parameters ξ and η.

A uniformly most powerful (UMP) test ϕ for testing $H_0{:}\eta=\eta_0$ against $H_1{:}\eta\neq\eta_0$ for some parameter η in exponential family of distributions can be derived as follows (Lehmann, 1959). Let the vector $\mathbf{X}$ be distributed according to the exponential family of distributions, i.e. Equation 3 is generalized to

$$P(\mathbf{X} = \mathbf{x}|\theta,\eta) = \omega(\theta,\eta)\exp\left\{\eta T(\mathbf{x}) + \sum_{i=1}^{k}\xi_i R_i(\mathbf{x})\right\}. \tag{3}$$

The UMP test for testing the hypothesis $H_0{:}\eta=\eta_0$ vs. $H_1{:}\eta\neq\eta_0$ with size α is given by $\phi(\mathbf{x})$, satisfying

$$\phi(\mathbf{x})=\begin{cases}0 & \text{when } c_1(r)<T(\mathbf{x})<c_2(r),\\ \gamma_i(r) & \text{when } T(\mathbf{x})=c_i(r) \text{ for i = 1,2,}\\ 1 & \text{otherwise,}\end{cases} \tag{4}$$

where $R(\mathbf{x})=r$ is given and the functions $c(r)$ and $\gamma(r)$ are determined by solving the following equations

$$\begin{aligned}&E\left[\phi(\mathbf{X})\middle|r,\eta=\eta_0\right]=\alpha,\\ &E\left[T(\mathbf{X})\phi(\mathbf{X})\middle|r,\eta=\eta_0\right]=\alpha E\left[T(\mathbf{X})\middle|r,\eta=\eta_0\right].\end{aligned} \tag{5}$$

So, H_0 is accepted when the test statistic T takes values between c_1 and c_2, and randomization is applied in the case of T=c_i, for i=1,2.

This approach can be generalized to the framework of the RM as follows. Let x_i be the score on a dichotomously scored item i, where a 1 denotes a correct or keyed response and 0 otherwise. Further, $\mathbf{x}=(x_1,\ldots,x_k)$ is a realization of a vector $\mathbf{X}=(X_1,\ldots,X_k)$ of responses to k items, θ is a person parameter, and β_i are item parameters. Then, the Rasch model as defined in Equation 1 can be written as

$$\begin{aligned}P(\mathbf{X}=\mathbf{x}|\theta)&=\underbrace{\left\{\prod_{i=1}^{k}\left[1+\exp\{\theta-\beta_i\}\right]^{-1}\right\}}_{\mu(\theta)}\times\underbrace{\exp\left\{-\sum_i\beta_i x_i\right\}}_{h(\mathbf{x})}\times\exp\left\{\theta\sum_i x_i\right\}\\ &=\mu(\theta)h(\mathbf{x})\exp\{\theta R(\mathbf{x})\},\end{aligned} \tag{6}$$

where $R(\mathbf{x})=\sum_i x_i$ is the raw score of the response vector $\mathbf{x}$; R is a minimal sufficient statistic for the ability parameter θ.

Klauer (1995) constructed uniformly most powerful person fit tests for three specific alternative models in a fixed format setting; these models are generalizations of the RM testing the assumptions of local stochastic independence, uni-dimensional abilities across subtests of the total test and the invariance of item discriminations across the items of the test. These alternative models are member of the two-dimensional exponential family. Below, three different alternative models will be introduced.

Non-Invariant Abilities across Subtests of the Total Test

An assumption of the RM is the invariance of the ability parameter θ. When the test is split into two subtests A_1 and A_2, the abilities on the two subtests will be the same, that is, $\theta_1=\theta_2$. The model used as an alternative model with invariant abilities contains a multidimensional ability parameter. Consider the simplest (two-dimensional) case, where the test is divided into two subtests A_1 and A_2 and the ability parameter can be written as $\theta=(\theta_1,\theta_2)$; thus the examinee has ability θ_1 on subtest A_1 and ability θ_2 on subtest A_2. Let $R_j(\mathbf{x})$ the raw score obtained from the subtest A_j for $j=1,2$, $R(\mathbf{x})$ the raw score of the total test, and $\mu(\theta,\eta)$ and $h(\mathbf{x})$ suitable functions. Then, after some algebra and for known item parameters, the model with a multidimensional ability parameter can be written as,

$$P(\mathbf{X}=\mathbf{x}|\theta,\eta)=\mu(\theta_2,\theta_1-\theta_2)h(\mathbf{x})\exp\{(\theta_1-\theta_2)R_1(\mathbf{x})+\theta_2R(\mathbf{x})\}. \tag{7}$$

$R_1(\mathbf{x})$ and $R(\mathbf{x})$ are sufficient statistics for $\eta=\theta_1-\theta_2$ and $\theta=\theta_2$ respectively. When $\eta=0$, that is $\theta_1=\theta_2$, the model becomes the RM. Positive values of η indicate that $\theta_1>\theta_2$, this can be the case when an examinee has pre-knowledge about one subtest and therefore the score on that subtest is much higher than the score on the rest of the test. Thus, the parameter η describes the size and direction of the differences of the ability parameters. Nonfitting response behavior like guessing, carelessness, sleeping and fumbling can also be the cause of non-invariant abilities.

Local Stochastic Dependence

Another assumption made for the RM is the assumption of local stochastic independence. The following model of Jannarone (1986) represents a violation of local stochastic independence

$$P(\mathbf{X}=\mathbf{x}|\theta,\eta)=\mu(\theta,\eta)h(\mathbf{x})\exp\left\{\eta\sum_{i=1}^{k-1}x_ix_{i+1}+\theta R(\mathbf{x})\right\}, \tag{8}$$

where again $\mu(\theta,\eta)$ and $h(\mathbf{x})$ are normalizing functions. Therefore $L(\mathbf{X}) = \sum_{i=1}^{k-1} X_i X_{i+1}$ is a sufficient statistic for parameter η. The RM is obtained for η=0, positive values of η occur when for example an item provides extra information that is useful for answering the next item. Thus parameter η describes the size and direction of the violation of the local stochastic independence assumption.

Person-Specific Item Discrimination

The third assumption of the RM is that the item discriminations are invariant across the items of the test. When η is used as a parameter indicating the overall level of the item discrimination, an alternative model is

$$P(\mathbf{X}=\mathbf{x}|\theta,\eta) = \underbrace{\prod_{i=1}^{k}\left[1+\exp\{\eta(\theta-\beta_i)\}\right]^{-1}}_{\mu(\theta,\eta)} \exp\left\{\sum_{i=1}^{k} x_i\eta(\theta-\beta_i)\right\}, \tag{9}$$

where again the item parameters are considered known.

For η=1 the RM is established, when 0<η<1 the overall level of item discrimination is less then the overall discrimination level in the RM, and when η>1 the overall level of item discrimination is higher than the overall level of discrimination in the RM. Thus, the parameter η describes the size and direction of the violation of invariant item discriminations.

Rewriting Equation 8 to the form of Equation 2, results in

$$P(\mathbf{X}=\mathbf{x}|\theta,\eta) = \mu(\theta,\eta)\exp\left\{-\eta\sum_{i=1}^{k}\beta_i x_i + \theta R(\mathbf{x})\right\}, \tag{10}$$

where η is absorbed in θ, i.e. $\theta := \eta\theta$. Therefore, the statistics $M(\mathbf{X}) = -\sum_{i=1}^{k}\beta_i X_i$ and $R(\mathbf{x})$, are sufficient statistics for the person specific parameters η and θ.

ERIC
Full Text Provided by ERIC

Limitations for the Two Parameter Logistic Model and Adaptive Testing

The two-parameter logistic model (2PL) is a more general model than the RM. Let α_i be the item discrimination of item i. Then the probability of a correct response to item i according to the 2PL is given by

$$P(X_i = 1|\theta) = \frac{\exp\{\alpha_i(\theta-\beta_i)\}}{1+\exp\{\alpha_i(\theta-\beta_i)\}}, \tag{11}$$

and the joint distribution of the observed response pattern **x** can be written as

$$\begin{aligned} P_i(\mathbf{X}=\mathbf{x}|\theta) &= \left[\prod_i \left(1+\exp\{\alpha_i(\theta-\beta_i)\}\right)^{-1}\right]\left[\exp\left\{-\sum_i \alpha_i\beta_i x_i\right\}\right]\left[\exp\left\{\theta\sum_i \alpha_i x_i\right\}\right] \\ &= \mu(\theta)h(\mathbf{x})\exp\{\theta W(\mathbf{x})\}, \end{aligned} \tag{12}$$

where $W(\mathbf{X})$ is the weighted score according to **x** with α_i as weights, where α_i are unknown item parameters to be estimated. In the 2PL, the total weighted score W is not a sufficient statistic for the ability parameter θ, because W depends on α_i. Further, the conditional distribution of the sample, given W, depends on the unknown item parameters α_i. For known item parameters α_i and β_i, W is a minimal sufficient statistic, although Andersen (1977) has shown that only for certain values of α_i, conditioning on W leads to a non-trivial likelihood. In that case, W can be used for the construction of UMP tests.

In an adaptive testing situation it is difficult to construct sufficient statistics. Adaptive testing is the limiting case of multistage testing, where each test consists only one item. Glas (1988) showed for the RM in a two-stage testing design that the conditional probability of the sample **X** given the raw score R, is not independent of θ and therefore R is not a sufficient statistic for parameter θ.

Because the absence of a minimal sufficient statistic, it is not sure whether a UMP test for the RM exists in an adaptive testing situation. Even though a uniformly most powerful test probably does not exist, the notion of using statistics related to the parameters of an alternative model as the basis of a test is intuitively appealing. This point will be resumed below.

ERIC Full Text Provided by ERIC

Let the design, **I**, be the observed sequence an examinee responded to (a vector of item numbers). In an adaptive test, item selection is based on the responses on previous items. Therefore, the observed design **I** is dependent on the observed vector of responses **X**. The missing data are the responses on all the non-selected items. Ignorability of the process that causes missing data concerns the consistency of the estimates of the model parameters. In adaptive testing unobserved responses are missing at random (MAR, Rubin, 1976, Mislevy, 1986), therefore, for the estimation of model parameters the process that causes missing data can be ignored. However, the data are not observed at random (OAR, Rubin, 1976). This means that the set of possible response patterns given the design, $\{\mathbf{X}|\mathbf{I}\}$, is only a subset of the possible response patterns in a fixed design. In a fixed design the probability of the observed sequence of items, $g(\mathbf{I})$, equals one, because every examinee responded to the same items. As a result $f(\mathbf{X},\mathbf{I}) = f(\mathbf{X}|\mathbf{I})g(\mathbf{I}) = f(\mathbf{X}|\mathbf{I})$, where $f(\mathbf{X},\mathbf{I})$ is the joint probability of **X** and **I**, and $f(\mathbf{X}|\mathbf{I})$ the conditional probability of **X** given **I**. In adaptive testing $g(\mathbf{I}) < 1$, because the items administered are different for every examinee, therefore, $f(\mathbf{X},\mathbf{I}) = f(\mathbf{X}|\mathbf{I})g(\mathbf{I}) \neq f(\mathbf{X}|\mathbf{I})$.

Method

This simulation study consists of two parts. The first part concerns the ignorability of the process that causes missing data and the impact of the fact that the data in adaptive testing are not OAR. Furthermore, the power of the proposed tests was investigated.

Model-Fitting Response Vectors

For each simulee, responses to dichotomous items were generated in the following way. The procedure starts with randomly drawing a true ability θ from the standard normal distribution. $P_i(\theta)$ was computed according to the 2PL given in Equation 11. Then, a random number y was drawn from U(0,1) and when $P_i(\theta) > y$ the response was set to 1, 0 otherwise. The first three items of the test were selected with item parameter β around 0, and the ability parameter was estimated using weighted maximum likelihood estimation (Warm, 1989) based on the responses to the first three items. The next item selected was the item with maximum information given the estimated ability $\hat{\theta}$. Again, a response was generated, the ability was estimated on the basis of previous item responses and the item with maximum information was selected. This procedure was repeated until the test consisted of 20 items.

Types of Nonfitting Response Vectors

Two different types of nonfitting response vectors were simulated. The first type of nonfitting vectors were generated with a two-dimensional ability parameter. Two values θ_1 and θ_2 were drawn from the bivariate standard normal distribution. The correlations $\rho=0.6$ and $\rho=0.8$ were used. During the first half of the test $P_i(\theta_1)$ was used and during the second half $P_i(\theta_2)$ was used to generate the responses.

The second type of nonfitting patterns were generated with violations against local stochastic independence. Item scores were generated according to a generalized version of the model proposed by Jannarone (1986). Let $\delta_{i,i+1}$ be a parameter modeling association between item, that is, the model can be written as

$$P\left(X_i = x_i, X_{i+1} = x_{i+1} | \theta\right) \propto \exp\left[x_i(\alpha_i\theta - \beta_i) + x_{i+1}(\alpha_{i+1}\theta - \beta_{i+1}) + x_i x_{i+1}\delta_{i,i+1}\right]. \quad (13)$$

The fact that $\delta_{i,i+1}$ models association is verified as follows. From the model in Equation 13, the four possible realizations of (X_i,X_{i+1}) have the following probabilities

$$P\left(X_i = 0, X_{i+1} = 0|\theta\right) \propto 1,$$
$$P\left(X_i = 1, X_{i+1} = 0|\theta\right) \propto \exp[\alpha_i\theta - \beta_i],$$
$$P\left(X_i = 0, X_{i+1} = 1|\theta\right) \propto \exp[\alpha_{i+1}\theta - \beta_{i+1}], \text{ and}$$
$$P\left(X_i = 1, X_{i+1} = 1|\theta\right) \propto \exp\left[(\alpha_i\theta - \beta_i) + (\alpha_{i+1}\theta - \beta_{i+1}) + \delta_{i,i+1}\right].$$

Above it was stated that a test could be based on statistics related to the parameters of an alternative model. One could, for instance, construct a Lagrange Multiplier test (Aitchison and Silvey, 1958), which is based on the derivative of the logarithm of the right hand side of Equation 13 with respect to $\delta_{i,i+1}$. The test boils down to evaluating the difference between L and its expected value relative to its variance, all under the null model $\delta_{i,i+1} = 0$. Since computing the variance is quite complicated here, this approach will not be pursued, and we will proceed as follows. The conditional probability of item i+1, given the response to item i can be written as

$$P\left(X_{i+1}=x_{i+1}\middle|X_i=x_i,\theta\right)=\frac{P\left(X_i=x_i,X_{i+1}=x_{i+1}|\theta\right)}{P\left(X_i=x_i,X_{i+1}=1|\theta\right)+P\left(X_i=x_i,X_{i+1}=0|\theta\right)}. \qquad (14)$$

Given the response to item i, the probability of a correct response to item i+1, $P\left(X_{i+1}=1\middle|X_i,\theta\right)$, was computed. If $P\left(X_{i+1}=1\middle|X_i,\theta\right)>y$, where y~$U(0,1)$, the response to item i+1 was set to 1, 0 otherwise. For generating response vectors, the values δ=0.2 and δ=0.4 were used.

Generating distributions under H_0

For each response vector the realizations of *T*, *L* and *X* were computed. To generate the distribution of the proposed statistics under the null model for each simulee, *m* response vectors (replications) were generated according to the null model (RM or 2PL). By generating the distribution of the statistics under the null model, it was possible to determine whether the observed response vector was classified as fitting or nonfitting.

The *m* replications were generated using two different designs: a fixed design and a stochastic adaptive design. First, the replications were generated according to a fixed design. That is, for each replication the same items and the same test length were used. Let **I** be the observed sequence of items the simulee responded to. In the fixed design approach the conditional distribution of the observed response pattern **X** given the observed sequence of items **I**, $f(\mathbf{X}|\mathbf{I})$, was generated. Second, in the stochastic design approach the replications were generated according to a stochastic adaptive design: given $\hat{\theta}$, *m* response patterns were generated according to the adaptive procedure described above. The weighted conditional distribution $f(\mathbf{X}|\mathbf{I})g(\mathbf{I})$ with $g(\mathbf{I})$ the probability of **I**, was taken into account.

For determining the distribution of *T*, the probability of each possible combination of (T_1,T_2) was determined as

$$P(t_1,t_2)=P(T_1=t_1,T_2=t_2)=\tfrac{1}{m}\left[\text{number of replications with }(t_1,t_2)\right].$$

The sum of the probabilities of all replications with $P(X_1, X_2) < P(t_1, t_2)$, and $P(X_1, X_2) > P(t_1, t_2)$ were determined, where

$$\begin{aligned} S_0 &= \textstyle\sum_{t_1,t_2} P(t_1,t_2) \quad \text{for all replications with } P(X_1,X_2) < P(t_1,t_2), \\ S_1 &= \textstyle\sum_{t_1,t_2} P(t_1,t_2) \quad \text{for all replications with } P(X_1,X_2) > P(t_1,t_2), \text{ and} \\ S_2 &= 1 - H_0 - H_1. \end{aligned}$$

Then the probability $p^* = S_1 + uS_2$ was determined, with $u \sim U(0,1)$. The observed response vector was classified as nonfitting the model when $p^* \geq 1 - \alpha$ or when $p^* \leq \alpha$.

For extreme values of L and X, that is $L>c_L$ and $X>c_X$, with c_L and c_X the (1-α)100%-percentile of the m generated values of L and X respectively, the observed response vector **x** can be classified as nonfitting the model.

The power of the test statistics T, L and X was defined as the percentage of response patterns classified as non-fitting. The mean absolute bias was defined as the mean of the absolute distances between θ and $\hat{\theta}$; MAB(θ) = $\sum_j \frac{1}{m}\left|\theta_j - \hat{\theta}_j\right|$, with j=1,...,$m$. The MAB was taken into account to investigate the differences between the true ability θ and the final estimated ability $\hat{\theta}$.

An empirical item bank of 1,000 items calibrated using the 2PL was used, where the item parameters were estimated from real examination data in the Netherlands. For comparing the distributions using a fixed and a stochastic design for the Rasch model, an infinite item bank with items fitting the RM was used, that is, it was assumed that the optimal item was always present. The abilities were standard normally distributed: θ~N(0,1). Let n be the number of response vectors simulated (simulees), and let m be the number of replications per simulee generated (replications).

Results

Distribution of the Statistics

The first study was designed to assess the adequacy of using fixed design generation of statistics in the case of stochastic design data. In Table 1, for the Rasch model with n=300 and m=1,000, and using an infinite item pool, the distribution of significance probabilities of the statistics

T, L, and X under the fixed and the stochastic generating approach are shown. Table 1 also shows the values of Pearsons' chi-squared test X^2. The expected percentages in all the cells are 10%, and the expected value of X^2 is 9 (degrees of freedom). It can be seen that generation of the distribution of the statistics using a fixed design does not give a good description of the distribution under the null-hypothesis. This results in high X^2 values: for all three statistics, $X^2 >> 9$. In the case of the stochastic design, the distribution of X and L are satisfying according to the X^2 values of 9.133 and 3.467, respectively. However, the distribution of T is not completely satisfactory. The probability of 0.04 in the left tail is small compared to the expected probability of 0.10; the result of this is a high X^2 value of 24.8. Note, that the probabilities of the statistics in the tails of the distribution are most important, because the tails contain the fit values of aberrant response vectors.

Table 2 shows the distribution of the statistics for the 2PL with n=500 and m=1,000. Again the fixed design does not give satisfactory results; the X^2 values were highly significant for all three distributions (not tabulated). This can also be seen in the tails: the left tail probabilities of X, L, and T are 0.036, 0.050, and 0.032, respectively, the right tail probabilities of X and L, 0.033 and 0.007, respectively. Thus, these probabilities are substantially smaller than 0.10. Furthermore, note that the right-tail probability of T is 0.298, which is too high. Table 2 also shows the distribution of the statistics under the stochastic design. In the case of the stochastic design for X, L, and T the probabilities in the left tail were also too small: 0.045, 0.070, and 0.045. The probability in the right tail of the distribution of T was again too high: 0.185.

The estimated ability $\hat{\theta}$ is a point estimator and the precision of the estimation is different for every $\hat{\theta}$. When it was assumed that ability is normally distributed with mean $\hat{\theta}$ and variance $I(\hat{\theta})$, $\hat{\hat{\theta}} \sim N(\hat{\theta}, I(\hat{\theta}))$, the uncertainty of $\hat{\theta}$ is taken into account. When for the generation of the distribution m values of $\hat{\hat{\theta}}$ were randomly drawn from $N(\hat{\theta}, I(\hat{\theta}))$, the distributions of the statistics did not improve. These results are also shown in Table 2: again, the X^2 values were highly significant (not tabulated). The probabilities in the left tail for both X and T were too small, 0.030 and 0.035, respectively, and the probability in the right tail for T (0.175) was too high.

Power Studies

The second study was designed to investigate the power of the statistics. Table 3 shows the power of the statistics under the violations of the 2PL and the MAB for a 20 item test. Generally, there was hardly any power for all three statistics. This can be seen in the first column of Table 3, for example, for ρ=0.8 the test against multi-dimensional abilities *T*, classified only 8% of the true nonfitting response as aberrant; thus, the power of *T* was 0.08. However, for ρ=0.6 the statistic *T* classified 24% of the true nonfitting response vectors as aberrant. The test against local dependence *L*, detected only 4% of the true nonfitting simulees for δ=0.4. However, the estimated ability $\hat{\theta}$, was rather robust against these model violations. The values of the MAB under the model violations were not much higher than under the null model. The MAB increased with 0.11 (from 0.34 to 0.45) for the multi-dimensional abilities with ρ=0.6.

Table 4 shows the power of the statistics under the model violations, and MAB for a 40 item test. Again, the test statistics had hardly any power. However, increasing the test length from 20 to 40 items resulted for a two-dimensional ability parameter in a 23% and 25% detection rate for ρ=0.8 and ρ=0.6 respectively. And the test against local dependence detected 27% of the true nonfitting simulees for δ=1. Again $\hat{\theta}$was rather robust against the violations of the model; in the worst case the MAB increased with 0.17 (from 0.25 to 0.42).

Discussion

The results with respect to the power of a person-fit test in a CAT suggests that tests against local independence in the 2PL have little power, whereas the power for testing against noninvariant abilities was relatively larger but still low. The low power in CAT compared to conventional testing is not surprising because the overall variance of the item difficulties in CAT is smaller than in conventional testing and it has been shown (Reise and Due, 1991) that the smaller the variance of the item difficulties, the lower the power to detect nonfitting respondents.

From a practical point of view the relatively low power of a person test may not be much of a problem if the θ value is robust against some specific violations of the model. In that case, it is not necessary to classify a simulee as nonfitting because the θ provides a reasonable good description of the testing behavior. In the context of a conventional test administration, there is some evidence that low power goes together with small bias of θ. Using simulated data, Meijer and Nering (1996) showed that nonfitting responses may lead to biased estimation of θ, but that the bias of θ depends on the θ level and the type and severeness of the nonfitting response behavior. For example, they found that, for a 50 item test with 70% of the items fitting the 2PL, and random response behavior on 30% of the items, the bias was approximately 0 for θ=0, whereas for θ=2 and θ=–2 the absolute bias was approximately .9. The corresponding detection rates were .25 for θ=0 and .55 for θ=2 and θ=–2. Thus, in that study the low detection rates at θ=0 did not seem to be much of a problem because the bias was approximately 0. If similar results apply for CAT applications, it may be interesting to investigate the trade off between test power and robust estimation of θ.

Nering (1995) generated nonfitting responses by changing a 1 score into a 0 score and vice versa. Results showed that only for nonfitting response behavior in the beginning of the test, the detection rates were acceptable. Using a .05 two-tailed error rate, when the location of the response manipulation occurred within the first 5 items the detection rate was between .30 and .70 .

One limitation of the present study was that true item parameters were used. In a real testing situation item parameters must be estimated, and additional research is needed to determine what influence this estimation process will have on $\hat{\theta}$ values estimated from different procedures when nonfitting responses are present.

References

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics 29*, 813-828.

Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

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

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

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

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. (1996). Ability estimation for nonfitting item score pattems. *Applied Psychological Measurement* (accepted).

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

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

Nering, M.L. (1995). The distribution of person-fit within the CAT environment. *Applied Psychological Measurement* (in press).

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications*. New York: Springer-Verlag.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley.

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

Reise, S. P. & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217-226.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581-592.

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

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, Recent Developments and Applications*. New York: Springer-Verlag.

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

Author Note

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

Table 1

Distribution of significant probabilities for the fixed and stochastic computation of the null-distribution of *X*, *L* and *T*.

For RM, *n*=300, *m*=1,000, *k*=20.

		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	**X^2**
fixed design												
	X	.183	.123	.187	.100	.113	.077	.070	.063	.037	.047	74.467
	L	.027	.030	.063	.163	.217	.217	.130	.107	.047	.000	169.933
	T	.037	.037	.050	.063	.070	.093	.070	.150	.150	.280	153.333
stochastic design												
	X	.087	.070	.087	.107	.090	.113	.127	.127	.093	.100	9.133
	L	.103	.097	.097	.097	.103	.123	.107	.103	.080	.090	3.467
	T	.040	.113	.070	.100	.113	.110	.113	.153	.103	.083	24.800

Table 2

Distribution of significant probabilities for the fixed and stochastic computation of the null-distribution of *X*, *L* and *T*.

For 2PL, *n*=500, *m*=1,000, *k*=20.

		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
fixed design											
	X	.036	.065	.090	.136	.187	.187	.123	.083	.061	.033
	L	.050	.071	.106	.142	.164	.168	.151	.100	.043	.007
	T	.032	.049	.053	.065	.049	.082	.104	.116	.154	.298
stochastic design											
	X	.045	.070	.095	.110	.138	.108	.128	.085	.125	.093
	L	.070	.130	.090	.148	.135	.113	.105	.087	.085	.065
	T	.045	.045	.073	.083	.085	.085	.123	.135	.143	.185
stochastic design, with $\hat{\hat{\theta}} \sim N\left(\hat{\theta}, I\left(\hat{\theta}\right)\right)$											
	X	.030	.063	.118	.120	.153	.145	.113	.110	.070	.080
	L	.085	.103	.105	.085	.115	.155	.125	.087	.098	.043
	T	.035	.055	.060	.095	.078	.098	.115	.108	.183	.175

Table 3

Power and MAD(θ) under various model violations, for 2PL, *n*=100, *m*=1000, *k*=20.

		10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	MAD(θ)
base line: ρ=1 and δ=0												0.34
	X	.045	.070	.095	.110	.138	.108	.128	.085	.125	.093	
	L	.070	.103	.090	.148	.135	.113	.105	.087	.085	.065	
	T	.045	.045	.073	.083	.085	.085	.123	.135	.143	.185	
ρ=0.8 and δ=0												0.41
	X	.110	.060	.120	.050	.100	.110	.140	.120	.080	.110	
	L	.080	.100	.100	.110	.100	.130	.130	.080	.080	.090	
	T	.080	.070	.110	.100	.050	.100	.110	.180	.100	.100	
ρ=0.6 and δ=0												0.45
	X	.180	.130	.090	.090	.100	.110	.070	.060	.030	.140	
	L	.060	.080	.080	.090	.150	.070	.140	.130	.090	.110	
	T	.240	.060	.040	.110	.070	.080	.080	.120	.100	.100	
ρ=0 and δ=0.2												0.35
	X	.080	.090	.090	.120	.070	.200	.120	.070	.050	.100	
	L	.040	.060	.090	.090	.090	.120	.140	.080	.140	.150	
	T	.060	.040	.120	.080	.070	.130	.120	.140	.100	.140	
ρ=0 and δ=0.4												0.41
	X	.060	.090	.050	.070	.140	.100	.130	.070	.110	.170	
	L	.040	.090	.050	.120	.100	.100	.160	.140	.100	.100	
	T	.060	.060	.080	.070	.070	.130	.080	.120	.150	.180	
ρ=0 and δ=1												0.45
	X	.160	.110	.080	.120	.110	.070	.090	.110	.070	.080	
	L	.010	.070	.040	.070	.050	.070	.120	.180	.200	.190	
	T	.120	.050	.100	.060	.090	.080	.120	.110	.100	.170	

Table 4

Power and MAD(θ) under various model violations, for 2PL, *n*=100, *m*=1000, *k*=40.

		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**	**MAD(θ)**
base line: ρ=1 and δ=0												0.25
	X	.080	.130	.070	.120	.110	.120	.180	.070	.090	.020	
	L	.090	.040	.080	.090	.110	.130	.130	.160	.090	.080	
	T	.090	.060	.110	.120	.150	.070	.090	.070	.090	.150	
ρ=0.8 and δ=0												0.37
	X	.150	.070	.100	.100	.200	.050	.100	.100	.080	.050	
	L	.070	.110	.100	.050	.120	.100	.140	.120	.110	.080	
	T	.230	.130	.060	.070	.080	.120	.070	.060	.070	.110	
ρ=0.6 and δ=0												0.42
	X	.150	.120	.070	.080	.120	.150	.080	.090	.080	.060	
	L	.060	.100	.080	.130	.130	.100	.090	.040	.100	.170	
	T	.250	.110	.060	.070	.080	.050	.060	.110	.060	.150	
ρ=0 and δ=0.2												0.25
	X	.060	.020	.160	.130	.160	.170	.110	.020	.060	.100	
	L	.060	.090	.090	.090	.110	.110	.100	.080	.120	.150	
	T	.100	.040	.070	.120	.090	.080	.130	.090	.130	.150	
ρ=0 and δ=0.4												0.25
	X	.030	.080	.130	.100	.130	.130	.130	.120	.090	.060	
	L	.060	.050	.100	.020	.100	.100	.160	.170	.090	.150	
	T	.080	.030	.080	.100	.100	.160	.130	.060	.080	.180	
ρ=0 and δ=1												0.40
	X	.150	.110	.080	.100	.100	.080	.120	.080	.080	.100	
	L	.010	.020	.010	.040	.070	.100	.140	.170	.170	.270	
	T	.130	.110	.080	.140	.050	.080	.040	.080	.160	.130	

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

RR-98-01	C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*
RR-97-07	H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*
RR-97-06	H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*
RR-97-05	W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*
RR-97-04	W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*
RR-97-03	W.J. van der Linden, *Multidimensional Adaptive Yesting with a Minimum Error-Variance Criterion*
RR-97-02	W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*
RR-97-01	W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*
RR-96-04	C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*
RR-96-03	C.A.W. Glas, *Testing the Generalized Partial Credit Model*
RR-96-02	C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*
RR-96-01	W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*
RR-95-03	W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*
RR-95-02	W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*
RR-95-01	W.J. van der Linden, *Some decision theory for course placement*
RR-94-17	H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*
RR-94-16	H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*
RR-94-15	H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*
RR-94-13	W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*
RR-94-12	R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*
RR-94-10	W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

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

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

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

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

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

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

ERIC
Full Text Provided by ERIC

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

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

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

NOTICE

REPRODUCTION BASIS

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

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