DOCUMENT RESUME

ED 450 126 TM 032 316

AUTHOR	van Krimpen-Stoop, Edith M. L. A.; Meijer, Rob R.
TITLE	Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admissions Council, Newtown, PA.
REPORT NO	RR-00-01
PUB DATE	2000-00-00
NOTE	31p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; Goodness of Fit; Item Response Theory; Scores; Test Items
IDENTIFIERS	*Person Fit Measures; *Polytomous Items

ABSTRACT

Item scores that do not fit an assumed item response theory model may cause the latent trait value to be estimated inaccurately. For computerized adaptive tests (CAT) with dichotomous items, several person-fit statistics for detecting nonfitting item score patterns have been proposed. Both for paper-and-pencil (P&P) test and CATs, detection of person misfit with polytomous items has hardly been explored. In this simulation study, the theoretical and empirical null distributions of a person-fit statistic for polytomous items are compared for P&P tests and CATs. Results show that the empirical distribution of this statistic was close to the standard normal distribution, for both P&P tests and CATs. Also statistics that are especially designed for a CAT are proposed. In these statistics observed and expected item scales are compared using cumulative sum (CUSUM) procedures. Results show that the critical values of the CUSUM were symmetric around zero and similar across latent trait values. Moreover, the results show that for the CUSUM procedure fixed critical values for all examinees can be used. (Contains 5 tables and 40 references.) (Author/SLD)

Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items

Research Report
00-01

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

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items

Edith M.L.A. van Krimpen-Stoop

Rob R. Meijer

Abstract

Item scores that do not fit an assumed item response theory model may cause the latent trait value to be inaccurately estimated. For computerized adaptive tests (CAT) with dichotomous items, several person-fit statistics for detecting nonfitting item score patterns have been proposed. Both for paper-and-pencil (P&P) tests and CATs, detection of person misfit with polytomous items is hardly explored. In this study, the theoretical and empirical null distributions of a person-fit statistic for polytomous items are compared for P&P tests and CATs. Results showed that the empirical distribution of this statistic was close to the standard normal distribution, for both P&P tests and CATs. Also, statistics that are especially designed for a CAT are proposed. In these statistics observed and expected item scores are compared using cumulative sum (CUSUM) procedures. Results showed that the critical values of the CUSUM were symmetric around zero and similar across latent trait values. Moreover, the results showed that for the CUSUM procedure fixed critical values for all examinees can be used.

Key words: appropriateness measurement, computer adaptive testing, cumulative sum, item response theory, person fit, polytomous item response models.

Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items

The aim of a computerized adaptive test (CAT) is to construct an optimal test for each examinee. This is realized by estimating the examinee's ability-level (θ) after administration of each item and selecting the next item based on the current ability estimate ($\hat{\theta}$). The θ-estimation procedure, the item selection procedure and the stopping rule of a CAT are all based on the assumption that the item scores of an examinee fit the assumed item response theory (IRT) model. It is questionable, however, whether the assumed IRT model gives a good description for each examinee's test behavior. For those examinees for whom this is not the case, the ability estimate as a measure of true θ may be inadequate, and as a result the construction of an optimal test may be difficult. There are all sorts of causes that may invalidate $\hat{\theta}$. For example, knowledge of the correct answers due to test preview on achievement tests, faking on biodata questionnaires or personality tests, randomly guessing on all items in the test in order to become familiar with the questions, or lack of motivation in item-pretesting situations. To detect examinees with invalid $\hat{\theta}$, person-fit statistics have been proposed.

Most person-fit research has been conducted for paper-and-pencil (P&P) tests with dichotomously scored items (e.g., Drasgow, Levine, & Williams, 1985; Meijer, 1994; Tatsuoka, 1984). Recently, some studies investigated the assessment of person fit in CATs with dichotomously scored items (e.g., Nering, 1997; van Krimpen-Stoop & Meijer, 1999b, 1999c). However, there are no studies known to the authors that deal with person-fit research in a CAT with polytomously scored items. This may be explained by the fact that only a few studies investigated the use and implementation of CATs with polytomous items (see Dodd, De Ayala, & Koch, 1995, for an overview). In the present study we will investigate the use of person-fit statistics for polytomously scored CATs. Also, some results for person-fit statistics in P&P tests with polytomous items are discussed.

This study is organized as follows. First, a short overview of item response theory for polytomous items with ordered score categories is given. Second, a short overview of research in the context of CAT with polytomous items is given. Third, existing person-fit statistics that are designed for polytomous items in P&P tests are described and new statistics that can be used in a CAT and that are based on theory from Statistical Process Control are proposed. Fourth, simulation studies are conducted in which the theoretical and empirical distributions of an existing person-fit statistic are compared both for P&P tests and for CAT. Finally, a simulation study investigating the critical values of the newly proposed statistics is conducted.

Polytomous Item Response Models

Models for unidimensional polytomous items with ordered score categories are considered here; that is, models in which the item responses are scored into more than two ordered categories. Examples of such items are Likert-type attitude items or achievement items with partially correct scoring. Let x_i be the realization of X_i, the score on item i and let $\mathbf{x} = (x_1, ..., x_N)$ denote the observed score pattern on an N-item test. Furthermore, let the responses to item i be categorized into $m + 1$ ordered score categories $j = 0, 1, ..., m$ where higher scores reflect a higher θ level.

According to Mellenbergh (1995; see also, Molenaar, 1983), three families of models for ordered polytomous items can be distinguished where the distinction between these models is based on three different methods to split an ordinal polytomous response variable into a set of dichotomies (see Agresti, 1990). In all three methods a polytomous response variable with $m + 1$ categories is split into m dichotomies. The models in the first family are called the adjacent-category models, where the $m + 1$ ordinal response variable is split into m adjacent-category pairs. The probability of obtaining a score j is determined conditional on obtaining a score $j - 1$ or j: $P(X_i = j | X_i = j - 1 \vee X_i = j)$. Examples of such models are the partial credit model (PCM; Masters 1982), the rating scale PCM (Andrich, 1978), and the generalized PCM (Muraki,1992).

The second family consists of the cumulative-probability models, where the $m + 1$ ordinal response variable is split into m cumulative probabilities. Here, the probability is determined of obtaining a score in category j or higher: $P(X_i \geq j)$. Examples of such models are the graded response model (Samejima,1969) and the rating scale graded response model (Muraki, 1990).

The third family consists of the continuation-ratio models, where the $m + 1$ ordinal variable is split into m continuation ratios, and the probability of obtaining a score j or higher, conditional on obtaining a score $j-1$ or higher is of interest: $P(X_i \geq j | X_i \geq j-1)$. An example is the sequential model (Tutz, 1990). For recent developments in polytomous item score models, see, for example Hemker (1996) and Akkermans (1998).

Let $P_{ij}(\theta)$ denote the probability that an examinee with ability θ obtains a score j on item i. Although the present study is restricted to the PCM (Masters, 1982), the theory and applications discussed can easily be generalized to other ordinal polytomous models. In the PCM the item parameters δ_{ik} for $k = 1, ..., m$, are often described as item-step difficulties where δ_{ik} is the point on the θ-axes where the probabilites of obtaining score k and $k-1$ intersect (i.e., $P_{ik}(\theta) = P_{i,k-1}(\theta)$). Let $\boldsymbol{\delta}_i = (\delta_{i1}, ..., \delta_{im})$ denote the vector with the individual step difficulties for item i and let $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, ..., \boldsymbol{\delta}_N)$ denote the vector of vectors $\boldsymbol{\delta}_i$. The probability of scoring $X_i = x_i$ on item i conditional on θ, according to the PCM (Masters, 1982) is defined as

$$P_{ij}(\theta) = P_i(X_i = j | \theta, \boldsymbol{\delta}_i) = \frac{\exp\left[j\theta - \sum_{k=0}^{j} \delta_{ik}\right]}{\sum_{h=0}^{m} \exp\left[h\theta - \sum_{k=0}^{h} \delta_{ik}\right]}, \tag{1}$$

such that $\sum_{j=0}^{m} P_{ij}(\theta) = 1$ and $\delta_{i0} \equiv 0$.

Adaptive Testing and Polytomously Scored Items

Most CAT research has been conducted for dichotomous items. The few studies that have used polytomous items used the PCM, the graded response model (Samejima, 1969), the nominal response model (Bock, 1972,), the rating scale model (Andrich, 1978), or the successive intervals model (Rost, 1988).

Polytomous CAT research investigated characteristics of the item pool, the item selection criterion, the θ-estimation procedure, and the stopping rule. Interesting results were, for example, that compared to CAT with dichotomous item scores, the size of the item pool may be substantially smaller to get an accurate estimate of θ (see e.g., Dodd, Koch, & De Ayala, 1993, and Koch & Dodd, 1989). Item pools should be not too small, however, in order to secure, for example, content validity.

As in dichotomous CAT, item selection in polytomous CAT is often based on maximum item information and in most cases maximum likelihood estimation is used for estimating θ. However, the maximum likelihood estimate can only be determined when the response to the first item is not in the lowest or the highest category of the item. An alternative may be to use Warm's (1989) estimation procedure (e.g. van Krimpen-Stoop & Meijer, 1999c).

When the maximum likelihood estimate is determined after the first response, the estimate will be very unstable with a high standard error. To overcome this problem, in most cases a systematic procedure to estimate θ is used, until item scores in two different item categories are observed, and after this, maximum likelihood is used to estimate θ (see e.g., Koch & Dodd, 1989, and Dodd, Koch, & De Ayala, 1989). One systematic procedure is the fixed stepsize procedure, in which the new preliminary estimate of θ is increased/decreased by a constant when the response was in the upper/lower half of the response scale. Also, the variable stepsize procedure can be used, where the new preliminary estimate of θ is increased/decreased by half times the highest/lowest step difficulty when the response was in the upper/lower half of the response scale. Research showed that the use of variable stepsize leads to better results compared with fixed stepsize, in terms of fewer cases of nonconvergence of the θ estimate (see e.g., Koch & Dodd, 1989).

For the stopping rule, a number of alternatives can be used. The test can be stopped when a certain number of items has been administered (fixed test length), when the accuracy in the estimation of θ is within a prespecified standard error of θ (standard error rule), or when there are no items available in the item pool that have a minimum level of information conditional on the current estimate of θ (minimum information rule). For a comparison of the minimum information stopping rule and the standard error stopping rule see, for example, Dodd et al. (1989).

Person-Fit Analysis

In person-fit analysis the fit of an individual item score pattern is investigated to detect misfitting item score patterns. In the few studies in which person fit was applied in the context of polytomous items, the polytomous items were dichotomized and person-fit statistics were used for dichotomous item scores (see e.g., Zickar & Drasgow, 1996). A disadvantage is that part of the information contained in the polytomous item is lost, because each pair of adjacent categories of the polytomous item can be seen as a single dichotomous item (see e.g., Molenaar, 1983). By dichotomizing the scores the length of the tests is actually decreased, which is unfavorable for the assessment of person fit (Reise & Due, 1991). Also, compared with the dichotomized version of the item, the item information function of a polytomously scored item is higher at the peak of the function and the information is also distributed across a wider range of θ, which may also enhance the assessment of person fit (Reise & Due, 1991). Finally, the dichotomization that is chosen for a polytomous score variable has a substantial effect on the measurement outcome θ, unless specific conditions on the item parameters hold (Jansen & Roskam, 1986, Roskam & Jansen, 1989).

For tests with dichotomous or polytomous items, misfitting item score patterns consist of many incorrect scores to easy or unpopular items and many correct scores to difficult or popular items (e.g., Meijer & Sijtsma, 1999). In the dichotomous case, X_i is 0 or 1, the expected score $E(X_i|\theta)$ of item i equals the probability of a correct response, and a weighted function of the residual

$$f\left(X_i - E\left(X_i|\theta\right)\right), \tag{2}$$

is used to determine person fit. Also for polytomous items, the expected score on item i can be determined and observed and expected scores can be compared. Because in this paper $P_{ij}(\theta)$ is defined as the probability of obtaining score j on item i, the expected score $E(X_i|\theta)$, according to the general definition of the expectation (see e.g., Lindgren, Chapter 4), can be written as

$$E\left(X_i|\theta\right) = \sum_{j=0}^{m} jP_{ij}\left(\theta\right),$$

and $X_i \in \{0, 1, ..., m\}$.

Existing Person-Fit Statistics

An often used person-fit statistic for dichotomously scored items is the log-likelihood statistic l (Levine & Rubin, 1979, Drasgow, Levine & Williams, 1985). Drasgow, Levine & Williams (1985) also proposed a standardized log-likelihood statistic for polytomous items. Assuming local independence between all items, the likelihood of score pattern x, can be written as

$$L(\mathbf{x}\,|\theta,\boldsymbol{\delta}) = \prod_{i=1}^{N} P_{ix_i}(\theta),$$

and the log-likelihood l is defined as the natural logarithm of L and can be written as

$$l(\mathbf{x}\,|\theta,\boldsymbol{\delta}) = \ln\left[L(\mathbf{x}\,|\theta,\boldsymbol{\delta})\right] = \sum_{i=1}^{N} \ln P_{ix_i}(\theta).$$

Because l is dependent on $\hat{\theta}$, Drasgow et al. (1985) proposed to use the standardized version of l, denoted as l_z:

$$l_z(\mathbf{x}\,|\theta,\boldsymbol{\delta}) = \frac{l(\mathbf{x}\,|\theta,\boldsymbol{\delta}) - E(l|\theta)}{\left[var(l|\theta)\right]^{1/2}},$$

where $E(l|\theta)$ denotes the expected value of l

$$E(l|\theta) = \sum_{i=1}^{N}\sum_{j=0}^{m} P_{ij}(\theta)\ln P_{ij}(\theta)$$

and $var(l|\theta)$ the variance of l

$$var(l|\theta) = \sum_{i=1}^{N}\left[\sum_{j=0}^{m}\sum_{h=0}^{m} P_{ij}(\theta)\,P_{ih}(\theta)\ln P_{ij}(\theta)\ln\frac{P_{ij}(\theta)}{P_{ih}(\theta)}\right].$$

In practice, θ is unknown and $\hat{\theta}$ should be used to determine l_z. Large negative values of l_z indicate a low probability of obtaining score pattern x ; thus, large negative values of l_z indicate misfitting item score patterns. Drasgow et al. (1985) found that for P&P tests the empirical distribution of l_z using $\hat{\theta}$ was reasonably close to the standard normal distribution for long tests (tests with more than 80 items).

Another person-fit statistic that can be used for polytomous items was proposed by Wright & Masters (1982). Wright & Masters (1982) proposed to use the standardized weighted mean squared residual

$$v = \frac{\sum_{i=1}^{N}\left(X_i - E(X_i|\theta)\right)^2}{\sum_{i=1}^{N} var(X_i|\theta)},$$

where a transformation of v was used to correct for kurtosis

$$t = \left(v^{1/3} - 1\right) \tfrac{3}{q} + \tfrac{q}{3},$$

with

$$\begin{aligned} E\left(X_i|\theta\right) &= \sum_{j=0}^{m} j P_{ij}\left(\theta\right), &(3)\\ var\left(X_i|\theta\right) &= \sum_{j=0}^{m} \left(j - E\left(X_i|\theta\right)\right)^2 P_{ij}\left(\theta\right), \text{ and} &(4)\\ q^2 &= \frac{\sum_{i=1}^{N}\left[\left(\sum_{j=0}^{m}\left(j - E\left(X_i|\theta\right)\right)^4 P_{ij}\left(\theta\right)\right) - var\left(X_i|\theta\right)^2\right]}{\left[\sum_{i=1}^{N}\left[var\left(X_i|\theta\right)\right]^{1/2}\right]^2}. \end{aligned}$$

Wright & Masters (1982, pp. 108-109) claim that t is standard normally distributed when the PCM holds. Some research has been conducted with this statistic using dichotomous data, where the PCM becomes the Rasch (1960) model and the statistic t is equivalent to the statistic proposed by Wright & Stone (1979, Chapter 4). For example, Rogers & Hattie (1987) showed that the empirical distribution was far off the expected theoretical distribution and, as a result, using critical values based on the theoretical distribution, t was insensitive to misfitting item score patterns. Also, Hoijtink (1986) showed that the distribution of the dichotomous version of t was far from standard normal in the case of the Rasch model.

Cumulative Sum Procedures

In a cumulative sum (CUSUM) procedure, originally proposed by Page (1954), sums of statistics are accumulated, but only if they exceed 'the goal value' by more than d units. Let Z_t be the value of a standard normally distributed statistic Z obtained from a sample of size n at time point t. Furthermore, let d be the reference value. Then, a two-sided CUSUM procedure can be written in terms of C_t^+ and C_t^-, where

$$\begin{aligned} C_t^+ &= \max\left[0, \left(Z_t - d\right) + C_{t-1}^+\right], \text{ and}\\ C_t^- &= \min\left[0, \left(Z_t + d\right) + C_{t-1}^-\right], \end{aligned}$$

with starting values $C_0^+ = C_0^- = 0$. Note that the sums are accumulating on both sides concurrently. Thus, as soon as $|Z_t| > d$, Z_t values are accumulated in C^+ and C^-. Let h denote some threshold. The process is 'out-of-control' when $C^+ > h$ or $C^- < -h$ and 'in-control' otherwise.

One assumption underlying the CUSUM procedure is that the Z_t-values are asymptotically standard normally distributed; the values of d and h are based on this assumption. The value of d is usually selected as one-half of the mean shift (in Z_t-units) one wishes to detect; for example, $d = 0.5$ is the appropriate choice for detecting a shift of one times the standard deviation of Z_t. In practice, CUSUM-charts with $d = 0.5$ and $h = 4$ or $h = 5$ are often used (for a reference of the underlying rationale of this choice, see Montgomery, 1997, p.322). Setting these values for d and h results in a significance level of approximately $\alpha = 0.0027$ (two-sided). Note that in person-fit research α is fixed and critical values are derived from the null distribution of the statistic. In this study, we will also use a fixed α and will derive critical values from simulations.

Both van Krimpen-Stoop & Meijer (1999b) and Bradlow, Weiss, & Cho (1998) proposed to use statistical process control techniques to detect person misfit in a CAT. Van Krimpen-Stoop & Meijer (1999b) proposed statistics to be used in a CUSUM procedure to investigate person fit in an on-line application or after complete administration of a CAT with dichotomously scored items. These statistics were based on the responses to single items resulting in a sample size of 1 at each t. Because the theoretical distribution of these statistics is a Bernoulli distribution, and not a standard normal distribution, it was necessary to determine critical values to classify a score pattern as nonfitting by means of a simulation study. The critical values were found to be stable across θ values. Van Krimpen-Stoop & Meijer (1999a) also proposed CUSUM-based statistics using the responses to disjoint subsets of items which resulted in a sample size of $n > 1$ at each t. These statistics followed a distribution that was close to the standard normal distribution when n was not too small (10 or more items in each disjoint subset). Thus for these statistics a theoretical distribution can be used to determine the critical values. Van Krimpen-Stoop & Meijer (1999a) found that the use of theoretically determined critical values resulted in empirical Type I errors that were close to the nominal ones. A limitation was, however, that the subsets of items should not be too small or too large. For detailed information see, van Krimpen-Stoop & Meijer (1999a).

CAT and CUSUM Procedures

Sums of consecutive negative or positive residuals can be investigated using a CUSUM procedure. Let i_k denote the kth item in the CAT; that is, k is the stage of the CAT. Further, let the statistic T_k be a function of the residuals at stage k, N the final test length, and let, without loss of generality, the reference value d be equal to 0. For each examinee, at each stage k of a CAT, the CUSUM procedure can be determined as

$$\begin{aligned} C_k^+ &= \max\left[0, T_k + C_{k-1}^+\right], && (5)\\ C_k^- &= \min\left[0, T_k + C_{k-1}^-\right], \text{ and} && (6)\\ C_0^+ &= C_0^- = 0, && (7) \end{aligned}$$

where C^+ and C^- are sensitive to series of positive and negative values of T_k, respectively. Let UB and LB be some appropriate upper and lower bound, respectively. Then, when $C^+ > UB$ or $C^- < LB$ the item score pattern can be classified as not fitting the model, otherwise, the item score pattern can be classified as fitting the model.

In the polytomous case, T_k can be written as a function of the residuals as in Equation 2. In Equation 2, the value of the statistic is determined given the true value of θ. In practice, however, this true value is unknown and as an alternative an estimate of θ can be used. In a CAT, two alternative estimates of θ can be chosen. First, during administration of the test at each stage k, θ is estimated based on the responses to the previous administered items (denoted as $\hat{\theta}_{k-1}$) and this updated estimate can be used to compute the value of T. Second, the final estimate of θ (denoted as $\hat{\theta}_N$) can be used to compute T. An advantage of using the updated estimate $\hat{\theta}_{k-1}$ is that the fit can be investigated during test administration, although $\hat{\theta}_{k-1}$ may be more inaccurate than $\hat{\theta}_N$. Due to the use of the final estimate $\hat{\theta}_N$, the fit can no longer be investigated during the test, because $\hat{\theta}_N$ needs to be computed first and this is done at the end of the test.

Statistics

Two simple statistics are the unweighted residual between the observed and expected score, corrected for test length

$$T_k^1 = \tfrac{1}{N}\left[X_{i_k} - E\left(X_{i_k}|\theta\right)\right],$$

and the weighted residual, corrected for test length and the variance of the item score i

$$T_k^2 = \frac{1}{N}\left[\frac{X_{i_k} - E\left(X_{i_k}|\theta\right)}{\left[var\left(X_{i_k}|\theta\right)\right]^{1/2}}\right],$$

where $E\left(X_{i_k}|\theta\right)$ and $var\left(X_{i_k}|\theta\right)$ are defined in Equations 3 and 4, respectively. Note that all kinds of other functions of the residual can be taken. This study, however, is restricted to the statistics T^1 and T^2.

To determine upper and lower bounds in a CUSUM procedure it is assumed that the statistic computed at each stage is asymptotically standard normally distributed. However, the null distribution of T^1 and thus T^2 are far from standard normal: in the dichotomous case, T^1 follows a Bernoulli distribution with parameter $P_k(\theta)$, and in the polytomous case, T^1 follows a multinomial distribution with m observations and parameter vector $\left(P_{i_k1}\left(\theta\right), ..., P_{i_km}\left(\theta\right)\right)$, where m is the highest ordered response category. As a result, setting $d = 0.5$ and the upper and lower bound to $h = 5$ and $h = -5$, respectively, is not appropriate in this context. Therefore, in this study, the numerical values of the upper and lower bound are investigated through simulation, with for example $\alpha = 0.05$ and $d = 0$. (See also van Krimpen-Stoop & Meijer, 1999a for similar research with dichotomous items).

This study is limited to the use of statistics based on the responses to single items, thus a sample size of 1 at each time point. Constructing a substantial number of disjoint subsets of items of 10 or more in a polytomous CAT or P&P test is difficult, because the test length of a polytomous CAT is in general smaller than the length of a dichotomous CAT, due to higher information of the polytomously scored items (see e.g., De Ayala, 1992). A disadvantage of the use of statistics based on responses to single items is the lack of theoretically determined critical values, and it is therefore necessary to determine critical values by means of a simulation study.

Simulation Studies

Purpose

This simulation study was designed to investigate whether the empirical null distribution of l_z using $\hat{\theta}$ was in agreement with the standard normal distribution for short polytomous P&P tests and CATs. Drasgow, Levine & Williams, 1985 showed that for long P&P tests (80 items or more) the empirical distribution of l_z is close to the standard normal distribution. However, it is unknown how well this theoretical distribution holds for shorter P&P tests and CATs.

Second, the numerical values of the upper and lower thresholds of the CUSUM procedures for statistics T^1 and T^2, across θ-levels were examined. In the case that these critical values are similar across θ values, in practice, one fixed UB and LB can be used for all examinees. This eases the use of these statistics.

Method

Item Pool

To be consistent with earlier research on polytomous CAT, an item pool consisting of 60 three-step items from Koch & Dodd (1989) that fit the PCM was used. In Table 1 the values of the item parameters are given.

P&P tests

Two P&P tests were constructed, one 20-item test and one 30-item test. The 20-item test was constructed using the first 20 items of item pool, whereas for the 30-item test the first 30 items were used. For each test, six datasets of 1,000 response vectors were simulated. Five datasets were simulated at five different θ levels: $\theta = -2, -1, 0, 1,$ and 2. One dataset was simulated in which 1,000 θs were drawn from $N(0;1)$. The simulation procedure was analogous to the procedure for dichotomously scored items in van Krimpen-Stoop & Meijer (1999c).

For each item score pattern, l_z was determined using $\widehat{\theta}_N$. These 1,000 values of l_z constituted the empirical distribution of l_z in each dataset. θ was estimated using the maximum likelihood procedure proposed by Masters (1982). For all simulated distributions, the empirical Type I errors were determined as the percentage of item score patterns that obtained a value of the statistic below the critical value of the standard normal distribution at one-sided significance level $\alpha = .005$, .01, .015, .02, and .025. Also, the first three moments of the simulated distributions of l_z were computed and compared with the moments of the standard normal distribution.

CAT

Three CATs were constructed consisting of 10, 20, or 30 items. For all CATs, six datasets of $1,000$ adaptive item score patterns were simulated. Five datasets were simulated at five θ levels: $\theta = -2, -1, 0, 1,$ and 2. For the sixth dataset, $1,000$ θs were drawn from $N(0;1)$.

The item selection criterion that was used was maximum item information where the item information function of an m-step PCM item is defined as (Samejima, 1969)

$$\begin{aligned} I_i(\theta) &\equiv \sum_{j=0}^{m} \left[\tfrac{\partial}{\partial\theta} P_{ij}(\theta)\right]^2 / P_{ij}(\theta) \\ &= \sum_{j=0}^{m} j^2 P_{ij}(\theta) - \left[\sum_{j=0}^{m} j P_{ij}(\theta)\right]^2 . \end{aligned}$$

Maximum likelihood estimation (Masters, 1982) was used to estimate θ, and the fixed stepsize procedure with stepsize equal to 0.5 was used until item scores in two different categories were obtained. A fixed test length stopping rule was used, where final test length N was set to $10, 20$, or 30.

For each item score pattern, the empirical distribution of l_z was determined similar to the procedure for P&P tests. For all simulated distributions, the empirical Type I errors were determined and the first three moments of the simulated distributions of l_z were computed as described above.

Also, for each dataset and each simulee, statistics T^1 and T^2 were computed in the CUSUM procedure described in Equations 5 through 7, where three different θ values were used to determine $E(X_i|\theta)$: the value of true θ, the value of the final θ estimate, $\hat{\theta}_N$, and the updated θ estimate, $\hat{\theta}_{k-1}$. For each simulee,

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

were determined, resulting in 1,000 values of $\max C^+$ and $\min C^-$ for each statistic and each dataset. Then, for each dataset and for both statistics, the upper bound, UB, was determined as the value of $\max C^+$ for which 2.5% of the simulees had higher $\max C^+$-

values and the lower bound, LB, was determined as the value of $\min C^-$ for which 2.5% of the simulees had lower $\min C^-$-values. That is, a two-sided test at $\alpha \leq 0.05$ was conducted, where $P(\max C^+ \geq UB) = P(\min C^- \leq LB) = 0.025$. So, for each dataset two bounds (the upper and lower bounds) were determined for both T^1 and T^2.

Results

Empirical Distribution of l_z

In Tables 2 and 3 the first three moments of the empirical distributions of $l_z\left(\hat{\theta}_N\right)$, and the empirical Type I errors at five levels of (one-sided) α are given for the P&P tests and the 20- and 30-item CATs, respectively.

Table 2 (P&P tests) shows that the mean of l_z was slightly larger than expected under the standard normal distribution, for all datasets and both test lengths. The variance of l_z was close to 1 as expected under the standard normal distribution, for most datasets and tests, provided that $\theta \neq |2|$. Furthermore, the skewness of the distribution of l_z was found to be negative for most datasets; for the 20-item P&P test and $\theta = -2$ the skewness was positive (.901). However, the empirical Type I errors were close to the nominal ones, for most datasets and tests. For the 20-item test, the empirical Type I errors were somewhat smaller than the nominal ones, whereas for the 30-item test, the empirical error rates were slightly larger than the nominal error rates.

Table 3 (CAT) shows that, for all datasets and both CATs, the mean and variance of l_z were found to be deviant from 0 and 1, respectively. On average (across all datasets and both CATs), the mean and variance were .21 and .76, respectively (not tabulated). Also, the skewness of l_z was negative for all datasets and both CATs. Although the first three moments of the distribution were deviant from expected, the empirical Type I errors were only slightly smaller than the nominal ones for both CATs and all datasets. This might be explained by the negative skewness. As a result, the person-fit statistics were only slightly conservative in classifying misfitting item score patterns as aberrant.

Critical Values of CUSUM

In Tables 4 and 5 the numerical values of UB and LB of the CUSUM procedure using statistic T^1 and T^2, respectively, are given for the 10-, 20-, and 30-item CATs.

Table 4 shows, that using true θ to calculate T^1, for all CATs and all datasets, the values of UB and LB were almost symmetrical around 0 and similar across different θ values. Moreover, when true θ was used, the numerical values of UB and LB obtained from the datasets $\theta \sim N(0,1)$ approximated the bounds of the datasets with fixed θ values. When $\hat{\theta}_N$ was used to determine T^1, the numerical values of UB and LB were asymmetric around 0 and differed across θ values. However, for all CATs, the values of UB and LB obtained using $\theta \sim N(0,1)$ and $\hat{\theta}_N$ were similar to those obtained when true θ was used, probably due to the fact that the value of $\hat{\theta}_N$ was close that of θ. When $\hat{\theta}_{k-1}$ was used to determine T^1, the values of UB and LB were found to be symmetrical around 0, but different across θ. But, again, for all CATs, the values of UB and LB obtained from the dataset $\theta \sim N(0,1)$ using $\hat{\theta}_{k-1}$ were close to those obtained when true θ was used, probably due to accurate estimation of θ.

The results for statistic T^2 in Table 5 show that, when true θ was used, the values of UB and LB were asymmetric around 0 and differed across θ values, for all CATs and all datasets. Furthermore, using θ, the bounds obtained from the dataset $\theta \sim N(0,1)$ were quite different from those obtained in the datasets with fixed θ. For both using $\hat{\theta}_N$ or $\hat{\theta}_{k-1}$, the numerical values of the bounds were asymmetric and differed across θ values for all CATs and all datasets.

Discussion

In this study, the empirical distribution of an existing person-fit statistic for polytomously scored items, l_z, was investigated. It was shown that, although the first three moments of the empirical distribution of l_z were slightly deviant from the expected values under the standard normal distribution, the empirical Type I errors were close to the nominal ones, for most datasets and most P&P tests. For CATs, the first three moments of the empirical distribution were more deviant from those of the standard normal distribution than for P&P tests. However, the empirical Type I errors were slightly smaller than the nominal Type I errors. Therefore, l_z is a slightly conservative person-fit statistic when the critical values of the standard normal distribution are used, for both P&P tests and CATs with partial credit items.

Interesting was that the results of the empirical distribution of l_z differed from the results found in van Krimpen-Stoop & Meijer (1999c) where P&P tests and CATs with dichotomous items were used. The better fit between the empirical and theoretical distribution for short tests in this study, especially the finding that the empirical Type I errors were close to the nominal error rates, may be explained by realizing that each pair of adjacent categories of the polytomous item can be seen as a single dichotomous item. This effect is comparable with using longer tests, which also results (e.g., van Krimpen-Stoop & Meijer, 1999c) in higher agreement between empirical and theoretical distributions.

Also, the use of two CUSUM procedures for the assessment of person fit in polytomous CAT was explored: the numerical values of the critical values were investigated. It was shown that the critical values of the CUSUM procedure using statistic T^1 were symmetric around 0. Moreover, determining bounds using $\theta \sim N(0,1)$ were largely in agreement with the bounds for different values of $\hat{\theta}_N$ or $\hat{\theta}_{k-1}$ and when true θ was used. The CUSUM using statistic T^2 was found to be less stable than the CUSUM using T^1. Even when the true value of θ was used to determine T^2, thus for the null model of fitting response behavior, the critical values were asymmetric around 0 and were different across θ values. Therefore, it is recommended to perform person-fit analysis with the CUSUM procedure using statistic T^1 and not T^2. The UB and LB obtained from the $\theta \sim N(0,1)$ dataset can be used as critical values at significance level $\alpha = .05$. In the case of examinees with θ values in the tails of the distribution (i.e. $\theta = \pm 2$), the classification of score patterns as either fitting or nonfitting may be slightly conservative: the empirical Type I error rate tends to be slightly smaller than the nominal Type I error rate for examinees with $\theta = \pm 2$, thus, slightly less than expected fitting score patterns are classified as misfitting.

A disadvantage of the CUSUM procedure is that critical values have to be determined by means of simulations, which may sometimes be difficult to realize for different item pools and different test lengths. However, an advantage of the CUSUM procedure compared to l_z is that it is possible to investigate the fit of an individual item score pattern during test administration. Also, by examining the graphical plot of the CUSUM, that is the plot of the values of C^+ and C^- against the stage of the CAT, it is possible to track "where-it-went-wrong". Suppose, for example, the situation of an examinee who is unfamiliar with the use of a computer, and during administration he/she becomes familiar with it. Then, it is plausible that the CUSUM passes the lower bound about halfway of the CAT, and reaches a stable level after the examinee is getting more familiar with the computer. On the other hand, when an examinee has preknowledge of a number of difficult items, the CUSUM may pass the upper bound after the response to these items.

This study furthermore showed that for simulees with high or low θ values, it is difficult to classify an item score pattern as fitting or misfitting solely on the basis of the outcome of a person-fit statistic as l_z or the CUSUM procedure because it is difficult to identify proper critical values of a statistic for these simulees. This is not only the case for polytomously scored CATs and P&P tests but also for dichotomously scored tests (see e.g., van Krimpen-Stoop & Meijer, 1999c, 1999b).

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.

References

Agresti, A. (1990). *Categorical data analysis*. Wiley, New York.

Akkermans, L. M. W. (1998). *Studies on statistical methods for polytomously scored test items*. Unpublished doctoral dissertation, University of Twente, Enschede.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.

Bradlow, E. T., Weiss, R. E., and Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association*, 93, 910–919.

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

Dodd, B. G., De Ayala, R. J., and Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19, 5–22.

Dodd, B. G., Koch, W. R., and De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: efects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, 53, 61–77.

Dodd, B. G. Koch, W. R. and De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement*, 13, 129–143.

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

Hemker, B. T. (1996). *Unidimensional IRT models for polytomous items, with results for Mokken scale analysis*. Unpublished doctoral dissertation, University of Utrecht, Utrecht.

Hoijtink, H. (1986). *Detecting aberrant response patterns in the unidimensional scaling model of Rasch (Heymans Bulletin HB-86-792-SW)*. University of Groningen, Groningen, The Netherlands.

Jansen, P. G. W. and Roskam, E. E. (1986). Latent trait models and the dichotomization of graded responses. *Psychometrika*, 51, 69–91.

Koch, W. R. and Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, 2, 335–357.

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

Lindgren, B. W. (1993). *Statistical Theory (4th ed.)*. Chapman and Hall, New York.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.

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. and Sijtsma, K. (in press). A review of methods for evaluating the fit of item score patterns on a test. *Applied Psychological Measurement*.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19, 91–100.

Molenaar, I. W. (1983). *Item steps (Heymans Bulletin HB-83-630-EX)*. University of Groningen, Groningen, The Netherlands.

Montgomery, D. C. (1997). *Introduction to statistical quality control (3rd ed.)*. John Wiley and Sons, New York.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, 14, 59–71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176.

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

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

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

Rogers, H. J. and Hattie, J. A. (1987). A monte carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement*, 11, 47–57.

Roskam, E. E. and Jansen, P. G. W. (1989). Conditions for Rasch-dichotomizationability of the unidimensional polytomous Rasch model. *Psychometrika*, 54, 317–332.

Rost, J. (1988). Measuring attitudes with a threshold model drawing on a traditional sclaing concept. *Applied Psychological Measurement*, 12, 397–409.

Samejima, F. (1969). Estimation of latent ability using response models of graded scores. *Psychometrika, Monograph Supplement No. 17*, 1969.

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

Tutz, G. (1990). Sequential item response models with an ordered response. *Brittish Journal* *of Mathematical and Statistical Psychology*, 43, 39–55.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (1999a). Cusum-based person-fit statistics for adaptive testing. Research Report RR 99-05, University of Twente, Enschede, The Netherlands.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (1999b). Detecting person misfit in adaptive testing using statistical process control techniques. In van der Linden, W. J. and Glas, C. A. W., (Eds.), *Computerized adaptive testing: Theory and practice*. Kluwer-Nijhoff, Boston, MA.

van Krimpen-Stoop, E. M. L. A. and Meijer, R. R. (1999c). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Pychological Measurement*, 23, 327–345.

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

Wright, B. D. and Masters, G. N. (1982). *Rating scale analysis*. MESA Press, Chicago.

Wright, B. D. and Stone, M. H. (1979). *Best test design*. Mesa Press, Chicago.

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

Table 1

Item Parameter Values of the Item Pool From Koch & Dodd (1989)

Item Number	δ_1	δ_2	δ_3	Item Number	δ_1	δ_2	δ_3
1	-0.50	0.00	0.50	31	-1.00	0.00	1.00
2	-0.35	0.00	0.35	32	-1.35	0.00	1.35
3	-0.75	0.00	0.75	33	-1.25	0.00	1.25
4	0.00	-0.75	0.75	34	0.00	-1.25	1.25
5	-0.75	0.75	0.00	35	-1.25	1.25	0.00
6	-0.50	0.00	0.50	36	-1.00	0.00	1.00
7	-0.35	0.00	0.35	37	-1.35	0.00	-1.35
8	-0.75	0.00	0.75	38	-1.25	0.00	1.25
9	0.00	-0.75	0.75	39	0.00	-1.25	1.25
10	-0.75	0.75	0.00	40	-1.25	1.25	0.00
11	1.30	1.80	2.30	41	0.50	1.50	2.50
12	0.80	1.55	2.30	42	0.50	-1.75	2.50
13	1.30	1.60	2.30	43	0.70	2.00	2.70
14	1.30	1.90	2.30	44	0.80	1.90	2.50
15	1.00	1.40	2.00	45	0.80	1.40	2.50
16	0.50	0.90	1.50	46	0.50	0.90	2.50
17	1.55	0.80	2.30	47	1.75	0.50	2.50
18	0.80	2.30	1.55	48	0.50	2.50	1.75
19	1.40	1.00	2.00	49	1.40	0.80	2.50
20	1.00	2.00	1.40	50	0.80	2.50	1.40
21	-2.30	-1.80	-1.30	51	-2.50	-1.50	-0.50
22	-2.30	-1.55	-0.80	52	-2.50	-1.75	-0.50
23	-2.30	-1.60	-1.30	53	-2.70	-2.00	-0.70
24	-2.03	-1.90	-1.30	54	-2.50	-1.90	-0.80
25	-2.00	-1.40	-1.00	55	-2.50	-1.40	-0.80
26	-1.50	-0.90	-0.50	56	-2.50	-0.90	-0.50
27	-2.30	-0.80	-1.55	57	-2.50	-0.50	-1.75
28	-1.55	-2.30	-0.80	58	-1.75	-2.50	-0.50
29	-2.00	-1.00	-1.40	59	-2.50	-0.80	-1.40
30	-1.40	-2.00	-1.00	60	-1.40	-2.50	-0.80

Table 2

Mean (M), Variance (V), Skewness (S), and Type I Errors of the Simulated Distributions of $l_{z,pol}$ (Using $\hat{\theta}$) for the 20- and 30-item P&P Tests

Test				Type I Errors				
and θ	M	V	S	.005	.010	.015	.002	.025
20-item P&P Test								
N(0,1)	.058	.639	-.496	.005	.008	.010	.012	.015
θ=-2.0	.133	.249	.901	.000	.000	.000	.000	.000
-1.0	.042	.260	-.609	.000	.001	.002	.003	.003
0.0	.105	.763	-.568	.006	.008	.012	.017	.022
1.0	.088	.908	-.564	.011	.016	.020	.026	.028
2.0	.087	.485	-.374	.000	.002	.003	.004	.005
30-item P&P Test								
N(0,1)	.117	.927	-.592	.010	.016	.021	.025	.028
θ=-2.0	.002	.574	-.677	.002	.012	.014	.016	.017
-1.0	.100	.838	-.473	.008	.010	.012	.015	.021
0.0	.058	1.037	-.597	.015	.022	.028	.031	.035
1.0	.098	.839	-.480	.010	.013	.014	.017	.020
2.0	.021	.536	-.463	.001	.001	.004	.007	.007

Table 3

Mean (M), Variance (V), Skewness (S), and Type I Errors of the Simulated Distributions of $l_{z,\mathrm{pol}}$ (Using $\hat{\theta}$) for 20- and 30-item CATs

Test				Type I Errors				
and θ	M	V	S	.005	.010	.015	.020	.025
20-item CAT								
N(0,1)	.317	.795	-.188	.002	.003	.006	.006	.007
θ=-2.0	.078	.601	-.086	.001	.002	.004	.005	.008
-1.0	.390	.783	-.304	.001	.003	.003	.004	.005
0.0	.247	.800	-.132	.000	.003	.006	.007	.008
1.0	.345	.762	-.440	.003	.003	.007	.008	.011
2.0	.067	.744	-.408	.005	.007	.010	.015	.016
30-item CAT								
N(0,1)	.195	.865	-.325	.005	.006	.013	.015	.018
θ=-2.0	.079	.576	-.444	.002	.007	.007	.008	.009
-1.0	.225	.899	-.518	.007	.012	.014	.017	.018
0.0	.271	.838	-.233	.003	.007	.008	.009	.012
1.0	.182	.941	-.435	.005	.009	.013	.015	.016
2.0	.094	.568	-.346	.001	.001	.004	.006	.008

Table 4

Boundaries of CUSUM Using T_k^1 for 10-, 20-, and 30-item CATs at $\alpha = .05$ (two-sided)

Test	θ		$\hat{\theta}_N$		$\hat{\theta}_{k-1}$	
and θ	UB	LB	UB	LB	UB	LB
10-item CAT						
N(0,1)	.68	-.63	.56	-.61	.84	-.85
θ=-2	.60	-.58	.65	-.24	.81	-.70
-1	.62	-.60	.58	-.39	1.00	-.77
0	.64	-.63	.33	-.56	.67	-.86
1	.67	-.65	.49	-.69	.49	-.92
2	.60	-.69	.27	-.61	.66	-.73
20-item CAT						
N(0,1)	.48	-.45	.50	-.60	.63	-.76
θ=-2	.47	-.44	.60	-.13	.78	-.34
-1	.46	-.51	.45	-.28	.73	-.51
0	.45	-.46	.22	-.51	.45	-.79
1	.44	-.47	.17	-.63	.29	-.83
2	.43	-.46	.16	-.60	.40	-.70
30-item CAT						
N(0,1)	.37	-.36	.41	-.53	.49	-.69
θ=-2	.34	-.33	.47	-.11	.60	-.29
-1	.37	-.35	.16	-.50	.32	-.73
0	.36	-.37	.32	-.24	.57	-.43
1	.38	-.35	.12	-.57	.18	-.75
2	.32	-.34	.11	-.50	.22	-.57

Table 5

Boundaries of CUSUM Using T_k^2 for 10-, 20-, and 30-item CATs at $\alpha = .05$ (two-sided)

Test	θ		$\hat{\theta}_N$		$\hat{\theta}_{k-1}$	
and θ	UB	LB	UB	LB	UB	LB
10-item CAT						
N(0,1)	.78	-.87	.67	-.67	.93	-.91
θ=-2	1.23	-.29	.86	-.26	1.26	-.65
-1	.76	-.49	.55	-.39	1.01	-.77
0	.45	-.81	.34	-.62	.63	-.92
1	.38	-.98	.30	-.76	.50	-1.01
2	.41	-1.17	.32	-.68	.71	-.83
20-item CAT						
N(0,1)	.59	-.84	.54	-.66	.94	-.84
θ=-2	1.06	-.17	.79	-.14	1.12	-.32
-1	.54	-.42	.42	-.28	.80	-.51
0	.27	-.70	.23	-.56	.43	-.85
1	.20	-.88	.18	-.70	.30	-.91
2	.22	-.96	.19	-.68	.41	-.83
30-item CAT						
N(0,1)	.54	-.78	.51	-.61	.60	-.78
θ=-2	.86	-.15	.68	-.14	1.00	-.30
-1	.19	-.63	.16	-.57	.32	-.82
0	.40	-.34	.32	-.24	.61	-.44
1	.15	-.84	.13	-.65	.19	-.85
2	.15	-.81	.13	-.60	.23	-.74

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

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

RR-99-03 B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

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

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

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

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

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

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

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

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

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

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