LSAC RESEARCH REPORT SERIES

Detection of Person Misfit in Computerized Adaptive Testing with Polytomous Items

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

Law School Admission Council
Computerized Testing Report 00-03
April 2000

A Publication of the Law School Admission Council

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

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

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), several person-fit statistics for detecting nonfitting item score patterns for dichotomously scored tests have been proposed. Both for paper-and-pencil (P&P) tests and CAT, the detection of person misfit with polytomous items is hardly explored. In this study, the theoretical and empirical distributions of a person-fit statistic for polytomous items are investigated both for P&P tests and CAT. Results showed that the distribution of this statistic was close to the standard normal distribution, for both P&P tests and CAT.

Introduction

One of the aims of a computerized adaptive test (CAT) is to construct an optimal test for each examinee. This is done by estimating the examinee's ability-level (θ) after administration of each item and, based on this current ability estimate ($\hat{\theta}$), the next item is selected using an item selection procedure. The θ-estimation procedure, the item selection procedure and the stopping rule of the CAT are all based on the assumption that the item scores of the examinee fit the assumed 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 current ability estimate as a measure of true θ may be inadequate, and as a result the construction of an optimal test may become 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, or randomly guessing on all items on the test in order to become familiar with the questions.

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). Only a few studies investigated the use and implementation of CATs with polytomous items (see Dodd, De Ayala, & Koch, 1995 for an overview). Recently, some studies investigated the assessment of person fit in CATs with dichotomously scored items (e.g., Nering, 1997; van Krimpen-Stoop & Meijer, 1999a, 1999b). However, there are no studies known to the authors that deal with person-fit research in a CAT with polytomously scored items. In the present study we will explore the use of person-fit statistics both for P&P tests and for CAT.

This study is organized as follows. First, a short overview of item response theory for polytomous items with ordered score categories is given. Second, existing person-fit statistics that are designed for polytomous items (Drasgow et al., 1985) are described. Third, research in the context of CAT with polytomous items is discussed. Finally, a small simulation study is conducted in which the empirical distributions of an existing person-fit statistic are compared with the theoretical distributions both for P&P tests and for CATs.

Polytomous Item Response Models and Person Fit

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

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 the use of three methods to split an ordinal polytomous response into a set of dichotomies (see Agresti, 1990); for a polytomous item of $m + 1$ categories the response 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: $P(X_i = j \mid X_i = j - 1 \vee X_i = j)$. 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: $P(X_i \geq j \mid X_i \geq j - 1)$. An example is the sequential model (Tutz, 1990).

Let $P_{ij}(\theta)$ denote the probability that an examinee with ability θ obtains a score j on item i. Although the present is restricted to the PCM (Masters, 1982), the theory and applications discussed can easily be generalized to other ordinal polytomous models, provided that the term $P_{ij}(\theta)$ is the probability of obtaining a score j on item i, conditional of θ. In the PCM the item parameters δ_{ij} for $j = 1, ..., m$, are often described as item-step difficulties of item i and category j. δ_{ij} is the point on the θ axes where the probability curves for score j and $j - 1$ intersect (i.e., $P_{ij} = P_{i,j-1}$). Let $\delta_i = (\delta_{i1}, ..., \delta_{im})$ denote the vector of the individual step difficulties for item i and let $\delta = (\delta_1, \delta_2, ..., \delta_N)$ denote the vector of the individual step difficulties for all items $i = 1, ..., N$ and all categories $j = 1, ..., m$. The probability of scoring $X_i = x_i$ on item i conditional on θ, according to the PCM (Masters, 1982) is defined as

$$P_{ix_i}(\theta) = P_i(X_i = x_i \mid \theta, \delta_i) = \frac{\exp\left[x_i\theta - \sum_{j=0}^{x_i}\delta_{ij}\right]}{\sum_{h=0}^{m}\exp\left[h\theta - \sum_{j=0}^{h}\delta_{ij}\right]}, \tag{1}$$

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

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. Thus, by dichotomizing the scores, the length of the tests is actually decreased, which is unfavorable for the assessment of person fit (Reise & Due, 1991). Compared with the dichotomized 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.

For tests with dichotomous items, misfitting item score patterns consist of many incorrect scores to easy items and many correct scores to difficult items (e.g., Meijer & Sijtsma, 2001). Because in the dichotomous case the probability of a correct response on item i equals the expected score, a high expected score means that the item is easy and a low expected score means that the item is difficult. When item score patterns consist of many incorrect scores to easy items and many correct scores to difficult items, the residual of the observed score and the expected score becomes very large. Therefore, most existing person-fit statistics for dichotomous items are functions of these residuals.

For tests with polytomous items, the misfitting item score patterns consist of correct answers to difficult items and incorrect answers to easy items. However, because it is also possible to score the item "partially correct," one can determine the expected score on item i and compare this expected score with the observed score. Because in this paper P_{ij} is defined as the probability of obtaining score j on item i, the expected score on item i (μ_i) can, according to the general definition of the expectation (see, e.g., Lindgren, 1993, Chapter 4), be written as

$$\mu_i(\theta) = \sum_{j=0}^{m} jP_{ij}(\theta) .$$

So now the difference between the observed score X_i and the expected score μ_i can be used: $(X_i - \mu_i)$. When an item score pattern consists of many high scores to items with a low expected score, and many low scores to items with a high expected score, the residuals become large, indicating aberrant response behavior. When, for example, an examinee with low ability has preknowledge of some of the most difficult items, large residuals can be expected. On the other hand, when an examinee is careless on almost all the items in the test because the test is only taken to become familiar with the items being administered, the observed score will in general be low, whereas the expected score is higher for almost all items in the test. Therefore, it is also in the polytomous case convenient to construct a person-fit statistic that is based on the difference between the observed and the expected item scores.

Person-Fit Statistics for Polytomous Items

Most person-fit statistics, for both dichotomous and polytomous items, can be written as a linear function of the observed item scores (see, e.g., Snijders, 2001). That is, for suitable functions $w_i(\theta)$, $i = 0, ..., N$,

the statistic can be written as

$$\sum_{i=1}^{N} X_i w_i(\theta) - w_0(\theta), \tag{2}$$

where X_i is either dichotomously or polytomously scored. A general centered form (i.e., a statistic with expectation zero) in which statistics can be expressed is

$$W_N(\theta) = \sum_{i=1}^{N} [X_i - \mu_i(\theta)]\, w_i(\theta), \tag{3}$$

where μ_i is defined as the expected score of item i. This results in a function of the residual between the observed score X_i and the expected score μ_i. In the dichotomous case, X_i is either 0 or 1 and μ_i is equal to the probability of a correct response. Because $X_i^2 = X_i$ holds for the dichotomous case, statistics that are based on the squared residuals $(X_i - \mu_i)^2$ can also be written in the form of Equation 2; however, this is not possible in the polytomous case. Examples of person-fit statistics for dichotomously scored items that are, after centering, of the form of Equation 3 are the log-likelihood statistic l_0 (Levine & Rubin, 1979; Drasgow et al., 1985). Drasgow et al. also proposed a standardized log-likelihood statistic for polytomous items. Assuming local independence between all items, the likelihood of score pattern $\mathbf{X}$ can be written as

$$L(\mathbf{X} \mid \theta, \delta) = \prod_{i=1}^{N} P_{ix_i}(\theta),$$

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

$$l(\mathbf{X} \mid \theta, \delta) = \ln(L(\mathbf{X} \mid \theta, \delta)) = \sum_{i=1}^{N} \ln P_{ix_i}(\theta).$$

Because l is confounded with $\hat{\theta}$, Drasgow, Levine, and Williams (1985) proposed to use the standardized version of l, denoted as $l_{z,pol}$:

$$l_{z,pol}(\mathbf{X} \mid \theta, \delta) = \frac{l(\mathbf{X} \mid \theta, \delta) - l}{\sigma_l},$$

where μ_l denotes the expected value of l and σ_l^2 the variance of l with

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

and

$$\sigma_l^2 = \sum_{i=1}^{N} \left[\sum_{j=0}^{m} \sum_{h=0}^{m} P_{ij}(\theta) P_{ih}(\theta) \ln P_{ij}(\theta) \ln \frac{P_{ij}(\theta)}{P_{ih}(\theta)} \right].$$

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

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

$$v = \frac{\sum_{i=1}^{N} (X_i - \mu_i)^2}{\sum_{i=1}^{N} \sigma_i^2},$$

where a transformation of v was used to correct for kurtosis

$$t = (v^{1/2} - 1)\frac{3}{q} + \frac{q}{3},$$

with

$$\mu_i = \sum_{j=0}^{m} jP_{ij}(\theta), \tag{4}$$

$$\sigma_i^2 = \sum_{j=0}^{m} (j - \mu_i)^2 P_{ij}(\theta), \text{ and} \tag{5}$$

$$q^2 = \frac{\sum_{i=1}^{N} \left[\left(\sum_{j=0}^{m} (j - \mu_i)^4 P_{ij}(\theta) \right) - \sigma_i^4 \right]}{\left(\sum_{i=1}^{N} \sigma_i \right)^2}.$$

Wright and Masters (1982, pp. 108–109) claimed that t was standard normally distributed when the model holds. Some research has been done with the dichotomous version of this statistic, where the PCM becomes the Rasch model and the statistic t is equivalent to the statistic proposed by Wright and Stone (1979, chapter 4). For example, Rogers and Hattie (1987) showed that 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) has shown that the distribution of the dichotomous version of t was far from standard normal in the case of the Rasch model.

It should be noted that the statistics $l_{z,pol}$, t, and v, can not be written in the general form presented in Equations 2 or 3: $l_{z,pol}$ is written in terms of P_{iX_i} instead of X_i, and t and v are functions of the squared residuals $(X_i - \mu_i)^2$ which cannot be written in terms of $(X_i - \mu_i)$ because $X_i^2 \neq X_i$ in the polytomous case.

Statistics Corrected for $\hat{\theta}$

As described above, most person-fit statistics can be written in the general form of Equation 2 or 3. It can be shown (Snijders, 2001), that, provided that θ is known,

$$Z_N(\theta) = \frac{W_N(\theta)}{\left[\sum_{i=1}^{N} w_i^2(\theta) \sigma_i^2(\theta) \right]^{1/2}} \sim AN(0, 1),$$

where W_N is as defined in Equation 3, for both dichotomously and polytomously scored items, and σ_i^2 is the variance of score X_i. In practice, θ is unknown and an alternative $\hat{\theta}$ can be used to determine the values of P_{ij}. However, replacing θ by $\hat{\theta}$ affects the null distribution of W_N and thus also of Z_N (Molenaar & Hoijtink, 1990; Snijders, 2001). Snijders derived the null distribution of a family of statistics of the form in Equation 3 where $\hat{\theta}$ was used as an estimate of θ. If, for suitable functions r_0 and r_i, the estimator $\hat{\theta}$ satisfies the following equation

$$r_0(\hat{\theta}) + \sum_{i=1}^{N} \left[X_i - \mu_i(\hat{\theta}) \right] r_i(\hat{\theta}) = 0, \tag{6}$$

then the statistic

$$T_N(\hat{\theta}) = \left[W_N(\hat{\theta}) - c_N(\hat{\theta}) r_0(\hat{\theta})\right] / \tau(\hat{\theta}) \tag{7}$$

is standard normal distributed for $N \to \infty$, where

$$\tau^2(\hat{\theta}) = \sum_{i=1}^{N} \tilde{w}_i^2(\hat{\theta}) \sigma_i^2(\hat{\theta}),$$

$$\tilde{w}_i(\hat{\theta}) = w_i(\hat{\theta}) - r_i(\hat{\theta}) \frac{\left[\sum_{i=1}^{N} \mu_i'(\hat{\theta}) w_i(\hat{\theta})\right]}{\left[\sum_{i=1}^{N} \mu_i'(\hat{\theta}) r_i(\hat{\theta})\right]},$$

$$\mu_i'(\hat{\theta}) = \frac{\partial}{\partial \theta} \mu_i(\theta) \mid_{\theta = \hat{\theta}} = \sum_{j=0}^{m} j \frac{\partial}{\partial \theta} P_{ij}(\theta) \Big|_{\theta = \hat{\theta}},$$

and μ_i and σ_i^2 are the expected score and the variance of X_i, respectively. An example of a suitable weight w_i that can be used in W_N is

$$w_i(\theta) = \ln \frac{\mu_i(\theta)}{m - \mu_i(\theta)}, \text{ where } w_i(\theta) \begin{cases} > 0 & \text{if } \mu_i(\theta) > 0.5m \\ \leq 0 & \text{if } \mu_i(\theta) \leq 0.5m \end{cases}.$$

When the items are dichotomously scored, this statistic becomes the standardized log-likelihood statistic l_z for dichotomous items proposed by Drasgow et al. (1985).The weights w become small in absolute value when the expected item score is high or low compared with half of the maximum score m. A large residual between expected and observed item score becomes even larger in combination with a large (positive or negative) weight w. Therefore, values of the statistic in the left and right tail area of the distribution indicate misfitting item score patterns.

For the PCM, however, the maximum likelihood estimator does not satisfy Equation 6, which complicates the construction of a statistic corrected for the use of $\hat{\theta}$ according to the Snijders (2001) theory.

CAT Using Polytomous IRT Models

Most research with respect to computer adaptive testing has been conducted for dichotomous items. Some studies, however, have been devoted to polytomous items. Polytomous models that have been used in CATs are, for example, the PCM (Masters, 1982), the graded response model (Samejima, 1969), the nominal response model of Bock (1972), the rating scale model (Andrich, 1978), and the successive intervals model (Rost, 1988).

So far, the research that has been done on polytomous CAT is restricted to the use and the characteristics of the four most important parts of a CAT: the item pool, the item selection criterion, the θ estimation procedure, and the stopping rule. Research has shown that, compared with dichotomous CAT, the size of the item pool can be substantially smaller to get an accurate estimate of θ (see e.g., Dodd, Koch, & De Ayala, 1993, and Koch & Dodd, 1989). However, it is advisable to use an item pool that consists of more than 30 items in order to secure, for example, content validity and item exposure. 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. 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. After this, maximum likelihood is used to estimate θ (see e.g., Koch & Dodd, 1989, and Dodd, Koch, & De Ayala, 1989). An alternative may be to use Warm's estimation procedure (e.g., van Krimpen-Stoop & Meijer, 1999b). Another systematic procedure is the fixed stepsize procedure, in which the new preliminary estimate of θ is increased/decreased by a

constant (e.g., .4 or .5) 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 $\hat{\theta}$ 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 led to better results compared with the fixed stepsize, in terms of fewer cases of nonconvergence of the θ estimate. 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 more items available in the item pool that have a minimum level of information conditional on the current estimate of θ (minimum information rule). In Dodd et al. (1989), the minimum information stopping rule and the standard error stopping rule are compared; in De Ayala (1992), the fixed test length rule is used. For a more extensive review of research on polytomous CAT, see Dodd et al. (1995).

Simulation Studies

Purpose

Because little is known about the empirical distribution of $l_{z,pol}$ in the context of short P&P tests and in CATs where $\hat{\theta}$ is used as an estimate of θ, this study is designed to investigate whether the empirical distribution of $l_{z,pol}$ is in concordance with the theoretical distribution (standard normal distribution). This is done for both P&P tests and for CATs.

Method

P&P Tests

Two P&P tests were constructed using $N = 20$ and $N = 50$ three-step items that fit the PCM. The item parameters of the first 20 and 50 items presented in Table 1 in Koch and Dodd (1989) were used to construct the 20-item and 50-item P&P tests, respectively.

For each test, eight datasets of 1,000 response vectors were simulated. Seven datasets were simulated at seven different θ levels: $\theta = -1.5, -1.0, -.5, 0, .5, 1.0$, and 1.5. The eighth dataset was simulated in which 1,000 θs were drawn from $N(0, 1)$.

Values of $l_{z,pol}$ were calculated using $\hat{\theta}$ in each dataset for each response vector. These 1,000 values of $l_{z,pol}$ were used to obtain the empirical distribution of $l_{z,pol}$ for each dataset. The thetas (θ) were 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. The empirical Type I error rates were compared with the nominal Type I error rates. Also, the first four moments of the simulated distributions of $l_{z,pol}$ were computed and compared with the moments of the standard normal distribution. Then $l_{z,pol}$ was calculated using true θ for each response vector to examine the distribution of $l_{z,pol}$ without the presence of estimation errors.

CAT

An item bank of 240 three-step items fitting the PCM was used to simulate two different CATs with $N = 20$ and $N = 50$; the 60 items in Table 1 in Koch and Dodd (1989) were quadrupled to 240 items. For both CATs, eight datasets of 1,000 adaptive item score patterns were simulated. Seven datasets were simulated at seven θ levels: $\theta = -1.5, -1.0, -.5, 0, .5, 1.0$, and 1.5. For the eighth 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)

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

$$= \sum_{j=0}^{m} j^2 P_{ij}(\theta) - \left[\sum_{j=0}^{m} j P_{ij}(\theta) \right]^2 .$$

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 20 or 50.

For each response vector, the values of $l_{z,pol}$ were calculated using $\hat{\theta}$. The 1, 000 values of $l_{z,pol}$ were used to obtain the empirical distribution of $l_{z,pol}$ for each dataset. $l_{z,pol}$ was also calculated using true θ to determine the distributions of $l_{z,pol}$ without the presence of estimation errors. 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 α = .005, .01, .015, .02, and .025. The empirical Type I error rates were compared with the nominal Type I error rates. Also, the first four moments of the simulated distributions of $l_{z,pol}$ were computed and compared with the moments of the standard normal distribution.

Results

In Tables 1 and 2, the first four moments of the simulated distribution of $l_{z,pol}$ are given for the tests (P&P and CAT) of 20 and 50 items, respectively.

TABLE 1
Distributional characteristics of the simulated distribution of $l_{z,pol}$ for the 20-item P&P test and the 20-item CAT at five critical values

					Empirical Type I Error at α				
	Mean	Variance	Skewness	Kurtosis	0.005	0.010	0.015	0.020	0.025
P&P, using θ									
$\theta \sim N$ (0, 1)	0.02	1.00	–0.52	0.23	0.012	0.021	0.025	0.030	0.032
θ = –1.5	0.02	0.99	–0.69	0.39	0.014	0.021	0.030	0.040	0.047
–1.0	–0.03	1.03	–0.58	0.15	0.014	0.022	0.030	0.040	0.045
–0.5	0.02	0.98	–0.47	0.15	0.011	0.017	0.020	0.029	0.033
0.0	0.01	1.08	–0.57	0.29	0.016	0.028	0.040	0.043	0.047
0.5	0.01	1.10	–0.57	0.33	0.014	0.020	0.032	0.037	0.044
1.0	0.00	0.99	–0.43	–0.09	0.010	0.016	0.022	0.034	0.038
1.5	–0.02	1.01	–0.59	0.39	0.016	0.022	0.031	0.040	0.044
P&P, using θ									
$\theta \sim N$ (0,1)	0.08	0.63	–0.35	0.64	0.003	0.009	0.010	0.010	0.010
θ = –1.5	0.05	0.17	–0.03	1.63	0.000	0.000	0.000	0.000	0.000
–1.0	0.01	0.24	–0.44	0.48	0.000	0.000	0.000	0.001	0.001
–0.5	0.05	0.44	–0.29	0.17	0.001	0.001	0.001	0.003	0.004
0.0	0.08	0.78	–0.51	0.32	0.006	0.009	0.014	0.019	0.022
0.5	0.12	1.01	–0.59	0.46	0.009	0.016	0.019	0.026	0.029
1.0	0.11	0.94	–0.46	0.07	0.007	0.014	0.018	0.027	0.031
1.5	0.07	0.78	–0.53	0.46	0.007	0.009	0.012	0.016	0.021
CAT, using θ									
θ ~N (0,1)	–0.01	1.12	–0.44	0.36	0.014	0.019	0.032	0.039	0.046
θ = –1.5	0.04	1.00	–0.11	–0.29	0.006	0.011	0.015	0.018	0.023
–1.0	–0.03	0.98	–0.33	–0.14	0.007	0.018	0.023	0.028	0.031
–0.5	–0.03	1.07	–0.34	0.19	0.008	0.017	0.024	0.035	0.039
0.0	–0.00	1.02	–0.12	–0.12	0.005	0.016	0.019	0.020	0.029
0.5	–0.01	1.09	–0.42	0.08	0.009	0.020	0.023	0.035	0.042
1.0	–0.01	0.99	–0.38	0.08	0.009	0.015	0.021	0.029	0.036
1.5	–0.00	0.98	–0.26	–0.04	0.008	0.016	0.022	0.028	0.029
CAT, using θ									
$\theta \sim N$(0, 1)	–0.06	3.04	**	**	0.021	0.024	0.029	0.032	0.043
θ = –1.5	–0.02	2.76	**	**	0.015	0.018	0.023	0.026	0.027
–1.0	0.06	0.88	–0.35	–0.11	0.005	0.008	0.013	0.018	0.021
–0.5	0.08	1.00	–0.34	0.05	0.006	0.012	0.019	0.026	0.031
0.0	0.10	0.96	–0.12	–0.12	0.004	0.007	0.009	0.016	0.020
0.5	0.09	1.01	–0.49	0.31	0.006	0.014	0.018	0.025	0.031
1.0	0.09	0.86	–0.31	0.00	0.002	0.009	0.011	0.015	0.017
1.5	0.01	1.84	**	**	0.010	0.016	0.022	0.025	0.031

**: no plausible values

TABLE 2
Distributional characteristics of the simulated distribution of $l_{z,pol}$ for the 50-item P&P test and the 50-item CAT at five critical values

					Empirical Type I Error at α				
	Mean	Variance	Skewness	Kurtosis	0.005	0.010	0.015	0.020	0.025
P&P, using θ									
$\theta \sim N(0, 1)$	0.00	0.98	–0.42	0.67	0.008	0.010	0.014	0.018	0.025
$\theta = -1.5$	0.01	1.05	–0.54	0.27	0.014	0.024	0.029	0.035	0.042
–1.0	–0.05	1.05	–0.51	0.34	0.017	0.023	0.034	0.039	0.043
–0.5	0.01	1.01	–0.37	0.28	0.012	0.019	0.023	0.029	0.031
0.0	0.03	0.97	–0.26	–0.01	0.008	0.013	0.020	0.025	0.028
0.5	–0.05	1.01	–0.14	0.10	0.010	0.015	0.019	0.026	0.034
1.0	0.03	0.93	–0.45	0.28	0.010	0.012	0.021	0.028	0.031
1.5	–0.04	1.03	–0.35	–0.07	0.009	0.015	0.025	0.034	0.039
P&P, using θ									
$\theta \sim N(0, 1)$	0.07	0.86	–0.30	0.25	0.006	0.008	0.009	0.012	0.014
$\theta = -1.5$	0.07	0.63	–0.39	–0.02	0.001	0.003	0.003	0.005	0.011
–1.0	0.02	0.82	–0.43	0.26	0.007	0.010	0.014	0.023	0.028
–0.5	0.07	0.94	–0.36	0.18	0.007	0.010	0.019	0.023	0.028
0.0	0.12	0.93	–0.26	0.02	0.007	0.013	0.017	0.019	0.021
0.5	0.03	1.00	–0.16	0.12	0.008	0.013	0.014	0.020	0.028
1.0	0.11	0.85	–0.37	0.03	0.006	0.009	0.015	0.019	0.020
1.5	0.05	0.83	–0.33	0.02	0.005	0.009	0.010	0.012	0.018
CAT, using θ									
$\theta \sim N(0, 1)$	0.04	1.01	–0.41	0.01	0.009	0.018	0.025	0.029	0.035
$\theta = -1.5$	–0.00	0.97	–0.31	–0.05	0.006	0.012	0.016	0.022	0.035
–1.0	–0.01	1.07	–0.41	0.22	0.014	0.020	0.027	0.035	0.036
–0.5	–0.04	1.01	–0.21	–0.02	0.012	0.018	0.021	0.027	0.035
0.0	–0.04	1.07	–0.32	0.02	0.009	0.017	0.023	0.031	0.036
0.5	0.02	1.02	–0.46	0.12	0.015	0.021	0.031	0.035	0.039
1.0	–0.06	1.03	–0.55	0.59	0.014	0.019	0.025	0.032	0.040
1.5	0.07	0.96	–0.43	0.04	0.009	0.012	0.014	0.019	0.025
CAT, using θ									
$\theta \sim N(0, 1)$	0.12	0.90	–0.36	0.03	0.007	0.012	0.016	0.019	0.025
$\theta = -1.5$	0.05	0.79	–0.24	–0.19	0.001	0.003	0.009	0.012	0.017
–1.0	0.06	1.01	–0.38	0.03	0.009	0.015	0.026	0.029	0.034
–0.5	0.04	1.01	–0.23	–0.07	0.011	0.016	0.019	0.023	0.030
0.0	0.04	1.06	–0.34	0.05	0.008	0.014	0.021	0.025	0.029
0.5	0.10	1.00	–0.46	0.10	0.010	0.020	0.028	0.032	0.036
1.0	0.02	0.95	–0.51	0.50	0.009	0.012	0.016	0.022	0.031
1.5	0.10	0.77	–0.38	0.17	0.004	0.006	0.009	0.010	0.012

Tables 1 and 2 show that, for almost all tests and all datasets, the mean and variance of $l_{z,pol}$ using $\hat{\theta}$ or θ were close to the expected 0 and 1, respectively. The distributions are slightly negatively skewed and have a small positive kurtosis. Exceptions are the distributions of $l_{z,pol}$ for both P&P tests and both CATs with $\theta = -1.5$ and -1.0 where $\hat{\theta}$ was used; here, the variances of $l_{z,pol}$ are unequal to 1. Also, the item pool may consist of too few items that are suitable for examinees with low θ values. However, due to the fact that for almost all tests and for all datasets the mean and variance were close to the expected values under the standard normal distribution, the empirical Type I error rates were close to the nominal Type I error rates, also.

Discussion

In this study, the empirical distributions of existing person-fit statistics for polytomously scored items were investigated. It was shown that the empirical distribution of $l_{z,pol}$ was close to the standard normal distribution for both P&P tests and CAT, for θ around zero. This can be explained by the fact that the item pool consisted mainly of items of medium difficulty. To be able to distinguish normal from aberrant examinees, it is therefore important to have an item pool that consists not only of items of medium difficulty, but also contains easy and difficult items.

It was interesting that the results differed from the results found in van Krimpen-Stoop and Meijer (1999a) where a CAT with dichotomous items was used. The better fit between the empirical and theoretical distribution in this study may be explained by the fact that, by using polytomous items, more information for each item is available. This effect is comparable with using longer tests, which also results in higher agreement between empirical and theoretical distributions. Future research may concentrate on CAT-specific statistics for polytomous scored items.

Another alternative might be to design a CAT algorithm that verifies model assumptions for an individual examinee. Instead of selecting items only on the basis of information, items could be selected to have a wider range of difficulties. This could result in a somewhat inefficient use of CAT (larger item banks,

longer tests), but it might improve the accuracy of individual measurement by determining an individual examinee's fit to the test model.

References

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

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.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of the partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5*, 17–34.

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

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

Dodd, B. G., Koch, W. R. & 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., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

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.

Koch, W. R., & 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., & 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.). New York: Chapman and Hall.

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., & Sijtsma (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

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.

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

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.

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.

Rogers, H. J., & 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.

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

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

Snijders, T. (2001). Asymptotic null distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66*, 331–342.

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. *British Journal of Mathematical and Statistical Psychology, 43*, 39–55.

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

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (l999b). The null distribution of a person-fit statistic in fixed- and computerized adaptive testing. *Applied Psychological Measurement, 23*, 327–345.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

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